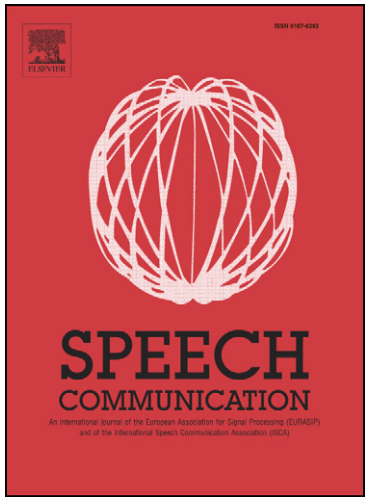
Signal subspace approach for psychoacoustically motivated speech enhancement

Adam Borowicz, Alexandr Petrovsky



Signal subspace approach for psychoacoustically motivated speech enhancement

Adam Borowicz, Alexandr Petrovsky

PII: S0167-6393(10)00152-4
DOI: 10.1016/j.specom.2010.09.002
Reference: SPECOM 1932

To appear in: *Speech Communication*

Received Date: 2 December 2009
Revised Date: 21 July 2010
Accepted Date: 10 September 2010

Please cite this article as: Borowicz, A., Petrovsky, A., Signal subspace approach for psychoacoustically motivated speech enhancement, *Speech Communication* (2010), doi: 10.1016/j.specom.2010.09.002

Signal subspace approach for psychoacoustically motivated speech enhancement

Adam Borowicz*, Alexandr Petrovsky

Department of Real-Time Systems, Bialystok Technical University, Wiejska Str. 45A, 15-351 Bialystok, Poland

Abstract

In this paper we deal with the perceptually motivated signal subspace methods for speech enhancement. We focus on extended spectral-domain-constrained (SDC) estimator. It is obtained using Lagrange multipliers method. We present an algorithm for a precise computation of the Lagrange multipliers, allowing for a direct shaping the residual noise power spectrum. In addition the SDC estimator is presented in a new, possibly more effective form. As a practical implementation of the estimator we propose perceptually constrained signal subspace (PCSS) method for speech enhancement. The approach utilizes masking phenomena for residual noise shaping and is optimal for the case of coloured noise. Also, less demanding approximate version of this method is derived. Finally comparative evaluation of the most known subspace-based methods is performed using objective speech quality measures and listening tests. Results show that the PCSS method outperforms other methods providing high noise attenuation and better speech quality.

Keywords:
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, speech coding, aids for the hearing impaired or analysis of low quality recordings.

*Corresponding author.
Email address: borowicz@wi.pb.edu.pl (Adam Borowicz)

An objective of speech enhancement is to improve performance of these systems in the presence of environmental noise by increasing speech intelligibility and quality. Often a single-channel application is considered, which makes the design much more difficult. Nevertheless, there is a need for an efficient noise reduction algorithms operating at extremely low signal to noise ratios (SNRs). Most of existing single-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 this phenomenon including perceptually motivated approaches (Gustafsson et al., 1998), (Petrovsky et al., 2004, 14 p.), but their optimality in a sense of linear estimation is not clear.

On the other hand a signal subspace approach for speech enhancement is an interesting generalization of the spectral weighting methods. This technique has been originally proposed in (Ephraim and Van Trees, 1995). The speech estimation is considered there as a constrained optimization problem, where the speech distortions are minimized subject to the residual noise power level (defined in a particular domain). Two linear estimators have been proposed: time-domain-constrained (TDC) and spectral-domain-constrained (SDC). Unlike DFT-based methods, signal subspace approaches decompose a signal space into speech subspace and noise subspace using Karhunen-Loeve transform (KLT). Then the spectral weighting is performed in the signal subspace only. The components projected onto the noise subspace are simply nulled which results in a significantly better performance when compared to the conventional frequency domain methods where a full-band spectrum must be processed.

Unfortunately an efficient implementation of the KLT-based methods is a difficult task and substantial simplifications are often taken into account. For instance, the conventional approaches (Ephraim and Van Trees, 1995), assume that the noise is white, in the case of coloured noise it is suggested to whiten the noisy speech signal first. In a such case optimality of the estimators is not guaranteed, because the distortions of the whitened speech are minimized rather than the distortions of the clean speech. Other methods (Rezayee and Gazor, 2001), (Mittal and Phamdo, 2000) deal with the coloured noise problem by approximating a noise covariance matrix, which in fact also results in suboptimal estimators.

Similarly as in (Ephraim and Van Trees, 1995) other signal subspace approaches perform the residual noise shaping in the eigendomain using a

generalised Wiener rule. Such a technique depends on erroneous signal-to-noise estimates and is not optimal from the perceptual point of view (the resulting residual noise might not be masked correctly). However a major difficulty with an incorporating psychoacoustics into the KLT-based methods is the fact that the hearing properties (i.e. masking effects) are unexplained in the eigendomain. In (Jabloun and Champagne, 2003) appropriate transformations have been proposed to convert the masking threshold to the KLT domain and vice versa. In that method, a psychoacoustically motivated weighting rule is used, but the problem of the coloured noise is solved similarly to (Rezayee and Gazor, 2001).

Extended approaches (Hu and Loizou, 2003), (Lev-Ari and Ephraim, 2003) exploit a joint diagonalisation of the noise and clean speech covariance matrices which allows for a derivation of the optimal estimators for the coloured noise. Unfortunately closed form expressions (Lev-Ari and Ephraim, 2003) for these estimators are rather impractical. In fact they involve Lagrange multipliers, which must be carefully set to obtain desired filter. However in general, analytical expressions for these multipliers are unknown. In (Hu and Loizou, 2003) Lagrange multipliers were simply set to fixed value which results in the conventional Wiener-like spectral weighting rule.

Major motivation of our work is to exploit the masking properties in the noise-independent signal subspace approach. We propose perceptually constrained signal subspace (PCSS) method for speech enhancement based on the extended SDC estimator. The solution is presented in a new form which makes the implementation of the estimator more reliable. Unlike the other approaches, our method exploits a perceptually motivated residual noise shaping and imposes the constraints strictly in the frequency domain using discrete Fourier transform (DFT) basis vectors. Namely, the residual noise levels are set just below a masking threshold to attenuate only audible noise components. Since the Lagrange multipliers are involved in the expression for the extended SDC estimator, they must be precisely set for a given set of the residual noise levels. However, we found that these multipliers are independent each other and can be computed numerically. In addition we present an approximate version of the PCSS method as a low-complexity alternative.

The paper is organized as follows. In Section 2 a signal subspace approach for speech enhancement is introduced. In Section 3 we derive the perceptually constrained signal subspace method, also the problems of the residual noise shaping and Lagrange multipliers estimation are discussed. The approximate

version of the PCSS method is derived in Section 4. Implementation details are given in Section 5. The proposed method is evaluated and compared to most known subspace-based approach in Section 6. Finally, we conclude the paper in Section 7.

2. Signal subspace approach for speech enhancement

The noisy speech model used in the signal subspace approach assumes that the clean speech and noise are additive. Let $\mathbf{x} = \mathbf{y} + \mathbf{n}$ denotes k-dimensional noisy speech vector, where $\mathbf{y}$ and $\mathbf{n}$ are zero-mean random vectors representing the clean speech and noise signal respectively. Since the speech and noise are assumed to be uncorrelated, the covariance matrix of the noisy speech process $\mathbf{R_x}$ can be written as

$$\mathbf{R_x} = \mathbf{R_y} + \mathbf{R_n}, \tag{1}$$

where $\mathbf{R_n}$ and $\mathbf{R_y}$ are the covariance matrices of the noise and clean speech process, respectively. It is also assumed that the matrix $\mathbf{R_n}$ is positive definite. Let $\hat{\mathbf{y}} = \mathbf{Hx}$ be a linear estimator of the clean speech. The effective filter $\mathbf{H}$ is found by minimizing an average speech distortion power and constraining a residual noise power level. The residual error vector is defined as follows:

$$\epsilon = \hat{\mathbf{y}} - \mathbf{y} = (\mathbf{H} - \mathbf{I})\mathbf{y} + \mathbf{Hn} = \epsilon_{\mathbf{y}} + \epsilon_{\mathbf{n}}, \tag{2}$$

where $\epsilon_{\mathbf{y}}$ and $\epsilon_{\mathbf{n}}$ are interpreted as the speech distortion vector and residual noise vector respectively. The average speech distortion power is given by

$$\overline{\epsilon_{\mathbf{y}}^2} = \frac{1}{k}\mathrm{tr}E\left\{\epsilon_{\mathbf{y}}\epsilon_{\mathbf{y}}^{\#}\right\} = \frac{1}{k}\mathrm{tr}\left\{(\mathbf{H} - \mathbf{I})\,\mathbf{R_y}\,(\mathbf{H} - \mathbf{I})^{\#}\right\}, \tag{3}$$

where $E\{.\}$ is an expectation operator, $\mathrm{tr}\{.\}$ is matrix trace and superscript $\#$ denote the transpose of a real matrix or the conjugate transpose of a complex matrix. The constraints can be defined in time- or spectral-domain giving the TDC or SDC estimator, respectively (Lev-Ari and Ephraim, 2003). In fact, the TDC estimator is a special case of the SDC estimator. Therefore, we give a brief description of the second one only. In this case the optimization problem is formulated as follows:

$$\min_{\mathbf{H}} \overline{\epsilon_{\mathbf{y}}^2}$$
$$\text{subject to: } E\left\{\left|\mathbf{v}_i^{\#}\epsilon_{\mathbf{n}}\right|^2\right\} \leq \alpha_i, \quad i = 1, \ldots, k, \tag{4}$$

where $\{\mathbf{v}_i,\, i = 1, \ldots, k\}$ is a set of k-dimensional real or complex vectors. Originally, in (Lev-Ari and Ephraim, 2003) the matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]$ was restricted to any orthogonal or unitary matrix.

The solution of (1) is found using Lagrange multipliers method. The Lagrangian is given by

$$L(\mathbf{H}, \bar{\mu}) = \overline{\epsilon_{\mathbf{y}}^2} + \sum_{i=1}^{k} \mu_i \left(\mathbf{v}_i^{\#} \mathbf{H} \mathbf{R}_{\mathbf{n}} \mathbf{H}^{\#} \mathbf{v}_i - \alpha_i \right). \tag{5}$$

Let's define $\mathbf{M} = k\,\mathrm{diag}\,\{\mu_1, \mu_2, \ldots \mu_k\}$ and $\mathbf{L} = \mathbf{V}\mathbf{M}\mathbf{V}^{\#}$. From $\nabla_H L(\mathbf{H}, \bar{\mu}) = 0$ we obtain

$$\mathbf{L}\mathbf{H}\mathbf{R}_{\mathbf{n}} + \mathbf{H}\mathbf{R}_{\mathbf{y}} = \mathbf{R}_{\mathbf{y}}. \tag{6}$$

Above equation can be solved iteratively as proposed in (Hu and Loizou, 2003). An explicit solution is based on a factorisation which transforms jointly both matrices $\mathbf{R}_{\mathbf{y}}$ and $\mathbf{R}_{\mathbf{n}}$ to a diagonal matrix. Such a transform was found in (Hu and Loizou, 2003) where the signal KLT was replaced with a non-orthogonal transform. In (Lev-Ari and Ephraim, 2003) an equivalent solution using the whitening approach was presented and explicit forms for the TDC and SDC estimators were derived. Namely, the signal KLT was replaced with the KLT of the whitened clean speech. Therefore the eigen-decomposition of the whitened clean speech covariance matrix is considered instead of the matrix $\mathbf{R}_{\mathbf{y}}$, i.e.

$$\mathbf{R}_{\tilde{\mathbf{y}}} = E\left\{\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\right\} = \mathbf{R}_{\mathbf{n}}^{-0.5}\mathbf{R}_{\mathbf{y}}\mathbf{R}_{\mathbf{n}}^{-0.5} = \mathbf{U}\Lambda\mathbf{U}^{\#}, \tag{7}$$

where $\tilde{\mathbf{y}}$ is whitened clean speech vector, $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$ denote an orthogonal matrix of eigenvectors and $\Lambda = \mathrm{diag}\,\{\lambda_1, \ldots, \lambda_k\}$ is a diagonal matrix of corresponding eigenvalues (a subscript $\tilde{\mathbf{y}}$ is omitted here for clarity). Let $\mathbf{Q} = \mathbf{R}_{\mathbf{n}}^{-\mathbf{0.5}}\mathbf{U}$ and $\mathbf{G} = \mathbf{Q}^{\#}\mathbf{H}(\mathbf{Q}^{\#})^{-\mathbf{1}}$. Substituting these relations in (6) we have

$$\mathbf{Q}^{\#}\mathbf{L}(\mathbf{Q}^{\#})^{-\mathbf{1}}\mathbf{G} + \mathbf{G}\Lambda = \Lambda. \tag{8}$$

The work (Lev-Ari and Ephraim, 2003) proposes the following implementation:

$$\mathbf{H} = \mathbf{R}_{\mathbf{n}}^{\mathbf{0.5}}\mathbf{U}\tilde{\mathbf{H}}\mathbf{U}^{\#}\mathbf{R}_{\mathbf{n}}^{-\mathbf{0.5}}. \tag{9}$$

The columns of the matrix $\tilde{\mathbf{H}}$ are defined as follows:

$$\mathbf{h}_l = \mathbf{T}\lambda_l \left(\mathbf{M} + \lambda_l \mathbf{I}\right)^{-1} \mathbf{T}^{-1}\mathbf{e}_l, \quad l = 1, \ldots, k, \tag{10}$$

where $\mathbf{T} = \mathbf{U}^{\#}\mathbf{R}_{\mathbf{n}}^{-\mathbf{0.5}}\mathbf{V}$ and $\mathbf{e}_l$ denote a unit-vector for which the $l-$th element is one and all other components are zeros.

Note that a direct implementation of the filter (9) is rather impractical. Even though the signal subspace dimension is small, the computation of the matrix $\mathbf{H}$ is still expensive since it requires a knowledge of a full-rank matrix $\mathbf{U}$. In addition, since $\tilde{\mathbf{H}}$ is not a diagonal matrix, the subspace decomposition is not evident. The equation (9) is called "closed form" expression (Lev-Ari and Ephraim, 2003), but in fact it involves the set of Lagrange multipliers which control a trade-off between speech distortion and residual noise and should be carefully set to obtain desired (possibly psychoacoustically motivated) residual noise shaping. Although setting the multipliers to fixed value gives relatively good results, it might not be optimal from the perceptual point of view. Usually, the residual noise constraints are defined in the eigendomain while the masking properties are explained in the frequency domain (i.e. DFT domain). In a such case it is difficult to use any psychoacoustically motivated rule for the noise shaping. At last, the extended SDC estimator requires prewhitening, which is computationally costly and might be ineffective for non-stationary noises. Note that the whitening and unwhitening transformations depend on a time-varying noise characteristics. Commonly they can be simply computed from the noise covariance matrix. However in practice this matrix is unknown and must be estimated. In Section 5 we will provide some details on how it can be done.

3. PCSS method

Taking into account the problems summarized in the previous Section we propose a novel perceptually constrained signal subspace method based on the extended SDC estimator. We selected the extended SDC estimator because it performs an optimal decorrelation in transform domain and its efficiency does not depend on the noise type. Optimality of the KLT transform is especially important for a cancellation of the musical tones effect. Although the main processing is performed in the KLT domain of the whitened speech, the constraints on the residual noise spectrum can be defined in other domain not necessarily related to the KLT. This possibility was suggested in (Lev-Ari and Ephraim, 2003), but it was not examined in practice until now.

3.1. Novel interpretation of the SDC estimator

As mentioned before, a direct implementation of the filter (9) is rather impractical, in addition the decomposition onto the signal and noise subspace is not evident. However, if the matrix $\mathbf{R}_{\tilde{\mathbf{y}}}$ is a positive semi-definite, the column-vectors $\mathbf{h}_l$ corresponding to zero eigenvalues have all elements equal to zero. Therefore, (9) can be rewritten as follows:

$$\mathbf{H} = \mathbf{R}_{\mathbf{n}}^{0.5}\mathbf{U}\sum_{l=1}^{r}\mathbf{h}_l\mathbf{u}_l{}^{\#}\mathbf{R}_{\mathbf{n}}^{-0.5}, \tag{11}$$

where r denote the signal subspace dimension. The parameter r is usually estimated as the number of the strictly positive eigenvalues according to the following rule:

$$r = \arg\max_{1\geq l\geq k}\left\{\lambda_l > \theta\right\}. \tag{12}$$

In practice, the threshold θ is usually set to some small positive number to avoid numerical problems. Greater values of θ lead to reducing the residual noise, however a special care must be taken because low-power speech components can be also cancelled. In our experiments we simply set this parameter to be 3 times greater than absolute value of the minor eigenvalue, but not smaller than 2^{-52}.

Recalling definition of (10) the expression for the effective filter can be simplified to

$$\mathbf{H} = \sum_{l=1}^{r}\mathbf{V}\lambda_l\left(\lambda_l\mathbf{I}+\mathbf{M}\right)^{-1}\mathbf{V}^{\#}\bar{\mathbf{q}}_l\mathbf{q}_l^{\#}, \tag{13}$$

where $\bar{\mathbf{q}}_l$ is the l-th column vector of the matrix $(\mathbf{Q}^{\#})^{-\mathbf{1}}$ and $\mathbf{q}_l$ is the l-th column vector of the matrix $\mathbf{Q}$. Note that to compute these vectors we require only the l-th eigenvector of the matrix $\mathbf{U}$ (and whitening/unwhitening transforms). The problem has unique solution if and only if $\lambda_l\mathbf{I}+\mathbf{M}$ is non-singular.

As can be seen in (13) the proposed approach does not require a complete set of eigenvectors. This fact is especially important if the eigenvalues are estimated using any iterative technique like PASTd algorithm (Yang, 1995). Moreover, it is clearly visible that the noisy components which are projected onto the noise subspace are nulled. Although both solutions are equivalent, our interpretation may avoid many unnecessary numerical operations. Namely, the computational load of our method is data-dependent. In the

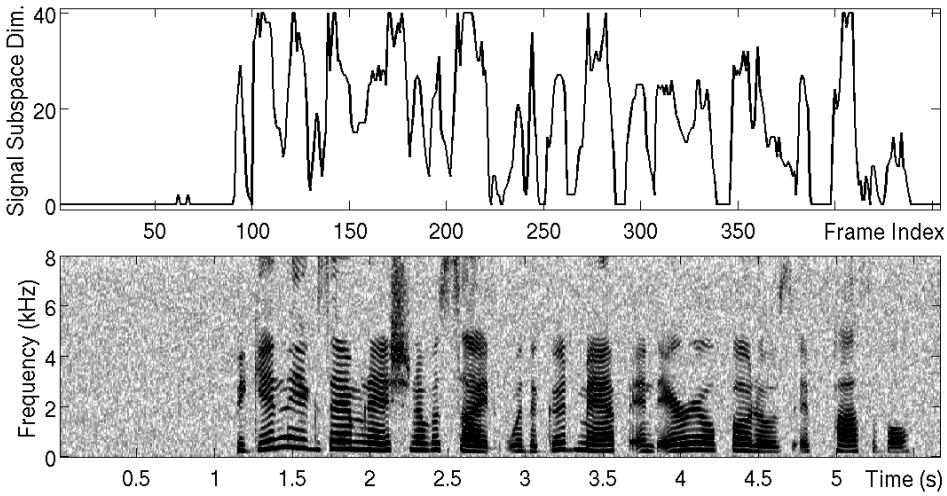


Figure 1: An example estimation of the signal subspace dimension (top) for typical speech signal (bottom).

worst-case the complexity of our solution is approximately the same as that in the work (Lev-Ari and Ephraim, 2003) but the number of the eigenvalues varies over time. As can be seen in Fig. 1 the situation when $r < k$ is very common for typical speech sample. Therefore in average-case our solution outperforms the conventional one.

Also note that, the collection of the non-orthogonal subspace projectors $\{\bar{\mathbf{q}}_l \mathbf{q}_l^{\#},\, l = 1, 2, \ldots, r\}$ can be interpreted as the r-channel filter bank. Such an interpretation is especially useful in parallel processing, especially when FPGA implementation is considered. In addition, if the matrix $\mathbf{V}$ is the DFT related, the in-channel filters can be efficiently implemented using fast Fourier transform (FFT) algorithm. Such a realisation is presented in Fig. 2. Note that the matrices $\mathbf{G}_l = \lambda_l \left(\lambda_l \mathbf{I} + \mathbf{M}\right)^{-1}$ can be viewed as the frequency domain weighting filters.

3.2. Perceptually motivated constraints

It was verified empirically that the Wiener-like weighting rule (Ephraim and Van Trees, 1995) makes enhanced spectrum similar to that of the clean speech. Unfortunately, such a technique is weakly correlated with the human auditory perception. If the constraints are defined in eigendomain, it is difficult to use any psychoacoustically motivated weighting rule (defined usually in the frequency domain) for shaping the residual noise spectrum.

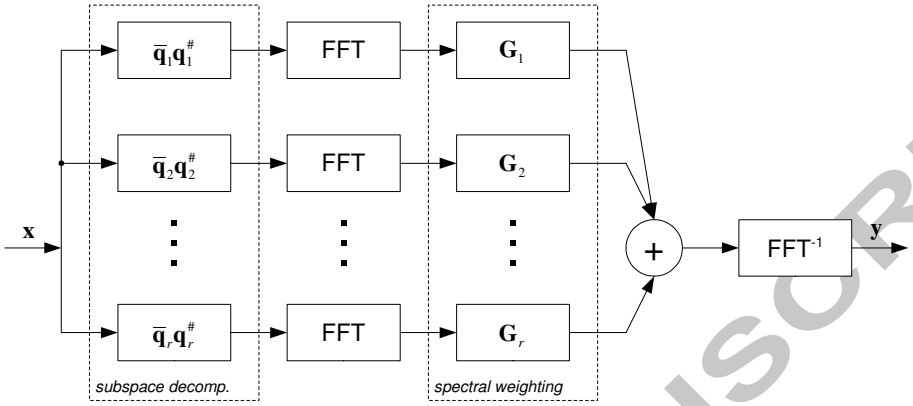


Figure 2: Direct realisation of the extended SDC estimator using the FFT algorithm.

According to well known IND rule (Gustafsson et al., 1998), if any frequency component of the residual noise is greater than the masking threshold, it is audible and clean speech is deteriorated by noise. In the opposite situation, when the frequency component is below the threshold, we have 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.

Although frequency-domain representation of the residual noise spectrum can be obtained using the appropriate transformation (Jabloun and Champagne, 2003), we propose a simpler solution. One possible choice of $\mathbf{V}$ is a unitary matrix. Thus the residual noise spectrum $\{\alpha_i,\, i = 1, 2, \ldots, k\}$ can be defined directly in the frequency domain using sinusoidal vectors

$$\mathbf{v}_i^{\#} = k^{-1/2} \left[e^{-j\omega_i \cdot 0}, e^{-j\omega_i \cdot 1}, \ldots, e^{-j\omega_i \cdot (k-1)} \right], \tag{14}$$

where

$$\omega_i = 2\pi \left(i - 1 \right) / k, \quad i = 1, 2, \ldots, k. \tag{15}$$

The vector $\mathbf{v}_i^{\#}$ is interpreted here as the i-th row of normalized DFT matrix. Since the masking threshold is also defined in the frequency domain, the frequency-to-eigendomain transformation (Jabloun and Champagne, 2003) is not necessary. Taking into account these considerations, we propose the following IND-based rule for shaping residual noise power spectrum

$$\alpha_i = \min \left(\phi_t(\omega_i), \alpha_{i,\max} \right), \quad i = 1, \ldots, k, \tag{16}$$

where $\phi_t(\omega_i)$ denote the masking threshold of the clean speech and

$$\alpha_{i,\max} = \sum_{l=1}^{r} \left| \mathbf{v}_i^{\#} \bar{\mathbf{q}}_l \right|^2, \tag{17}$$

is a maximum possible residual noise level for the i-th spectral bin (no attenuation-case).

3.3. Lagrange multipliers estimation

An interesting aspect of the Lagrange multipliers method is that the values of the multipliers at the solution point usually have some significance. In our optimization problem they control a trade-off between the residual noise and speech distortions and should be carefully set to obtain desired filter. As we mentioned before, in the case of coloured noise, an explicit derivation of the Lagrange multipliers for a given set of residual noise levels seems to be a difficult task. According our best knowledge a such expression is currently unknown. However as we will show they can be computed numerically after all.

If the constraints in (4) are satisfied with equality, the residual noise levels can be written as follows:

$$\alpha_i = \mathbf{v}_i^{\#} \mathbf{R}_{\mathbf{n}}^{0.5} \mathbf{U} \mathbf{G} \mathbf{G}^{\#} \mathbf{U}^{\#} \mathbf{R}_{\mathbf{n}}^{0.5} \mathbf{v}_i, \quad i = 1, 2, \ldots, n. \tag{18}$$

It can be observed that

$$\mathbf{G}\mathbf{G}^{\#} = \sum_{l=1}^{r} \mathbf{g}_l \mathbf{g}_l^{\#}, \tag{19}$$

where $\mathbf{g}_l$ is the l-th column vector of the matrix $\mathbf{G}$. Thus, substituting (19) into (18) we have

$$\alpha_i = \sum_{l=1}^{r} \left| \mathbf{v}_i^{\#} \frac{\lambda_l}{k\mu_i + \lambda_l} \bar{\mathbf{q}}_l \right|^2, \quad i = 1, 2, \ldots, k. \tag{20}$$

In general, $\mathbf{v}_i^{\#} \bar{\mathbf{q}}_l \neq 0$, for all i, l. If we assume that residual noise levels are defined in the eigendomain (i.e. $\mathbf{V} = \mathbf{U}$) and the noise is white with the variance $\sigma_{\mathbf{n}}^2$, i.e. $\bar{\mathbf{q}}_l = \sqrt{\sigma_{\mathbf{n}}^2} \mathbf{u}_l$, then above equation can be simplified to

$$\alpha_i = \left(\frac{\lambda_l}{k\mu_i + \lambda_l} \right)^2 \sigma_{\mathbf{n}}^2 \tag{21}$$

It results in the following expression for the Lagrange multipliers

$$\mu_i = \frac{\lambda_i}{k}\left[\left(\alpha_i/\sigma_{\mathbf{n}}^2\right)^{-0.5} - 1\right]. \tag{22}$$

Note that, substituting these relations to (9) and using an appropriate rule for the noise shaping we can obtain the conventional SDC estimator for white noise (Ephraim and Van Trees, 1995). However in our case we make no assumption about noise nature as well as the constraints domain. Therefore the Lagrange multipliers must be estimated directly. Taking into account the relation (20) it is easy to see that the estimation of the i-th multiplier is equivalent to finding the root of the following function:

$$g_i(\mu_i) = \sum_{l=1}^{k}\left|\mathbf{v}_i^{\#}\frac{\lambda_l}{k\mu_i + \lambda_l}\bar{\mathbf{q}}_l\right|^2 - \alpha_i. \tag{23}$$

As we will show further, it can be found numerically for a given residual noise level.

4. Approximate solution

In this Section we present an approximate version of the PCSS method. The proposed approach does not require a whitening (unwhitening) step, however it exploits the same perceptually motivated residual noise shaping rule and imposes the constraints strictly in the frequency domain using discrete Fourier transform (DFT) basis vectors.

Let $\mathbf{U}_{\mathbf{y}}\Lambda_{\mathbf{y}}\mathbf{U}_{\mathbf{y}}^{\#}$ be eigen-decomposition of the matrix $\mathbf{R}_{\mathbf{y}}$. In the case of the white noise i.e. $\mathbf{R}_{\mathbf{n}} = \sigma_{\mathbf{n}}^2 I$, where $\sigma_{\mathbf{n}}^2$ is the noise variance, both matrices $\mathbf{R}_{\mathbf{y}}$ and $\mathbf{R}_{\mathbf{n}}$ can be diagonalised jointly using the matrix $\mathbf{U}_{\mathbf{y}}$, which makes the solution of (6) trivial. In the case of coloured noise, the following approximation has been suggested (Rezayee and Gazor, 2001):

$$\mathbf{R}_{\mathbf{n}} \approx \mathbf{U}_{\mathbf{y}}\hat{\Lambda}_{\mathbf{n}}\mathbf{U}_{\mathbf{y}}^{\#}, \tag{24}$$

where $\hat{\Lambda}_{\mathbf{n}}$ is a diagonal matrix with entries defined as follows:

$$\hat{\lambda}_{\mathbf{n},i} = \mathbf{u}_{\mathbf{y},i}^{\#}\mathbf{R}_{\mathbf{n}}\mathbf{u}_{\mathbf{y},i}, \quad i = 1, \ldots, k. \tag{25}$$

Substituting (24) to (6) and denoting the sub-optimal filter by $\hat{\mathbf{H}}$ we obtain

$$\mathbf{L}\hat{\mathbf{H}}\mathbf{U}_{\mathbf{y}}\hat{\Lambda}_{\mathbf{n}}\mathbf{U}_{\mathbf{y}}^{\#} + \hat{\mathbf{H}}\mathbf{R}_{\mathbf{y}} = \mathbf{R}_{\mathbf{y}}. \tag{26}$$

Let $\hat{\mathbf{G}} = \mathbf{U}_{\mathbf{y}}^{\#}\hat{\mathbf{H}}\mathbf{U}_{\mathbf{y}}$, then the above equation can be written as follows:

$$\mathbf{U}_{\mathbf{y}}^{\#}\mathbf{L}\mathbf{U}_{\mathbf{y}}\hat{\mathbf{G}}\hat{\Lambda}_{\mathbf{n}} + \hat{\mathbf{G}}\Lambda_{\mathbf{y}} = \Lambda_{\mathbf{y}}. \tag{27}$$

Note that

$$\mathbf{U}_{\mathbf{y}}^{\#}\mathbf{L}\mathbf{U}_{\mathbf{y}}\hat{\mathbf{g}}_l\hat{\lambda}_{\mathbf{n},l} + \hat{\mathbf{g}}_l\lambda_{\mathbf{y},l} = \lambda_{\mathbf{y},l}\mathbf{e}_l, \tag{28}$$

where $\hat{\mathbf{g}}_l$ denote the l-th column vector of $\hat{\mathbf{G}}$. Thus we have

$$\hat{\mathbf{g}}_l = {\mathbf{U}_{\mathbf{y}}}^{\#}\lambda_{\mathbf{y},l}\left(\hat{\lambda}_{\mathbf{n},l}\mathbf{L} + \lambda_{\mathbf{y},l}\mathbf{I}\right)^{-1}\mathbf{U}_{\mathbf{y}}\mathbf{e}_l. \tag{29}$$

Recalling the definition of $\hat{\mathbf{G}}$, the closed form expression for the sub-optimal linear filter is given by

$$\hat{\mathbf{H}} = \sum_{l=1}^{k}\mathbf{V}\lambda_{\mathbf{y},l}\left(\lambda_{\mathbf{y},l}\mathbf{I} + \mathbf{M}\hat{\lambda}_{\mathbf{n},l}\right)^{-1}\mathbf{V}^{\#}\mathbf{u}_{\mathbf{y},l}\mathbf{u}_{\mathbf{y},l}^{\#}. \tag{30}$$

Note that, (30) has unique solution if and only if $\hat{\lambda}_{\mathbf{n},l}\mathbf{M}+\lambda_{\mathbf{y},l}\mathbf{I}$ is non-singular. It is worthwhile to note that although presented approximate solution is sub-optimal for the coloured noise, it is optimal for the white noise. Thus it is an an interesting, low-complexity alternative for whitening-based approaches.

For instance, if the constraints are defined in the KLT domain i.e. $\mathbf{V} = \mathbf{U}_{\mathbf{y}}$, the filter (30) simplifies to the sub-optimal SDC estimator (Jabloun and Champagne, 2003). In a such case the Lagrange multipliers can be easily computed. On the other hand if we set all multipliers to a fixed value say $\mu_l = \mu/k$, we obtain the sub-optimal TDC estimator (Rezayee and Gazor, 2001). Otherwise the multipliers should be estimated in similar way as in the exact PCSS method. Particularly, if the constraints in (4) are satisfied with an equality, the residual noise levels can be written as follows:

$$\alpha_i = \mathbf{v}_i^{\#}\hat{\mathbf{H}}\mathbf{R}_{\mathbf{n}}\hat{\mathbf{H}}^{\#}\mathbf{v}_i. \tag{31}$$

Substituting (31) into the right side of the inequality (4) and using the approximation (24) we obtain

$$\alpha_i = \sum_{l=1}^{k}\left|\mathbf{v}_i^{\#}\frac{\lambda_{\mathbf{y},l}}{\hat{\lambda}_{\mathbf{n},l}k\mu_i + \lambda_{\mathbf{y},l}}\mathbf{u}_{\mathbf{y},l}\right|^2\hat{\lambda}_{\mathbf{n},l}. \tag{32}$$

Thus, in this case, we have k independent one-dimensional equations and the Lagrange multipliers can be computed numerically for a given set of the residual noise levels in a similar way as before.

5. Implementation

In our implementation we use a frame-by-frame algorithm. Namely, we divide the signal into the frames of length N_f with overlap of N_o samples. Each frame is partitioned into $m = N_f - k$ smaller overlapping k-dimensional vectors. Let's define t-th in-frame vector as follows:

$$\mathbf{x}_t = \begin{bmatrix} x(\ell(N_f - N_o) + t + 1) \\ x(\ell(N_f - N_o) + t + 2) \\ \vdots \\ x(\ell(N_f - N_o) + t + k) \end{bmatrix}, \tag{33}$$

where ℓ is the frame index and $x(.)$ are the noisy speech samples. The sequence of these vectors can be considered as a trajectory in k-dimensional Euclidean space. Such a sequence is arranged into so called trajectory matrix of size k-by-m

$$\mathbf{X}^{(\ell)} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \end{bmatrix}. \tag{34}$$

The outer product of the trajectory matrix is then used to compute the sample covariance matrix of noisy speech

$$\mathbf{C_x}^{(\ell)} = \frac{1}{m}\mathbf{X}^{(\ell)}(\mathbf{X}^{(\ell)})^T. \tag{35}$$

This estimate is the basis for the computation of the noise eigenstructures (in the speech pauses only) and the KLT of the whitened signal, respectively:

$$\begin{aligned} \mathbf{C_n} &\approx \mathbf{U_n}\Lambda_\mathbf{n}\mathbf{U_n}^\#, \\ \mathbf{C}_{\tilde{\mathbf{y}}} &= \mathbf{C_n}^{-0.5}\mathbf{C_x}\mathbf{C_n}^{-0.5} - \mathbf{I} \approx \mathbf{U}\Lambda\mathbf{U}^\#. \end{aligned} \tag{36}$$

In the above we omitted the frame index ℓ for clarity. In order to avoid the numerical problems the square roots of the matrices are calculated using the eigenstructures $\mathbf{U_n}, \Lambda_\mathbf{n}$ of the noise covariance matrix. A simplified processing scheme is depicted in Fig. 3. First we compute the effective filter $\mathbf{H}$ 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 $\hat{\mathbf{Y}}^{(\ell)}$. The enhanced vectors are obtained from the matrix $\hat{\mathbf{Y}}^{(\ell)}$ using a diagonal averaging technique (Vetter et al., 1999). Finally, the frames are multiplied by Hanning window and synthesized using overlap-add method.

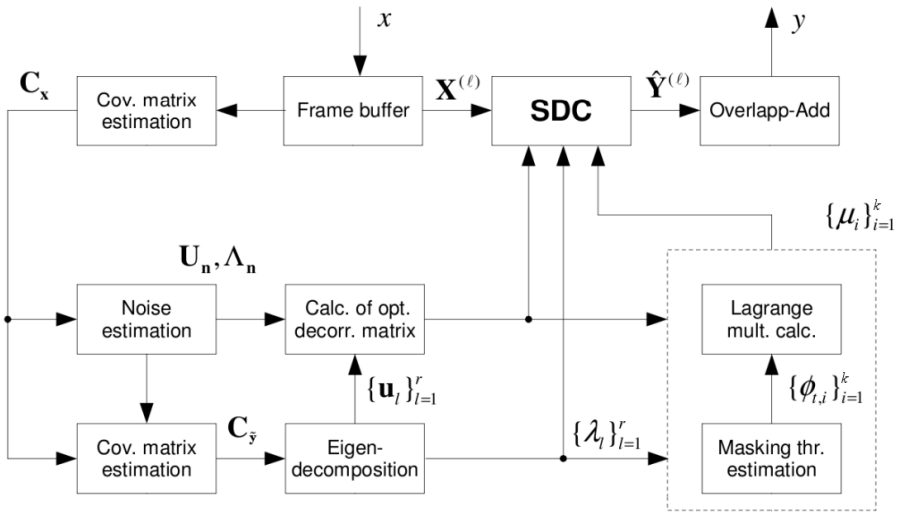


Figure 3: Block diagram of the PCSS method.

As can be seen in Fig. 3, for the computation of the effective filter we need the set of the non-orthogonal projectors, whitened clean speech eigenvalues and Lagrange multipliers. In our implementation the 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. Since a first-derivative of (23) is negative for $\mu_i \geq 0$, i.e.

$$\frac{dg_i(\mu_i)}{d\mu_i} = -2k \sum_{l=1}^{k} \frac{1}{k\mu_i + \lambda_l} \left| \mathbf{v}_i^{\#} \frac{\lambda_l}{k\mu_i + \lambda_l} \bar{\mathbf{q}}_l \right|^2 < 0, \tag{37}$$

the relation (23) is a monotonically decreasing function in $\langle 0; \infty)$. Thus only the second problem can arise. If $\min\left(\phi_t(\omega_i), \alpha_{i,\max}\right) \approx 0$, then $g_i(\mu_i) = 0$ for $\mu_i \to \infty$. Such a situation occurs in the speech pauses or when the power of the noise signal is very low. Since $\mathbf{R_n}$ is assumed to be positive definite, the maximum residual noise is always greater than zero, for $r > 0$. If it is not the case, the matrix $\mathbf{R_n}$ could be regularized by adding some small positive constant to the estimated eigenvalues.

Simply, each μ_i can be initialised to zero, but the number of iterations can be reduced by setting $\mu_i = \mu_{i-1}$, for $i > 2$ in the first iteration. It was supported by the observation that the constraints are defined on the smoothed spectrum, so the functions, have similar shapes and properties. In our experiments the solution was found in the 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.

A direct method for estimation of $\phi_t(\omega_i)$ from the noisy data does not exist. Commonly used methods operate on the critical-band energies which are obtained by appropriate grouping the power spectral components of the clean speech. Thus in fact the power spectrum of the clean speech is needed here. From the definition, it is given by

$$\phi_y(\omega_i) = E\left[\left|\mathbf{v}_i^{\#}\mathbf{y}\right|^2\right] = \mathbf{v}_i^{\#}\mathbf{R}_{\mathbf{y}}\mathbf{v}_i. \tag{38}$$

Therefore, the clean speech covariance matrix should be estimated first. Typically it can be obtained from $\mathbf{R_y} = \mathbf{R_x} - \mathbf{R_n}$, which is equivalent to a spectral subtraction technique. On the other hand we can use the whitened speech covariance matrix. In order to minimize musical tones effect we propose to reconstruct the power spectrum of the clean speech from the signal subspace only. Using the eigen-decomposition (7) we can rewrite (38) as follows:

$$\phi_y(\omega_i) = \mathbf{v}_i^{\#}(\mathbf{Q}^{-1})^{\#}\Lambda(\mathbf{Q}^{-1})\mathbf{v}_i = \sum_{l=1}^{r}\left|\mathbf{v}_i^{\#}\bar{\mathbf{q}}_l\right|^2\lambda_l, \tag{39}$$

The resulting estimates for $i = 1, 2, \ldots, k$ are used as the input data for Johnston's psychoacoustic model (Johnston, 1988). Note that the parameters $\mathbf{Q}$ and Λ in (39) are calculated using the eigenstructures of the estimated covariance matrix (36).

In our implementation we use the following settings: $N_f = 400$, $N_o = 200$, $k = 40$ and assume the sampling rate at 16 kHz. Due to fact, that the filter is estimated in relatively long frames, the computational complexity is effectively reduced. Such an approach is only applicable when the processed signal is stationary within the frame. In the case of highly non-stationary processes it can lead to the significant estimation errors. However, in the proposed method, the stationarity period (frame length) is about 16 ms and seems to be short enough even for the speech signal.

6. Experiments

The proposed PCSS method and its approximate version (denoted here as PCSSa) were implemented and tested using MATLAB software. For comparative purposes we also implemented a conventional SDC estimator for the

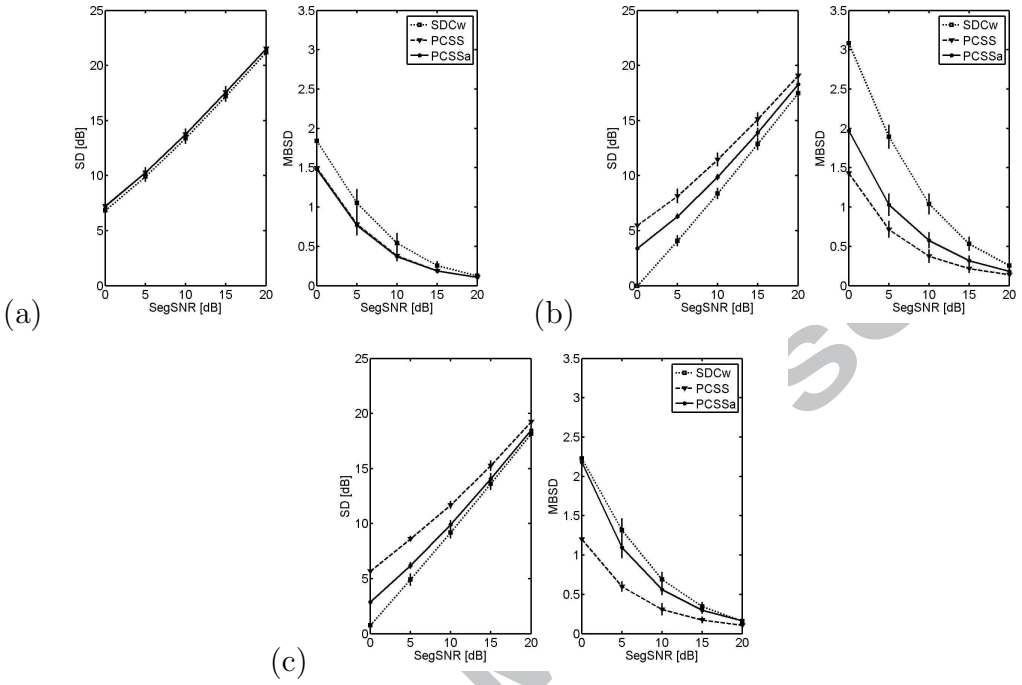


Figure 4: Objective evaluation of speech distortions for white noise (a), car engine noise (b), jet cockpit noise (c) using SNR-based measure (SD) and perceptual measure (MBSD); vertical lines denote standard deviations.

white noise (Ephraim and Van Trees, 1995). The set of eight speech sentences uttered by both male and female speakers was selected from TIMIT database (Garofolo et al., 1993). The sentences were about 5-8 s long. Three types of noise signal were selected: white noise, car engine 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 the SNR based and perceptual measures for an objective performance evaluation of the implemented algorithms. The speech distortion measure (SD) is defined as the segmental signal to noise ratio where the noise is interpreted as a difference between the original and enhanced speech. A modified Bark spectral distortion (MBSD) measure (Yang et al., 1998) was used for the evaluation of the audible speech distortion.

Figure 4 shows the evaluation results in different noise environments. Generally, the worst performance was reported for the conventional SDC estimator (SDCw). Even in the case of the white noise the both PCSS

methods offer improvements comparing to the SDCw method. Although the performance gain for the SegSNR measure is not stunning (about 3%), it is evident for the MBSD measure. It confirms our thesis that the perceptually motivated constraints (7) are more robust than the Wiener-like rule used in (Ephraim and Van Trees, 1995). The best results were obtained for the exact (whitening-based) PCSS method, but this is done at the cost of increased computational complexity. The approximate version of this method (PCSSa) is much simpler and gives similar results for the white-like noises. In the case of the coloured noise we must accept a trade-off between complexity and speech quality. As expected, the relative performance gain of the PCSSa method strongly depends on the noise type. The best results are obtained for the car noise. Our observations show that the optimal KLT basis of the car noise used in the tests is similar to the KLT basis of the long-term estimate of the clean speech covariance matrix, thus the approximation (16) produces near diagonal matrix. However it is not the case for the F16-jet cockpit noise, which was confirmed in our experiments too. As can be seen, the presented method outperforms the SDCw approach, but the relative improvement is much smaller for coloured noises. This suggests that an approximation error may be significant in some situations.

The MSBD measure is relatively highly correlated with subjective speech distortion measures, however it is less useful in the evaluation of the noise shaping capabilities. Therefore the informal listening test was also carried out in order to compare the residual noise artefacts generated by each method. For this purpose we were used a similar test as in the work (Jabloun and Champagne, 2002). The evaluation was carried out in a group of five listeners from 25 to 40 years of age. Three noisy speech signals corresponding to three previously used noise types were enhanced using a given method. We selected the signals at the lowest SegSNR of 0 dB to make sure the residual noise will be audible in all cases. The subjects were asked to score a difference between the residual noise characteristics of the enhanced signals using 5 level scale. The following choices were allowed:

0: Completely different.

1: Different.

2: Do not know.

3: Almost similar.

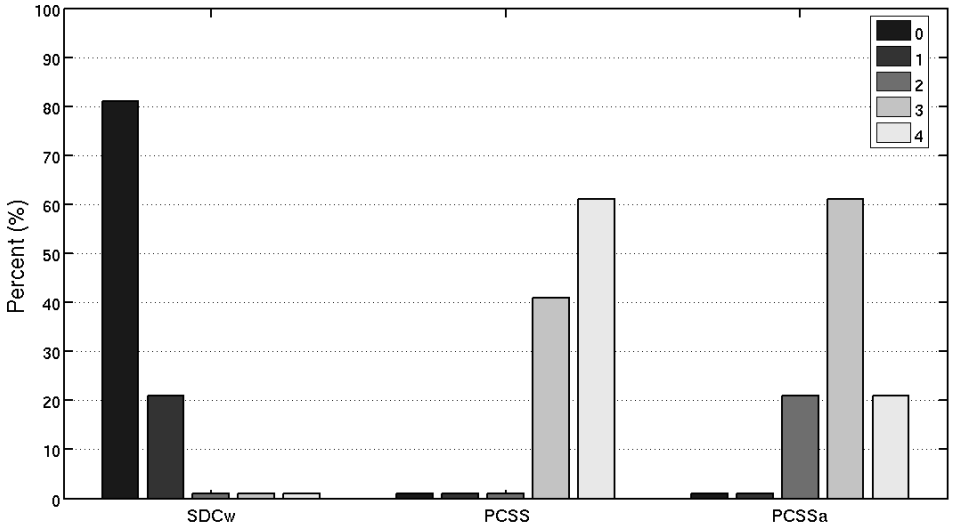


Figure 5: Subjective evaluation of the noise shaping capabilities for the selected speech enhancement methods.

4: Similar.

The same test was performed for all three methods. The percentage results are depicted in Fig. 5. As can be seen, for the conventional SDC estimator, the scores in range of 0-1 suggest that residual noise is different for every signal. For the PCSS method we obtained the average score of 3.6 which indicates that the residual noise characteristics is almost identical for all noises. The approximate version of the PCSS method gives slightly lower scores, however it is clear that its performance does not depend on noise conditions as strong as for the SDC method.

The subjective tests were also confirmed by spectrogram analysis. As can be seen in Fig. 6, the conventional SDC approach generates annoying low-band residual noise. Note that it is completely different from well known musical noise which is typical for most of the DFT-based methods. Also the approximate PCSS method generates the similar residual signal, but at lower level. We verified that this effect is especially audible in speech pauses only. During the voice activity, it is rather inaudible due to the masking phenomenon.

7. Conclusion

We proposed the perceptually constrained signal subspace approach for speech enhancement. An extended SDC estimator has been presented in a new form making an implementation of the subspace approach more practical. The residual noise constraints are defined strictly in the frequency domain using DFT-related vector basis and perceptual criteria. Experiments show that the proposed method outperform other approaches providing perceptually optimal residual noise shaping and lower speech distortions. Also approximate solution have been derived as a low-complexity alternative for the PCSS method. Unlike the exact method, the approximate approach 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.

8. Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

9. References

Ephraim, Y., Van Trees, H., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 3 (4), 251–266.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., 1993. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology (NIST), CD-ROM.

Gustafsson, S., Jax, P., Vary, P., 1998. A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristic. In: Proc. ICASSP. Vol. 1. pp. 397–400.

Hu, Y., Loizou, P., 2003. A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans. Speech Audio Process. 11 (4), 334–341.

Jabloun, F., Champagne, B., May 2002. A perceptual signal subspace approach for speech enhancement in colored noise. In: Proc. ICASSP. Vol. 1. Orlando, USA, pp. 569–572.

Jabloun, F., Champagne, B., 2003. Incorporating the human hearing properties in the signal subspace approach for speech enhancemnt. IEEE Trans. Speech Audio Process. 11 (6), 700–708.

Johnston, J., February 1988. Transform coding of audio signals using perceptual noise criteria. IEEE J. on Selected Areas in Comm. 6, 314–323.

Lev-Ari, H., Ephraim, Y., 2003. Extension of the signal subspace enhancement to colored noise. IEEE Signal Process. Lett. 10 (4), 104–106.

Mittal, P., Phamdo, N., 2000. Signal/noise KLT based approach for enhancing speech degraded by colored noise. IEEE Trans. Speech Audio Process. 8 (2), 159–167.

Petrovsky, A., Parfieniuk, M., Borowicz, A., May 2004, 14 p. Warped DFT based perceptual noise reduction system. In: Proc. AES 116th. Berlin, Germany.

Rezayee, A., Gazor, S., 2001. An adaptive KLT approach for speech enhancement. IEEE Trans. Speech Audio Process. 9 (2), 87–95.

Vetter, R., Virag, N., Renevey, P., Vesin, J., 1999. Single channel speech enhancement using principal component analysis and mdl subspace selection. In: Proc. EUROSPEECH. pp. 2411–2414.

Yang, B., 1995. Projection approximation subspace tracking. IEEE Trans. Signal Process. 43 (1), 95–107.

Yang, W., Benbouchta, M., Yantorno, R., 1998. Performance of a modified bark spectral distortion measure as an objective speech quality measure. In: Proc. ICASSP. Seattle, USA, pp. 541–544.

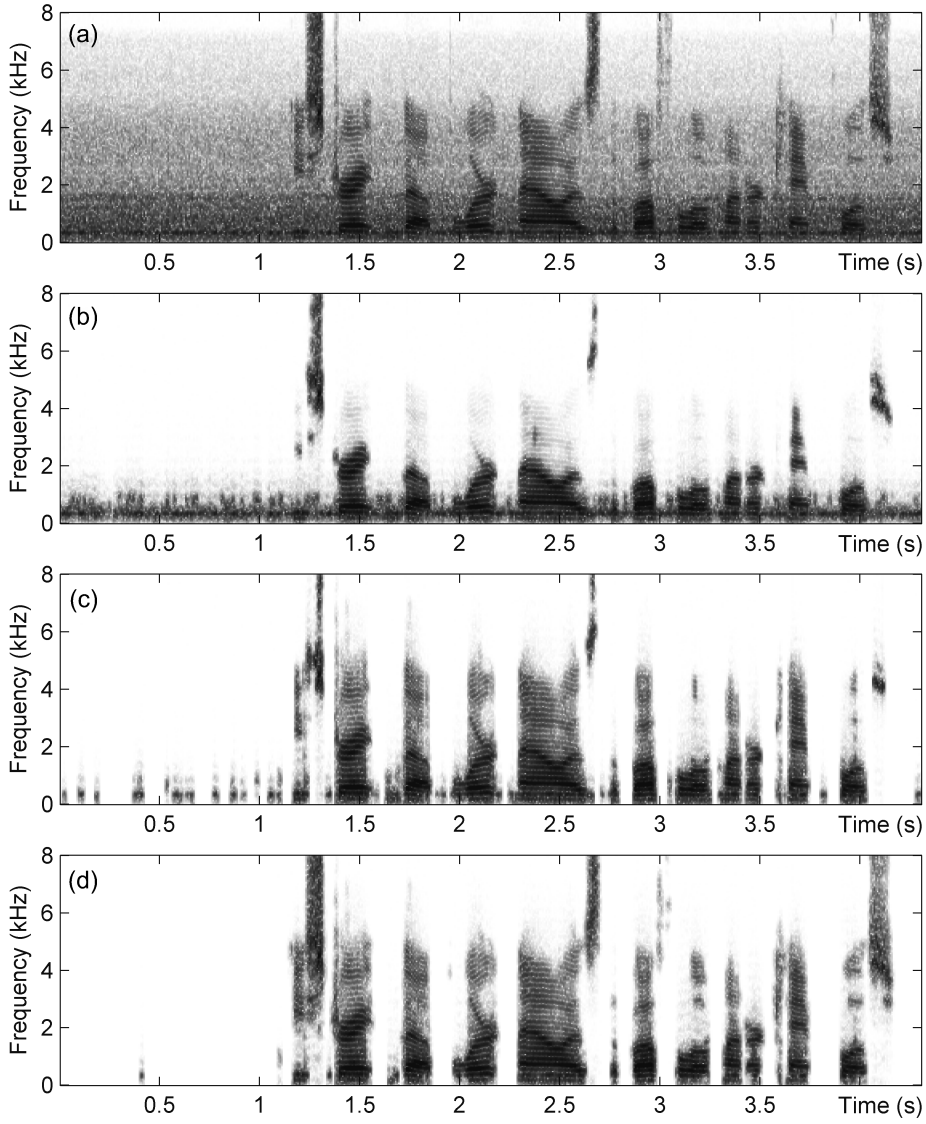


Figure 6: Speech spectrograms: (a) noisy speech signal (car-engine noise at SegSNR = 5 dB), enhanced signals using (b) SDCw, (c) PCSSa, (d) PCSS.

Highlights:

- Extended spectral-domain-constrained estimator for speech enhancement is considered.
- Lagrange multipliers are directly computed allowing precise residual noise shaping.
- The residual noise is shaped optimally according to the masking phenomenon.
- A novel perceptually constrained signal-subspace (PCSS) method is proposed.
- Finally comparative evaluation of the selected subspace-based methods is performed.