Noise Reduction in the Time Domain using Joint Diagonalization

Nørholm, Sidsel Marie; Benesty, Jacob; Jensen, Jesper Rindom; Christensen, Mads Græsbøll

Published in:
2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

DOI (link to publication from Publisher):
10.1109/ICASSP.2014.6854969

Publication date:
2014

Document Version
Early version, also known as pre-print

NOISE REDUCTION IN THE TIME DOMAIN USING JOINT DIAGONALIZATION

Sidsel Marie Nørholm[1], *Jacob Benesty*[1,2], *Jesper Rindom Jensen* [1], *and Mads Græsbøll Christensen*[1]

[1] Audio Analysis Lab, AD:MT, Aalborg University, {smn, jrj,mgc}@create.aau.dk
[2] INRS-EMT, University of Quebec, benesty@emt.inrs.ca

ABSTRACT

A new filter design based on joint diagonalization of the clean speech and noise covariance matrices is proposed. First, an estimate of the noise is found by filtering the observed signal. The filter for this is generated by a weighted sum of the eigenvectors from the joint diagonalization. Second, an estimate of the desired signal is found by subtraction of the noise estimate from the observed signal. The filter can be designed to obtain a desired trade-off between noise reduction and signal distortion, depending on the number of eigenvectors included in the filter design. This is explored through simulations using a speech signal corrupted by car noise, and the results confirm that the output signal-to-noise ratio and speech distortion index both increase when more eigenvectors are included in the filter design.

Index Terms— Noise reduction, speech enhancement, single channel, time-domain filtering, joint diagonalization.

1. INTRODUCTION

Noise reduction is known to be a very difficult problem in speech applications and has been studied for more than five decades. More than ever, good noise reduction algorithms are required with the revolution of electronic communications. The methods used so far for solving the task of noise reduction can be categorized in three major groups: spectral-subtraction, subspace, and statistical-model-based methods [1]. The first group of methods work by subtracting the spectrum of the noise from the spectrum of the noisy signal (see, e.g, [2–4]). They are some of the first methods used, but one of their major drawbacks is that they introduce musical noise in the reconstructed signal, which might be even worse for the listener than the original noise. The subspace methods divide the noisy signal into two subspaces, one containing the desired signal and noise and one containing only noise, based on, for example, the singular value decomposition of the covariance matrix of the noisy signal (see, e.g., [5–7]). The statistical-model-based methods use the statistics of the signal to design suitable filters for noise reduction, sometimes based on models of the desired signal with parameters that have to be estimated beforehand. Traditionally, subspace methods and statistical-model-based methods have been seen as separate methodologies, although attempts to bridge the gap have been made in, for example, [8] where finite impulse response filters are made based on the singular value decomposition of a Hankel matrix of the signal.

One of the milestones in the statistical-model-based methods was the introduction of the Wiener filter in speech processing [1, 9], which was later also derived in a subspace-based framework [10]. The design of the Wiener filter is dependent on the covariance matrix of the observed signal as well as an estimate of the noise statistics. Much work has been done in order to make good estimates of the noise. In [11–13], algorithms are presented which are capable of tracking the noise statistics even during voice activity. However, the algorithm still suffers from a large tracking delay, which is minimized using a Bayesian method for noise estimation in [14]. Even though good estimates of the noise can now be obtained, the Wiener filter still introduces a large amount of signal distortion, which was shown in [15] to be an important factor for the intelligibility of the enhanced speech.

Alternatively, a model of the signal can be assumed, and thereby filtering with less distortion can be obtained, for example, by use of the Capon filter [16] or the linearly constrained minimum variance (LCMV) filter [17], applied to speech enhancement in [18, 19]. In principle, the Capon and LCMV filters are distortionless, but this feature depends on correctly estimated model parameters and the validity of the assumed model. When the covariance matrices and model parameters have been estimated, it is not possible to change the Wiener, Capon and LCMV filters in order to trade noise reduction for less speech distortion or vice versa. In relation to speech intelligibily, a low degree of signal distortion might be the most important factor, but dependent on the application, noise reduction might be more important. Therefore, a filter designed according to the specific need of noise reduction relative to signal distortion is convenient and, hence, one such approach is presented in this paper.

The proposed filter design belongs to the group of statistical-model-based filters, but with its starting point in the ideas behind the subspace methods. The idea is to perform a

This work was funded by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF 1337-00084

joint diagonalization of the speech and noise covariance matrices and use the eigenvectors corresponding to the least significant eigenvectors to build a filter. The number of eigenvectors used in this process determines the amount of noise reduction but also the distortion of the signal. The noise reduction can, therefore, be exchanged for less signal distortion, which makes it a very flexible filter that can be designed to meet the specific need of the user. Since both signal and noise covariance matrices are used, the method is also dependent on the noise statistics being properly estimated using, for example, one of the aforementioned methods.

In Section 2, we introduce the signal model used and the basic theory behind the filter design. This is used in Section 3 to deduce the filter. In Section 4, the effect of the number of eigenvectors used are illustrated through simulations with a speech signal and in Section 5 the presented method is related to former work.

2. FUNDAMENTALS

The problem of speech enhancement considered in this work, which is sometimes also referred to as noise reduction, is that of recovering the desired signal, $x(k)$, with k being the discrete-time index, from noisy observations [1, 20, 21]:

$$y(k) = x(k) + v(k), \tag{1}$$

where $v(k)$ is the additive noise, which is here unwanted and assumed to be uncorrelated with $x(k)$. Moreover, all signals are considered to be real, zero mean, and stationary.

The signal model given in (1) can be put into a vector form by considering the L most recent successive time samples of the noisy signal, i.e., $\mathbf{y}(k) = \mathbf{x}(k) + \mathbf{v}(k)$, where $\mathbf{y}(k) = [y(k)\ y(k-1)\ \cdots\ y(k-L+1)]^T$ is a vector of length L, the superscript $(\cdot)^T$ denotes matrix/vector transpose, and $\mathbf{x}(k)$ and $\mathbf{v}(k)$ are defined similarly to $\mathbf{y}(k)$. Since $x(k)$ and $v(k)$ are uncorrelated by assumption, the covariance matrix (of size $L \times L$) of the noisy signal can be written as

$$\mathbf{R}_\mathbf{y} = E\left\{\mathbf{y}(k)\mathbf{y}^T(k)\right\} = \mathbf{R}_\mathbf{x} + \mathbf{R}_\mathbf{v}, \tag{2}$$

where $E\{\cdot\}$ denotes mathematical expectation, and $\mathbf{R}_\mathbf{x} = E\left\{\mathbf{x}(k)\mathbf{x}^T(k)\right\}$ and $\mathbf{R}_\mathbf{v} = E\left\{\mathbf{v}(k)\mathbf{v}^T(k)\right\}$ are the covariance matrices of $\mathbf{x}(k)$ and $\mathbf{v}(k)$, respectively. The noise covariance matrix, $\mathbf{R}_\mathbf{v}$, is assumed to be full rank, i.e., equal to L, whereas the speech covariance matrix, $\mathbf{R}_\mathbf{x}$, can either be rank deficient or full rank.

The objective, considered herein, is then to estimate the desired signal sample, $x(k)$, from the observation vector, $\mathbf{y}(k)$. This should be done in such a way that the noise is reduced as much as possible with little or no distortion of the desired signal. The proposed filtering method takes advantage of the joint diagonalization technique [22], where the symmetric, positive definite matrix $\mathbf{R}_\mathbf{v}$ and the symmetric matrix $\mathbf{R}_\mathbf{x}$ can be jointly diagonalized as follows:

$$\mathbf{B}^T\mathbf{R}_\mathbf{x}\mathbf{B} = \mathbf{\Lambda}, \quad \text{and} \quad \mathbf{B}^T\mathbf{R}_\mathbf{v}\mathbf{B} = \mathbf{I}_L, \tag{3}$$

where $\mathbf{B}$ is a full-rank, square matrix (of size $L \times L$), $\mathbf{\Lambda}$ is a diagonal matrix whose elements are real and nonnegative, and $\mathbf{I}_L$ is the $L \times L$ identity matrix. Furthermore, $\mathbf{\Lambda}$ and $\mathbf{B}$ are the eigenvalue and eigenvector matrices, respectively, of $\mathbf{R}_\mathbf{v}^{-1}\mathbf{R}_\mathbf{x}$, i.e., $\mathbf{R}_\mathbf{v}^{-1}\mathbf{R}_\mathbf{x}\mathbf{B} = \mathbf{B}\mathbf{\Lambda}$. Since $\mathbf{R}_\mathbf{v}$ is positive definite and $\mathbf{R}_\mathbf{x}$ is at least positive semidefinite, the eigenvalues of $\mathbf{R}_\mathbf{v}^{-1}\mathbf{R}_\mathbf{x}$ are non-negative and can be ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L \geq 0$. The corresponding eigenvectors are denoted by $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_L$. Due to the assumption in (2), the noisy signal covariance matrix can be diagonalized as $\mathbf{B}^T\mathbf{R}_\mathbf{y}\mathbf{B} = \mathbf{\Lambda} + \mathbf{I}_L$.

3. NOISE REDUCTION FILTERING

The idea is to design a filter from the eigenvectors corresponding to the least significant eigenvalues of $\mathbf{R}_\mathbf{v}^{-1}\mathbf{R}_\mathbf{x}$. The filter will, thereby, pass a large part of the noise whereas most of the desired signal is blocked, and the output from the filter can, therefore, be seen as an estimate of the noise. The estimate of the desired signal, $\widehat{x}(k)$, is then obtained by subtraction of the filter output, $z(k)$, from the noisy observation, $y(k)$.

First, we apply the filter $\mathbf{h} = [h_0\ h_1\ \cdots\ h_{L-1}]^T$ of length L to the noisy observation vector, $\mathbf{y}(k)$, to get the filter output:

$$z(k) = \mathbf{h}^T\mathbf{x}(k) + \mathbf{h}^T\mathbf{v}(k). \tag{4}$$

We then choose $\mathbf{h} = \sum_{i=P+1}^{L} \beta_i \mathbf{b}_i$, where β_i, $i = P+1, \ldots, L$, are arbitrary real numbers. Now, we estimate the desired signal as

$$\begin{aligned}\widehat{x}(k) &= y(k) - z(k) \\ &= x(k) - \sum_{i=P+1}^{L} \beta_i \mathbf{b}_i^T\mathbf{x}(k) + v(k) - \sum_{i=P+1}^{L} \beta_i \mathbf{b}_i^T\mathbf{v}(k).\end{aligned} \tag{5}$$

By minimizing the power of the residual noise, $\left[v(k) - \sum_{i=P+1}^{L} \beta_i \mathbf{b}_i^T\mathbf{v}(k)\right]^2$, or, alternatively, the power of the speech distortion, $\left[x(k) - \sum_{i=P+1}^{L} \beta_i \mathbf{b}_i^T\mathbf{x}(k)\right]^2$, we find that

$$\beta_i = \mathbf{i}_L^T\mathbf{R}_\mathbf{v}\mathbf{b}_i = \frac{1}{\lambda_i}\mathbf{i}_L^T\mathbf{R}_\mathbf{x}\mathbf{b}_i. \tag{6}$$

Substituting (6) into (5), we obtain

$$\begin{aligned}\widehat{x}(k) = x(k) &- \sum_{i=P+1}^{L} \frac{1}{\lambda_i}\mathbf{i}_L^T\mathbf{R}_\mathbf{x}\mathbf{b}_i\mathbf{b}_i^T\mathbf{x}(k) \\ &+ v(k) - \sum_{i=P+1}^{L} \mathbf{i}_L^T\mathbf{R}_\mathbf{v}\mathbf{b}_i\mathbf{b}_i^T\mathbf{v}(k).\end{aligned} \tag{7}$$

The variance of $\widehat{x}(k)$, found by using the relations in (3), is

$$\begin{aligned}\sigma_{\widehat{x}}^2 = \sigma_x^2 &- \sum_{i=P+1}^{L} \frac{1}{\lambda_i}\left(\mathbf{i}_L^T\mathbf{R}_\mathbf{x}\mathbf{b}_i\right)^2 \\ &+ \sigma_v^2 - \sum_{i=P+1}^{L} \left(\mathbf{i}_L^T\mathbf{R}_\mathbf{v}\mathbf{b}_i\right)^2,\end{aligned} \tag{8}$$

where σ_x^2 and σ_v^2 are the variances of the desired signal and noise before filtering.

The performance of the filters can be evaluated by means of the output signal-to-noise ratio (oSNR) and the speech distortion index, υ_{sd} [21]:

$$\text{oSNR} = \frac{\sigma_{x,\text{nr}}^2}{\sigma_{v,\text{nr}}^2} = \frac{\sigma_x^2 - \sum_{i=P+1}^{L} \frac{1}{\lambda_i} \left(\mathbf{i}_L^T \mathbf{R}_{\mathbf{x}} \mathbf{b}_i\right)^2}{\sigma_v^2 - \sum_{i=P+1}^{L} \left(\mathbf{i}_L^T \mathbf{R}_{\mathbf{v}} \mathbf{b}_i\right)^2}, \quad (9)$$

where $\sigma_{x,\text{nr}}^2$ and $\sigma_{v,\text{nr}}^2$ are the variances after noise reduction of the desired signal and noise, and

$$\upsilon_{\text{sd}} = \frac{E\{[\mathbf{h}^T \mathbf{x}(k)]^2\}}{E\{x^2(k)\}} = \frac{1}{\sigma_x^2} \sum_{i=P+1}^{L} \frac{1}{\lambda_i} \left(\mathbf{i}_L^T \mathbf{R}_{\mathbf{x}} \mathbf{b}_i\right)^2. \quad (10)$$

The smaller P is, the more noise reduction is obtained, but in the same instant the distortion of the desired signal is increased. The choice of the value of P will, therefore, be a compromise between noise reduction and signal distortion.

4. SIMULATIONS

In this section, the filter design is evaluated through simulations, and the influence of the choice of P on the output SNR and the speech distortion index is investigated.

The filter is tested on a known speech signal with car noise from the AURORA database [23] added to give an average input SNR (iSNR) of 10 dB. The covariance matrices of clean speech and noise are estimated from segments of 230 samples from the desired signal and noise vectors, respectively, and they are updated for each new sample. The speech signal used in the present simulations is a female speaker uttering the sentence "Why were you away a year Roy?" sampled with a frequency of 8000 Hz. The frequency response of a filter is plotted in Fig. 1 along with the spectrum of the corresponding desired signal. The filter is designed based on $L = 70$, and $P = 20$. The filter has zeros located close to the unit circle at the frequencies of the major components of the desired signal, and, therefore, it will primarily pass the noise, which is also the purpose of the filter. Based on the same settings, the speech signal is estimated, and a small part is shown in Fig. 2 where the original speech signal and the noisy observation are plotted as well.

The noise removed from the noisy observation is dependent on the value of P. If a smaller value of P is chosen, more noise can be removed, but, simultaneously, the signal will be more distorted. The influence of the value of P on the output SNR and speech distortion index is depicted in Figs. 3a and 3b. The filter length in this case is $L = 110$. For comparison, the performance of the Wiener filter ($\mathbf{h}_{\text{w}}$) and three subspace filters from [7] ($\mathbf{h}_{\text{ls}}$, $\mathbf{h}_{\text{mv}}$, $\mathbf{h}_{\text{mls}}$) are plotted as well. The filters from [7] are based on a decomposition of the Hankel matrix of the signal. With a segment length of 230, we construct a Hankel matrix of the observed signal of size 151 by 80, whereas the Hankel matrix of the noise is replaced by the Cholesky factorization of the noise covariance matrix, as also proposed in [7]. The rank (according to P) is varied from 1 to 71, since restrictions in the method limits the range of the chosen rank. Using the proposed filter, both output SNR and speech distortion index are decreasing with P, as was depicted in Section 3. Therefore, the filter can be designed according to the need for a high output SNR or a low speech distortion, which makes the filter very flexible. This is also the case for the filters proposed in [7], but here the range of possible combinations of output SNR and speech distortion is smaller. Given $P = 1$ the proposed filter has an extra gain in SNR of approximately 3 dB relative to the filters in [7], whereas the speech distortion is comparable. At the other extreme it is possible to lower the speech distortion by approximately 20 dB keeping the output SNR comparable to $\mathbf{h}_{\text{ls}}$. The Wiener filter is independent of the value of P which leaves no possibility for a trade-off between output SNR and signal distortion.

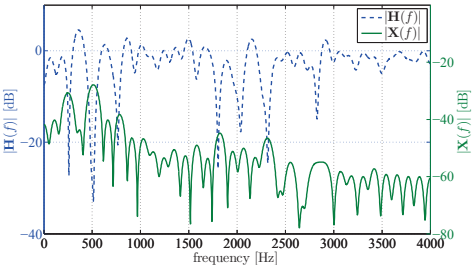

Fig. 1: Spectrum of clean speech vector, $|\mathbf{X}(f)|$, and frequency response of the corresponding filter, $|\mathbf{H}(f)|$. The iSNR is 10 dB.

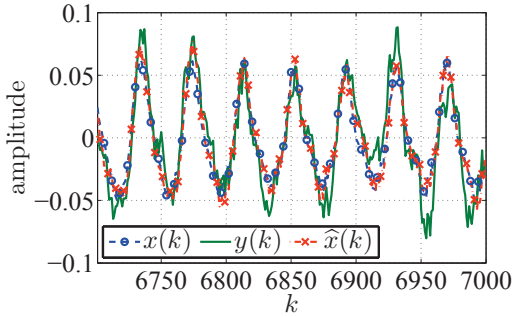

Fig. 2: Desired signal, $x(k)$, noisy observation, $y(k)$, and estimated signal, $\widehat{x}(k)$. The iSNR is 10 dB.

The effect of choosing different values of P is shown in Figs. 4a-4d. Figs. 4a and 4b show the spectrograms of the clean speech signal and the speech signal added noise, respectively, whereas Figs. 4c and 4d show the spectrograms of the reconstructed speech signal with $P = 10$ and $P = 100$. Using $P = 10$, much noise has been removed, when comparing to the noisy speech signal from Fig. 4b, but a high degree of signal distortion has been introduced as well, which can be seen especially in the left and lower right part of the spectro-

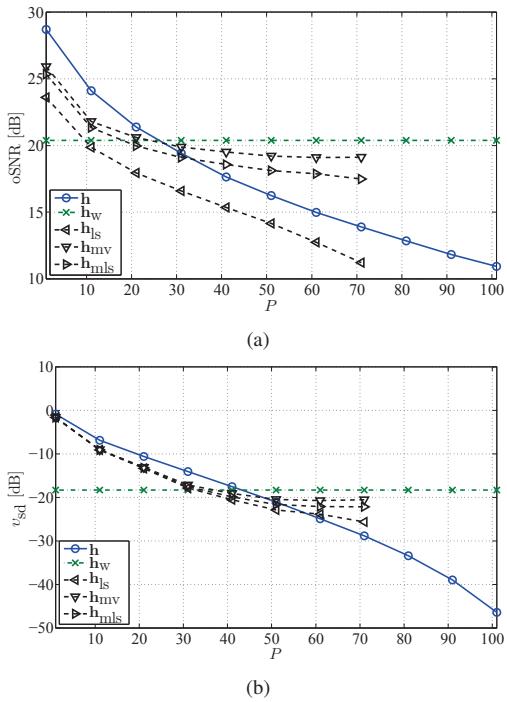

Fig. 3: (a) Output SNR and (b) speech distortion index as a function of P for a speech signal with full rank covariance matrix.

grams when comparing to Fig. 4a. Using $P = 100$ both noise reduction and signal distortion are not as prominent as when $P = 10$. The speech signal in the aforementioned areas of high signal distortion in Fig. 4c is much more well preserved in Fig. 4d, but, as is seen in the background of the figure, the price is a higher noise level.

5. DISCUSSION

A novel subspace-based filter was designed by use of the joint diagonalization of covariance matrices of desired signal and noise. The speech distortion and gain in SNR are dependent on the low rank approximation of the signal made in the filter design through the choice of P. Related methods are presented in [6–8]. Since the proposed method uses the joint diagonalization instead of the singular value decomposition, it can cope with both white and colored noise, therefore, there is no need for preprossesing in the case of colored noise as in [6]. Compared to [7, 8], the enhancement problem is here stated and solved as a linear filtering problem, and, compared to [7], the proposed filter has a larger interval for the choice of the rank, and was found to have a broader range of flexibility for the trade-off between output SNR and speech distortion index in the case of speech in car noise. It should be noted that while [10] also uses the joint diagonalization it is only as a computational tool since the resulting filter is the traditional Wiener filter.

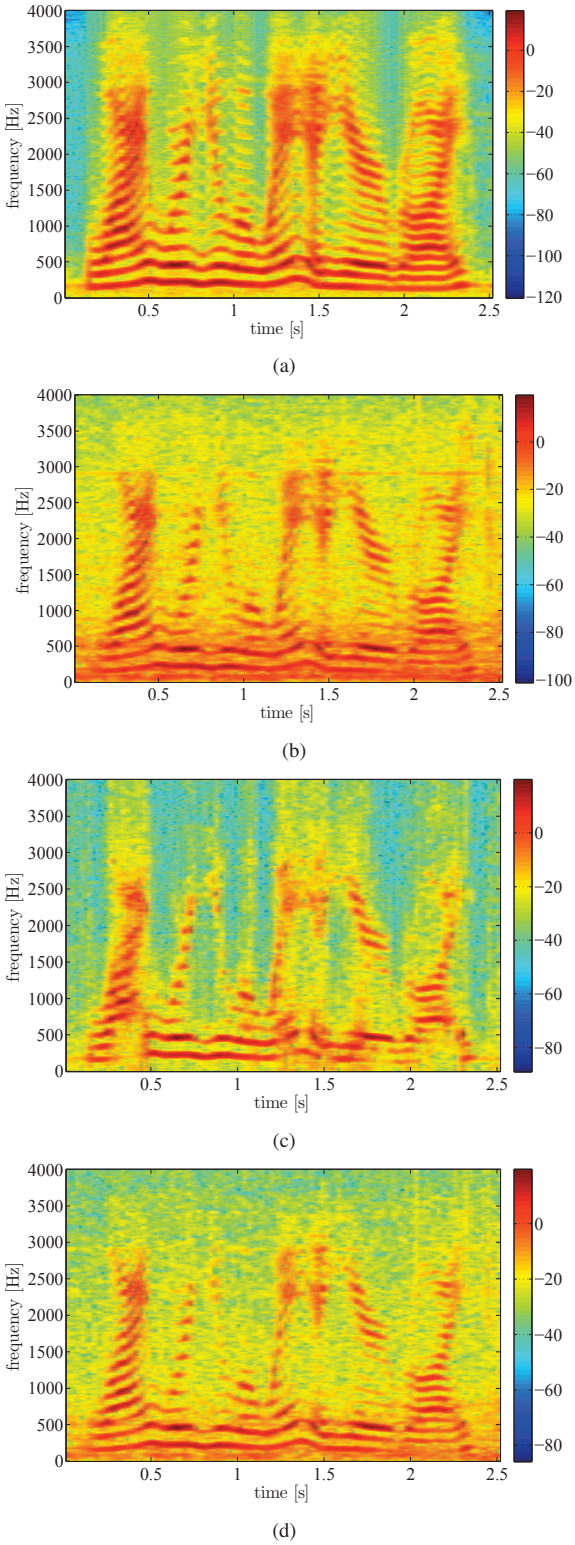

Fig. 4: Spectrograms of (a) the original and (b) noisy speech signals and the reconstructed speech signals with (c) $P = 10$ and (d) $P = 100$.

6. REFERENCES

[1] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1979, vol. 4, pp. 208–211.

[4] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2002, vol. 4.

[5] K. Hermus, P. Wambacq, and H. V. hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP J. on Advances in Signal Processing*, vol. 2007, no. 1, pp. 15, Sep. 2007, Article ID 045821.

[6] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[7] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Processing*, vol. 2007, no. 1, pp. 24, Jun. 2007, Article ID 092953.

[8] P. C. Hansen and S. H. Jensen, "FIR filter representations of reduced-rank noise reduction," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1737–1741, 1998.

[9] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[10] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[12] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[13] L. Lin, W. H. Holmes, and E. Ambikairajah, "Subband noise estimation for speech enhancement using a perceptual Wiener filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2003, vol. 1, pp. 80–83.

[14] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4266–4269.

[15] G. Kim and P. C. Loizou, "Why do speech-enhancement algorithms not improve speech intelligibility?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4738–4741.

[16] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[17] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[18] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. on Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[19] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[20] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*, Springer-Verlag, 2009.

[21] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters – A Theoretical Study*, Number VII. Springer, 1 edition, 2011.

[22] J. N. Franklin, *Matrix Theory*, Prentice-Hall, 1968.

[23] D. Pearce and H. G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Int. Conf. Spoken Language Process.*, Oct. 2000.