An Approximate Solution for Perceptually Constrained Signal Subspace Speech Enhancement Method

Adam Borowicz, Alexander Petrovsky

Department of Real-Time Systems, Białystok Technical University, Białystok, Poland
borowicz@wi.pb.edu.pl, palex@it.org.by

Abstract

In this paper we present a low-complexity version of perceptually constrained signal subspace method (PCSS) for speech enhancement. An approximate solution is presented in a new form which provides perceptually optimal residual noise shaping. The proposed approach does not require a whitening transformation and is sub-optimal for coloured noise. A comparative evaluation of selected methods is performed using objective speech quality measures and informal listening tests. The results show that the approximate method outperforms conventional one and gives comparable results as the exact solution in common situations.

Index Terms: speech enhancement, KLT, psychoacoustics

1. Introduction

A noise reduction problem arises in a wide range of speech processing applications including mobile radio devices, speech recognition systems and speech coding. Often one-channel application is assumed, which makes the design much more difficult. Nevertheless, there is a need for efficient noise reduction algorithms operating at extremely low signal to noise ratio (SNR). Most of existing one-channel speech enhancement systems work in the frequency domain using spectral weighting technique. Unfortunately, these methods suffer from self-generated distortions known as "musical tones". Many methods have been proposed to eliminate these phenomena including perceptually motivated approaches [1], [2].

A signal subspace approach for speech enhancement is an interesting generalisation of conventional frequency domain methods. This technique has been originally proposed in [3]. The approach decomposes a noisy signal space into a signal subspace and a noise subspace using Karhunen-Loeve transform (KLT). Speech enhancement is performed in the signal subspace only. The components projected onto the noise subspace are excluded which results in significantly better performance when compared to frequency domain methods. However, further improvements are also possible. Recently several attempts have been made to incorporate masking properties of the human ear in the signal subspace approach [4] [5]. These methods use frequency-to-eigendomain transformation which provides a way to compute masking energies in the KLT-domain. Then a conventional weighting rule is modified according to hearing properties. On the other hand it is possible to exploit masking phenomena more explicitly constraining a residual noise energy strictly in the frequency domain [6], but such a approach is computationally expensive, especially for coloured noise. In this paper we present an approximate solution, which does not require a whitening step. The proposed method exploits perceptually motivated residual noise shaping and imposes the constraints strictly in the frequency domain using discrete Fourier transform (DFT) basis vectors. Finally we present a comparative evaluation of selected subspace-based methods using objective measurement and informal listening tests.

2. Perceptually constrained signal subspace method

Let \( x = y + n \) denotes \( k \)-dimensional noisy speech vector, where \( y \) and \( n \) are zero-mean random vectors representing clean speech and noise signal respectively. Since the speech and noise are assumed to be uncorrelated, the covariance matrix of the noisy speech process can be written as

\[
R_x = R_y + R_n,
\]

where \( R_n \) and \( R_y \) are the covariance matrices of the noise and clean speech process, respectively. It is also assumed that the matrix \( R_n \) is positive definite. Let \( \hat{y} = Hx \) be a linear estimator of the clean speech. The effective filter \( H \) is derived as a solution of constrained optimization problem by minimizing an average speech distortion power and constraining a residual noise energy. The estimation error vector is defined as

\[
\epsilon = y - \hat{y} = (H - I)y + Hn = \epsilon_y + \epsilon_n,
\]

where \( \epsilon_y \) and \( \epsilon_n \) are the speech distortion and residual noise vector, respectively. The average speech distortion power is given by

\[
\bar{\epsilon}_y^2 = \frac{1}{k} \text{tr} E \{ \epsilon_y \epsilon_y^H \} = \frac{1}{k} \text{tr} \{ (H-I) R_y (H-I)^H \},
\]

where \( E\{\} \) is an expectation operator, \( \text{tr}\{\} \) is a matrix trace and the superscript \# denote the transpose of a real matrix or the conjugate transpose of a complex matrix. The residual noise energy can be defined in spectral domain giving a spectral-domain-constrained (SDC) estimator. The SDC estimator allows defining perceptually motivated constraints strictly in the frequency domain. Namely, the optimisation problem can be formulated as follows

\[
\min_H \bar{\epsilon}_y^2
\]

subject to:

\[
E \{ |v_i^# |^2 \epsilon_n^2 \} \leq \alpha_i, \quad i = 1, \ldots, p.
\]

where \( p \geq k \) and \( \{ v_i, i = 1, \ldots, p \} \) is a set of \( k \)-dimensional sinusoidal vectors defined by

\[
v_i^# = k^{-1/2} \left[ e^{-j\omega_i 0}, e^{-j\omega_i 1}, \ldots, e^{-j\omega_i (k-1)} \right],
\]
where
\[ \omega_i = 2\pi (i - 1)/p, \quad i = 1, 2, \ldots, p. \]

(6)

It is easy to see that for \( p = k \), the column vectors \( \mathbf{v}_i \) can be arranged in a normalised inverse DFT matrix, i.e. \( \mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k] \). In fact, \( \mathbf{V} \) may be extended to any \( k \)-by-\( p \) real or complex matrix of rank \( k \), so that it is possible to use non-orthogonal or overcomplete vector bases. For instance, warped discrete Fourier transform (WDFT) matrix [7] can be used, which allows defining constraints directly in Bark or ERB domain. However, such an approach is computationally expensive and its implementation is a challenge.

Here, we consider the DFT-based constraints as a simplified solution. As was mentioned, in this case the masking properties of the human ear can be used more directly. Namely, the residual noise levels \( \alpha_i \) can be set according to well known psychoacoustically motivated rule [1], [2]

\[ \alpha_i = \min (\phi_i(\omega_i), \alpha_{i,\text{max}}), \quad i = 1, \ldots, k, \]

(7)

where \( \phi_i(\omega_i) \) denote the masking threshold of the clean speech and \( \alpha_{i,\text{max}} \) is the maximum possible residual noise level. The proposed approach is motivated by the following observation: If any frequency component of the residual noise is greater than the masking threshold, it is audible and the clean speech is deteriorated by the noise, in the opposite situation (when the frequency component is below the threshold) we have an unnecessary attenuation of the clean speech. Thus, ideally, these components should be placed just below the masking threshold of the clean speech signal to make the noise inaudible and avoid unnecessary attenuation.

The solution of (4) is found using Lagrange multipliers method. Namely, an optimal linear filter \( \mathbf{H} \) is defined as

\[ \mathbf{H} = \mathbf{M}^\# \mathbf{R}_n + \mathbf{H}_y = \mathbf{R}_y, \]

(8)

where \( \mathbf{M} = k \text{ diag} \{\mu_1, \mu_2, \ldots, \mu_k\} \) is a diagonal matrix constructed from Lagrange multipliers. Let \( \mathbf{U}_y \mathbf{\Lambda}_y \mathbf{U}_y^\# \) be eigen decomposition of the matrix \( \mathbf{R}_y \). Note that if the noise is white and both matrices \( \mathbf{R}_y \) and \( \mathbf{R}_n \) can be diagonalised jointly using \( \mathbf{U}_y \). In the case of coloured noise, the signal must be prewhitened first, which is computationally expensive [8], [6]. Therefore, we use the following approximation [9]:

\[ \mathbf{R}_n \approx \mathbf{U}_y \mathbf{\Lambda}_n \mathbf{U}_y^\#, \]

(9)

where \( \mathbf{\Lambda}_n \) is a diagonal matrix with entries defined as follows

\[ \lambda_{n,i} = \mathbf{u}_y^\# \mathbf{R}_n \mathbf{u}_{y,i}, \quad i = 1, \ldots, k. \]

(10)

Assuming that \( p = k \) and substituting (9) to (8) we obtain a suboptimal filter

\[ \mathbf{H} = \sum_{i=1}^{k} \mathbf{V} \lambda_{y,i} (\lambda_{y,i} I + \mathbf{M} \hat{\mathbf{\Lambda}}_{n,i})^{-1} \mathbf{V}^\# \mathbf{u}_{y,i} \mathbf{u}_{y,i}^\#, \]

(11)

where the parameter

\[ r = \arg \max_{1 \leq i \leq k} \{ \lambda_{y,i} > 0 \}, \]

(12)

is interpreted as a signal subspace dimension (we assume that the eigenvalues are ordered as \( \lambda_{y,1} \geq \lambda_{y,2} \geq \cdots \geq \lambda_{y,k} \)). It is clearly visible that a noisy signal is decomposed using \( r \) orthogonal projectors \( \mathbf{u}_{y,i} \mathbf{u}_{y,i}^\# \) and then processed using frequency domain filters. The result is summed up, giving the clean speech estimate. Note that the noisy speech components projected onto the noise subspace are null which is consistent with a signal subspace interpretation.

Figure 1: Block diagram of the proposed method.

3. Implementation

A simplified processing scheme is depicted in Fig. 1. In the proposed method we use a batch-mode algorithm which operates on relatively long signal frames. Namely, we divide the signal \( x \) into frames of length \( N_f \) with overlap of \( N_o \) samples. Each frame is partitioned into \( m = N_f - k \) smaller overlapping \( k \)-dimensional column vectors. These vectors are then arranged into so called trajectory matrix \( \mathbf{X}^{(f)} \) of size \( k \times m \)

\[ \mathbf{X}^{(f)} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m], \]

(13)

where \( f \) is a frame index. The outer product of the trajectory matrix is used to compute the sample covariance matrix of the noisy speech \( \mathbf{C}_x \approx \mathbf{R}_x \). This estimate is a basis for computation of the remaining parameters. In our experiments the noise covariance matrix \( \mathbf{C}_n \approx \mathbf{R}_n \) was estimated from the initial frames of the speech sentences, but in a practical implementation one can use voice activity detector (VAD) to collect the noise statistics during the speech pauses. Due to complexity reasons, first we compute the effective filter (14) and then all in-frame vectors are processed using the same matrix. The result is stored in the trajectory matrix of the enhanced speech, say \( \mathbf{Y}^{(f)} \). The enhanced vectors are obtained from this matrix using diagonal averaging. Finally, the frames are multiplied by a Hanning window and synthesized using overlap-add method. In our implementation we use the following settings: \( N_f = 400 \), \( N_o = 200 \), \( k = 40 \) and assume the sampling rate at 16 kHz.

In the presented system, the masking threshold \( \phi_y(\omega_i) \) of the clean speech must be estimated. Most psychoacoustic models operate on critical-band energies which are obtained by appropriate grouping the power spectral components of the clean speech. Thus in fact, we need the estimate of the clean speech power spectrum. From definition, it is given by

\[ \phi_y(\omega_i) = E \left[ |\mathbf{v}_{y,i}^\# \mathbf{v}_i|^2 \right] = |\mathbf{v}_{y,i}^\# \mathbf{R}_y \mathbf{v}_i|. \]

(14)

Since \( \mathbf{R}_y \) may be a positive semi-definite (14) can be written as follows

\[ \phi_y(\omega_i) = \sum_{i=1}^{r} |\mathbf{v}_{y,i}^\# \mathbf{u}_{y,i}^2 \lambda_{y,i}|. \]

(15)

We reconstruct the power spectrum from the signal subspace only, which makes the estimate more accurate. Further, (15) is used as an input for Johnston’s psychoacoustic model [10].
The values of Lagrange multipliers at a solution point usually have some significance. In our optimization problem they control a trade-off between residual noise and speech distortions and they must be precisely set. Particularly, if the constraints in (4) are satisfied with equality, the residual noise levels can be written as follows

\[ \alpha_i = v_i^\theta \hat{H} y_i \hat{H}^\theta v_i. \]  

As the Lagrange multipliers are involved in \( \hat{H} \) it is difficult to express them explicitly. However, if \( V \) is orthogonal, an estimation of the \( i \)-th multiplier is equivalent to finding the root of the equation

\[ g_i(\mu_i) = \sum_{l=1}^{r} \left| v_i^\theta \lambda_{y,l}^i \frac{\lambda_{y,l}^i}{\lambda_{n,l}^i k \mu_i + \lambda_{y,l}^i} u_{y,l} \right|^2 \lambda_{n,l}^i - \alpha_i. \]  

In the proposed system Lagrange multipliers are calculated iteratively using Newton’s method. It is known that this method can be unstable near a local extremum or a horizontal asymptote. First-derivative of (17) can be expressed as follows

\[ \frac{d g_i(\mu_i)}{d \mu_i} = -2 \sum_{l=1}^{r} \frac{\left( \lambda_{n,l}^i \lambda_{y,l}^i \right)^2}{\left( \lambda_{n,l}^i k \mu_i + \lambda_{y,l}^i \right)^2} \left| v_i^\theta u_{y,l} \right|^2 \frac{\lambda_{n,l}^i}{\lambda_{y,l}^i}. \]  

Since (18) is negative for \( \mu_i \geq 0 \), the relation (17) is monotonically decreasing function in \( (0; \infty) \). Thus only the second problem is important. Namely if \( \alpha_i \approx 0 \) then \( g_i(\mu_i) = 0 \) for \( \mu_i \to \infty \). Recalling (7), we see that residual noise level can not be smaller than masking threshold which is always greater than absolute threshold of hearing. Also note that for \( \phi_i(\omega_i) > \alpha_{i,max} \) corresponding spectral component is not suppressed. More formally it can be written in terms of unconstrained minimization as follows

\[ \mu_i = \begin{cases} \arg \min_{\mu_i \geq 0} g_i^2, & \phi_i(\omega_i) < \alpha_{i,max} \\ 0, & \text{otherwise} \end{cases} \]  

In our experiments the solution was found in an acceptable number of iterations (5-20). It is worthwhile to note that since the spectrum \( \{\alpha_i, i = 1, 2, \ldots, k\} \) is symmetric and \( k \) is even, only \( k/2 + 1 \) Lagrange multipliers have to be estimated.

Table 1: Averaged results of the objective measurement for different noise environments.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>SegSNR (dB)</th>
<th>MBSD (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDCw</td>
<td>PCSS</td>
</tr>
<tr>
<td>white</td>
<td>13.71</td>
<td>14.08</td>
</tr>
<tr>
<td>car</td>
<td>8.56</td>
<td>10.36</td>
</tr>
<tr>
<td>pink</td>
<td>9.22</td>
<td>10.08</td>
</tr>
<tr>
<td>f16</td>
<td>9.34</td>
<td>10.31</td>
</tr>
<tr>
<td>avg</td>
<td>10.21</td>
<td>11.21</td>
</tr>
</tbody>
</table>

4. Experiments

The proposed method was implemented and tested using MATLAB software. For comparative purposes we also implemented the SDC estimator [3] for white noise, and exact whitening-based PCSS method [6]. The set of eight speech sentences uttered by both male and female speakers was selected from TIMIT database. The sentences were about a 5 - 8 s long. Four types of noise signal were selected: white noise, car engine noise, pink noise and F16-jet cockpit noise. These signals were artificially added to the clean speech sentences such that the segmental SNR (SegSNR) was between 0 dB and 20 dB. We used SNR based and perceptual measures for objective performance evaluation of the implemented algorithms. The SNR based speech distortion measure is defined as the segmental signal to noise ratio where the noise is interpreted as a difference between original and enhanced speech. The higher the value of this factor, the more accurate estimation. Also modified Bark spectral distortion (MBSD) measure [11] was used for evaluation of the audible difference between the clean and enhanced speech. The lower the value of this factor, the better quality of the enhanced speech.

Table 1 shows the averaged results for all methods in different noise environments. Due to limited size of the paper, a more detailed results (at different SNRs) are presented for car noise only (see Fig. 2). Generally, the worst performance was reported for the conventional SDC estimator (SDCw). Even in the case of white noise the both PCSS methods offer improve-
The proposed method is a low-complexity alternative for the PCSS method presented in [6]. Unlike the exact method the approximate solution does not require whitening, thus the number of operations per frame can be significantly reduced. The experiments show that a degradation due to approximation depends on noise type and can be neglected for white-like noises. Nevertheless our method outperforms conventional approach in all cases giving better noise attenuation and less speech distortions. Although the best results have been obtained for whitening-based PCSS method the proposed approach is easier to implement and offers acceptable speech quality. Only minor drawback is a presence of the residual noise in speech pauses. This problem indicates the direction of future work.

6. Acknowledgements

This work was supported by the Polish Ministry of Science and Higher Education under Grant MNiSW No. N519 030 32/3775.

7. References