MINIMUM SUBSPACE NOISE TRACKING FOR NOISE POWER SPECTRAL DENSITY ESTIMATION

Mahdi Triki and Kees Janse
Digital Signal Processing Group, Philips Research Laboratories, Eindhoven, The Netherlands
Email: {mahdi.triki,kees.janse}@philips.com

ABSTRACT

Speech enhancement is the processing of speech signals in order to improve one or more perceptual aspects. If the statistics of the clean signal and the noise process are explicitly known, enhancement could be 'optimally' accomplished (minimizing a distortion measure between the clean and the estimated signals). In practice however, these statistics are not explicitly available, and the overall enhancement accuracy critically depends on the estimation quality of the unknown statistics. The estimation of noise (and speech) statistics is particularly a critical issue and a challenging problem under non-stationary noise conditions.
In this paper, we investigate the noise floor estimation using subspace decomposition. We examine the speech DFT rank limited assumption. We propose a new noise PSD estimation scheme (called Minimum Subspace Noise Tracking (MSNT)). The proposed scheme can be interpreted as a combination of the subspace structure and the minimum statistics tracking. Experimental investigation of the MSNT tracking performance and comparison with the state of the art is also presented.

Index Terms— single microphone speech enhancement; noise floor estimation; non-stationary noise; subspace methods

1. INTRODUCTION

Speech enhancement aims at improving the performance of audio communication in a noisy environment. Several practical methods have already been proposed. Among them, the group of frequency domain methods has been relatively successful due to their implementation simplicity and their capability of handling noise non-stationarity to some extent. These schemes recover the clean signal by applying a gain filter. The design of these filters relies on the knowledge of the clean and noise signal statistics. In practice however, these statistics are not explicitly available and should be estimated. The accuracy of the overall enhancement approach critically depends on the estimation quality of the unknown statistics. Particularly, an overestimation of the spectral noise variance leads to over-suppression and to more speech distortion; while an underestimation leads to a high level of residual noise.

Joint clean speech and noise Power Spectral Density (PSD) estimation is an underdetermined problem. In fact using a unique observation, we aim tracking both the clean speech and noise statistics. A classic trick to overcome the underdeterminacy problem is to exploit speech pauses. The key observation is that the speech signal is not present everywhere. Then, the noise PSD can be estimated and updated during speech absence. Typically, a voice activity detector (VAD) is used to identify speech pause periods. Unfortunately, these algorithms have some problems in low Signal-to-Noise Ratios (SNRs), especially when the noise is non-stationary [1]. Consistent accuracy can also not be achieved since VADs rely on a threshold level (difficult to set in an arbitrary environment).
A tractable alternative of VAD based schemes is provided by Minimum Statistics (MS) [3, 2]. The MS algorithm is based on the observation that even during speech activity the short term power spectrum of the noisy signal frequently decays to values which are representative of the noise power level. Then, by tracking the minimum of the smoothed noisy power spectrum within a finite window, an estimate of the noise floor can be obtained. The search memory (for local minima tracking) needs to be long enough to bridge any period of speech activity. It is assumed that a sliding window of $1.5 - 2$ seconds is large enough to bridge high power speech segments. This implies that a sudden increase of the noise PSD will be detected only with a 2-seconds delay. That means also that the tracking of impulsive noise type is not possible. These facts constitute the major drawbacks of the MS scheme.
Recently, a subspace decomposition based scheme was proposed for noise floor estimation [4, 5]. The subspace considered herein characterizes the time evolution of the noisy Discrete Fourier Transform (DFT) coefficients. The basic observation is that in such a domain the speech signal can be described with a low rank model, when the noise is full rank. Therefore, a noise subspace can be identified, and the noise PSD is still updated even when speech is constantly present. Simulations show that the Subspace Noise Tracking (SNT) approach achieves better tracking capability, but is still suffering from some problems in low SNR [4, 5].

In this paper, we propose a new noise PSD estimation scheme (called Minimum Subspace Noise Tracking (MSNT)) exploiting the limited-rank structure of the clean speech signal. The proposed scheme can be interpreted as a combination of subspace structure and minimum statistics tracking. Experimental comparison of MS, SNT, and MSNT performances is also investigated.

Notations: Upper- and lower-case boldface letters denote matrices and vectors, respectively. Upper- and lower-case normal letters represent scalar constants and processes, respectively. Either as a subscript or as an argument t, n and f refer respectively the time, frame, and frequency indexes.

2. SUBSPACE DECOMPOSITION FOR SPEECH SIGNAL IN DFT DOMAIN

Classic noise floor estimation (either based on VAD or MS) hinges on the assumption that a speech signal is not constantly present. The received signal in the pause frames is used to update the noise PSD estimate. Herein, we exploit further speech signal structures in order to get information on noise statistics even when speech is present. We focus on the time evolution of the speech *DFT coefficients*.

The sampled time domain signal is divided into overlapping blocks that are windowed by a smooth function, such as a Hanning window. Each windowed block is transformed into the frequency domain using a DFT. We use $s(n, f)$ and $v(n, f)$ to denote the

complex DFT coefficients of the clean speech and the noise signals, respectively. f represents the frequency index and n the time-frame index (figure 1).

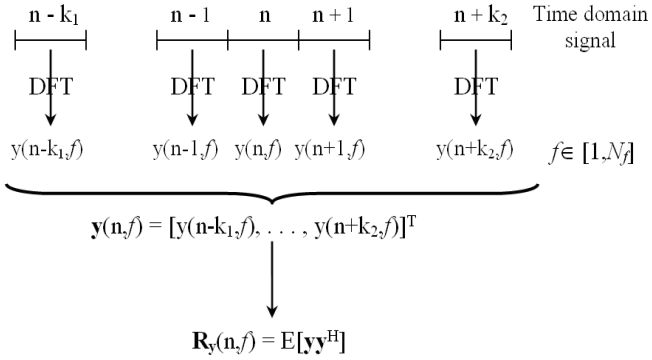


Fig. 1. Subspace decomposition in DFT domain.

We define correlation matrices in the DFT domain (for each DFT coefficient) as shown in figure 1: we collect DFT coefficients per frequency bin f that originate from the time frame $n-k_1$ up to frame $n+k_2$ and we form a vector $\mathbf{y}(n,f)$ of size $K = k_1 + k_2 + 1$. The noisy speech correlation matrix (at the frequency bin f and the time frame n) is:

$$\mathbf{R}_y(n,f) = E\left[\mathbf{y}(n,f)\mathbf{y}^H(n,f)\right].$$

Assuming an additive noise model, we split the noisy speech correlation matrix $\mathbf{R}_y(n,f)$ into:

$$\mathbf{R}_y(n,f) = \mathbf{R}_s(n,f) + \mathbf{R}_v(n,f). \quad (1)$$

We assume that $\mathbf{R}_v(n,f) = \sigma_v^2(n,f)\mathbf{I}_K$ ($\mathbf{I}_K$ is the $K \times K$ dimensional identity matrix). This assumption holds for noise with a small enough correlation time. The noise could be either white or colored. However, the previous assumption is only valid if the DFT coefficients (in $v(n,f)$) are computed from time domain frames that are not overlapping. In case of overlapping frames, this assumption will be violated. This noise coherence artifact could be alleviated by applying a pre-whitening transform (see [4] for details).

Contrary to the noise correlations, the matrix $\mathbf{R}_s(n,f)$ is assumed to be rank limited (further, we will investigate the validity of such assumption). We denote Q the rank of this matrix ($Q < K$). In such a case, the eigen decomposition of the received signal covariance matrix $\mathbf{R}_y$ can be expressed as:

$$\mathbf{R}_y = \mathbf{U}\left(\mathbf{\Lambda}_s + \sigma_v^2\mathbf{I}_K\right)\mathbf{U}^H \quad (2)$$

where $\mathbf{\Lambda}_s = \text{diag}\,(\lambda_{s,1}, \cdots, \lambda_{s,Q}, 0, \cdots, 0)$ is a diagonal matrix containing the eigenvalues of $\mathbf{R}_s$, and $\mathbf{U}$ is an unitary matrix containing the corresponding eigenvectors. All quantities in the previous equation depend on (n,f). As the treatment is identical for each DFT coefficient, the DFT index (n,f) is suppressed for better readability.

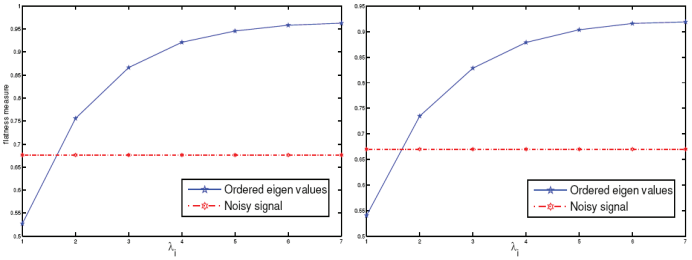
Next, we consider the experimental validation of the rank-limited assumption. In [4], the authors compute the signal subspace dimension Q that is needed to describe at least 95% of the energy of the speech signal. It turned out that, for frequency bins containing speech energy, the effectively needed dimension is $Q = 3.5$ on average ($K = 7$). Thus on average, the three lowest eigenvalues contain less than 5% of the speech energy. Then, we investigate the contribution of the clean speech signal on the i^{th} ordered eigenvalue. Due to the bursty nature and the non-whiteness of the speech signal, the relative contribution of the speech signal can be quantified using the flatness measure:

$$FM = \frac{\text{harmonic average}}{\text{arithmetic average}} = \frac{\left(\frac{1}{K}\sum_{k=1}^{K}\lambda_k^{-1}\right)^{-1}}{\frac{1}{K}\sum_{k=1}^{K}\lambda_k} \quad (3)$$

We use FM as a measure of:

- Non-whiteness: we consider the evolution of the ordered eigenvalues with the frequency index (we average over frames) (figure 2.a).
- Non-stationarity: we consider the evolution of the ordered eigenvalues with the time index (we average over frequencies) (figure 2.b).

As a reference, we plot the flatness measure of the noisy signal (contaminated with a stationary additive white noise) with respect to the frame (figure 2.a) and frequency (figure 2.b) indexes. The flatness measure of the noisy signal is plotted as a reference.



(a) averaged over frame index (b) averaged over frequency index

Fig. 2. Flatness of the ordered eigenvalues ($\lambda_1 \geq \cdots \geq \lambda_7$).

One can remark that the relative contribution of the speech signal decreases with increasing eigenvalue index. The noise PSD could be tracked based on the DFT eigenvalue information (the smaller the eigenvalue, the more consistent the information).

As stated in (2), the clean speech DFT coefficient lives in a low dimension subspace. The signal subspace can be computed using an eigen decomposition of the DFT covariance matrix; and the eigenvalues in the noise subspace can be exploited to update and track the noise PSD (even in presence of the speech signal). Hendriks et al. estimate the noise PSD by averaging the eigenvalues in the noise subspace, i.e,

$$\widehat{\sigma}_v^2 = \frac{1}{K-Q}\sum_{q=Q+1}^{K}\widehat{\lambda}_{y,q}, \quad (4)$$

where $\widehat{\lambda}_{y,Q+1} \cdots \widehat{\lambda}_{y,K}$ represent the $K-Q$ smallest eigenvalues of $\widehat{\mathbf{R}}_y(n,f)$ and Q is the assumed dimension of the signal subspace (model order). The performance of the proposed approach (called Subspace Noise Tracking (SNT) hinges on the estimation of the model order Q. The model order selection is particularly challenging since $\widehat{\mathbf{R}}_y$ is estimated using few data-samples. Typically, the model order and the data size have almost the same magnitude. In [4], the authors use a Bayesian approach to classify the noisy eigenvalues between the signal and noise subspaces (assuming slowly varying noise). The model order corresponds to the cardinality of the eigenvalues belonging to the signal subspace. The noise PSD is updated by averaging the eigenvalue classified to the noise subspace (as in (4)).

3. MINIMUM SUBSPACE NOISE TRACKING FOR NOISE PSD ESTIMATION

The SNT approach focuses on the noise PSD estimation under non-stationary conditions. The structure of the DFT coefficients time evolution is exploited to enhance noise tracking. The signal vs. noise subspace decomposition is first performed. The noise PSD is then updated using the projection of the noisy DFT coefficient onto the noise subspace.

As we have noticed in the previous section, a key parameter in the SNT approach is the signal subspace dimension (model order). The model order selection is a challenging problem (VAD can be interpreted as a simplified model order selection). The order selection is a difficult detection problem due to the:

- Bursty nature of the speech signal: Speech signals are highly non-stationary. In addition, the spectral characteristics (power, sparseness, flatness...) between voiced and unvoiced frames are quite different.
- Speech eigenvalue distribution: there is no clear distinction between the noise and the signal subspace. Despite the dependence on the speech signal decreases with the eigenvalue index, it cannot be neglected, especially at high SNR.

In the SNT scheme, the classification strategy leads to a systematic model order underestimation (especially in unvoiced frames and at low SNR). This fact leads to noise PSD overestimation. Simulations illustrate that such estimation error is difficult to alleviate as it is a function of the unknown speech signal. Subjective tests show that model order selection inaccuracy decreases the speech intelligibility.

In this paper, we propose exploiting the DFT evolution structure in a different way. We take a more 'pessimistic' attitude in the sense that:

- We consider only the minimum eigenvalue to update the noise PSD. The use of the minimum eigenvalue is motivated by three main issues. First, as shown in figure 2 the minimum eigenvalue is less depending on the speech signal and provides more consistent information. The second motivation is related to complexity. In fact, several approaches are introduced to efficiently estimate the minimum eigenvalue (with no need to perform a complete eigenvalue decomposition). Finally, if the noise varies slowly with time, adaptive schemes can be proposed to increase the computational efficiency of the minimum eigenvalue estimation (exploiting the fact that the current noise PSD estimation gives a good initialization).
- We assume that (for a given frequency and using a sufficiently long memory time) at least one DFT covariance matrix is rank limited. Exploiting the observation that the minimum power level reaches the noise level, we propose a 'Minimum Statistics'-like approach to update the noise PSD. Notice that the rank-limited assumption is not assumed for each time-frequency bin and model order selection is no-longer needed.

Compared to the MS approach, the Minimum Subspace Noise Tracking (MSNT) exploits further the structure of the speech signal time variation (not only in terms of presence or absence). In the previous section, we have shown that often the DFT covariance matrix is rank limited. In these frames, noise PSD information is available (even in presence of a speech signal). Thus, the MSNT does not need a large memory to perform an acceptable steady state performance. Hence, it provides better noise tracking.

The MSNT leads also to a biased noise PSD estimate (like MS and SNT). The bias is mainly due to the minimum statistic based tracking, i.e.,

$$E\left\{\min\left(.\right)\right\} \leq \min\left(E\left\{.\right\}\right).$$

The bias depends mostly on the search memory M, and may be corrected using a multiplicative bias compensation factor. The compensation factor may be trained over a speech data degraded with white noise with a known variance $\sigma_v^2 = 1$. Simulations show that, compared to SNT [4], the MSNT bias is less dependent on the unknown speech signal. Therefore, it is more robust to the input SNR and to the speaker characteristics.
The MSNT scheme is summarized in the table below

Minimum Subspace Noise Tracking Algorithm.	
#	**Computation**
Initialization	
1	Tracking bias $B\left(M\right)$ training.
2	Pre-whitening transform $\widehat{\mathbf{R}}_{pre}$ computation.
Iteration	
	for $f = 1 : N_f$ do
1	DFT covariance matrix $\widehat{\mathbf{R}}_y(n,f) = \frac{1}{n_2+n_1+1}\sum_{i=n-n1}^{n+n2}\mathbf{y}(i,f)\mathbf{y}^H(i,f)$
2	Pre-whitening transform $\widehat{\mathbf{R}}_{y,pre}(n,f) = \mathbf{R}_{pre}^{-\frac{1}{2}}(f)\widehat{\mathbf{R}}_Y(n,f)\mathbf{R}_{pre}^{-\frac{1}{2}}(f)$ $\widehat{\mathbf{R}}_{y,pre}(n,f) = \frac{\mathrm{tr}\left[\mathbf{R}_{pre}(f)\right]}{K}\widehat{\mathbf{R}}_{y,pre}(n,f)$
3	Minimum eigenvalue estimation $\hat{\lambda}_{min}(n,f) = \min\left\{\hat{\lambda}_{y,1}\cdots\hat{\lambda}_{y,K}\right\}$
4	Noise PSD tracking $\widehat{\sigma}_v^2(n,f) = \frac{1}{B(M)}\min\left\{\hat{\lambda}_{min}(n-i,f)\right\}_{i=0:M-1}$
	end for

Table 1. MSNT scheme for non-stationary noise tracking.

4. EXPERIMENTAL RESULTS

In this section, we investigate the tracking and the misadjustment accuracy of Subspace Noise Tracking (SNT) and Minimum Subspace Noise Tracking (MSNT), respectively. We also compare the subspace based schemes to the Minimum Statistics (MS).

For better understanding of the advantage and drawbacks of the three tracking schemes (MS, SNT, MSNT), we consider a simple white Gaussian noise scenario. The noise non-stationarity is introduced by two abrupt changes in the noise level. Indeed, the input SNR varies suddenly from 0 dB to -10 dB, then back to 0 dB. The time-domain signal (sampled at 8 kHz) is divided into overlapping blocks. These blocks are segmented using Hanning smoothing windows (length = 32 ms, overlap = 87.5%). A smoothing factor $\alpha = 0.8$ is applied to the noise PSD estimate. Figure 3 illustrates the MS noise floor estimation accuracy (with appropriate bias compensation).

As expected, minimum statistics leads to good final misadjustment accuracy at the expense of a large estimation delay. This delay makes the MS unsuited for fast varying noise and leads to an annoying impulsive remaining noise.

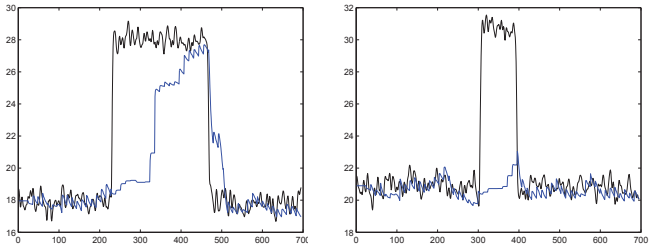


Fig. 3. Perfect(black) vs. MS(blue) white noise tracking.

Next, we add the curve of the noise PSD tracking using the SNT scheme (Fig.4).

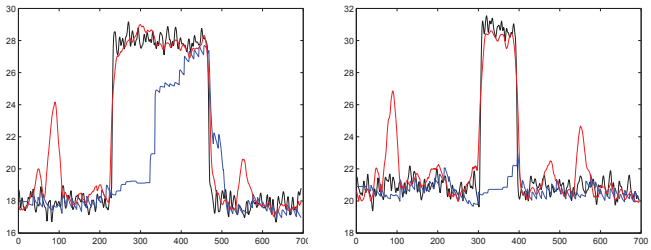


Fig. 4. Perfect(black), MS(blue) and SNT(red) white noise tracking.

We observe that the SNT performs a good tracking (comparing to MS). Indeed, exploiting the speech rank-limited property, SNT updates the noise PSD even in presence of the speech signal. On the other hand, we remark that SNT locally overestimates noise level. The origin of this artifact is the systematic model order selection error (especially in unvoiced frames and at low SNR). In such a case, some of the speech energy is considered as a part of the noise. These errors are difficult to predict and alleviate (as they are function of the unknown speech signal and the input SNR). Subjective tests show that such artifact reduces considerably the speech intelligibility (but with no significant consequence on the speech quality).
In Fig. 5 and Fig. 6, we add the curves of the noise PSD tracking using the MSNT scheme using respectively a search memory $M = 30$ and $M = 60$.

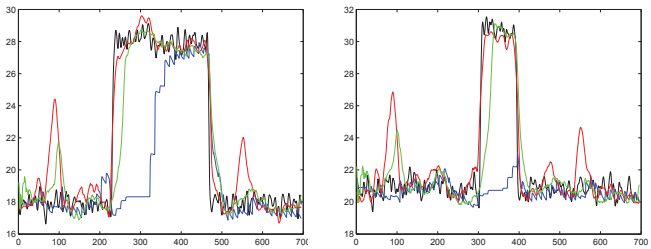


Fig. 5. Perfect(black), MS(blue), SNT(red) and MSNT(green) white noise tracking (M=30).

We notice that, depending on the search memory, the MSNT leads to a different tradeoff between the estimation delay and the noise overestimation. However comparing to MS, MSNT uses lower estimation delay to perform an equivalent final misadjustment. Intuitively, due to the speech rank-limited nature, we observe more often frames

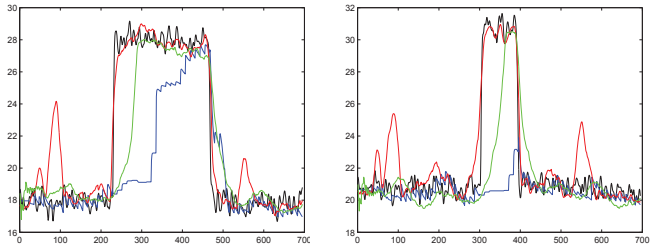


Fig. 6. Perfect(black), MS(blue), SNT(red) and MSNT(green) white noise tracking (M=60).

containing no-speech in *some* directions than frames containing no-speech in *all* directions.

Similar conclusions can be drawn while using different non-stationary noise sources (originated respectively from the NOIZEUS [6] database, a passing car and a passing train).
Subjective tests reveal that the MS performs better in terms of intelligibility, while SNT leads to a better comfort quality. MSNT leads to a compromise between the two perceptual criteria (depending on the search memory).

5. CONCLUDING RESULTS

We have introduced a new noise floor estimation scheme (called Minimum Subspace Noise Tracking (MSNT)). The proposed scheme exploits the speech subspace structure without an explicit model order selection (the update is performed via a local search approach). Comparing to SNT, the proposed approach seems to be advantageous in terms of consistency, complexity and adaptivity. Simulations show that MS leads to good final misadjustment accuracy at the expense of a large estimation delay; while the SNT performs good tracking accuracy except for occasional noise floor overestimation. Such artifact considerably reduces the speech intelligibility. The MSNT performs an intermediate tracking vs. final misadjustment (quality vs. intelligibility) tradeoff, and generally leads to an increase in non-stationary noise floor estimation accuracy (compared to both MS and SNT).

6. REFERENCES

[1] S.G. Tanyer, H. Ozer,"Voice Activity Detection in Nonstationary Noise," IEEE Trans. on Speech and Audio Processing, Jul. 2000.

[2] R. Martin,"Spectral Subtraction Based on Minimum Stastistics," In Proc. of Eur. Signal Processing Conf., Sep. 1994.

[3] R. Martin,"Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," IEEE Trans. on Speech and Audio Processing, Jul. 2001.

[4] R. C. Hendriks, J. Jensen and R. Heusdens, "Noise Tracking using DFT Domain Subspace Decompositions," IEEE Trans. on Audio, Speech, and Language Processing, Mar. 2008.

[5] R. C. Hendriks, J. Jensen and R. Heusdens, "DFT Domain Subspace Based Noise Tracking for Speech Enhancement," In Proc. of Interspeech, Aug. 2007.

[6] Y. Hu and P. Loizou,"Subjective comparison of speech enhancement algorithms," In Proc of IEEE Int. Conf Conf. on Acoustics, Speech, and Signal Processing, May 2006. http://www.utdallas.edu/ loizou/speech/noizeus/