Speech enhancement in the reconstructed phase-space

ISTVÁN PINTÉR

Kecskemét College, GAMF Faculty, Sándor Kalmár Institute of Information Technology
pinter.istvan@gamf.kefo.hu

Keywords: speech enhancement, signal subspace, reconstructed phase space, dimension embedding

The speech enhancement method, presented in this paper, is based on the concepts of reconstructed phase-space and dimension embedding. The proposed algorithm separates the speech from noise using a non-linear transformation in a transformed domain. Our recent results in case of uncorrelated, additive noise are presented in this paper.

1. Introduction

Speech enhancement is a long-standing problem in digital speech processing [1]. Several methods for noise suppression have been elaborated during the past three decades. The common assumption in most cases is the slow variation of noise parameters, corresponding to the linear speech model.

As an example, a system worth mentioning uses an auditory-model based filterbank with Wiener-filtering in sub-bands [2]. According to published results these methods give acceptable solutions in case of SNRs (signal to noise ratio) greater than 6...9 dB [3]. In case of either lower SNRs or nonstationary noise the speech enhancement methods are based on non-linear models. A non-linear model of human auditory system has been applied for noise suppression in [4], while the reconstructed phase-space representation of speech belongs to the class of non-linear signal models [5]. The latter is also the subject of the recent paper.

The structure of the paper is as follows. In the first part the optimal representation of the clean speech is reviewed, followed by the introduction of a noise suppression method based on the notion of speech subspace. The generalised version, working in the reconstructed phase space, is also introduced. Our numerical results, achieved by realisation of the algorithm are presented in the fourth section. The paper ends with the conclusions, acknowledgement and references.

2. Representation of the clean speech in the transformed domain and in the reconstructed phase space

The noise suppresion method, presented in this paper, is based on two assumptions. The first one is the existence of the optimal representation of the clean speech, the second one is that the concept of reconstructed phase space is suitable for speech processing problems.

Concerning the first assumption, a vector can be formed from α_n speech samples of the segment under pressing. If N denotes the number of samples in the segment, the resulting vector corresponds to a vector of N dimensional Euclidean-space. This vector $\underline{s}$ can be written as a linear combination using the $\{\underline{t}_n\}$ natural orthonormal basis, where coefficients are the speech samples: $\alpha_n = (s,\underline{t}_n)$ and the *n*th component of the N-dimensional $\underline{t}_n$ column vector is 1, the others are 0. According to experiences in solutions of practical problems in digital speech processing, there exists an orthonormal basis, so that by using this 'optimal' basis the speech vector can be represented with fewer components than N [6]. The optimality means that the speech vector in question can be given as

$$\hat{\underline{s}} = \sum_{n=0}^{L-1} a_n \cdot \underline{v}_n \tag{1}$$

where $\{\underline{v}_n\}$ denotes the optimal orthonormal basis and L<N holds. Moreover, the representation in (1) is optimal in the sense that the value of the criterion function below is

$$J(\underline{e}) = E\left\{\|\underline{e}\|^2\right\} = E\left\{\|\underline{s} - \hat{\underline{s}}\|^2\right\} = \sum_{n=L}^{N-1} \underline{v}_n^T \cdot E\left\{\underline{s} \cdot \underline{s}^T\right\} \cdot \underline{v}_n = \sum_{n=L}^{N-1} \underline{v}_n^T \cdot \underline{\underline{R}} \cdot \underline{v}_n \tag{2}$$

that is the mean square error is minimal (ideally L<N and $||\underline{e}||=0$). By the assumtion of $E\{\underline{s}\}=0$, we get $\underline{\underline{R}}=\underline{\underline{K}}$, which is the covariance matrix. The solution of (1) and (2) is the $\{\underline{v}_n\}$ eigenvector system of the covariance matrix, and the minimal value othe mean square error can be written using the corresponding eigenvalues as

$$J(\underline{e}) = \sum_{n=L}^{N-1} \lambda_n$$

where λ_n denotes the *n*th eigenvalue. The new representation of the speech vector $\underline{s}$ can be computed as a matrix-vector product using the matrix below

$$\underline{\underline{T}} = \left(\underline{v}_0^T, \underline{v}_1^T, \ldots, \underline{v}_i^T, \ldots, \underline{v}_{N-1}^T\right)^T \tag{3}$$

which has the eigenvectors in its rows corresponding to eigenvalues organized in desending order.

The second assumption goes for the representation of the speech in the reconstructed phase space. The concept of reconstructed phase space applies to the

motion equation of the discrete dynamical system, $\underline{x}_{n+1}=\underline{F}(\underline{x}_n)$, where $\underline{x}_n$ and $\underline{x}_{n+1}$ are D dimensional points in the phase space, $\underline{F}$ denotes a suitable mapping. The set $\{\underline{x}_n\}$ of phase-space points constitutes the so-called trajectory. This trajectory cannot be observed directly, only through the non-linear mapping $\underline{x}_n \rightarrow g(\underline{x}_n)$ – the resulting observable real number is the speech sample $\alpha_n = g(\underline{x}_n)$. By taking these samples in regular time intervals T_{MV}, one finally gets the speech sample sequence $\{\alpha_n\}$. It is provable that when the condition $M>2\cdot D+1$ holds, then from the number sequence α_n the vector sequence $\{\underline{y}_n\}$ can be reconstructed, which is equivalent of the original vector sequence $\{\underline{x}_n\}$. The method of the reconstruction is the so-called dimension embedding, which results a vector

$$\underline{y}_n(M,\tau)=(\alpha_n,\alpha_{n+\tau},\ldots,\alpha_{n+(M-1)\cdot\tau}) \tag{4}$$

where $\tau>0$ is the time lag (given by number of samples here), and $M>0$ denotes the embedded dimension. The equivalence mentioned above means that there exists an invertible, smooth mapping $\underline{h}:\underline{y}_n\,(M,\tau)\rightarrow\underline{x}_n$, by which the two vector sequence in question can be transformed into each other [7]. The values of the embedding dimension M and time lag τ can be determined experimentally, depending on the type of the speech technology application. According to relevant literature the value of the embedding window $M\cdot\tau\cdot T_{MV}$ is in the interval of 1...5 ms [8].

3. Noise suppression in the reconstructed phase space by using the sub-space method

The noise suppression algorithm, which can be given by using the concept of the reconstructed phase space, is in essence a generalisation of an earlier method published in the relevant literature, so the latter is reviewed first.

The basis of the method is the property of the speech described in Section 1, namely that the speech can be optimally represented. It means, that the N dimensional orthonormal basis is not necessary for the representation, but L<N dimensional orthonormal basis is enough, and ideally the mean square error value is zero. So, the N dimensional speech vector can be found in an L dimensional sub-space, titled as 'speech-subspace'.

The noise suppression algorithm determines an estimated, optimal speech vector from the noisy speech samples. Let's denote the noisy speech as

$$\underline{u}=\underline{s}+\underline{w} \tag{5}$$

where $\underline{w}$ denotes the additive noise vector, uncorrelated with speech. Starting from the noisy samples, it is necessary to give an estimate of the speech $\tilde{\underline{s}}$, so that the expectation value of the norm of the diference $\underline{s}-\tilde{\underline{s}}$ should be minimum, that is

$$E\left\{\left\|\underline{s}-\tilde{\underline{s}}\right\|^2\right\}\rightarrow\min \tag{6}$$

First of all – similarly to the above discussed problem – it is necessary to determine the optimal orthonormal basis for the speech, however in this case only the noisy speech vector $\underline{u}$ is known. By assuming that $E\{\underline{w}\}=0$, and using the previous assumption $E\{\underline{s}\}=0$, gives $E\{\underline{u}\}=0$. Additional assumption is that the zero-mean noise is white noise, if its covariance-matrix can be written as $\underline{\underline{K}}^{NOISE}=\sigma^2\cdot\underline{\underline{I}}$, where $\sigma>0$ and $\underline{\underline{I}}$ denotes the NxN identity matrix. Because the speech and noise are uncorrelated, the correlation matrix of the noisy speech can be written as a sum of correlation matrices of speech and noise, respectively, that is

$$\underline{\underline{K}}^{NOISY}=E\left\{\underline{u}\cdot\underline{u}^T\right\}=\underline{\underline{K}}^{SPEECH}+\underline{\underline{K}}^{NOISE} \tag{7}$$

holds. As it can also be proven, the eigenvectors of the noisy and clean speech are the same. The latter property makes it possible to determine the estimated speech vector, because the vectors of the orthonormal basis, necessary for the ideal representation of the speech, can be determined from the given noisy speech samples. In other words, the optimal basis $\{\underline{v}_n\}$ can be computed from the covariance matrix of the noisy speech, so it is not necessary to know the covariance matrix of the clean speech. Moreover, as a consequence of the summability of the covariance matrices (7), it is provable, that the covariance matrix of the noisy speech in the transformed domain is the diagonal matrix below:

$$\underline{\underline{K}}^{NOISY}\Big|_{\{\underline{v}_n\}}=diag\left(\lambda_0+\sigma^2\ \ldots\ \lambda_{L-1}+\sigma^2\ \ \sigma^2\ \ldots\ \sigma^2\right). \tag{8}$$

According to our assumption described in Section 1, the speech can optimally be represented in the subspace, spanned by the vectors $\underline{v}_0,\underline{v}_1,\ldots,\underline{v}_{L-1}$. In other words, in case of noisy speech, in this sub-space both speech and noise 'can be found', while in the orthogonal complement, that is in the sub-space, spanned by the vectors $\underline{v}_L,\underline{v}_{L+1},\ldots,\underline{v}_{N-1}$, only noise 'can be found'.

The noise suppression algorithm should be given in the form of a linear transformation $\underline{\underline{H}}$, that is

$$\tilde{\underline{s}}=\underline{\underline{H}}\cdot\underline{u}\,. \tag{9}$$

The estimation error is the remainder $\underline{r}=\underline{s}-\tilde{\underline{s}}$. The authors of [6] demonstrated, that the remainder signal

$$\underline{r}=\underline{r}^{SPEECH}-\underline{r}^{NOISE} \tag{10}$$

has two components. One of them correlated with the speech, while the other is correlated with the noise. Because of this the task is not only to minimize the speech-correlated component, but to suppress the noise-correlated component in a prescribed manner. This problem has been solved in [6] both in time domain and in spectral domain. Our results concerning the time domain have been published in [9]. In spectral domain it is also necessary to minimize the speech-correlated component, however it is possible to specify a noise suppression condition for every spectral component. Thus, the noise suppression problem can be formulated as a constrained optimization problem:

$$J\left(\underline{r}^{SPEECH}\right)\xrightarrow[\underline{\underline{H}}]{}\min \tag{11}$$

so that:

$$E\left\{\left|\underline{v}_n^T \cdot \underline{r}^{NOISE}\right|^2 \le \beta_n \cdot \sigma^2\right\} \quad n = 0,1,...L-1, \text{ and}$$
$$E\left\{\left|\underline{v}_n^T \cdot \underline{r}^{NOISE}\right|^2 = 0\right\} \quad n = L,...N-1. \tag{12}$$

The first condition is a component-wise specification for the remainder noise in the speech sub-space with conditions of $\beta_n > 0$, respectively, while the second condition is simply the zeroing the components in the noise sub-space. The optimal transformation matrix can be given by using the Karush-Kuhn-Tucker constraint optimization method as [6]:

$$\underline{\underline{H}}^{OPT} = \underline{\underline{V}} \cdot \underline{\underline{G}} \cdot \underline{\underline{V}}^T,$$
$$\underline{\underline{G}} = \mathrm{diag}\left(g_{0,0},..., g_{N,N}\right),$$
$$g_{n,n} = \begin{cases} \sqrt{\gamma_n} & n = 0,1,...,L-1 \\ 0 & n = L,...,N-1 \end{cases} \tag{13}$$

where the column vectors of the matrix $\underline{\underline{V}}$ are the eigenvectors. For the values of γ_n two methods can be found in [6]. In this paper the relationship below

$$\gamma_n = \exp\left(-\frac{\kappa \cdot \sigma^2}{\lambda_n^{SPEECH}}\right) \tag{14}$$

has been used. The degree of noise suppression can be set up with the experimental constant of $\kappa \ge 1$, also affecting the distorsion of the estimated speech.

The method described above can be generalised to the case of reconstructed phase space. Namely, the latter as a model background makes it possible to generate an M-dimensional data set from a given single noisy vector $\underline{u} = \underline{s} + \underline{w}$ by using the method of dimension embedding. Because of its construction, for the resulting trajectory matrix $\underline{\underline{U}}_{MxN}$ the following relationship holds

$$\underline{\underline{U}}_{MxN} = \underline{\underline{S}}_{MxN} + \underline{\underline{W}}_{MxN}. \tag{15}$$

Moreover, because for every corresponding sample the relationship $u_n = s_n + w_n$ holds, for the trajectory matrix-based covariance matrix we obtain:

$$\underline{\underline{K}}_{\underline{\underline{U}}_{MxN}} = \underline{\underline{K}}_{\underline{\underline{S}}_{MxN}} + \underline{\underline{K}}_{\underline{\underline{W}}_{MxN}}, \tag{16}$$

where

$$\underline{\underline{K}}_{\underline{\underline{W}}_{MxN}} = \sigma^2 \cdot \underline{\underline{I}}_{MxM}$$

Because of this, the noise suppression procedure above can also be applied for the estimation of the trajectory matrix $\underline{\underline{\tilde{S}}}$. Finally, from a given trajectory matrix estimate it is necessary to determine a speech vector estimate $\underline{\tilde{s}}$, which can be peformed based on the construction of the matrix $\underline{\underline{U}}$. This latter method differs from the original sub-space method not only in determining the data set necessary to determine the covariance matrix, but in the estimation of the speech sample as well, because the phase space-based method results several speech sample estimates for a given sample.

The trajectory matrix used in our work is based on a periodic extension of the noisy speech segment, so every speech sample has exactly M estimates, and the final estimate is their average. That is, in our case the weighting matrix for the final estimate is not necessary, while in other constructs it is needed [10]. More formally, the element $u_{i,j}$ of our trajectory matrix can be given as

$$u_{i,j} = u_{(j+i\cdot\tau) \bmod N}, \tag{17}$$

where N denotes the number of segment's samples, M denotes the embedding dimension, and τ denotes the time-lag. It is worthwile mentioning that our covariance matrix also differs from the empirical Toeplitz covariance matrix of [6] and from those published in [5] and [10].

4. Realisation of the speech enhancement algorithm and numerical results

In this work our goal was to demonstrate the method and the algorithm, so we have analysed only one long Hungarian sentence. The sentence was uttered by a native Hungarian male speaker, and the speech has been sampled with 8 kHz sampling frequency followed by a 16 bit linear quantisation. The resulting speech sample sequence was the 'clean' speech. However, even in this case the value of the global SNR was 45,8 dB (computed in active speech regions only). The noisy speech has been computed using these samples by artificially adding noise. The source of the noise samples is a part of the RSG-10 noise database [11]. Because of the different sampling frequencies, a suitable re-sampling was necessary before addition. The noise types, investigated in this work were the following: white noise, pink noise, high frequency channel noise. The noise level has been set up using the energy of the clean speech, computed in active speech regions.

The effectiveness of noise suppression has been charecterized by the number below

$$SRR = 10 \cdot \lg\left(E^{SPEECH} / E^{RESIDUAL}\right), \tag{18}$$

(Signal to Residual Ratio), where the nominator is the speech energy in the active speech region, and the denominator is the energy of the residual signal (computed on the same index-set as the nominator).

The noise suppression has been applied to a sequence of overlapped speech segments, using 50% overlap. The segment has been windowed using a Hanning-window before enhancement, and the final estimation has been computed by the overlap-add re-synthesis technique. The segment length, the embedding dimension, the time lag, the dimension of the speech sub-space and the value of the constant κ has been determined experimentally by many listening tests. The parameters γ_n, necessary for the spectral method, have been estimated as follows.

The value of σ^2 has been estimated by the first eigenvalue of the noise sub-space, while the values of λ_n^{SPEECH} have been estimated with the difference between the eigenvalues in the speech sub-space and the estimated value of σ^2. The computation of the eigenvalues and ei-

genvectors based on Jacobi's algorithm, and the noise suppression algorithm has been realised in C. *Table 1* contains the numerical results obtained – they correspond to those of published in the relevant literature [5,10].

SNR (dB)	SRR (dB)		
	White noise	High frequency channel noise	Pink noise
15	9,3	9,3	9,0
12	9,1	9,0	8,3
9	8,7	8,6	7,3
6	8,1	7,9	5,7
3	7,2	6,9	3,7
0	6,0	5,7	1,8
-3	4,5	4,2	0,0

Table 1.
The SRR values in cases of different SNRs and noise types (segment length: 800 sample, embedding dimension: 20, time-lag: 1 sample, speech sub-space dimension: 7, empirical constant: κ=5)

It is seen from Table 1, that SRR values are greater than SNRs, only if SNRs are lower than 6 dB. The reason is a property of the method itself discussed in Section 2, namely it is not only supresses the noise, but distortes the speech as well. The graphical illustration of the the algorithm can be seen in *Figure 1*.

It impressively demonstrates the noise suppression capability of the algorithm and also its speech distortion effect.

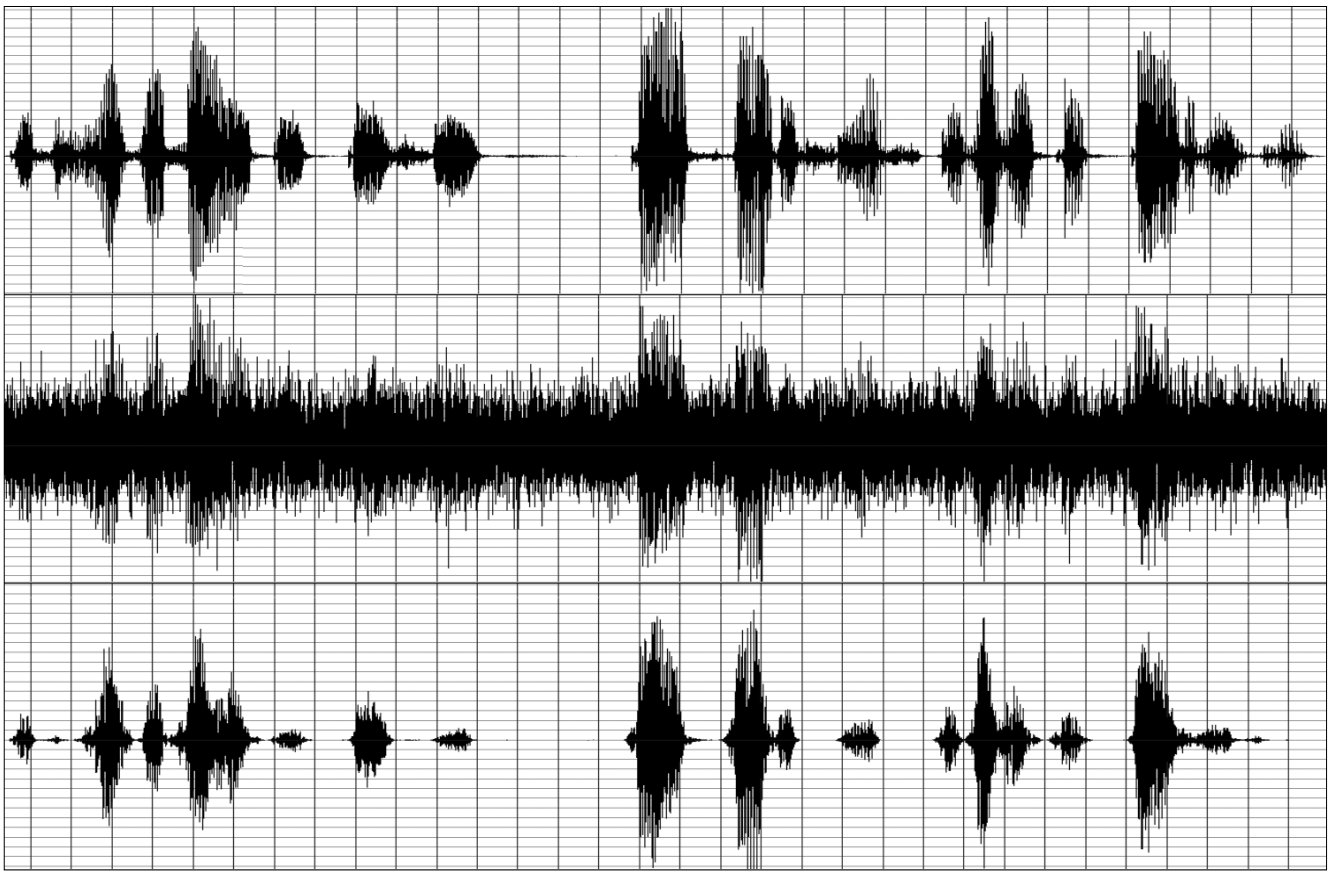
5. Conclusions

We discussed a speech enhancement algorithm, working in the reconstructed phase space.

The algorithm is based on dimension embedding, and assumes the separability the speech sub-space and the noise sub-space in the Euclidean space, determined by the covariance matrix of the data set after embedding. The Euclidean space in question is spanned by the eigenvectors of the covariance matrix above and the eigenvectors have been computed using Jacobi's algorithm. The data set has been determined after periodic extension of the speech segment, which differs from the published methods. That means, the so-called weighting matrix is not necessary for the estimation of the speech sample in our method.

The program has been tested using a Hungarian sentence by artificially added noise using three noise types and seven different noise levels. The enhancement capability has been determined numerically, the parameters have been set up experimentally by many listening tests. The best results have been achieved using the parameters as follows: about 100 ms segment length, 50% segment overlap, Hanning window, overlap-add re-

Figure 1.
Noise suppression in case of -3 dB and white noise in case of the same parameters as in Table 1. (upper trace: original utterance, middle trace: noise speech, lower trace: enhanced speech)



synthesis, 20 dimensional embedding space, 1 sample time lag, 7 dimensional speech sub-space.

The values correspond well to those published in the literature, not only in case of white noise, but in case of high frequency channel noise and pink noise. However, the algorithm is optimal only in case of white noise, for other noise types it is necessary to apply a whitening transformation.

Our further work is the automatic determination of the values of the embedding dimension, the time-lag and the dimension of speech sub-space, moreover the testing of the method using a large noisy speech database.

Acknowledgement

The author would like to thank Géza Gordos, Géza Németh and Péter Tatai for their kind help and encouragement in his speech processing algorithm development work.

Author

ISTVÁN PINTÉR received his MS degree in electrical engineering in 1983 and his PhD degree in informatics in 1997 from the Technical University of Budapest. In 1983 he joined the MIKI and in 1984 the GAMF (now College of Kecskemét, GAMF Faculty), where he is a professor. His research interests are digital speech processing (novel speech representations, speech enhancement), digital signal processing (discrete orthogonal transforms), pattern recognition (application of artificial neural networks). He has published several journal articles and conference papers in the areas above. He has received the Pollák-Virág award from HTE in 2007.

References

[1] J. S. Lim, A. V. Oppenheim:
Enhancement and bandwidth compression of noisy speech.
Proceedings of IEEE 67 (12), 1979, pp.1586–1604.

[2] Yang Gui, Kwan, H. K.:
Adaptive sub-band Wiener filtering for speech enhancement using critical-band gammatone filterbank,
Proceedings of 48th Midwest Symposium on Circuits and Systems, 2005, Vol. 1, pp.732–735.

[3] Haci Tasmaz, Ergun Ercelebi:
Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE-STSA estimation in various noise environments.
Digital Signal Processing,
(p.16, in press, available online 12 October 2007).

[4] T. F. Quatieri, R. B. Dunn:
Speech enhancement based on auditory spectral change.
Proceedings of International Conference on Acoustics, Speech and Signal Processing, Orlando, Florida, IEEE, 13-17 May 2002, pp.257–260.

[5] J. Sun, N. Zheng, X. Wang:
Enhancement of Chinese speech based on nonlinear dynamics.
Signal Processing 87, 2007, pp.2431–2445.

[6] Y. Ephraim, H. L. Van Trees:
A Signal Subspace Approach for Speech Enhancement.
IEEE Trans. on Speech and Audio Processing, Vol. 3, No.4., July 1995, pp.251–266.

[7] H. Kantz, T. Schreiber:
Nonlinear Time Series Analysis.
Cambridge University Press, 1997.

[8] G. Kubin, C. Lainscsek, E. Rank:
Indentification of Nonlinear Oscillator Models for Speech Analysis and Synthesis.
In: Chollet et al. (eds.):
Nonlinear Speech Modeling.
LN AI 3445, Springer Verlag, 2005, pp.74–113.

[9] I. Pintér:
Noise suppression using non-linear speech model.
Pollack Periodica, Vol. 2, Supplement, Akadémiai Kiadó, 2007, pp.121–133.

[10] M. T. Johnson, R. T. Povinelli:
Generalized phase space projection for nonlinear noise reduction.
Physica D 201, 2005, pp.306–317.

[11] http://spib.rice.edu/spib/select_noise.html