Borowicz *EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:5
DOI 10.1186/s13636-015-0051-z

EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

A signal subspace approach to spatio-temporal prediction for multichannel speech enhancement

Adam Borowicz

Abstract

The spatio-temporal-prediction (STP) method for multichannel speech enhancement has recently been proposed. This approach makes it theoretically possible to attenuate the residual noise without distorting speech. In addition, the STP method depends only on the second-order statistics and can be implemented using a simple linear filtering framework. Unfortunately, some numerical problems can arise when estimating the filter matrix in transients. In such a case, the speech correlation matrix is usually rank deficient, so that no solution exists. In this paper, we propose to implement the spatio-temporal-prediction method using a signal subspace approach. This allows for nullifying the noise subspace and processing only the noisy signal in the signal-plus-noise subspace. As a result, we are able to not only regularize the solution in transients but also to achieve higher attenuation of the residual noise. The experimental results also show that the signal subspace approach distorts speech less than the conventional method.

Keywords: Signal subspace; Spatio-temporal prediction; Speech enhancement

Introduction

Speech enhancement is important for many applications including mobile communications, speech coding, speech recognition, and hearing aids. The traditional objective of multichannel speech enhancement is to recover the source speech signal from the outputs of an array of microphones. It is usually achieved by using the beamforming techniques [1-3]. The key idea of beamforming is to process signals of a microphone array, so as to extract the sounds that come from only one direction. In this way, it is possible to dereverberate speech, but the background noise can be reducted as well by avoiding noise directions. Unfortunately, in order to work reasonably well in a reverberant environment, these techniques usually require knowing the impulse responses of the acoustic room or their relative ratios. These parameters can be fixed, provided the geometry of the microphone array is known, or estimated adaptively [4], which in general is a difficult task, however.

Recently, the objective of multichannel speech enhancement has been reformulated, so that noise reduction can be achieved without dereverberating speech. In opposition to the beamforming techniques, the knowledge about the geometry of the microphone array is not required, and the optimal filter depends only on the second-order statistics of the noisy signal.

In [5], the authors presented the most common techniques of multichannel noise reduction based on linear filtering. In such solutions, the noise-free speech is estimated by a linear transformation of the observation vector. The simplest approach is to minimize the mean square error (MSE) between the noise-free and filtered speech signals at a given microphone, which leads to a multichannel version of the classical Wiener filter. In this case, some noise is reduced at the cost of the increased speech distortion, but we cannot explicitly control the trade-off between these quantities.

Speech estimation can also be considered as a constrained optimization problem, where the speech distortions are minimized subject to the residual noise power. This approach is used by the single-channel methods [6] and was implemented in a similar way using a signal subspace technique in [5]. Unlike the frequency

Correspondence: a.borowicz@pb.edu.pl
Department of Computer Graphics and Digital Media, Faculty of Computer Science, Bialystok University of Technology, Wiejska str. 45A, 15-351 Bialystok, Poland

domain methods, which are based on the discrete Fourier transform (DFT), the signal subspace approach decomposes the vector space of noisy signals into the speech-plus-noise subspace and noise-only subspace using the Karhunen-Loeve transform (KLT). Then, spectral weighting is performed only in the signal-plus-noise subspace. The components projected onto the noise-only subspace are simply nullified, which results in significantly better performance when compared to the conventional DFT-based methods, where the full-band (and thus erroneous) spectrum must be processed. Unfortunately, also in this case, it is impossible to reduce the residual noise without introducing speech distortions. Several single-channel approaches [7-9] that exploit the masking effects are known to make the speech distortion or the residual noise inaudible, but introducing psychoacoustics into multi-channel speech enhancement is a challenging task. On the other hand, some hearing properties have been introduced in a beamforming technique [10], but the resulting improvement is not as great as in the single-channel case.

It seems that the major limitation of all these methods is that they use only temporal prediction. In fact, spatial correlations are implicitly embedded in the second-order statistics, or inter-channel correlation matrices, but are not explicitly used. Therefore, in [11,12], the authors proposed a novel technique based on the spatio-temporal prediction (STP). A DFT-based implementation of this technique has also been proposed [13,14], but in this case, the algorithm has been restricted to use only spatial prediction. It has been verified experimentally that the STP approach outperforms the classical beamforming techniques in terms of noise reduction [11]. In [5], it was proved analytically that by using the STP method, it is theoretically possible to reduce the residual noise without distorting the speech. However, a major drawback of the STP method is its numerical instability, as this approach assumes that speech correlation matrix is of full rank. Because this is not true for low power speech at transients, the solution must be regularized empirically in practice. Alternatively, under the uncertainty about the speech presence, the conditional estimators can be used [15]. Even if the speech correlation matrix is of full rank, the STP method requires many microphones to effectively reduce the residual noise.

In this paper, we propose a signal-subspace implementation of the STP method. By decomposing the signal vector space, we are able to limit processing to the signal-plus-noise subspace only. Thus, the numerical problems can be evaded in a more natural way. Since the noisy speech projected on the noise-only subspace can simply be nullified, the signal subspace approach allows for attenuating noise more, even for a small number of microphones. In addition, we have rederived the STP method using a notation slightly different from that in [5], in order to expose the possibility of denoising all microphone signals at once.

Signal model and linear filtering

Let us consider an array of N microphones with arbitrary geometry and a single speech source $s(k)$ located inside a reverberant enclosure. The observation signal at the nth microphone is given by:

$$y_n(k) = a_n(k) * s(k) + v_n(k) = x_n(k) + v_n(k), \tag{1}$$

where $*$ denotes convolution, a_n is the acoustic impulse response from the source to the nth microphone, and $x_n(k)$ and $v_n(k)$ are, respectively, the noise-free speech and the noise components received by the nth microphone. Such a mixing model is illustrated in Figure 1.

Usually data are processed in L-sample blocks. Thus, the signals can be represented using the vector-matrix notation as follows:

$$\mathbf{y}_n(k) = \left[y_n(k)\; y_n(k-1)\; \ldots\; y_n(k-L+1)\right]^T. \tag{2}$$

The estimate of the noise-free speech at the nth microphone can be obtained using a linear transformation of the observation vector:

$$\hat{\mathbf{x}}_n(k) = \mathbf{H}_n\mathbf{y}(k) = \mathbf{H}_n\left[\mathbf{x}(k) + \mathbf{v}(k)\right], \tag{3}$$

where:

$$\begin{aligned} \mathbf{y}(k) &= \left[\mathbf{y}_1^T(k)\; \mathbf{y}_2^T(k)\; \ldots\; \mathbf{y}_N^T(k)\right]^T, \\ \mathbf{x}(k) &= \left[\mathbf{x}_1^T(k)\; \mathbf{x}_2^T(k)\; \ldots\; \mathbf{x}_N^T(k)\right]^T, \\ \mathbf{v}(k) &= \left[\mathbf{v}_1^T(k)\; \mathbf{v}_2^T(k)\; \ldots\; \mathbf{v}_N^T(k)\right]^T. \end{aligned} \tag{4}$$

The vectors $\mathbf{x}_n(k)$ and $\mathbf{v}_n(k)$ denote the noise-free speech and the noise, respectively, and are defined similarly to Equation 2. $\mathbf{H}_n$ is a filtering matrix of size $L \times LN$. The estimation error is defined by:

$$\begin{aligned} \mathbf{e}(k) &= \hat{\mathbf{x}}_n(k) - \mathbf{x}_n(k) \\ &= \underbrace{(\mathbf{H}_n - \mathbf{U}_n)\mathbf{x}(k)}_{\mathbf{e}_x(k)} + \underbrace{\mathbf{H}_n\mathbf{v}(k)}_{\mathbf{e}_v(k)}, \end{aligned} \tag{5}$$

where:

$$\mathbf{U}_n = \left[\mathbf{0}_{L\times(n-1)L}\; \mathbf{I}_L\; \mathbf{0}_{L\times(N-n)L}\right], \tag{6}$$

is a selection matrix of size $L \times LN$. The terms $\mathbf{e}_x(k)$ and $\mathbf{e}_v(k)$ denote the speech distortion and the residual noise, respectively.

For completeness, we also define the correlation matrix of an arbitrary vector $\mathbf{a}$ as:

$$\mathbf{R}_{\mathbf{aa}}(k) = E\left\{\mathbf{a}(k)\mathbf{a}^T(k)\right\}, \tag{7}$$

where $E\{.\}$ is the expectation operator. Assuming that the speech and noise are short-term stationary

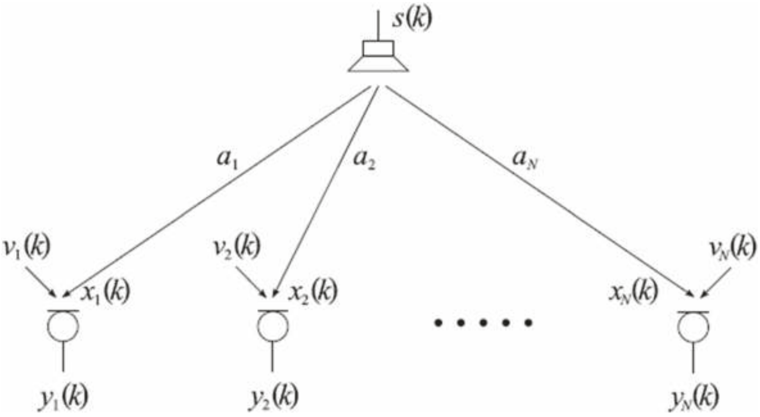

Figure 1 Multi-microphone signal model.

and uncorrelated processes, the correlation matrix of the noisy speech can be written as:

$$\mathbf{R}_{\mathbf{yy}}(k) = \mathbf{R}_{\mathbf{xx}}(k) + \mathbf{R}_{\mathbf{vv}}(k). \tag{8}$$

Unless otherwise stated, all equations hold for any arbitrarily chosen point in time. Therefore, for the sake of brevity, the time index k is often omitted in the rest of this paper.

Spatio-temporal prediction

The STP method is based on the assumption that the microphone signals can be predicted not only in the time domain but also in the space domain [11]. In particular, the signal $\mathbf{x}_m(k)$ can be predicted from the signal $\mathbf{x}_n(k)$ using a linear filter matrix $\mathbf{W}_{n,m}$ such that:

$$\mathbf{x}_m(k) = \mathbf{W}_{n,m}^T\mathbf{x}_n(k), \quad m = 1, 2, 3, \ldots, N, \tag{9}$$

with $\mathbf{W}_{n,n} = \mathbf{I}_L$. The prediction matrices can be concatenated so as to form the $L \times NL$ matrix:

$$\mathbf{W}_n = \left[\mathbf{W}_{n,1}\ \mathbf{W}_{n,2}\ \ldots\ \mathbf{W}_{n,N}\right], \tag{10}$$

and:

$$\mathbf{x}(k) = \mathbf{W}_n^T\mathbf{x}_n(k). \tag{11}$$

By substituting Equation 11 into Equation 5 and assuming that $\mathbf{H}_n\mathbf{W}_n^T = \mathbf{I}_L$, we can deduce that the residual noise can be minimized without distorting speech. Thus, the constrained optimization problem is formulated as follows:

$$\min_{\mathbf{H}_n} \mathrm{tr}\left\{E\left[\mathbf{e}_v(k)\mathbf{e}_v^T(k)\right]\right\} \text{ subject to } \mathbf{H}_n\mathbf{W}_n^T = \mathbf{I}_L. \tag{12}$$

The optimal filter matrix is found using the Lagrange multipliers method:

$$\mathbf{H}_n = \left(\mathbf{W}_n\mathbf{R}_{\mathbf{vv}}{}^{-1}\mathbf{W}_n^T\right)^{-1}\mathbf{W}_n\mathbf{R}_{\mathbf{vv}}{}^{-1}. \tag{13}$$

A solution exists if and only if $\mathbf{R}_{\mathbf{vv}}$ is positive definite, and the matrix $\mathbf{W}_n$ is of rank L. As noise signals are usually stationary and have smooth spectra, $\mathbf{R}_{\mathbf{vv}}$ has full rank and can be estimated using long-term averaging during speech pauses.

Unfortunately the prediction matrices $\mathbf{W}_{n,m}$ for $m \neq n$ are not known and have to be estimated. They can be found by solving the following minimization problem:

$$\min_{\mathbf{W}_{n,m}} E\left\{\left[\mathbf{x}_m(k) - \mathbf{W}_{n,m}^T\mathbf{x}_n(k)\right]^T\left[\mathbf{x}_m(k) - \mathbf{W}_{n,m}^T\mathbf{x}_n(k)\right]\right\}. \tag{14}$$

whose solution is given by:

$$\mathbf{W}_{n,m}^T = \mathbf{R}_{\mathbf{x}_m\mathbf{x}_n}\mathbf{R}_{\mathbf{x}_n\mathbf{x}_n}^{-1} \tag{15}$$

where $\mathbf{R}_{\mathbf{a}_i\mathbf{a}_j}$ stands for the (i,j)th $L \times L$ submatrix of the matrix $\mathbf{R}_{\mathbf{aa}}$. The correlation matrices of the clean speech are unknown, and the vectors $\mathbf{x}_n(k)$ cannot be observed directly, but by using Equation 8 we can write:

$$\mathbf{R}_{\mathbf{x}_n\mathbf{x}_m} = \mathbf{R}_{\mathbf{y}_n\mathbf{y}_m} - \mathbf{R}_{\mathbf{v}_n\mathbf{v}_m}, \quad m = 1, 2, \ldots, N. \tag{16}$$

Thus, finally, we obtain the following expression for the prediction matrices:

$$\mathbf{W}_{n,m}^T = \left(\mathbf{R}_{\mathbf{y}_m\mathbf{y}_n} - \mathbf{R}_{\mathbf{v}_m\mathbf{v}_n}\right)\left(\mathbf{R}_{\mathbf{y}_n\mathbf{y}_n} - \mathbf{R}_{\mathbf{v}_n\mathbf{v}_n}\right)^{-1}. \tag{17}$$

In order to obtain a full rank matrix $\mathbf{W}_{n,m}$, the matrices $\mathbf{R}_{\mathbf{x}_m\mathbf{x}_n}$ and $\mathbf{R}_{\mathbf{x}_n\mathbf{x}_n}$ have to be positive definite. In [5], the authors suggest to estimate the filter matrix (Equation 13) only when the speech source is active, using a voice activity detector (VAD), but this generally does not prevent the matrix $\mathbf{W}_{n,m}$ from being rank deficient. Moreover, such a technique can introduce discontinuity effects at transients or/and increased residual noise during silence intervals. For low-power speech signals, the covariance matrix of the clean speech is usually positive semi-definite, or at least ill-conditioned, which means that in practice the STP method is numerically stable only for high signal-to-noise ratios (SNRs). The simplest solution is to add some white noise to the speech signal, so that the inverses in

Equation 13 and Equation 17 can be replaced with pseudoinverses and properly regularized [16]. However, all these approaches are rather empirical and need a careful adjustment. Thus, we need a more robust solution, which can be applied also to low power speech signals, especially at low SNRs.

Signal subspace approach

In the conventional STP method, data are processed in the vector space of the noisy speech. The key idea of the signal subspace approach is to decompose that vector space into the signal-plus-noise and noise-only subspaces and to process data only in the signal-plus-noise subspace, while the projection of the noisy signal onto the noise-only subspace is simply nullified. The dimensionality of the signal-plus-noise or, simply, signal subspace is closely related to the rank of the speech correlation matrix. Thus, by introducing the signal subspace approach to the STP method, we are able to not only increase the attenuation of the residual noise during silence intervals but also to avoid the ill-conditioning issues.

Let us rewrite Equation 13 more compactly. Please notice that the prediction matrix can be alternatively written as:

$$\mathbf{W}_n = \mathbf{R}^{-1}_{\mathbf{x}_n\mathbf{x}_n}\mathbf{U}_n\mathbf{R}_{\mathbf{xx}}, \tag{18}$$

and then, by substituting the above into Equation 13, we obtain:

$$\mathbf{H}_n = \mathbf{R}_{\mathbf{x}_n\mathbf{x}_n}\left(\mathbf{U}_n\mathbf{R}_{\mathbf{xx}}\mathbf{R}^{-1}_{vv}\mathbf{R}_{\mathbf{xx}}\mathbf{U}^T_n\right)^{-1}\mathbf{U}_n\mathbf{R}_{\mathbf{xx}}\mathbf{R}^{-1}_{vv}. \tag{19}$$

Since $\mathbf{R}_{\mathbf{vv}}$ is positive definite, the matrices $\mathbf{R}_{\mathbf{xx}}$ and $\mathbf{R}_{\mathbf{vv}}$ can be jointly diagonalized [17,18], i.e.:

$$\mathbf{R}^{-1/2}_{\mathbf{vv}}\mathbf{R}_{\mathbf{xx}}\mathbf{R}_{\mathbf{vv}}{}^{-1/2} = \mathbf{V}\Lambda\mathbf{V}^T, \tag{20}$$

where $\mathbf{V}$ denotes the orthogonal matrix of the eigenvectors, and $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_{NL}\}$ is the diagonal matrix of the corresponding eigenvalues. We also assume that the eigenvalues in Λ are arranged in descending order, i.e. $\lambda_i \geq \lambda_j$ for any $i < j$. The matrix $\mathbf{V}$ can also be interpreted as the KLT matrix of the whitened clean speech. Alternatively, it can be obtained using the eigendecomposition of the whitened noisy speech correlation matrix:

$$\mathbf{R}^{-1/2}_{\mathbf{vv}}\mathbf{R}_{\mathbf{yy}}\mathbf{R}_{\mathbf{vv}}{}^{-1/2} = \mathbf{V}\left(\Lambda + \mathbf{I}\right)\mathbf{V}^T. \tag{21}$$

As shown in [17], the vector space of the noisy speech can be decomposed using the square matrix:

$$\mathbf{B} = \mathbf{V}^T\mathbf{R}^{1/2}_{\mathbf{vv}} \tag{22}$$

which has full rank but is not necessarily orthogonal. Please notice that applying $\mathbf{B}^{-T}$ to the noisy signal is equivalent to whitening data before performing the subspace decomposition, so that the resulting coefficients are perfectly decorrelated in the transform domain, i.e.:

$$E\left[\tilde{\mathbf{y}}(k)\tilde{\mathbf{y}}^T(k)\right] = \Lambda + \mathbf{I}, \tag{23}$$

where $\tilde{\mathbf{y}}(k) = \mathbf{B}^{-T}\mathbf{y}(k)$. Thus, our correlation matrices can be expressed as follows:

$$\begin{aligned}\mathbf{R}_{\mathbf{yy}} &= \mathbf{B}^T(\Lambda + \mathbf{I})\mathbf{B}\\ \mathbf{R}_{\mathbf{xx}} &= \mathbf{B}^T\Lambda\mathbf{B}\\ \mathbf{R}_{\mathbf{vv}} &= \mathbf{B}^T\mathbf{B}\end{aligned} \tag{24}$$

Let $\mathbf{Q}_n = \Lambda\mathbf{B}\mathbf{U}^T_n$. Substituting the relations given in Equation 24 into Equation 19 results in the optimal filter matrix:

$$\mathbf{H}_n = \mathbf{U}_n\mathbf{B}^T\left[\mathbf{Q}_n\left(\mathbf{Q}^T_n\mathbf{Q}_n\right)^{-1}\mathbf{Q}^T_n\right]\mathbf{B}^{-T}. \tag{25}$$

Since $\mathbf{R}_{\mathbf{vv}}$ is positive definite, and $\mathbf{R}_{\mathbf{xx}}$ can be semi-positive definite, the dimension of the signal-plus-noise subspace is equal to the number of non-zero eigenvalues of the correlation matrix of the whitened clean speech. Assume that $NL = L_s + L_v$, where L_s and L_v denote the dimensions of the signal-plus-noise and noise-only subspaces, respectively. Thus, for $L_s < NL$, we can rewrite Equation 25 as follows:

$$\mathbf{H}_n = \mathbf{U}_n\mathbf{B}^T\begin{bmatrix}\Sigma_n & \mathbf{0}_{L_s\times L_v}\\ \mathbf{0}_{L_v\times L_s} & \mathbf{0}_{L_v\times L_v}\end{bmatrix}\mathbf{B}^{-T}, \tag{26}$$

where:

$$\Sigma_n = \mathbf{Q}_{n,1:L_s}\left[\mathbf{Q}^T_{n,1:L_s}\mathbf{Q}_{n,1:L_s}\right]^{-1}\mathbf{Q}^T_{n,1:L_s} \tag{27}$$

can be viewed as a reweighting matrix, with $\mathbf{Q}_{n,1:L_s}$ denoting sub-matrix of $\mathbf{Q}_n$ consisting rows from 1 to L_s. As can be seen the noisy signal is transformed using a non-orthogonal matrix $\mathbf{B}^{-T}$. The denoising is achieved by 'reweighting' the coefficients in the signal-plus-noise subspace using the matrix Σ_n and simply nullifying the noise-only subspace. In opposition to the conventional signal subspace approach, the reweighting matrix is not diagonal here but symmetric and idempotent.

Finally, the filtered signal is brought back to the time domain using the inverse transform $\mathbf{B}^T$.

In practice, L_s can be estimated as the number of the strictly positive eigenvalues, according to the following rule:

$$L_s \approx \underset{1\geq l\geq NL}{\arg\max}\ \{\lambda_l > \theta\}, \tag{28}$$

where the threshold θ is a some small positive number.

It can be noticed that $\mathbf{Q}^T_n\mathbf{Q}_n$ is invertible as long as $L_s \geq L$. However, even when this condition is not in force (which is fairly common at transients or during silence intervals), the inverse can be easily regularized. For example, if $L_s = L$, $\mathbf{Q}_{n,1:L}$ is a square matrix, and

$\Sigma_n = \mathbf{I}$, which means that the filter performs nullifying the noise subspace without cleaning the signal-plus-noise subspace, or that the residual noise can be effectively reduced without distorting the speech.

Therefore, in order to regularize the solution, the best we can do is to use the following rule:

$$\Sigma_n = \begin{cases} \mathbf{Q}_{n,1:L_s}\left[\mathbf{Q}_{n,1:L_s}^T \mathbf{Q}_{n,1:L_s}\right]^{-1}\mathbf{Q}_{n,1:L_s}^T, & L_s > L \\ \mathbf{I}_{L_s}, & \text{otherwise.} \end{cases} \tag{29}$$

Please also notice that if $N = 1$ and $L_s = L$, then the filter matrix is simply the identity matrix. For $N > 1$, it is possible to arrange matrices $\mathbf{H}_n$, $n = 1, 2, \ldots, N$ into the single filter matrix:

$$\mathbf{H}_\mathrm{P} = \left[\mathbf{H}_1^T \ \mathbf{H}_2^T \ \ldots \ \mathbf{H}_N^T\right], \tag{30}$$

which can be used to estimate all noise-free microphone signals at once. Namely, the vector $\mathbf{x}(k)$ can be estimated as follows:

$$\mathbf{x}(k) \approx \hat{\mathbf{x}}(k) = \mathbf{H}_\mathrm{P}\mathbf{y}(k). \tag{31}$$

The filter matrix $\mathbf{H}_\mathrm{P}$ can also be written in a more convenient form:

$$\mathbf{H}_\mathrm{P} = \left[\mathbf{U} \circ \left(\mathbf{B}^T \Lambda \mathbf{B}\right)\right]\left[\mathbf{U} \circ \left(\mathbf{B}^T \Lambda^2 \mathbf{B}\right)\right]^{-1} \mathbf{B}\Lambda\mathbf{B}^{-T}, \tag{32}$$

where:

$$\mathbf{U} = \mathbf{I}_N \otimes \mathbf{J}_{L\times L}, \tag{33}$$

and the operators $\circ$ and $\otimes$ stand for the Hadamard and the Kronecker products, respectively, and $\mathbf{J}_{L\times L}$ is the $L \times L$ matrix of ones.

The proposed approach can be verified analytically in terms of noise reduction and speech distortion. The noise reduction factor can be defined for any filter matrix $\mathbf{H}_n$ as follows:

$$\xi_\mathrm{nr}\left(\mathbf{H}_n\right) = \frac{\mathrm{tr}\left\{E\left[\mathbf{U}_n \mathbf{v}\mathbf{v}^T \mathbf{U}_n^T\right]\right\}}{\mathrm{tr}\left\{E\left[\mathbf{H}_n \mathbf{v}\mathbf{v}^T \mathbf{H}_n^T\right]\right\}} = \frac{\mathrm{tr}\left\{\mathbf{U}_n \mathbf{R}_{\mathbf{vv}} \mathbf{U}_n^T\right\}}{\mathrm{tr}\left\{\mathbf{H}_n \mathbf{R}_{\mathbf{vv}} \mathbf{H}_n^T\right\}}. \tag{34}$$

It is expected that $\xi_\mathrm{nr}(k) \geq 1$: the larger this factor, the lower residual noise. Usually, the noise is reduced at the cost of attenuating speech. Therefore, in order to quantify this attenuation, we define the speech reduction factor:

$$\xi_\mathrm{sr}\left(\mathbf{H}_n\right) = \frac{\mathrm{tr}\left\{E\left[\mathbf{U}_n \mathbf{x}\mathbf{x}^T \mathbf{U}_n^T\right]\right\}}{\mathrm{tr}\left\{E\left[\mathbf{H}_n \mathbf{x}\mathbf{x}^T \mathbf{H}_n^T\right]\right\}} = \frac{\mathrm{tr}\left\{\mathbf{U}_n \mathbf{R}_{\mathbf{xx}} \mathbf{U}_n^T\right\}}{\mathrm{tr}\left\{\mathbf{H}_n \mathbf{R}_{\mathbf{xx}} \mathbf{H}_n^T\right\}} \tag{35}$$

and expect $\xi_\mathrm{sr}(\mathbf{H}_n) \geq 1$. The output SNR of the filter $\mathbf{H}$ can be expressed in the following way:

$$\mathrm{SNR}(\mathbf{H}) = \frac{\mathrm{tr}\left\{\mathbf{H}_n \mathbf{R}_{\mathbf{xx}} \mathbf{H}_n^T\right\}}{\mathrm{tr}\left\{\mathbf{H}_n \mathbf{R}_{\mathbf{vv}} \mathbf{H}_n^T\right\}} = \mathrm{SNR}\frac{\xi_\mathrm{nr}(\mathbf{H})}{\xi_\mathrm{sr}(\mathbf{H})}, \tag{36}$$

where the SNR stands for the input SNR.

For $L_s \geq L$, the proposed approach is theoretically equivalent to the time-domain implementation of the STP method. In order to analyse performance of the proposed implementation for $L_s < L$, we consider the case of the white noise, for which $\mathbf{R}_{\mathbf{v}_n\mathbf{v}_n} = \sigma_{\mathbf{v}_n}\mathbf{I}$. Because the inverse $\left(\mathbf{Q}_n^T\mathbf{Q}_n\right)^{-1}$ does not exist for $L_s < L$, we use Equation 29. Then, by replacing Σ_n in Equation 26 with the identity matrix and by substituting it to Equation 34 and Equation 35, we obtain:

$$\xi_\mathrm{nr}(\mathbf{H}_n) = \frac{L}{\sum_{i=1}^{L}\sum_{j=1}^{L_s}\mathbf{V}_{(n-1)L+i,j}^2} > 1 \tag{37}$$

$$\xi_\mathrm{sr}(\mathbf{H}_n) = 1. \tag{38}$$

Since $\xi_\mathrm{nr}(\mathbf{H}_n) > \xi_\mathrm{sr}(\mathbf{H}_n)$, we always have $\mathrm{SNR}(\mathbf{H}) > \mathrm{SNR}$, or an improvement of the SNR.

Simulations

Although a full evaluation of the proposed approach, including listening tests, is out of the scope of this article, we have conducted some experiments using objective measurements. In this section, we compare the performances of the conventional time-domain implementation of the STP method and of the proposed approach based on the signal subspace.

Implementation

Both methods have been implemented in MATLAB. Instead of recalculating the filter from sample to sample, we collect the microphone recordings in overlapped buffers and process them frame-by-frame in a similar way as in [8] or [19]. Namely, we divide the microphone signals into frames of length N_f with 50% overlap. Each frame is partitioned into $M = N_f - L + 1$ shorter overlapping L-dimensional vectors. The sequence of these vectors is arranged into the trajectory matrix of size L-by-M. The trajectory matrices for all microphones are concatenated together so as to form the noisy speech matrix $\mathbf{Y}(k)$ of size LN-by-M so that:

$$\mathbf{Y}(k) = \left[\,\mathbf{y}(k) \ \mathbf{y}(k-1) \ \cdots \ \mathbf{y}(k-M+1)\,\right]. \tag{39}$$

As all required parameters are estimated, the effective filter matrix $\mathbf{H}_n$ is computed, and then all in-frame vectors are processed using the same matrix, i.e. $\hat{\mathbf{Y}}(k) = \mathbf{H}_n\mathbf{Y}(k)$. The enhanced vectors are obtained from the matrix $\hat{\mathbf{Y}}(k)$ using the diagonal averaging technique [19]. Finally, the frames are multiplied by the

Hanning window and synthesized using the overlap-add method.

The correlation matrix of the noisy speech can be estimated according to:

$$\mathbf{R}_{\mathbf{yy}}(k) \approx \frac{1}{MN}\mathbf{Y}(k)\mathbf{Y}(k)^T, \tag{40}$$

being the outer product of the matrix $\mathbf{Y}(k)$. This estimate is the basis for computing both noise statistics and the KLT of the whitened signal (Equation 20). The matrix $\mathbf{R}_{\mathbf{vv}}$ is estimated only during speech pauses as:

$$\mathbf{R}_{\mathbf{vv}}(k) \approx \begin{cases} \alpha\mathbf{R}_{\mathbf{vv}}(k-1) + (1-\alpha)\mathbf{R}_{\mathbf{yy}}(k), & \text{if} \quad I(k) = 1 \\ \mathbf{R}_{\mathbf{vv}}(k-1), & \text{otherwise} \end{cases} \tag{41}$$

where $0 < \alpha < 1$ is the forgetting factor, and $I(k)$ is the VAD output of the kth frame. In our simulations, the VAD was not implemented, and the speech pause/activity regions were marked manually.

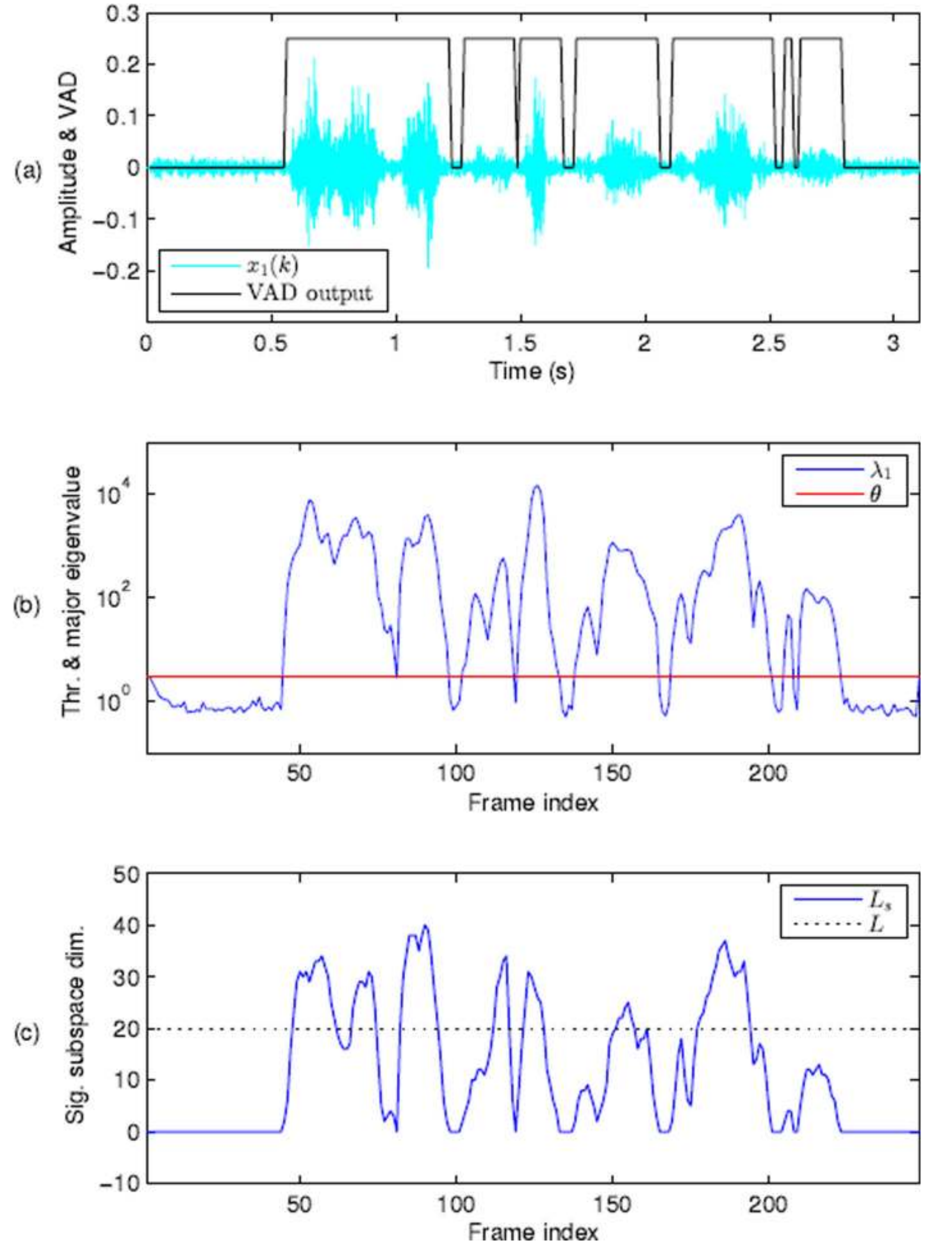

Figure 2 Estimation of the dimension of the signal-plus-noise subspace. (a) Example noisy speech signal at SNR = 10 dB. **(b)** The parameter θ and major eigenvalue of the whitened clean speech. **(c)** Estimate of the dimension of the signal-plus-noise subspace.

In most cases, the noise correlation matrix is positive definite, so that the computations of both whitening and unwhitening transformations ($\mathbf{R}_{\mathbf{vv}}^{-1/2}, \mathbf{R}_{\mathbf{vv}}^{1/2}$, respectively) should be numerically stable. The transformations can be calculated at once using the eigenstructures of the matrix $\mathbf{R}_{\mathbf{vv}} = \mathbf{V}_{\mathbf{v}}\Lambda_{\mathbf{v}}\mathbf{V}_{\mathbf{v}}{}^{T}$ in the following way:

$$\begin{aligned} \mathbf{R}_{\mathbf{vv}}^{-1/2} &= \mathbf{V}_{\mathbf{v}}\Lambda_{\mathbf{v}}^{-1/2}\mathbf{V}_{\mathbf{v}}{}^{T}, \\ \mathbf{R}_{\mathbf{vv}}^{1/2} &= \mathbf{V}_{\mathbf{v}}\Lambda_{\mathbf{v}}^{1/2}\mathbf{V}_{\mathbf{v}}{}^{T}, \end{aligned} \tag{42}$$

where $\mathbf{V}_{\mathbf{v}}$ denotes the orthogonal matrix of the eigenvectors, and $\Lambda_{\mathbf{v}}$ is the diagonal matrix of the corresponding eigenvalues.

In our experiments, we take $\alpha = 0.75$, $N_f = 400$, and $L = 20$. A proper choice of the value of the parameter θ seems to be crucial for the proposed implementation. In general, greater values of θ lead to cancellation of the residual noise, but a special care must be taken because low-power speech components can be also nullified. Therefore, the simplest solution is to fix this threshold, so that it is large enough to give $L_s = 0$ (or equivalently $\theta \gg \lambda_1$) during speech pauses. We found empirically that its value depends mainly on the bias of the estimator of the noise correlation matrix, i.e. on the forgetting factor α and the frame/window size N_f. In Figure 2c, we present the variability of the estimated dimension of the signal-plus-noise subspace for the parameter $\theta = 3$. Further experiments show that the optimal value of the parameter θ (in terms of speech distortion) does not depend on the input SNR. It can be observed that $L_s < L$ occurs fairly commonly, not only at transients, but also during speech activity.

In the case of the conventional implementation, all inverses in Equation 17 and Equation 13 were replaced with pseudoinverses. They were computed using singular value decomposition (SVD), and all singular values less than some tolerance were treated as zeros. In fact, that tolerance plays the same role as the parameter θ in the signal subspace approach. Thus, by setting it sufficiently large, it is possible to increase noise reduction. Unfortunately, the speech reduction factor is also increased. Additionally, we have found empirically that the optimal tolerance is SNR dependent. Therefore, during our simulations, all SVD-based pseudoinverses were computed using the default tolerance set by MATLAB.

Objective evaluation

The acoustic environment was simulated using the image method [20]. We assumed that the enclosure is rectangular with dimensions 6 × 5 × 2.8 (all dimensions and coordinates are in meters). A uniform linear array of eight microphones was placed along the x-axis, with spacing 0.1 and beginning from the first microphone at the position $(2.65, 4, 1)$. The locations of the microphones and the sound sources are shown in Figure 3. The source speech signal was sampled at 16 kHz. The signal was about 14-s-long and comprised of four short sentences uttered by male and female speakers (see Additional file 1). In order to represent general broadband signals the pink noise was chosen. The microphone signals were obtained by convolving the source speech signal with the generated

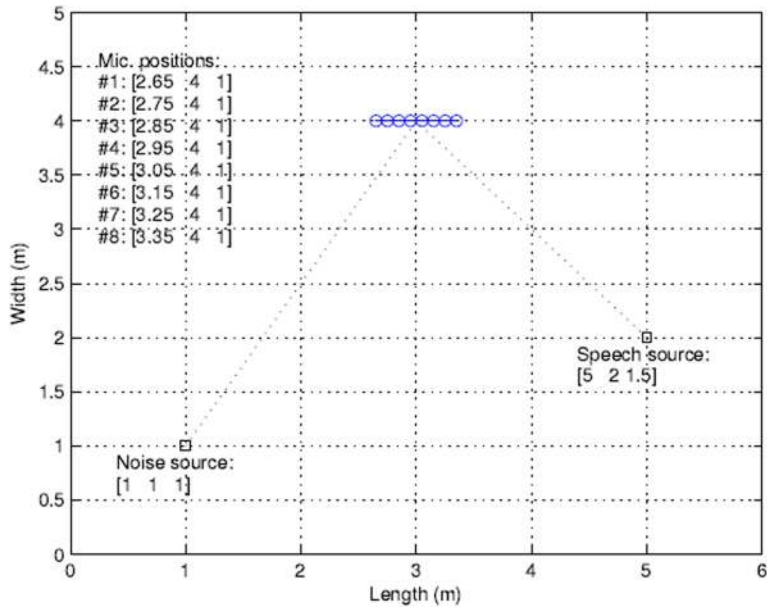

Figure 3 Floor plan of the simulated enclosure (all coordinates in meters).

impulse responses of a room, and by adding noise signals at SNRs ranged from -5 to 20 dB, in accordance with Equation 1. An example noisy speech sample is provided as the Additional file 2. In all experiments, we estimated the noise-free signal only at the first microphone, $n = 1$, which served as the reference microphone.

The SNR-based measures were used for evaluating the objective performance. The speech distortion measure (SD) was defined as the segmental signal-to-noise ratio, in which the noise was identified with the difference between the source signal and enhanced speech. The higher the value of this factor, the better the performance. The amount of reduced noise was measured using the noise attenuation (NA) factor defined as the mean ratio between the input noise power and output noise power.

Firstly, taking into consideration only on the first four microphones, we have evaluated the impact of the

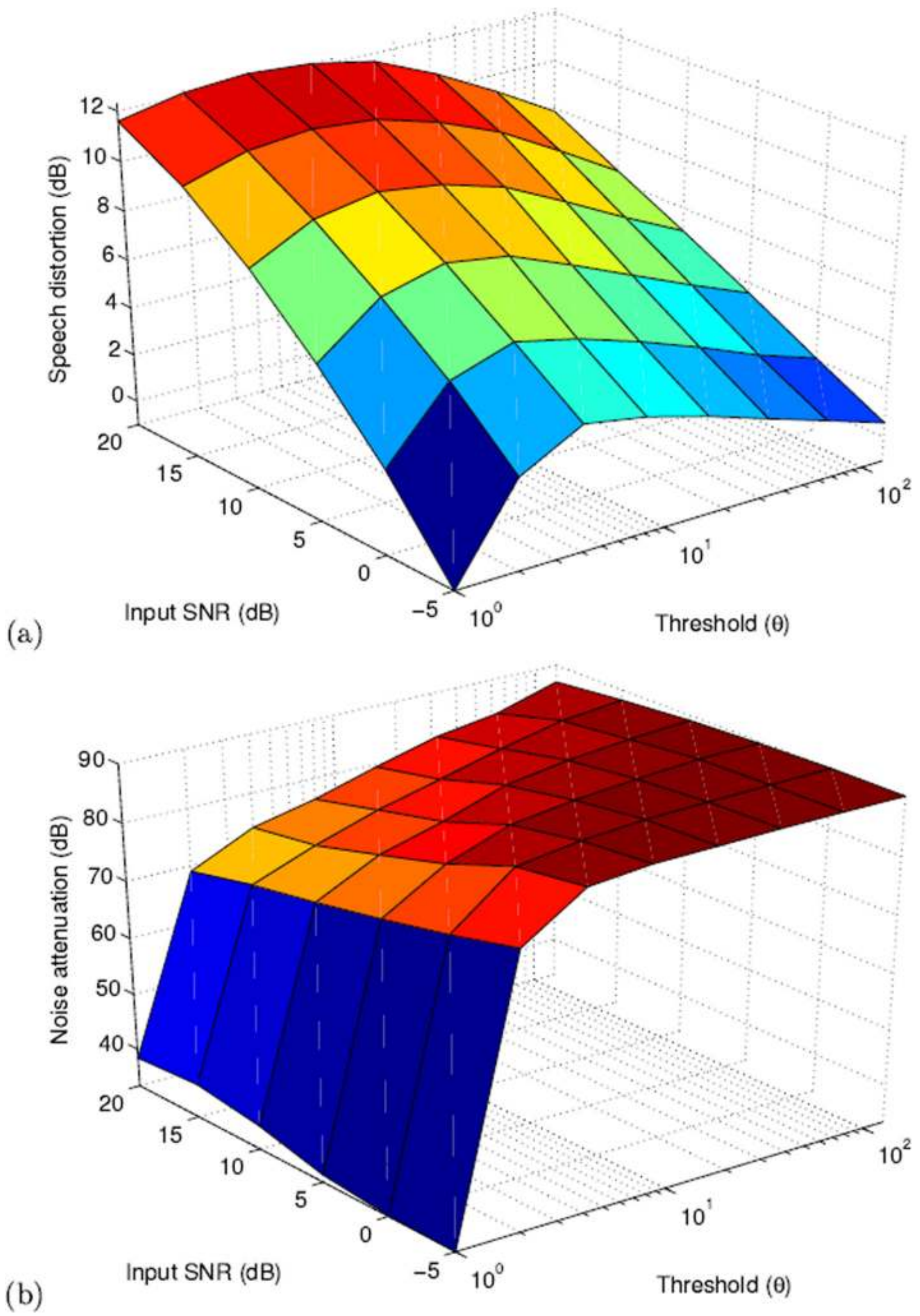

Figure 4 Adjustment of the parameter θ for $N = 4$. (a) Speech distortion measure (SD) and **(b)** noise attenuation factor (NA).

parameter θ on speech distortion and noise attenuation. The measured speech distortion, which is shown in Figure 4a, indicates rather weak influence of the parameter θ on the input SNR. The optimal value of θ is between 3 to 4 for all SNRs. On the other hand, the plot of the noise attenuation factor in Figure 4b, demonstrates that the higher the value of the θ, the higher noise attenuation.

The subsequent simulations were performed for $\theta = 3$ and $N = 2, 3, \ldots, 8$. For conciseness, we present in Figure 5 only the results of objective measurements of the systems with $N = 2, 4$, and eight microphones. Example recordings of the speech enhanced using conventional and proposed method are provided as Additional files 3 and 4, respectively.

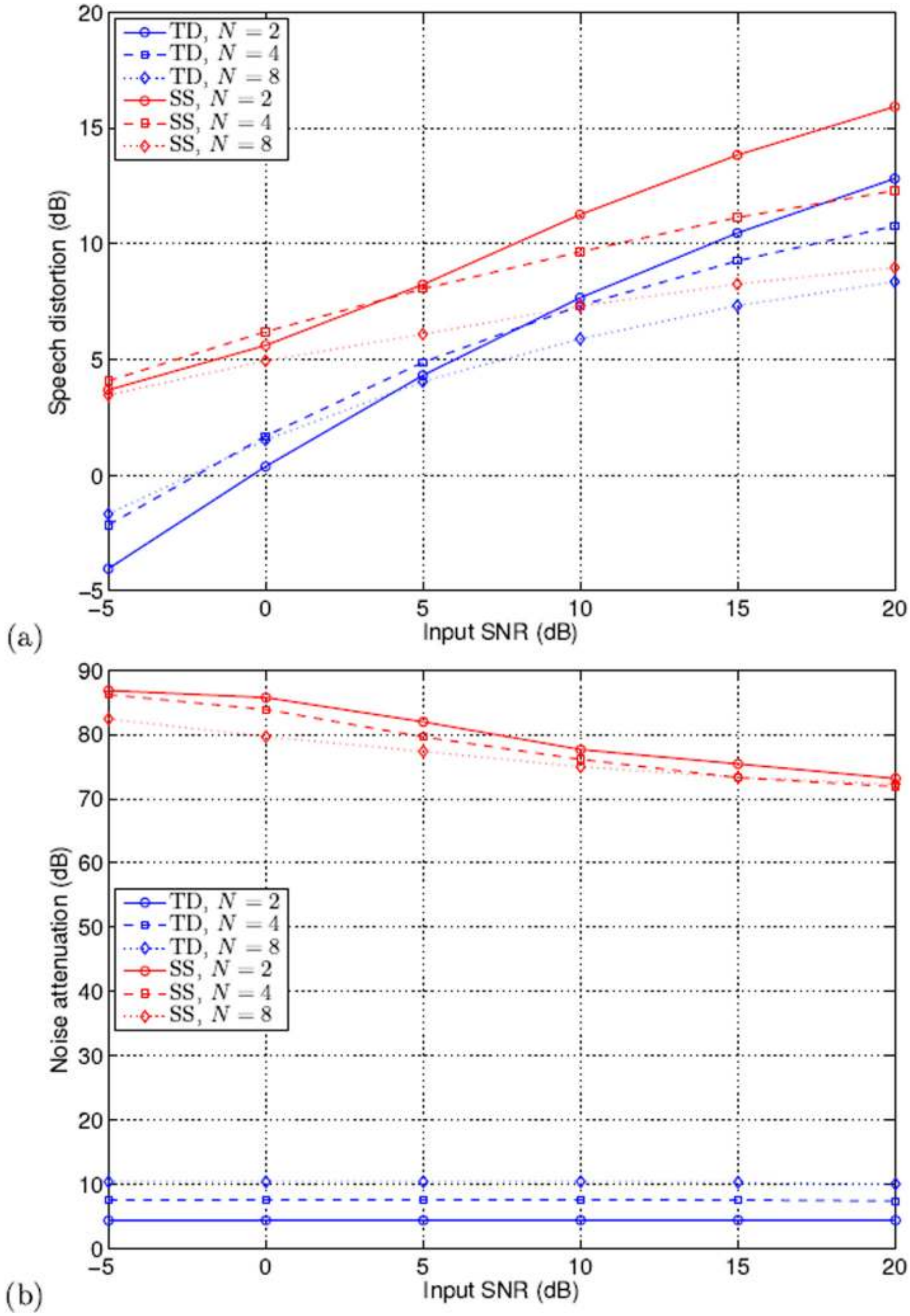

Figure 5 Objective measurement of the time-domain (TD) and signal subspace (SS) implementations. (a) Speech distortion and **(b)** noise attenuation factor.

It can easily be seen that the proposed method outperforms the conventional one, as it provides lower speech distortions and higher noise attenuation. Surprisingly, the speech distortion for the system with $N = 2$ microphones was lower than for the eight-microphone system, especially at high SNRs. A possible explanation of this phenomena is that for more microphones, the correlation matrix is larger, which makes the estimation less accurate. In practice, it makes sense to use more microphones only in the conventional time-domain method (in order to improve the noise attenuation). Figure 5a shows that the speech distortion can be also decreased but only at low SNRs.

Unlike the conventional method, the signal subspace approach does not require many microphones to work

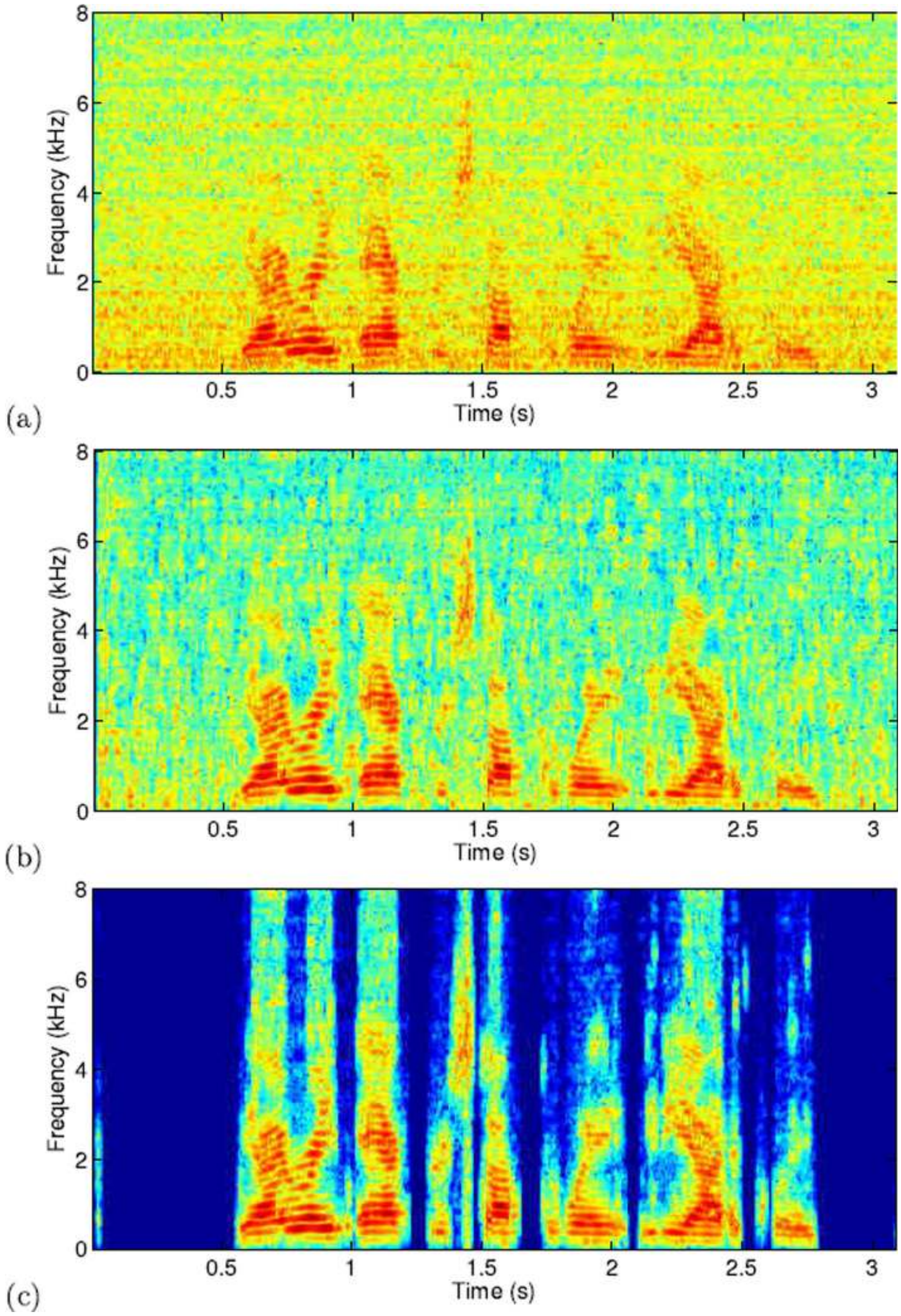

Figure 6 Speech spectrograms. (a) Noisy speech at microphone number 1 (input SNR = 10 dB). **(b)** Speech enhanced with time-domain STP method. **(c)** Speech enhanced with the signal subspace implementation of the STP method.

reasonably well. The proposed method removes the residual noise almost completely (NA $= 70$ to 90 dB) without introducing speech distortions or unnatural discontinuity effects at transients. This is not surprising, since the matrix Σ_n may contain only zeros during silence intervals, which is highly desirable in speech coding or automatic speech recognition (ASR) systems. On the other hand, complete cancellation of the noise is neither necessary nor desired in some applications, like mobile communication. In such cases, zero diagonal coefficients in Σ_n can be replaced with some small positive numbers.

The objective evaluation has been validated using spectrograms. Figure 6a shows the spectrogram of the noisy speech signal recorded at the first microphone, at SNR $=$ 10 dB. The enhancement results for the conventional and proposed methods with $N = 4$ are presented below. Once again, we see that the proposed method offers incomparably higher noise attenuation during both speech pauses and voice activity periods. Unlike the time-domain implementation, the signal subspace approach does not generate musical tones (random peaks in the time-frequency plane). However, one should remember that this is an idealized situation, because the VAD has not been implemented, and speech/pause frames were marked manually. In practice, the VAD is difficult to implement, and its performance generally depends on the input SNR. Therefore, we expect some performance drop in real applications.

Conclusions

We have shown that the STP method can be implemented using a signal subspace approach. The conditions for uniqueness of a solution have been provided. We proposed Equation 29 as a simple rule that can be used when the speech correlation matrix is rank deficient. It has been verified analytically that the proposed approach can reduce noise without distorting the speech (as long as the parameter L_s is not less than the true rank of $\mathbf{R_{yy}}$). In order to estimate the dimension of the speech-plus-noise subspace, we also used some sort of the thresholding technique. However, we have found empirically that, unlike in the conventional SVD-based regularization, a corresponding threshold (or the parameter θ) is not SNR dependent and can be adjusted to fixed value. The objective measurements show that the signal subspace approach outperforms the conventional one providing higher noise attenuation and lower speech distortion. We have also reported that the proposed implementation does not require as many microphones as its time-domain counterpart to work reasonably well.

Listening tests are usually difficult and time-consuming, thus they were not used to evaluate our approach.

In this article, we have introduced a novel notation that allows for estimating the speech signals at all microphones at once. This can potentially be useful if the system has to work as a preprocessor for a beamformer. Since the STP method relies only on the second-order statistics, it may find other applications in areas where multi-sensor data are processed, i.e. in the electroencephalography, as a means for enhancing EEG signals. These points have not been discussed here, but they are promising directions for future work.

Additional files

Additional file 1: Clean speech sample recorded at microphone number 1.

Additional file 2: Noisy speech sample recorded at microphone number 1 (at SegSNR = 0 dB).

Additional file 3: Speech signal enhanced using time-domain STP method.

Additional file 4: Speech signal enhanced using the proposed method.

Competing interests

The author declares that he has no competing interests.

Acknowledgements
This work was supported by the Polish National Science Centre under Decision No. DEC-2012/07/D/ST6/02454.

Received: 23 October 2014 Accepted: 20 January 2015
Published online: 10 February 2015

References

1. OL Frost, An algorithm for linearly constrained adaptive array processing. Proc. IEEE. **60**, pp. 926–935 (1972)
2. LJ Griffiths, CW Jim, An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. **AP-30**(1), 27–34 (1982)
3. S Gannot, D Burshtein, E Winstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001)
4. S Affes, Y Grenier, A signal subspace tracking algorithm for microphone array processing of speech. IEEE Trans. Speech Audio Process. **5**(5), 425–437 (1997)
5. Y Huang, J Benesty, J Chen, Analysis and comparison of multichannel noise reduction methods in a common framework. IEEE Trans. Audio, Speech, Lang. Process. **16**(5), 957–968 (2008)
6. Y Ephraim, HL Van Trees, A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. **3**(4), 251–266 (1995)
7. D Virette, P Scalart, C Lamblin, Analysis of background noise reduction techniques for robust speech coding. Proc. EUSIPCO. **3**, 297–300 (2002)
8. A Borowicz, A Petrovsky, Signal subspace approach for psychoacoustically motivated speech enhancement. Speech Comm. **53**(2), 210–219 (2011)
9. F Jabloun, B Champagne, Incorporating the human hearing properties in the signal subspace approach for speech enhancemnt. IEEE Trans. Speech Audio Process. **11**(6), 700–708 (2003)
10. A Borowicz, A Petrovsky, Incorporating auditory properties into generalised sidelobe canceller. Paper presented at the 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO). Bucharest, Romania, 27–31 August 2012
11. J Chen, J Benesty, Y Huang, A minimum distortion noise reduction algorithm with multiple microphones. IEEE Trans. Audio, Speech, Lang. Process. **16**(3), 481–493 (2008)
12. J Benesty, J Chen, Y Huang, *Microphone Array Signal Processing*. (Springer, Berlin, Germany, 2008)
13. B Cornelis, M Moonen, J Wouters, Comparison of frequency domain noise reduction strategies based on multichannel wiener filtering and spatial prediction. Paper presented at the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei, 19–24 April 2009

14. J Benesty, J Chen, EAP Habets, *Speech Enhancement in the STFT Domain. SpringerBriefs in Electrical and Computer Engineering*. (Springer, Berlin, Germany, 2012)
15. EAP Habets, A distortionless subband beamformer for noise reduction in reverberant environments. Paper presented at the Proc. IWAENC, Tel Aviv. Israel, August 2010
16. PC Hansen, The truncated SVD as a method for regularization. BIT. **27**, 534–553 (1987)
17. Y Hu, PC Loizou, A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans. Speech Audio Process. **11**(4), 334–341 (2003)
18. H Lev-Ari, Y Ephraim, Extension of the signal subspace enhancement to colored noise. IEEE Signal Process. Lett. **10**(4), 104–106 (2003)
19. R Vetter, N Virag, P Renevey, JM Vesin, Single channel speech enhancement using principal component analysis and MDL subspace selection. Paper presented at the Proc. EUROSPEECH. Budapest, Hungary, 5–9 September 1999
20. JB Allen, DA Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943 (1979)