KALMAN FILTERING BASED NOISE POWER SPECTRAL DENSITY ESTIMATION FOR SPEECH ENHANCEMENT

Ivo Batina, Jesper Jensen and Richard Heusdens

Information and Communication Theory Group,
Delft University of Technology,
2628 CD Delft, The Netherlands,
email:{i.batina,j.jensen,r.heusdens}@ewi.tudelft.nl

ABSTRACT

We propose a method for estimating the power spectral density (PSD) of nonstationary noise when a noisy speech signal is given. The method is based on the Kalman filtering technique. In contrast to the known noise statistics tracking methods that are based on time smoothing of the noisy speech periodogram, we use a Kalman filter based on a low order model of the noise power spectrum and update the noise estimate for the next frame according to the difference between the measurement of the noisy speech power spectrum and the current Kalman estimate of it. We derive a recursive estimation scheme of a low computational complexity, which makes the proposed method well suited for real time implementations. The method can be combined with any speech enhancement algorithm that requires a noise PSD estimate. Objective and subjective performance evaluations show that the proposed scheme exhibits a good noise tracking performance and that it achieves improvement in the quality of the enhanced speech as compared to the case where noise PSD estimate remains invariant across time. Listening test results indicate a statistically significant improvement in the quality of enhanced speech compared to the fixed PSD case.

1. INTRODUCTION

With the growth of mobile communication applications, the problem of reducing the background noise in noisy speech signals has become increasingly important. The class of speech enhancement techniques based on short-time spectral amplitude (STSA) estimation (see [1, 2]) have proved to be of particular practical interest due to their low complexity and relatively good performance. As most single-channel speech enhancement (SE) methods, STSA based techniques require a power spectral estimate of the noise process in order to extract a clean speech signal estimate from a noisy realization. As any SE scheme, the performance of STSA based techniques is much affected by the capability to track variations in the statistics of the noise [3], particularly under low signal-to-noise ratio (SNR) conditions and non-stationary noise environments. In [3] a recursive scheme for noise estimation, commonly known as the Minimum Statistics (MS) method is designed to be combined with STSA speech enhancement schemes. The method is based on tracking the noisy speech spectral minima without any distinction between speech activity and speech pause, enabling the algorithm to update the noise estimate even in the speech presence regions. Although the method generally works well for tracking of relatively slowly varying noise sources, a disvantage of the method is a tracking lag in the noise estimate [3, 4]. A similar method is described in [4], where the response of the noise estimator to the rise of the noise level is improved by periodogram smoothing in both time and frequency and speech presence probability estimation.

In this paper, we propose a method for noise power spectral density (PSD) that is based on the application of the Kalman filtering technique in the STSA context. To develop a Kalman filter for this application, we exploit a priori knowledge about the noisy speech by using a low-order model that describes the development across time of speech and noise power levels. The resulting estimator is then optimal (given the model) within the class of linear estimators. Instead of the noisy speech periodogram smoothing, as proposed in [3, 4], the Kalman filter based estimator uses a difference between estimate of the noisy speech PSD and the current measurement of the noisy spectrum, multiplied with the so-called Kalman gain, to obtain the noise PSD estimate in the next frame. The resulting algorithm is of a low computational complexity that can be further reduced by pre-computing Kalman gains.

2. STOCHASTIC MODEL OF THE NOISY SPEECH POWER SPECTRUM

The first step in development of our method is to derive a low-order model of the noisy speech that will be used in the estimator. We assume that the noise is additive such that the short-time Fourier transform (STFT) of the noisy speech signal can be written as

$$Y(k,l) = S(k,l) + N(k,l), \tag{1}$$

where $Y(k,l)$, $S(k,l)$, $N(k,l)$ denote STFT coefficients of the noisy speech signal, speech and the noise, respectively, k denotes the frequency bin index and l represents frame index. Furthermore, we assume that speech and noise are uncorrelated random processes and that STFT are Gaussian distributed (see e.g. [1, 2]). It can be then easily shown that the magnitude square of the noisy speech STFT coefficients are exponentially distributed random variables for all k and l with a probability density function (PDF) given by

$$f_{|Y(k,l)|^2}(x) = \frac{1}{\lambda_s(k,l)+\lambda_n(k,l)} \exp\left\{-\frac{x}{\lambda_s(k,l)+\lambda_n(k,l)}\right\}, \tag{2}$$

with $x \geq 0$. In (2), $\lambda_s(k,l) = \mathbb{E}\{|S(k,l)|^2\}$ and $\lambda_n(k,l) = \mathbb{E}\{|N(k,l)|^2\}$ are variances of the speech and the noise STFT coefficients given in (1).

Next, define the stochastic process

$$y(k,l) = \big(\lambda_s(k,l)+\lambda_n(k,l)\big)e(k,l), \tag{3}$$

where $e(k,l)$ is an exponentially distributed random variable with mean and variance equal to 1. It can easily be shown that the probability distribution of $y(k,l)$ is identical to that of $|Y(k,l)|^2$ given in (2). Equation (3) can be rewritten in the matrix form as

$$y(k,l) = Cx(k,l)e(k,l), \tag{4}$$

with

$$C = \begin{bmatrix} 1 & 1 \end{bmatrix} \quad \text{and} \quad x(k,l) = \begin{bmatrix} \lambda_s(k,l) \\ \lambda_n(k,l) \end{bmatrix}.$$

We assume that variations of the power spectrum variance across time can be modelled as

$$x(k,l+1) = A(k)x(k,l) + E(k)w(k,l), \tag{5}$$

This research was partly supported by Philips Research and the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs.

where

$$A(k) = \begin{bmatrix} a_n(k) & 0 \\ 0 & a_s(k) \end{bmatrix}, \qquad E(k) = \begin{bmatrix} e_n(k) \\ e_s(k) \end{bmatrix},$$

and $w(k,l) \in \mathcal{N}(0,1)$ is a standard normal distribution. The state transition matrix $A(k)$ is chosen to be diagonal, because for uncorrelated speech and noise process, variations in one process should not influence variations in the other. Since the variances of the speech and the noise processes can not be smaller than zero, the state x has to satisfy the following constraint

$$\lambda_s(k,l) \geq 0 \quad \text{and} \quad \lambda_n(k,l) \geq 0, \tag{6}$$

for all k and all l.

The numerical values of $A(k)$ and $E(k)$, which in this work remain constant across time, are obtained by (offline) numerical optimization. In this way, the proposed scheme exploits a priori knowledge of the speech production process (through $a_s(k)$ and $e_s(k)$) and can exploit any available priori knowledge about the noise process (through $a_n(k)$ and $e_n(k)$).

3. KALMAN FILTERING BASED NOISE POWER SPECTRAL DENSITY ESTIMATOR

Equations (4), (5) and (6) form the model of the noisy speech power spectrum. We note that this model does not fit in the classical Kalman filter setting because the output equation (3) is in a multiplicative form, rather than the well known additive form (see [5]). Therefore the standard Kalman filtering formulas can not be applied to the model described in section 2. In this subsection we derive the Kalman filtering equations for the model described in section 2.

The problem that we consider is to find the linear MMSE estimator for the system given by (4), (5) and (6). It is well known that the MMSE estimate can be expressed as the following conditional mean (see [5])

$$\hat{x}(k,l+1) = \mathbb{E}\big\{x(k,l+1)\big|\mathcal{Y}_l(k)\big\}, \tag{7}$$

where $\mathbb{E}(\cdot)$ denotes the statistical expectation operator and $\mathcal{Y}_l(k)$ is the set of observations of the noisy speech power spectrum defined by

$$\mathcal{Y}_l(k) := \{|Y(k,l)|^2 \; |Y(k,l-1)|^2 \; \cdots \; |Y(k,0)|^2\}.$$

We constrain our estimator to be from the class of linear estimators i.e. it can be expressed as

$$\hat{x}(k,l+1) = \sum_{i=0}^{l} P(k,i)|Y(k,l-i)|^2. \tag{8}$$

It can be shown that the one-step ahead prediction of $y(k,l)$ is

$$\hat{y}(k,l) = \mathbb{E}\big\{y(k,l)\big|\mathcal{Y}_{l-1}(k)\big\} = C\hat{x}(k,l), \tag{9}$$

Let us define an innovation process as

$$\tilde{y}(k,l) = |Y(k,l)|^2 - \hat{y}(k,l). \tag{10}$$

The innovation [5, 6] process has the following properties

1. Innovations are orthogonal to each other: $\mathbb{E}\{\tilde{y}(k,l)\tilde{y}(k,l-i)\} = 0$ for $i = 1, \cdots, l$.
2. The set of innovations up to frame index l can be obtained by linear transformation of observations $\mathcal{Y}_l(k)$.
3. The set of innovations up to frame index l contains the same information about power spectrum as $\mathcal{Y}_l(k)$.

Let $\hat{P}(k,i)$ denote a linear filter satisfying

$$\hat{x}(k,l+1) = \sum_{i=0}^{l} \hat{P}(k,i)\tilde{y}(k,l-i).$$

Because the innovations are orthogonal to each other and we seek the best linear estimator it follows from the orthogonality principle that innovations and estimation error obtained by the optimal estimator are orthogonal i.e.

$$\mathbb{E}\Big\{\Big(x(k,l+1) - \sum_{i=0}^{l} \hat{P}(k,i)\tilde{y}(k,l-i)\Big)\tilde{y}(k,l-j)\Big\} \equiv 0 \quad 0 \leq j \leq l. \tag{11}$$

From (11) and by using (4) it follows that

$$\hat{x}(k,l+1) = A(k)\sum_{i=0}^{l} \frac{\mathbb{E}\Big\{x(k,l)\tilde{y}(k,l-i)\Big\}}{\mathbb{E}\Big\{\tilde{y}^2(k,l-i)\Big\}}\tilde{y}(k,l-i). \tag{12}$$

By splitting the sum in (12), it can be rewritten as

$$\begin{aligned}\hat{x}(k,l+1) = {} & A(k)\frac{\mathbb{E}\Big\{x(k,l)\tilde{y}(k,l)\Big\}}{\mathbb{E}\Big\{\tilde{y}^2(k,l)\Big\}}\tilde{y}(k,l) \\ & + A(k)\sum_{i=1}^{l} \frac{\mathbb{E}\Big\{x(k,l)\tilde{y}(k,l-i)\Big\}}{\mathbb{E}\Big\{\tilde{y}^2(k,l-i)\Big\}}\tilde{y}(k,l-i). \end{aligned} \tag{13}$$

Define the Kalman gain as

$$K(k,l) = A(k)\frac{\mathbb{E}\Big\{x(k,l)\tilde{y}(k,l)\Big\}}{\mathbb{E}\Big\{\tilde{y}^2(k,l)\Big\}}. \tag{14}$$

Next, observe that substitution of (5) in (7) gives

$$\hat{x}(k,l+1) = A(k)\mathbb{E}\big\{x(k,l)\big|\mathcal{Y}_l(k)\big\}. \tag{15}$$

By applying the orthogonality principle on the estimate at l it can be shown that

$$\hat{x}(k,l) = \sum_{i=1}^{l} \frac{\mathbb{E}\Big\{x(k,l)\tilde{y}(k,l-i)\Big\}}{\mathbb{E}\Big\{\tilde{y}^2(k,l-i)\Big\}}\tilde{y}(k,l). \tag{16}$$

By substituting (14), (16) and (10) in (13) we obtain

$$\hat{x}(k,l+1) = A(k)\hat{x}(k,l) + K(k,l)\left(|Y(k,l)|^2 - C\hat{x}(k,l)\right). \tag{17}$$

It remains to derive a recursive expression for the Kalman gain $K(k,l)$ (14). By using (10), the model of the noisy speech power spectrum (4) and equation (9), we can rewrite (14) as

$$K(k,l) = A(k)Q_e(k,l)C^T\mathbb{E}\big\{\tilde{y}^2(k,l)\big\}^{-1}, \tag{18}$$

where

$$Q_e(k,l) = \mathbb{E}\big\{\left((x(k,l) - \hat{x}(k,l)\right)\left((x(k,l) - \hat{x}(k,l)\right)^T\big\}, \tag{19}$$

is the variance of the estimation error. By using (4), (9) and (10) we can compute

$$\mathbb{E}\big\{\tilde{y}^2(k,l)\big\} = C\left(2Q_e(k,l) + \hat{Q}(k,l)\right)C^T, \tag{20}$$

where

$$\hat{Q}(k,l) = \mathbb{E}\{\hat{x}(k,l)\hat{x}(k,l)^T\}. \tag{21}$$

Inserting (20) into (18)[1]

$$K(k,l) = A(k)Q_e(k,l)C^T \left(C\big(2Q_e(k,l) + \hat{Q}(k,l)\big)C^T\right)^{-1}. \quad (22)$$

From (19), (5) and (17) we obtain a recursive expression for the variance of the estimation error

$$\begin{aligned} Q_e(k,l+1) = \big(A(k) - K(k,l)C\big)Q_e(k,l)\big(A(k) - K(k,l)C\big)^T + \\ K(k,l)CQ_e(k,l)C^T K^T(k,l) + \\ C\hat{Q}(k,l)C^T + EQ_w(k,l)E^T. \end{aligned}$$

From (21), (19) and (5) it follows

$$\begin{aligned} \hat{Q}(k,l+1) = \big(A(k) + K(k,l)C\big)\hat{Q}(k,l)\big(A(k) + K(k,l)C\big)^T + \\ 2K(k,l)CQ_e(k,l)C^T K^T(k,l) + EQ_w(k,l)E^T. \end{aligned}$$

4. IMPLEMENTATION OF THE ALGORITHM

To implement the estimator presented in section 3 there are a number of issues that have to be addressed. The algorithm is implemented with pre-computed steady state Kalman gain in (17). To deal with the constraint (6) we use the methodology proposed in [7]. For each frame, the unconstrained estimate (16) is optimally projected onto the state constraint surface if the constraint is violated. In the case we consider, it is easy to show that the optimal projection is equivalent to setting the estimated variances to zero when the constraint (6) is violated.

Next, the function $a_s(k)$ is determined by an off-line estimation procedure on a clean speech signal that consists of four different speech utterances. The utterances, two from male and two from female speakers are taken from the TIMIT database and downsampled to 8 kHz. We obtain the magnitude square of the speech STFT coefficients $|S(k,l)|^2$ by using the discrete Fourier transform of the signal frames extracted with a Hanning window of 256 samples and use of an inter frame overlap of 50%. Given $|S(k,l)|^2$ of the speech signal at hand, we compute $a(k)$ by solving the following optimization problem

$$\min_{a_s(k)} \|\mathbb{E}\{|S(k,l)|^2\} - \hat{x}(k,l)\|_2 \quad (23)$$

for each k by using a nonlinear minimum search method. In order to improve the performance of the estimator in the regions with a high speech energy we do not update the estimate, when the a priori SNR (estimated using the decision directed approach in [1] with the Kalman estimate of the noise PSD), exceeds a certain threshold.

5. PERFORMANCE EVALUATION

The performance evaluation of the Kalman filtering based noise tracking algorithm consists of two parts. First we show the tracking capability of the algorithm for nonstationary white noise. We compare the performance of the noise estimator based on Kalman filter (17) with the minimum statistics (MS) noise estimator [3]. Second, we use different noise estimation strategies in in the log spectral amplitude (LSA) enhancement scheme [2] and perform an objective as well as subjective quality assessment of the enhanced speech samples.

The speech signal used in the objective evaluations is constructed from five different speech utterances. The speech utterances are taken from the TIMIT database and are outside the training set that was used in the off-line model estimation. The speech signal is sampled at 8 kHz and degraded by various noise types with SNR in the range $[-5, 10]$ dB. The noise signals are taken from the Noisex92 database. We include white Gaussian noise, babble noise, factory noise and car noise in the evaluations. The spectral analysis is implemented with the Hanning window of 256 samples (32 ms) and use of an inter frame overlap of 50%.

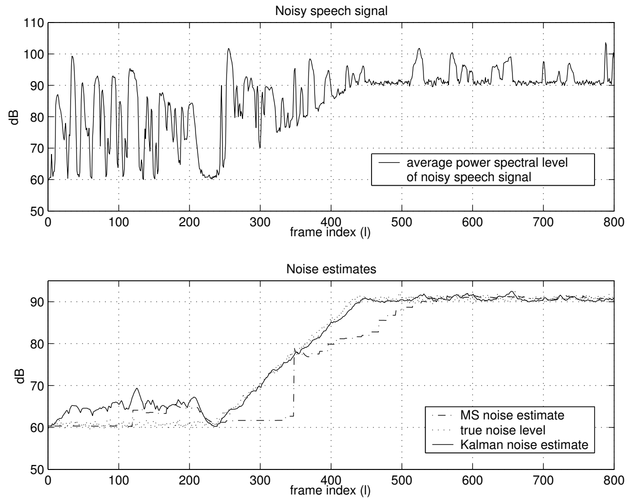

Figure 1: The mean over frequency of the noisy speech sample power spectrum and the noise PSD estimates obtained with different noise estimators.

To show the tracking capability of the noise estimator based on the Kalman filtering we degrade the speech signal with nonstationary white Gaussian noise. The speech signal is degraded at SNR 30 dB for the first 3.75 seconds of the signal. After that, the noise level rises with the constant rate of 0.15 dB/frame up to 0 dB SNR where it stays for the remaining part of the signal. To ease the visualisation of the results, we adopted the procedure used in [3, 4] and compute the average noise PSD estimate across frequency for each frame, for the proposed method as well as for the MS method. We emphasize that this frequency averaging is only done for presentation purposes. Neither the proposed method nor the MS noise estimator exploits the a priori knowledge that the noise source in this case is spectrally flat. It can be observed that the response of the proposed estimator is faster than the MS estimator. For the increasing noise power, the MS estimator lags behind with a delay of $D+V$ samples [3] where D is the size of the minimum search window and V is the size of the subwindow (see [3] for details). For the constant noise level of 0 dB both estimators give a good estimate of the noise level. Next, we turn to natural noise sources and consider speech signals degraded with babble, factory and car noise, at various SNR levels. We compare the performance of the LSA enhancement scheme [2] for different noise estimators. For objective quality assessment we use the Symmetric Itakura-Saito (I.S) distortion measure [8] and Segmental SNR measure [9]. Results are summarized in Tab. 1. In the first case no noise tracking is performed (NNT in Tab. 1). In this case we take a snapshot of the noise in the noise only region preceding the speech signal and keep this value as the noise level estimate for the whole duration of the signal. Next, we use MS noise estimator (MS in Tab. 1) and finally the proposed noise estimator (KF in Tab. 1). We also give values of the Symmetric I.S. and Segmental SNR for noisy speech sample (NS in Tab. 1). The results of the objective quality assesment show that both symmetric I.S. and Segmental SNR distortion measures indicates improvement of the performance when the noise estimator based on the Kalman filtering is used in the enhancement scheme.

For subjective evaluation an OAB listening test was performed with ten participants, the authors not included. We compare the per-

[1] For a stable matrix A and observable matrix pair (A,C) the Kalman gain (22) converges to a steady state value which can be pre-computed for the given matrices $A(k)$, $E(k)$ and $C(k)$ (see [5] for detailed treatment of this issues in the standard Kalman filter setup).

Segmental SNR												
Input SNR	Babble noise				Factory noise				Car noise			
[dB]	NS	NNT	KF	MS	NS	NNT	KF	MS	NS	NNT	KF	MS
-5	-6.88	-4.72	-4.32	-4.74	-7.13	-4.27	-3.35	-3.67	-6.40	1.90	2.98	2.10
0	-5.14	-2.44	-2.04	-2.44	-5.42	-1.94	-1.24	-1.32	-4.40	4.94	6.10	5.11
5	-2.66	0.21	0.28	0.24	-2.96	0.56	1.07	0.72	-1.83	7.62	8.96	8.07
10	0.38	2.82	3.13	3.09	0.09	3.04	3.56	3.30	1.30	9.64	10.86	10.45

Symmetric Itakura-Saito distortion measure												
Input SNR	Babble noise				Factory noise				Car noise			
[dB]	NS	NNT	KF	MS	NS	NNT	KF	MS	NS	NNT	KF	MS
-5	16526	12040	9516	10463	15272	3257	2101	2386	2516	126	58	76
0	6565	2829	1520	2306	5593	735	361	454	936	41	19	27
5	2217	745	295	509	1875	195	86	111	326	17	6	9
10	715	219	80	122	611	58	28	35	109	10	2	3

Table 1: Segmental SNR and Symmetric Itakura-Saito (I.S) distortion measure for babble, factory and car noise at various SNR levels using different noise PSD estimators

noise source	input SNR	P value	significant
babble noise	15	$9.06*10^{-8}$	yes
	5	$2.35*10^{-6}$	yes
factory noise	15	$5.12*10^{-9}$	yes
	5	$1.11*10^{-8}$	yes
car noise	15	$1.01*10^{-8}$	yes
	5	$2.35*10^{-8}$	yes

Table 2: Wilcoxon test results to verify the statistically significant difference between the methods used in the listening test

formance of the LSA enhancement scheme with no noise tracking with the performance of the LSA enhancement scheme when the Kalman filtering based noise tracking method is used. In this listening test we used babble and factory noise at 5 dB and 15 dB and car noise at 0 dB and 10 dB SNR. For each noise source and noise level we presented listeners two female and two male sentences. The listeners were presented first the noise free signal followed by the two different enhanced signal in the randomized order, and this was repeated three times for each series. For speech signals corrupted with babble noise the proposed method was preferred above the no noise tracking case in 71 % (15 dB) and 63 % (5 dB) of the cases, for factory noise in 88 % (15 dB) and 84 % (5 dB) cases and for car noise the Kalman filtering based noise tracking was preferred in 83 % (10 dB) and 94 % (0 dB). A statistical significance Wilcoxom test [10] was used to test a statistical difference between the two methods. The P-value of this test are given in Tab. 2. The results presented in Tab. 2 show that for all noise sources and SNRs used in the test the difference between methods is statistically significant.

6. CONCLUSIONS

We present an algorithm that provide an accurate estimate of the noise power level, that is suitable for the real time implementation and is of a low computational complexity. The method is based on the Kalman filtering technique. Since the model does not fit in the standard Kalman filtering setting, we derive a linear, recursive estimator for the stochastic model of the noisy speech spectrum. We perform objective and subjective evaluation of the proposed method. As results presented in section 5 show, the Kalman based noise estimator has good noise tracking capabilities and listening test showed preference over the case when there is no noise tracking. Although we apply this method in the LSA based speech enhancement context, the Kalman filtering based noise estimation method is very general and can be applied to any other speech enhancement system that requires a noise power spectral estimate, e.g., codebook-driven methods [11] and subspace based approaches [12].

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, April 1985.

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

[4] I. Cohen, "Noise spectrum estimation in adverse enviroments: Improved minima controller recursive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, September 2003.

[5] M. S. Grewal and A. P. Andrews, *Kalman filtering: theory and practice*, Prentice Hall, Englewood Cliffs, 1993.

[6] M. S., P. Rajasekaran, and R. Viswanathan, *An introduction to statistical processing with applications*, Prentice Hall, Englewood Cliffs, 1996.

[7] D. Simon and T. L. Chia, "Kalman filtering with state equality constraints," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 38, no. 1, pp. 128–136, January 2002.

[8] A. H. Gray, Jr. and J. D. Markel, "Distance measures for speech processing," *assp*, vol. ASSP-24, no. 5, pp. 380–391, October 1976.

[9] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, Macmillan, NewYork, 1993.

[10] D. J. Sheskin, *Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, 2004.

[11] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, April 1991.

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