AN EVALUATION OF NOISE POWER SPECTRAL DENSITY ESTIMATION ALGORITHMS IN ADVERSE ACOUSTIC ENVIRONMENTS

Jalal Taghia*, Jalil Taghia, Nasser Mohammadiha†, Jinqiu Sang‡, Vaclav Bouse§, and Rainer Martin¶

Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany
email: {jalal.taghia, rainer.martin}@rub.de

ABSTRACT

Noise power spectral density estimation is an important component of speech enhancement systems due to its considerable effect on the quality and the intelligibility of the enhanced speech. Recently, many new algorithms have been proposed and significant progress in noise tracking has been made.

In this paper, we present an evaluation framework for measuring the performance of some recently proposed and some well-known noise power spectral density estimators and compare their performance in adverse acoustic environments. In this investigation we do not only consider the performance in the mean of a spectral distance measure but also evaluate the variance of the estimators as the latter is related to undesirable fluctuations also known as musical noise. By providing a variety of different non-stationary noises, the robustness of noise estimators in adverse environments is examined.

Index Terms— Noise power estimation, speech enhancement

1. INTRODUCTION

A crucial part of speech enhancement in the frequency domain is noise power spectral density (PSD) estimation. The noise statistics can be used in noise reduction for instance to compute the a priori signal-to-noise ratio (SNR) and subsequently the spectral gain [1]. Thus, the performance of the overall enhancement system highly depends on the accurate estimation of the unknown noise statistics. When noise is non-stationary this task is difficult to solve since updating the noise PSD estimate during speech pause only is not sufficient to achieve an accurate and fast tracking of the noise PSD. Noise power estimation approaches must be robust with respect to speech activity and low SNR conditions.

In fact noise power estimation in adverse environments (such as low SNR conditions and non-stationary noise environments) is still a very challenging problem and there are fairly many approaches in the literature which shows a lot of interest in this topic. Unfortunately, there is no recent and comprehensive comparison of noise power estimators with respect to their capabilities of tracking the noise PSD in adverse environments. In [2] a comparison was done for several single and two channel noise estimation techniques which

* Jalal Taghia is with Ruhr-Universität Bochum, Germany.
† Jalil Taghia and Nasser Mohammadiha are with KTH Stockholm, Sweden.
‡ Jinqiu Sang is with ISVR Southampton, U.K.
§ Vaclav Bouse is with Siemens Audiologische Technik GmbH, Erlangen, Germany.
¶ This work was funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699. The authors performed this work while at Ruhr-Universität Bochum.

were proposed before 1997. In [3] different classes of speech enhancement algorithms have been compared based on a subjective evaluation, however, this work also includes different noise reduction approaches and is not directly related to the evaluation of noise power estimators. Therefore, regarding the recent progress in noise PSD tracking, the aim of this paper is to implement some state-of-the-art noise PSD tracking algorithms in a common framework and to propose a suitable procedure for the evaluation of their performance.

The remainder of this paper is organized as follows: In Section 2 we briefly describe the selected algorithms. Section 3 outlines the evaluation framework and performance measures. The experimental framework is described in Section 4 and followed by a discussion of experimental results in Section 5.

2. OVERVIEW OF ALGORITHMS

In our framework 8 algorithms are implemented and evaluated. One of state-of-the-art methods for noise power estimation is the minimum statistics algorithm (MS algorithm) introduced in [4] and extended in [5]. The MS algorithm tracks the minima values of a smoothed power estimate of the noisy signal. By compensating the inherent bias, an estimate of the noise PSD is derived.

The second considered state-of-the-art approach is the minima-controlled recursive averaging algorithm (MCRA algorithm) [6] which is a combination of minimum tracking and recursive averaging. In the MCRA algorithm, the speech presence probability estimation is determined by taking the ratio between the power in the current frame and its minimum within a specified time frame. This speech probability estimation is subsequently used for adjusting a smoothing parameter which controls the averaging of the noise PSD. The improved minima-controlled recursive averaging algorithm (IMCRA algorithm) in [7] uses a two step procedure which makes the minimum tracking more robust during speech activity and refines an initial speech presence detector. In addition to the IMCRA algorithm, two other methods belonging to this category are MCRA-MAP [8] and EMCRA [9] which have been recently introduced and which we consider in our evaluation.

Recently, a subspace decomposition based approach in the discrete Fourier transform (DFT) domain was proposed [10]. The method, called subspace noise tracking (SNT) algorithm, is based on an eigenvalue decomposition of correlation matrices constructed using time series of noisy DFT coefficients. SNT works very well in many noise environments. However, a critical point of the SNT algorithm is to find the number of eigenvalues in the signal subspace (i.e. signal dimension estimation) which is used to estimate the noise PSD in each DFT bin.

The two last methods considered in our evaluation, are based
on minimum mean-squared error (MMSE) estimation of the noise periodogram. In our paper, these algorithms are referred to as MMSE-Hendriks [11] and MMSE-Yu [12].

3. EVALUATION FRAMEWORK

There are two issues that we took into account to achieve an accurate and meaningful evaluation. First, our evaluation of the noise PSD estimation performance shall be independent from any specific speech enhancement system which means, for the implemented algorithms, we compare the results of the noise PSD estimation but not the quality of enhanced speech. The reason is that there are a variety of different approaches in employing the estimated noise PSD to derive enhanced speech, and we want to separate their effects in our evaluation. Second, we aim at the estimation of the noise power spectral density and thus we should reduce the impact of random fluctuations in the original noise periodogram. For instance, these instantaneous fluctuations will depend on the temporal alignment of the spectral analysis window. Furthermore, most noise reduction approaches require a smoothed noise estimate. Thus, it is necessary to obtain a suitable reference noise PSD for use in the performance measurements.

3.1. Performance Measures

We assume that in the DFT domain the noisy speech signal \( Y(k, i) \) is an additive mixture of clean speech coefficients \( X(k, i) \) and noise coefficients \( D(k, i) \) where \( k \) and \( i \) are the frequency bin and time frame index, respectively. It is also assumed that clean speech and noise are uncorrelated. To preserve the variance of noise in the transformation from the time domain to the DFT domain, the spectral analysis window is normalized to its energy. We denote the reference noise PSD and the estimated PSD by \( \text{PSD} \) and \( \hat{\text{PSD}} \) respectively. The first evaluation measure is the averaged log distance \( \text{LogErr}_{\text{mean}} \) between the estimated and reference noise PSD defined as follows

\[
\text{LogErr}_{\text{mean}} = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} \Delta_{dB}(k, i),
\]

where

\[
\Delta_{dB}(k, i) = 10 \log_{10} \left[ \frac{\sigma_{D}^{2}(k, i)}{\sigma_{\hat{D}}^{2}(k, i)} \right].
\]

\( I \) and \( K \) are the number of time frames and frequency bins (DFT length) respectively. \( \text{LogErr}_{\text{mean}} \) is expressed in dB and is the mean over all frequency bins and frame indices. \( \text{LogErr}_{\text{mean}} \) includes both overestimation and underestimation of the noise PSD but does not differentiate between a systematic error (bias) and random fluctuations. These fluctuations are important to be measured since they may introduce a possibly large variance to enhanced speech fluctuations. These fluctuations are important to be measured since they may introduce a possibly large variance to enhanced speech fluctuations. These fluctuations are important to be measured since they may introduce a possibly large variance to enhanced speech fluctuations. These fluctuations are important to be measured since they may introduce a possibly large variance to enhanced speech fluctuations.

Therefore, we propose to also use the variance of the logarithmic difference \( \text{LogErr}_{\text{var}} \) between the reference and estimated noise PSD over all frequency bins in order to evaluate the amount of fluctuations in the estimated noise PSD. Since it is not guaranteed that the logarithmic error \( \Delta_{dB}(k, i) \) is constant over frequency bins and time frames, we partition this quantity in the time-frequency plane and compute the variance \( \text{LogErr}_{\text{var}} \) as follows:

\[
\text{LogErr}_{\text{var}} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \text{LogErr}_{\text{var}}^{n,m}
\]

with

\[
\text{LogErr}_{\text{var}}^{n,m} = \frac{1}{I_{sub}K_{sub}} \sum_{i=(m-1)I_{sub}+1}^{mI_{sub}} \sum_{k=(n-1)K_{sub}+1}^{nK_{sub}} \left( \Delta_{dB}(k, i) - \mu_{n}^{i} \right)^{2}
\]

where

\[
\mu_{n}^{i} = \frac{1}{K_{sub}} \sum_{k=(n-1)K_{sub}+1}^{nK_{sub}} \Delta_{dB}(k, i).
\]

\( N \) and \( M \) are the numbers of blocks respectively over frequency bins and time frames and \( \text{LogErr}_{\text{var}}^{n,m} \) is the value of \( \text{LogErr}_{\text{var}} \) computed in \((n, m)\)-th block in the time-frequency plane. The number of frequency bins and time frames are indicated by \( K_{sub} \) and \( I_{sub} \). In (4), for simplicity, it is assumed that \( K \) and \( I \) are integer multiple of \( K_{sub} \) and \( I_{sub} \). To clarify the interpretation of \( \text{LogErr}_{\text{var}} \), it is noted that for example two noise estimators may have very close performance in terms of \( \text{LogErr}_{\text{mean}} \) but the one which gives the smaller variance \( \text{LogErr}_{\text{var}} \) would be more preferable due to lower tendency to produce musical noise.

To derive more detailed results on the capability of noise estimators for tracking the noise PSD, we evaluate their performance separately for voiced and unvoiced speech parts along with discarding the speech-absent parts. The voiced-unvoiced detection procedure is done on the clean speech signal, and then by detecting the time frame indices relevant to voiced and unvoiced parts, \( \text{LogErr}_{\text{mean}} \) is separately computed and denoted by \( \text{LogErr}_{\text{mean}}^{\text{voiced}} \) and \( \text{LogErr}_{\text{mean}}^{\text{unvoiced}} \), respectively. The voiced-unvoiced decision is achieved by considering the properties of voiced and unvoiced speech in cepstral domain as proposed in [13].

As we mentioned before, it is necessary to find an appropriate reference noise PSD. After evaluating several smoothing methods we decided to employ a recursive temporal smoothing of noise periodograms i.e.,

\[
\sigma_{\hat{D}}^{2}(k, i) = \alpha \sigma_{\hat{D}}^{2}(k, i-1) + (1-\alpha) |D(k, i)|^{2},
\]

where \( \alpha \) is the smoothing parameter and is set to 0.9 in the experiments. The smoothing parameter is chosen such that the smoothed noise power (reference noise PSD) can follow the power fluctuations of the selected non-stationary noise signals reasonably well.

4. EXPERIMENTAL FRAMEWORK

In this section the noise tracking performance of the mentioned algorithms in section 2 is evaluated. In total, 8 algorithms are considered. Their source code was either provided by the original authors, taken from publicly available sources, or was implemented according to their published description. In the case of MMSE-Yu algorithm we found it necessary to further optimize parameter settings in order to improve the performance for non-stationary noises. For this estimator we set parameters \( \alpha = 0.95, \beta = 0.15 \) and \( \psi = 0.216 \) which are explained in [12]. The sampling frequency of all signals used in this work is 8kHz. All methods were implemented in an DFT-based spectral analysis system using overlapping square-root periodic Hann windows. The window length as well as the DFT length is 256 samples, and the amount of overlap between the frames is considered separately based on recommendations reported by the authors of the algorithms. In fact it should be noted that the frame overlap has an important impact on the performance of noise tracking algorithms, and in our experiments we found it necessary to stick to the
recommended values. The frame overlap factor for MS, MMSE-Yu, MMSE-Hendriks, SNT algorithms is set to 50%, and for MCRA, IMCRA, EMCRA, MCRA-MAP algorithms to 75%. Having different frame overlap factors results in producing different numbers of frames in the frequency domain. Thus, to have the same number of frames for the evaluation of algorithms we sub-sample the reference and estimated noise frames for those algorithms which use more than 50% frame overlap. Clean speech signals are taken from the TIMIT database [14]. By concatenating sentences and removing the beginning and trailing silences, two sorts of clean speech signals are generated for our experiments; one female speech \( sp_{\text{female}} \) and one male speech \( sp_{\text{male}} \). Each clean speech signal has a length of 2 minutes and includes speech of 8 different speakers spoken in English. At the beginning of each clean speech data 0.1 seconds silence is included. In our experiments clean speech is degraded by 7 different types of noise signals taken from SOUND-IDEAS database [15]. We select babble noise (produced by a large crowd), car noise (inside a car during acceleration and deceleration), traffic-1 (heavy highway traffic without any horn) and traffic-2 (traffic in a city mainly involving horns). Moreover, to examine the performance of algorithms in the case of highly non-stationary noises, we consider two modulated versions of white Gaussian noise (WGN) named varying-step WGN and sinusoidal WGN. Varying-Step WGN is obtained by modulating WGN with a rectangular pulse signal which repetitively has an abrupt 10 dB change in the noise level for a few seconds and falls back to the previous level. Sinusoidal WGN is obtained through modulating WGN by the function

\[
g(n) = 1 + \sin \left( \frac{2\pi n}{f_s} \cdot f_{\text{mod}} \right)
\]

where \( n \) is the sample index, \( f_s \) the sampling frequency, and \( f_{\text{mod}} \) indicates the varying modulation frequency which linearly increases in 30 seconds from 0 Hz to 0.5 Hz. The power spectra of different noise signals (i.e. reference noise PSDs averaged over frequency bins) that we used for degrading clean speech signals to generate a noisy speech signal with 0 dB overall SNR are depicted in Fig. 1. In the experiments for each type of noises, we change the input

![Reference noise power](image1.png)

**Fig. 1.** Reference noise power used in degrading clean speech \( sp_{\text{female}} \) to produce noisy speech at overall SNR 0 dB; From top to bottom: WGN, Varying-Step WGN, Sinusoidal WGN, Babble noise, Car noise, Traffic1 noise, and Traffic2 noise.

![Experimental results for selected types of noises](image2.png)

**Fig. 2.** Experimental results for selected types of noises.
overall SNR from -5 dB to 20 dB in 5 dB steps. The first 3 seconds of the input signals are used for initialization of algorithms and so are excluded in the performance measurements. Moreover, in the computation of performance measures, the first 10 frequency bins as well as the 10 frequency bins below the Nyquist frequency are excluded in order to reduce the influence of DC-removal and anti-aliasing filter.

5. RESULTS AND DISCUSSION

By performing the experiments and having the results of algorithms for all types of noises, we could observe that although each of the performance measures \( \text{LogErr}_{\text{mean}} \), \( \text{LogErr}_{\text{voiced}} \) and \( \text{LogErr}_{\text{mean}} \) leads to different values, the rank order of algorithms based on these measures hardly differs. Therefore, because of space limitation we present results for the mean of both classes only. However, by comparing the values of \( \text{LogErr}_{\text{voiced}} \) and \( \text{LogErr}_{\text{mean}} \) we could see, for different types of noise, that most of algorithms have slightly better performance for unvoiced speech in comparison to voiced speech.

The results of our experiments in terms of the \( \text{LogErr}_{\text{mean}} \) and \( \text{LogErr}_{\text{var}} \) for some selected types of noises are depicted in Fig. 2. Our conclusions based on the depicted results and also the results for other types of noises are expressed in three parts. First of all, it should be mentioned that some of the noise estimators are very sensitive to the level of SNR. For instance IMCRA performs very well in low SNRs. However, one can observe that by increasing the SNR level \( \text{LogErr}_{\text{mean}} \) highly increases for some algorithms like IMCRA, MMSE-Yu, and ECRA while for some other noise estimators like MMSE-Hendriks and MS show a more robust performance. The other issue is the ability of algorithms in tracking noise with small amount of fluctuations. As we mentioned already, \( \text{LogErr}_{\text{var}} \) is used here to evaluate the noise tracking performance of algorithms in terms of their tendency to produce musical noise. For example, consider the case of Sinusoidal WGN for two algorithms IMCRA and MCRA-MAP for high SNR. In 20 dB SNR the performance of IMCRA in terms of \( \text{LogErr}_{\text{mean}} \) seems to be very close to MCRA-MAP or even slightly better. However, one can find that in terms of \( \text{LogErr}_{\text{var}} \) MCRA-MAP performs better. This effect is analyzed in Fig. 3 which compares the frequency-averaged noise PSDs of these two estimators as a function of time. In Fig. 3 it is clear that the estimated noise PSD provided by IMCRA frequently overestimates the reference noise power, i.e. it follows the speech power much more than the noise PSD estimated by MCRA-MAP.

For environments where noise PSD does not change rapidly in time (like WGN, car noise) most of algorithms perform similarly while for non-stationary noise a few methods show to be robust. The most robust noise estimator in our experiments using the two mentioned performance measures is MMSE-Hendriks. The SNT gives similar performance. The only exception is the case of traffic noise which mainly includes horn sounds (a deterministic noise type). Deterministic type noises can be described by low rank model and are confused with speech in the signal subspace. Therefore, here SNT does not follow the noise correctly.

6. REFERENCES