

Estimation of Rapidly Time-Varying Harmonic Noise for Speech Enhancement

Thomas Esch, *Student Member, IEEE*, Matthias Rüngeler, *Student Member, IEEE*, Florian Heese, *Student Member, IEEE*, and Peter Vary, *Fellow, IEEE*

Abstract—Robust speech enhancement relies on the estimation of stationary as well as non-stationary background noise. This contribution presents a novel approach for estimating the short-term power spectral densities (ST-PSDs) of rapidly time-varying harmonic noise as produced, e.g., by cars or motorcycles. The well-known Minimum Statistics algorithm is modified by frequency warping controlled by the fundamental frequency of the harmonic noise which is assumed to be known a priori. The resulting noise estimates are used for the enhancement of the noisy signal. A detailed description of the algorithm is given and it is shown by a thorough analysis that the new solution considerably outperforms two conventional noise ST-PSD estimation techniques.

Index Terms—Speech enhancement, noise reduction, adverse environments, non-stationary noise

EDICS Category: SPE-LANG

I. INTRODUCTION

THE problem of improving the quality of noisy speech is still an active field of research. The most popular method for enhancing degraded speech is based on the short-time Fourier transform (ST-FT). In the ST-FT domain, individual adaptive gains are applied to the noisy input coefficients for noise reduction. Therefore, most practical single-microphone speech enhancement algorithms require an estimate of the short-term noise power spectral density (ST-PSD). For this purpose, several approaches can be found in literature, e.g., the application of a voice activity detector (VAD) [3], the Minimum Statistics (MS) approach [4], a minimum mean square error (MMSE) noise ST-PSD tracking algorithm [5] and an estimator based on the probability of speech presence [6]. All methods have severe problems in tracking a sudden rise in noise energy.

In [1] and [2], we have presented an experimental approach which overcomes some of these limitations for *time-varying harmonic noise* by using dynamic frequency warping. The main focus of this paper is, besides a detailed description of the estimation algorithm, a thorough performance analysis.

The proposed algorithm can be applied to many noise estimators. This contribution considers as representative application the Minimum Statistics noise estimator [4]. Typical applications are hands-free devices inside a car or intercom systems for motorcycles where the engine is the main noise source. An example is depicted in Fig. 1 showing the noisy

Copyright (c) 2012 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org. Parts of this work have appeared in [1] and [2]. T. Esch, M. Rüngeler, F. Heese, and P. Vary are with the Institute of Communication Systems and Data Processing, RWTH Aachen University, Aachen 52074, Germany (e-mail: esch@ind.rwth-aachen.de; ruengeler@ind.rwth-aachen.de; heese@ind.rwth-aachen.de; vary@ind.rwth-aachen.de).

spectrogram of a speech signal disturbed by a real noise signal recorded inside a car. The strong spectral components of the harmonic noise signal are present at multiples of a fundamental frequency f_h which directly corresponds to the revolutions per minute (RPM) of the engine. In the following, it is assumed that the instantaneous fundamental frequency f_h (or RPM) of the harmonic noise is available a priori to the noise estimation algorithm, e.g., provided from the vehicle's onboard computer or estimated in advance. For the estimation, the noisy input signal (including speech) can be used or the fundamental frequency can be determined from an auxiliary microphone placed, e.g., near the engine. In the latter case, new degrees of freedom are offered for the placing of the second microphone in comparison to noise suppression techniques which rely, e.g., on the coherence between the two microphone signals.

The remainder of this contribution is organized as follows: In Sec. II, the noise estimation problem is formulated introducing all notations which are required in the sequel. Section III comprises the procedure of the proposed harmonic noise tracking algorithm in detail. Experimental results are shown in Sec. IV and conclusions are drawn in Sec. V.

II. PROBLEM FORMULATION

A clean speech signal $s(k)$ is assumed to be degraded by some additive noise $n(k)$. The resulting noisy signal $y(k)$ picked up by a microphone is given by:

$$y(k) = s(k) + n(k), \quad (1)$$

where k is the discrete time-sample index. Speech and noise signals are assumed to be uncorrelated. The aim of any noise suppression system is to estimate the clean speech having access only to the noisy microphone signal $y(k)$. It is desirable

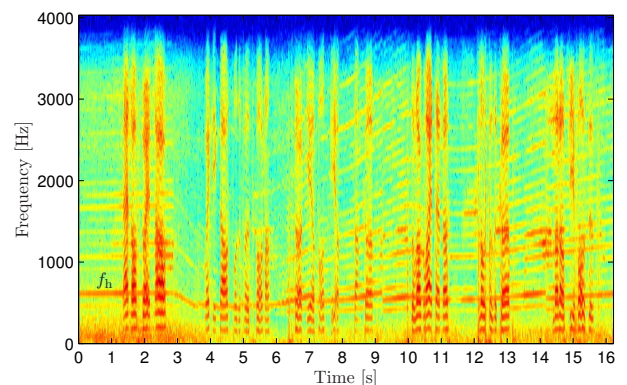


Fig. 1. Spectrogram of noisy input signal showing strong spectral components of harmonic noise at multiples of a time-varying fundamental frequency f_h .

to attenuate the noise signal as much as possible while keeping the distortions of the speech signal as low as possible at the same time.

A widely used method for single-microphone noise reduction is based on spectral weighting in the ST-FT domain using statistical noise suppression techniques. The microphone signal is segmented into overlapping frames of length L . After windowing and (if applied) zero-padding, these frames are transformed into the spectral domain via a Fast Fourier Transform (FFT) of length $M \geq L$. The spectrum of the noisy input signal is given by:

$$Y(\lambda, \mu) = S(\lambda, \mu) + N(\lambda, \mu), \quad (2)$$

where $S(\lambda, \mu)$ and $N(\lambda, \mu)$ represent the spectral FFT coefficients of speech and noise at frame λ and frequency bin μ . The noisy input coefficients $Y(\lambda, \mu)$ are modified by a gain function $G(\lambda, \mu)$ resulting in the speech estimates $\hat{S}(\lambda, \mu) = G(\lambda, \mu) \cdot Y(\lambda, \mu)$. In most cases, the corresponding weighting rules rely on statistical characteristics of the speech and the noise signal and provide weighting gains for each frame and frequency bin. Besides appropriate statistical models and adequate distortion measures, the estimation of the noise ST-PSD $\sigma_N^2(\lambda, \mu)$ is the most crucial task in any noise reduction system.

III. HARMONIC NOISE ESTIMATION

For noise estimation, conventional algorithms are usually based on the assumption that noise is stationary or at least quasi-stationary. However, realistic background noise can be rapidly time-varying and highly non-stationary. In this paper, speech signals are assumed to be disturbed by harmonic noise characterized by (strong) spectral components at multiples of a time-varying fundamental frequency f_h , cf. Fig. 1. As this fundamental frequency might change very fast over time (e.g., when the engine is accelerated or a gear is changed), conventional noise estimation techniques mostly fail in tracking time-varying spectral harmonics. In the following, a new modified Minimum Statistics algorithm is presented which achieves considerably better noise estimates.

1) *Original Minimum Statistics Approach*: The original Minimum Statistics (MS) approach [4] relies on two 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). With these assumptions, the concept of MS is based on tracking the minimum of the smoothed noisy ST-PSD within a sliding time window. As the true minimum is always smaller or equal to the mean noise power, a bias correction is applied.

In a first step, $|Y(\lambda, \mu)|^2$ is recursively smoothed over time. The smoothed ST-PSD $\bar{\sigma}_Y^2(\lambda, \mu)$ is given by:

$$\bar{\sigma}_Y^2(\lambda, \mu) = \bar{\alpha}(\lambda, \mu) \cdot \bar{\sigma}_Y^2(\lambda - 1, \mu) + (1 - \bar{\alpha}(\lambda, \mu)) \cdot |Y(\lambda, \mu)|^2, \quad (3)$$

where $\bar{\alpha}(\lambda, \mu) \in [0, 1]$ denotes a frame and frequency-dependent adaptive smoothing factor [4] and $|\cdot|$ the magnitude operator. Afterwards, the minimum $\bar{\sigma}_{Y,\min}^2(\lambda, \mu)$ of the most recent D values is tracked for each frequency bin *separately* by a sliding time window according to:

$$\bar{\sigma}_{Y,\min}^2(\lambda, \mu) = \min_{\tilde{\lambda} \in [\lambda - D + 1, \lambda]} \bar{\sigma}_Y^2(\tilde{\lambda}, \mu). \quad (4)$$

The duration of the time window for the minimum search should be equal to approximately 1.5 seconds [4]. The minimum value is multiplied by a bias correction factor $B(\lambda, \mu)$, which is mainly dependent on the variance of the noisy input periodogram. The final noise ST-PSD estimate is given by [4]:

$$\bar{\sigma}_N^2(\lambda, \mu) = B(\lambda, \mu) \cdot \bar{\sigma}_{Y,\min}^2(\lambda, \mu). \quad (5)$$

MS shows good estimation results in stationary and slowly changing noise conditions. However, due to the large window length D , it is not able to track a sudden rise in noise energy leading to an under-estimation of the noise power in this case.

2) *Novel Modified Minimum Statistics Approach*: In order to estimate the power of the time-varying harmonic noise components, the original Minimum Statistics approach is modified. The new concept is illustrated and compared with the original one in Fig. 2. The figure shows the temporal course of four harmonic oscillations over frequency including a sudden rise of the fundamental frequency f_h within the search window. To determine the noise ST-PSD in frame λ_0 at one particular frequency, the original MS algorithm tracks the minimum of the most recent D frames by considering entities *only* at *this* specific frequency, see Method ① in Fig. 2. In contrast, the modified method adaptively ‘looks back’ inclined according to the evolution of the harmonics in the time-frequency domain, see Method ② in Fig. 2. Following one specific harmonic over time, the short-term powers of the harmonic components are no longer fluctuating that much but are relatively stationary. Thus, tracking the minimum along the courses of the harmonics will lead to much better noise ST-PSD estimation results in this scenario. The realization of this modified MS approach is described in the following.

First, the smoothing procedure needs to be adapted according to the fundamental frequency $f_h(\lambda)$ of each frame. To achieve the tilted ‘viewing direction’, Eq. 3 is modified:

$$\hat{\sigma}_Y^2(\lambda, \mu) = \hat{\alpha}(\lambda, \mu) \cdot \hat{\sigma}_Y^2\left(\lambda - 1, \frac{\mu}{r_1(\lambda_0)}\right) + (1 - \hat{\alpha}(\lambda, \mu)) \cdot |Y(\lambda, \mu)|^2, \quad (6)$$

ensuring a recursive smoothing along the course of the harmonics based on the ratio $r_1(\lambda_0) = \frac{f_h(\lambda_0)}{f_h(\lambda_0 - 1)}$. In case $\mu/r_1(\lambda_0)$ is not an integer, linear interpolation is used to compute $\hat{\sigma}_Y^2(\lambda - 1, \mu/r_1(\lambda_0))$. If $r_1(\lambda_0) < 1$, elements are missing for $\mu > (M/2 - 1) \cdot r_1(\lambda_0)$. They are replaced by $\hat{\sigma}_Y^2(\lambda - 1, M/2 - 1)$. The adaptive smoothing factors $\hat{\alpha}(\lambda, \mu)$ in Eq. 6 are warped in the same way using the same ratio r_1 .

The resulting smoothed signal powers $\hat{\sigma}_Y^2(\lambda, \mu)$ of the most recent D frames are buffered in the matrix $\hat{\Sigma}_Y^2(\lambda)$:

$$\hat{\Sigma}_Y^2(\lambda) = \begin{pmatrix} \hat{\sigma}_Y^2(\lambda - D + 1, 0) & \dots & \hat{\sigma}_Y^2(\lambda, 0) \\ \hat{\sigma}_Y^2(\lambda - D + 1, 1) & \dots & \hat{\sigma}_Y^2(\lambda, 1) \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_Y^2(\lambda - D + 1, \mu) & \dots & \hat{\sigma}_Y^2(\lambda, \mu) \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_Y^2(\lambda - D + 1, \frac{M}{2} - 1) & \dots & \hat{\sigma}_Y^2(\lambda, \frac{M}{2} - 1) \end{pmatrix}. \quad (7)$$

In order to track the minimum according to the evolution of the harmonics, the entries of the matrix $\hat{\Sigma}_Y^2(\lambda)$ need to be modified as well. The first harmonic oscillation passes through $f_h(\lambda_0)$ in the current frame λ_0 and through $f_h(\lambda_0 - D + 1 + j)$ in frame $\lambda_0 - D + 1 + j$ with $0 \leq j < D$. For the estimation

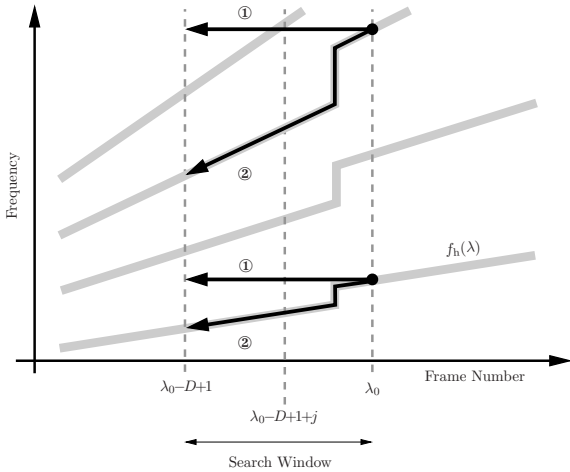


Fig. 2. ‘Direction of view’ of ① original Minimum Statistics and ② modified Minimum Statistics algorithm.

of the noise ST-PSD at frame λ_0 , the j -th column of the matrix $\hat{\Sigma}_Y^2(\lambda_0)$ is therefore compressed/expanded according to the ratio:

$$r_D(\lambda_0, j) = \frac{f_h(\lambda_0)}{f_h(\lambda_0 - D + 1 + j)}. \quad (8)$$

After transformation, the j -th column of the new modified matrix $\hat{\Sigma}_{Y,\text{mod}}^2(\lambda_0)$ comprises the noisy signal powers at the new positions $\tilde{\mu} = \frac{\mu}{r_D(\lambda_0, j)}$. The matrix $\hat{\Sigma}_{Y,\text{mod}}^2(\lambda_0)$ has the same structure and size as $\hat{\Sigma}_Y^2(\lambda)$ in Eq. 7. Linear interpolation is again used for this curve fitting problem. The proposed frequency warping technique is illustrated in Fig. 3 schematically. Visually speaking, the course of the harmonics is brought into a horizontal position within the time-frequency plain. Figure 3a) depicts the warping scheme for expansion ($r_D > 1$, i.e., f_h is increasing). In this case, all information from the past required for the interpolation is available (frequency bin $\mu = 0$ is mapped to frequency bin $\tilde{\mu} = 0$) and the harmonics within the search window are ‘tilted up’. In the case of compression ($r_D < 1$, i.e., f_h is decreasing), the harmonics are ‘tilted down’ in the spectrogram as can be seen in Fig. 3b). Here, the elements of the j -th column are missing for bins $\tilde{\mu} > M/2 - 1$ and replaced by $\hat{\sigma}_Y^2(\lambda_1 - D + 1 + j, M/2 - 1)$.

Afterwards, the original minimum tracking concept of the MS algorithm can be applied to the ‘warped’ spectrogram by tracking the minimum for each row of the modified matrix $\hat{\Sigma}_{Y,\text{mod}}^2(\lambda)$:

$$\hat{\sigma}_{Y,\text{mod},\text{min}}^2(\lambda, \mu) = \min \left(\hat{\sigma}_{Y,\text{mod}}^2(\lambda, \mu) \right), \quad (9)$$

where $\hat{\sigma}_{Y,\text{mod}}^2(\lambda, \mu)$ represents the μ -th row of $\hat{\Sigma}_{Y,\text{mod}}^2(\lambda)$. From the Minimum Statistics’ point of view, the harmonics in the time-frequency domain of $\hat{\Sigma}_{\text{mod}}^2(\lambda)$ appear more stationary over time than in the original approach. Finally, the bias is calculated and applied according to the original MS approach.

IV. PERFORMANCE RESULTS

The performance of the proposed noise estimation technique for harmonic noise environments is compared with the original Minimum Statistics (MS) approach [4] and the MMSE

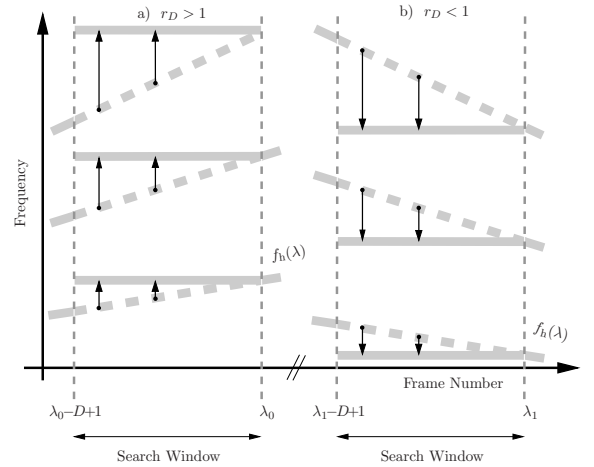


Fig. 3. Frequency warping of all frames within search window using linear interpolation for a) $r_D > 1$ (expansion) and b) $r_D < 1$ (compression).

based noise PSD tracking algorithm [5]. For the objective evaluation of the algorithms, the log-error distortion measure LogERR [5] is applied defined as:

$$\text{LogERR} = \frac{1}{M \cdot N} \sum_{\mu=0}^{M-1} \sum_{\lambda=0}^{N-1} \left| 10 \cdot \log_{10} \frac{|N(\lambda, \mu)|^2}{\hat{\sigma}_N^2(\lambda, \mu)} \right|, \quad (10)$$

where N represents the total number of evaluated frames and $\hat{\sigma}_N^2(\lambda, \mu)$ states the noise ST-PSD estimate of the respective investigated technique. The lower the value of LogERR, the better the performance. Additionally, the following system settings are applied: frame length $L = 160$ ($\hat{=} 20$ ms) and FFT length $M = 256$ (including zero-padding) using a frame overlap of 50% (Hann window) and a sampling frequency $f_s = 8$ kHz.

The performance of the different algorithms is evaluated in two scenarios. For a proof of concept, an artificially generated noise signal *without* speech components is used as input signal in the first scenario. The spectrogram of the noise signal is depicted in Fig. 4 showing six time-varying harmonic oscillations with increasing frequency. The results of the original Minimum Statistics approach are compared with the results of the proposed modified Minimum Statistics approach where the fundamental frequency f_h was provided with perfect knowledge in each frame. The measured LogERR values for this scenario are 7.41 dB for the original Minimum Statistics approach and 2.87 dB for the novel approach and clearly show the preference for the proposed noise estimation technique when applied to this specific input signal.

In the second scenario, four different (real) noise signals are used which were recorded inside a car during acceleration and deceleration phases, i.e., the signals contain a relatively large portion of time-varying harmonic engine noise. The fundamental frequency f_h which is required for the proposed technique was provided directly by the engine control unit. In order to investigate the influence of f_h estimation errors, additional simulations are carried out with a disturbed fundamental frequency $f_\Delta(\lambda) = f_h(\lambda) + \Delta(\lambda)$, where $\Delta(\lambda)$ represents a zero-mean, uniformly distributed noise signal causing a relative deviation of $d_r = \frac{1}{N} \sum_{\lambda=0}^{N-1} \frac{|f_h(\lambda) - f_\Delta(\lambda)|}{f_h(\lambda)}$.

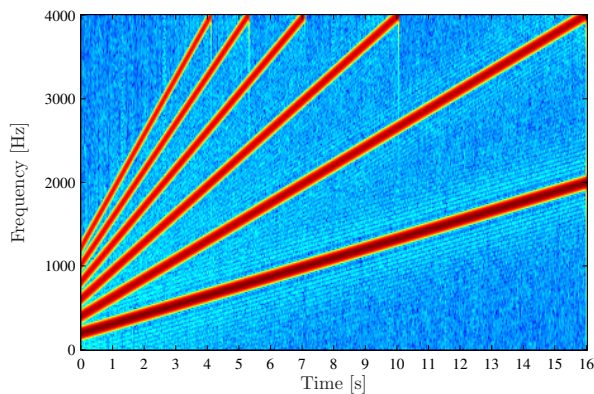


Fig. 4. Spectrogram of artificially generated noise signal consisting of six time-varying harmonic oscillations.

The recordings are each added to three male and two female speech sequences (each with a length of 8 seconds) taken randomly from the NTT speech database. The input SNR is adjusted to realistic values varying between -10 dB and 15 dB with an increment of 5 dB.

Figure 5 depicts the averaged results ($N = 16000$) for the LogERR measure plotted over the input SNR for all investigated methods. It can be seen that the results for the two conventional methods, [4] and [5], are quite similar at low input SNR values with slightly lower LogERR values for the MMSE based noise PSD tracking algorithm [5] at higher input SNR values. The best noise estimation performance however is provided again by the proposed modified Minimum Statistics approach if the actual fundamental frequency f_h is available. In this case, the original MS technique and the MMSE based noise PSD tracking algorithm are clearly outperformed with gains of up to 1.75 dB. If f_h can not be provided accurately by the engine control unit or can not be properly estimated in advance, the performance decreases. However, small deviations of f_h up to 15%¹ are acceptable and still yield good results as depicted in Fig. 5. If f_h has to be estimated in advance, frequency doubling or halving errors can occur. However, these errors will have no influence on the proposed algorithm as only the ratio of the fundamental frequencies within consecutive frames has to be known, see Eq. 8.

As mentioned before, the proposed algorithm can be applied in a noise reduction system to enhance the overall speech quality, e.g., in a speech communication device. Investigations are carried out in [1] and [2] based on a Wiener filter suppression rule. In summary, the results here also showed clearly perceivable improvements due to the modified Minimum Statistics algorithm leading to a more comfortable listening condition when compared to state-of-the-art noise estimation techniques. For the evaluations in [1] and [2], noise attenuation and speech distortion measures have been used. In addition, strong quality improvements could be verified by informal subjective listening tests.

In principle, the application of the proposed ‘warping’ technique according to the fundamental frequency f_h is not restricted to the Minimum Statistics algorithm and can be applied to any noise estimation method which takes into

¹The application of our own single-channel fundamental frequency estimator achieved SNR-dependent estimation errors less than 15%. The description of this estimation algorithm would exceed the scope of the paper.

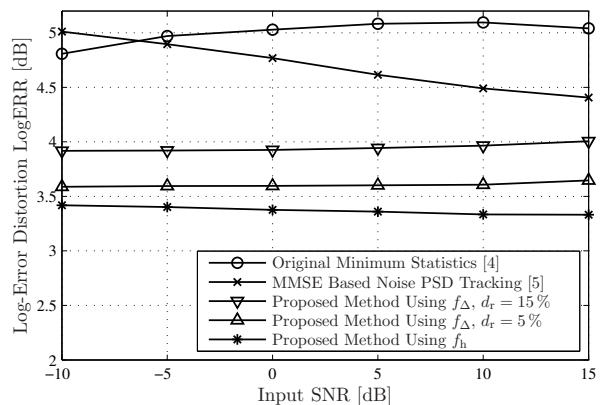


Fig. 5. Log-error distortion between estimated and true noise short-term PSD plotted over input SNR.

account noisy or enhanced coefficients from previous frames.

As the modified noise estimation technique is adapted to the fundamental frequency, the stationary noise components (e.g., wind or tyre noise) are suppressed by this approach satisfactorily only in the case of white noise. Therefore, we proposed in [2] to add a second stage of conventional noise suppression to reduce the remaining stationary noise components.

V. CONCLUSIONS

In [1] and [2], we presented some experiments for estimating the short-term noise PSD in rapidly time-varying harmonic noise environments. This approach performs frequency warping according to the harmonics’ fundamental frequency in a first step. Then, the well-known concept of Minimum Statistics is applied to the modified spectrum to track the harmonic noise. In this contribution, a detailed description of the novel estimation algorithm is given. Although the complexity is moderately increased, the proposed method considerably outperforms two conventional noise estimation techniques as shown in a thorough performance analysis.

REFERENCES

- [1] 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, Oct. 2010.
- [2] —, “Combined Reduction of Time Varying Harmonic and Stationary Noise Using Frequency Warping,” in *Conference Record of Asilomar Conference on Signals, Systems, and Computers (ACSSC)*, Pacific Grove, CA, USA, Nov. 2010.
- [3] J. Sohn, N. S. Kim, and W. Sung, “A Statistical Model-Based Voice Activity Detection,” *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
- [4] 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, Jul. 2001.
- [5] R. C. Hendriks, R. Heusdens, and J. Jensen, “MMSE Based Noise PSD Tracking with Low Complexity,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, Mar. 2010.
- [6] T. Gerkmann and R. C. Hendriks, “Noise Power Estimation Based on the Probability of Speech Presence,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2011.