New Insights into Subspace Noise Tracking

Mahdi Triki
Philips Research Laboratories, Eindhoven, The Netherlands
mahdi.triki@philips.com

Abstract

Various speech enhancement techniques (e.g. noise suppression, dereverberation) rely on the knowledge of the statistics of the clean signal and the noise process. 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 respect, subspace-based approaches have been shown to provide a good tracking vs. final misadjustment tradeoff. Subspace-based techniques hinge critically on both rank-limited and spherical assumptions of the speech and the noise DFT matrices, respectively. In [3, 4], the speech rank-limited assumption was experimentally tested and validated. In this paper, we will investigate the structure of nuisance sources. We will discuss the validity of the spherical assumption for a variety of nuisance sources (environmental noise, reverberation), and preprocessing (overlapping segmentation).

Index Terms: noise floor estimation; non-stationary noise; subspace methods; noise suppression; speech dereverberation.

1. Introduction

Speech enhancement aims at improving the performance of audio communication in noisy environments. 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 over-estimation of the spectral noise variance leads to over-suppression and to more speech distortion; while an under-estimation 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 at 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 the speech pauses. The noise is typically updated either using a Voice Activity Detector (VAD) to identify speech pauses and the minimum statistics (MS) of the noisy signal [2].

Recently, a subspace decomposition based scheme was proposed for noise floor estimation [3]. 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 still suffers from some problems at low SNRs [3]. In [4], 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). Compared to SNT, the proposed approach is 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 over-estimation. Such artifacts considerably reduce the speech intelligibility. The MSNT provides 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).

Subspace-based techniques hinge critically on both rank-limited and spherity assumptions of the speech and the noise DFT matrices, respectively. In [4], the speech rank-limited assumption was experimentally tested and validated. Similar results were reported in [3]. In this paper, we will investigate the noise structure, and we will discuss the validity of the spherical assumption for a variety of nuisance sources (environmental noise, reverberation, etc.) and preprocessing (overlapping segmentation).

The remainder of this paper is organized as follows. After a brief introduction to subspace-based noise tracking techniques in Section 2, the spherical noise structure will be studied and commented on in Section 3. Robustness to room reverberation will be investigated in Section 4, and finally, a discussion and concluding remarks are made in Section 5.

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 to the time, frame, and frequency indices.

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 overlap-
ping 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 (Fig. 1).

\[
\begin{align*}
\text{Time domain signal} & \quad \text{DFT} \quad \text{DFT} \quad \text{DFT} \quad \text{DFT} \\
\text{y}(n-k_2,f) & \quad y(n-1,f) \quad y(n,f) \quad y(n+1,f) \quad y(n+k_2,f) \\
\end{align*}
\]

\[
y(n,f) = [y(n-k_2,f), \ldots, y(n+k_2,f)]^T
\]

\[
R_y(n,f) = E[yy^H]
\]

Figure 1: Subspace decomposition in the DFT domain.

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

\[
R_y(n,f) = E[y(n,f)y^H(n,f)].
\]

Assuming an additive noise model, we split the noisy speech correlation matrix \( R_y(n,f) \) into:

\[
R_y(n,f) = R_s(n,f) + R_v(n,f). \quad (1)
\]

We assume that the matrix \( R_s(n,f) \) is rank limited. This assumption was experimentally tested and validated in [4, 3]. A relation can also be established between the rank-limited assumption and standard audio signal representations such as damped sinusoidal [5] and amplitude modulation [6] models. If we assume the noise DFT correlation matrix to be spherical, i.e., \( R_v(n,f) = \sigma_v^2 I_K \) (where \( K \) is the \( K \) dimensional identity matrix), the eigendecomposition of the received signal covariance matrix \( R_y \) can be expressed as:

\[
R_y = U \Lambda_s + \sigma_v^2 I_K U^H \quad (2)
\]

where \( \Lambda_s = \text{diag} \{ \lambda_s, 1, \ldots, \lambda_s, 0, \ldots, 0 \} \) is a diagonal matrix containing the eigenvalues of \( R_s \), and \( U \) is an unitary matrix containing the corresponding eigenvectors. \( Q \) denote the rank of the speech correlation matrix \( (Q < K \) due to the rank-limited assumption). Therefore, the noise PSD can still be updated even when speech is constantly present. 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 omitted for better readability.

A critical assumption of the subspace-based approaches is that \( R_v = \sigma_v^2 I_K \). This assumption holds as long as the correlation time of the noise is small enough [8] and the DFT coefficients are computed from non-overlapping frames. Simulation results have shown that a variety of noise signals (e.g. train, car, babble, etc) satisfy the previous assumption [3]. However, this structure does not hold for all type of signals that may be considered as undesirable. For instance, an interfering speaker (speech signal) or fan noise (periodic signal) are rank-limited and do not meet the spherical assumption. The remainder of this paper investigates the noise structural assumption. Particular attention will be paid to the effect of overlapping segmentation and late reverberation.

3. Noise Structure under Overlapping Processing

Let \( v(t) \) and \( w(t) \) denote the noise and the segmentation window in the time-domain, respectively. \( L \) and \( N_f \) represent respectively the frame shift and length. The \((p,q)\) entry of the noise DFT coefficient covariance matrix at the frequency bin \( f \) has to be estimated from a limited number of samples:

\[
\hat{R}_v(p,q,f) = \frac{1}{n_1 + n_2 + 1} \sum_{i=n-n_1}^{n+n_2} v(i+p,f)w^H(i+q,f). \quad (3)
\]

The choice of the parameters \((n_1, n_2)\) is subject to a trade-off between the tracking delay and the estimation accuracy, and should account for the non-stationarity of speech and noise signals. On the other hand, the noise DFT coefficient at time-frequency bin \((i,f)\) is expressed as

\[
v(i,f) = \sum_{t} w(t-iL)v(t) e^{-j2\pi(t-iL)f}. \quad (4)
\]

Using a variable substitution, one can show that

\[
\hat{R}_v(p,q,f) = \sum_{t_1,t_2} w(t_1-pL)w(t_2-qL)e^{-j2\pi(t_1-t_2+(p-q)L)f} \times \frac{1}{n_1 + n_2 + 1} \sum_{i=n-n_1}^{n+n_2} v(i+p,f)v(t_2+iL). \quad (5)
\]

If we assume that

\[
\frac{1}{n_1 + n_2 + 1} \sum_{i=n-n_1}^{n+n_2} v(t_1+iL)v(t_2+iL) \approx \sigma_s^2(n) \delta_{t_1,t_2}, \quad (6)
\]

where \( \delta_{t_1,t_2} \) denotes the Kronecker delta operator, (5) can be simplified as

\[
\hat{R}_v(p,q,f) \approx \sigma_s^2(n)e^{-j2\pi(p-q)Lf} \sum_{t} w(t-pL)w(t-qL). \quad (7)
\]

Hence, the side effect due to the frames overlapping can be alleviated by applying a pre-whitening transform. As the pre-whitening is defined up to scaling factor, it depends exclusively on the segmentation process (window shape, overlap) and not on the instantaneous noise characteristics. This can be performed as:

\[
\tilde{R}_y(p,q,f) = R_{prv}(p,q,f) \hat{R}_v(p,q,f) R_{prv}^H(f) \quad (8)
\]

where the \((p,q)\) entry of the matrix \( R_{prv}(f) \) are given by

\[
R_{prv}(p,q,f) = e^{-j2\pi(p-q)Lf} \sum_{t} w(t-pL)w(t-qL) \quad (9)
\]

Notice, that now \( \sigma_s^2(n,f) \) gets estimated in a transformed (whitened) domain. To recover an unbiased estimate, the noise PSD should be corrected using the scaling factor \( \frac{1}{K} \). Note also that the pre-whitening transform preserves the rank-limited property of the clean speech signal.
We investigate the effect of the pre-whitening transform on various stationary and non-stationary noise sources. Synthetic stationary and non-stationary white Gaussian noises have been generated, and real-world noise sources (originating from passing car, passing train, and street noise [7]) have been recorded and sampled at 8 kHz. The noise signals were segmented using a square root Hanning window (length=256, overlap=87.5 %). The dimension of the DFT covariance matrix was set to \( K = 7 \) (for the estimation of the DFT correlation matrix, we set \( n_1 = n_2 \)). Ideally the output of noise pre-whitening transform should be spherical (multiple of identity). The Flatness Measure (FM) was used to validate the spherical requirement:

\[
FM = \frac{\text{harmonic average}}{\text{arithmetic average}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{\lambda_k} \tag{10}
\]

The flatness measure was applied to the eigenvalues of the transformed noise DFT covariance matrix \( \mathbf{R}_{\text{pre}}(n, f) \), then averaged over time and frequency indices. The table 1 summarizes the FM for the various noise sources we have considered and various parameters \( n_1 \). As a reference, we have added the FM of a clean speech signal.

<table>
<thead>
<tr>
<th>noise source</th>
<th>( n_1 = 10 )</th>
<th>( n_1 = 20 )</th>
<th>( n_1 = 50 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>stat. white</td>
<td>0.64</td>
<td>0.84</td>
<td>0.94</td>
</tr>
<tr>
<td>non-stat. white</td>
<td>0.63</td>
<td>0.81</td>
<td>0.91</td>
</tr>
<tr>
<td>train</td>
<td>0.6</td>
<td>0.79</td>
<td>0.88</td>
</tr>
<tr>
<td>street</td>
<td>0.6</td>
<td>0.77</td>
<td>0.86</td>
</tr>
<tr>
<td>car</td>
<td>0.57</td>
<td>0.75</td>
<td>0.84</td>
</tr>
<tr>
<td>speech</td>
<td>0.05</td>
<td>0.08</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 1: Flatness measure of pre-whitened noise sources. One may observe that the spherical assumption is valid for a number of non-white noise sources (train, car, street). Indeed, for non-overlapping segmentation, the spherical assumption implies that (i) the noise realization in two different frames are not correlated (ii) and the noise energy is slowly varying over time. These requirements are met by white processes, but not only. For instance, filtering a white process by a short FIR filter still closely verifies the above specifications. In general, noise sources with sufficiently low correlation time [8] satisfy these assumptions.

More challenging, however, is the estimation of the DFT covariance matrix: the parameter \( n_1 \) sets a tradeoff between the validity of the spherical noise and the rank-limited speech assumptions. The larger \( n_1 \), the better the noise covariance is estimated and the closer to identity. But also, as more speech is averaged over time frames, the less often it resides in a low-dimension subspace.

### 4. Speech Structure under Room Reverberation

The quality of audio signals captured in a real-world environment is invariably degraded by acoustical interference. This interference can be broadly classified in two distinct categories: additive (originating from surrounding sounds) and convolutive (due to wave reflections from surrounding walls and objects). If convolutive interference (commonly referred to as reverberation) is present, no-speech free frames are present and the classical noise PSD tracking techniques (based on VAD and/or MS) provide biased estimations. On the other hand, dereverberation techniques often assume that the received signal is either noise-free or that noise statistics are available [9, 10]. In such a case, decoupling the two problems is not possible as the performance of each block depends critically on the accuracy of the other. Moreover, it has been shown that the speech PSD can be segregated in the DFT domain, and noise statistics can be updated even in the presence of (clean) speech. With an eye on noise PSD tracking under reverberant environment, we investigate, in this section, the (rank-limited) structure of the speech signal corrupted by convolutive interference.

Reverberation is the results of a multiple number of echoes of which the repetition speed is too high to be perceived as separate from one another. Generally, a reverberant signal \( x(t) \) is expressed as a convolution of a clean speech \( s(t) \) and an impulse response \( h(t) \) characterizing the room reflections, i.e.,

\[
x(t) = \sum_{\tau} h(\tau)s(t-\tau) \tag{11}
\]

In the following analysis, we model the room impulse response as one realization of the non-stationary stochastic process [11]

\[
h(t) = b(t)e^{-t/\delta} \tag{12}
\]

where \( b(t) \) is a centered stationary white Gaussian noise, and \( \delta \) is a constant characterizing the sound energy decay in the room. The \((p, q)\) entry of the reverberant signal DFT coefficient covariance matrix at the frequency bin \( f \) is expressed as in (5), where the reverberant signal \( x(t) \) replaces the noise process \( v(t) \). Under the previous notations and assumptions, the second term in (5) gets expressed as:

\[
\frac{1}{n_1+n_2+1} \sum_{k=n-n_1}^{n+n_2} x(t_1+iL)x(t_2+iL) = \sum_{iL}^{f} h(\tau_1)h(\tau_2)\sigma_{s,n}^2(\tau_1-\tau_2) = \sum_{\tau_2}^{f} \sigma_{s,n}^2(\tau_1)h(\tau_1+\tau_2) \leq \sigma_{s,n}^2(\tau_1-t_1)
\]

where \( \sigma_{s,n}^2(\tau_1-t_1) = \frac{1}{n_1+n_2+1} \sum_{k=n-n_1}^{n+n_2} s(t_1+iL)s(t_2+iL) \) is assumed to be locally stationary. The second equality is derived using a variable substitution, and the third proportionality exploits the whiteness of the channel impulse response \( h(t) \). Indeed,

\[
\sum_{\tau_1} h(\tau_1)h(\tau_1+\tau_2) = e^{-\tau_2/\delta} \sum_{\tau_1} e^{-\tau_1/2\delta} h(\tau_1)h(\tau_1+\tau_2).
\]

Because \( \sum_{\tau_1} e^{-\tau_1/2\delta} \) behaves as an exponentially weighted averaging operator, the whiteness of \( b(t) \) causes the whiteness of \( h(t) \). One may note that only the autocorrelations of the channel impulse responses matter: both impulsive response (clean speech) and white decaying response (as modeled in (12)) generate equivalent DFT coefficient covariance matrices, which have been shown to be rank-limited [3, 4]. Moreover, early reverberation could be modeled as a short FIR filter. It only alters the input color (a constant frequency-dependent scaling). As the processing is performed independently for each frequency bin, color modification has no effect on the speech/noise subspaces. In summary, we have shown that both early and late reverberations do not alter the rank-limited structure of the speech signal.
To validate the previous analysis, a set of room impulse responses were synthetically generated using the image method [12]. A clean speech signal was convolved with the generated impulse responses (having various reverberation times). The signals were segmented using a rectangular window (length=32, no-overlap). The dimension of the DFT covariance matrix was set to $K = 7$ (for the estimation of the DFT correlation matrix, we set $n_1 = n_2 = 10$). Fig. 2 shows the flatness measure of the reverberant signal as function of the reverberation time $T_{60}$. The flatness measure of a white noise signal is plotted as a reference.

![Figure 2: Flatness measure of the reverberant speech as a function of the reverberation time $T_{60}$.](image)

The curves in Fig. 2 confirm that the speech signal is still rank-limited (in the DFT domain) even if corrupted by reverberation. Subspace methods seem then transparent to convolutive interference and are not biased by the presence of echo during pause frames (contrary to VAD and MS). In such, subspace methods allow better decoupling of speech enhancement and deconvolution problems, and present a good potential for blind equalization and dereverberation problems. Indeed, the problem of blind channel identification and equalization has motivated intensive interest in communication and signal processing society. Most of the earlier approaches to blind identification are based on the use of higher order statistics, which are known to suffer from many drawbacks. They usually require a large number of data samples and a heavy computational load, making them unattractive for practical applications. More recently, it has been shown that the second order statistics contain sufficient information for the identification and equalization of FIR channels. Second-order equalization techniques have been successfully introduced, investigated, and applied to communication [13] and speech dereverberation [6] applications. These techniques, however, depend critically on the estimation of the second-order statistics of the reverberant signal. If additive noise is present, noise correlations should be identified and compensated for. On the other hand, the estimation of the noise PSD is more challenging under reverberant conditions. State-of-the-art techniques (e.g. VAD and MS) are generally biased by reverberation and overestimate the noise statistics. The noise overcompensation leads to non-perfect equalization (in addition to a reduced algorithmic stability). Contrarily, we have shown that speech signal is still residing in a rank-limited subspace, even when corrupted by reverberation. The noise statistics tracked and updated as shown in Section 2 are not biased by reverberation, and leads then to better equalization performance.

5. Concluding Remarks

Subspace based techniques hinge critically on both rank-limited and spherical assumptions of speech and the noise DFT matrix, respectively. In [3, 4], the speech rank-limited assumption was experimentally tested and validated. In this paper, we will investigate the structure of nuisance sources. We have observed that the spherical assumption is valid for a number of non-white noise sources (train, car, street), and that correlations due to overlapping segmentation could be alleviated using simple pre-whitening transform. This transformation depends exclusively on the segmentation process (window shape, overlap) and not on the signal characteristics. Moreover, we have shown that speech signal is still rank-limited (in the DFT domain) even when corrupted by reverberation. Subspace methods are not biased by the presence of echo during pause frames (contrary to VAD and MS), and seem appropriate for blind channel equalization and dereverberation applications.

6. References