SPEECH ENHANCEMENT USING THE SPARSE CODE SHRINKAGE TECHNIQUE

I. Potamitis, N. Fakotakis, G. Kokkinakis

Wire Communications Laboratory, Electrical and Computer Engineering Dept.,
University of Patras, 261 10 Rion, Patras, Greece, Tel:+30 61 991722, Fax:+30 61 991855
e-mail: potamitis@wcl.ee.upatras.gr

ABSTRACT

Our work introduces the sparse code shrinkage (SCS) technique as a speech enhancement algorithm that aims at improving the quality of speech perception. SCS is a fairly new statistical technique originally presented to the applied mathematics and image denoising community, but, to our knowledge, its potential for speech enhancement has not yet been exploited. Its application on speech denoising gives rise to a conceptual framework which is quite different from the techniques dominating speech enhancement domain. SCS originates in applying Independent Component Analysis (ICA) to a large ensemble of clean speech frames, revealing their underlying basis of statistically independent functions. Projecting the frames composing a noisy speech signal on this basis, facilitates the application of Bayesian denoising to each of the resulting independent components individually. The maximum *a-posteriori* (MAP) formulation leads to a soft threshold function optimally adapted to the statistics of each independent component which effectively reduces white and coloured Gaussian noise. Subsequently, an inverse transformation from the ICA-transformed domain back to the time domain reconstructs the enhanced signal.

1. INTRODUCTION

The primary objective of noise compensation methods as applied in the context of speech processing is to reduce the effect of any signal which is alien to and disruptive of the message conveyed among participants in a communicative event (whether humans or ASR machines). Depending on the application, speech enhancement methods aim at speech quality improvement and/or speech or speaker recognition. The key difference is that in the latter case, the complexity of the effort undertaken by the recognizer is relaxed by a pre-processing transformation from the time domain to a domain with more desirable properties as regards the recognition process. When speech quality and intelligibility is the issue, it is essential that we respect the specific idiosyncrasies of human speech hearing and, therefore, reconstruct the time-domain signal. Due to the polymorphic manifestations and detrimental effect of noise, speech enhancement remains an open challenge.

Comprehensive assessments of noise compensation methods that belong to different speech processing strategies can be found in [1]. The subtractive and the attenuating type of filters are well-established, one-channel noise compensation methods that predominate in speech enhancement literature. Namely: a) spectral subtraction (SS) [2], b) least mean square (LMS), adaptive filtering [3], c) filter-based parametric approaches [4], d) model-based, short-time spectral magnitude estimation [5], e) hidden Markov model (HMM)-based speech enhancement techniques [6]. In what follows we foreground the main presuppositions of these techniques from which our approach departs:

a) Most aforementioned algorithms, are based on a transformation such as Discrete Fourier Transform [2][6], Discrete Cosine Transform [7], Karhunen-Loeve Transform [5], which facilitates the estimation of the clean speech model parameters. The transformation itself is more or less determined on an ad hoc basis. SCS is based on a data driven transformation kernel adapted to the structure of clean speech data.

b) Most methods focus on the distorted short-time amplitude of the speech signal leaving the phase unprocessed, on the assumption that the human ear does not perceive phase distortion [8]. On the other hand, experiments carried out in [9] experimentally demonstrated that as regards enhancement aiming at improving speech quality, phase has its share of importance. In the SCS technique the speech signal is processed uniformly, meaning that there is no inherent need that compels us to concentrate on amplitude and ignore phase processing.

c) The noise reduction process in the subtractive types of algorithms introduces a trade-off between distortion of spectral balance of the processed speech signal and noise suppression factor. In very noisy situations the enhanced speech can be even more disturbed than the corrupted speech signal in terms of intelligibility. SCS is not a subtractive technique and musical noise is perceivable at very low SNRs.

d) Even if the statistics of the degrading noise are known, SS and some versions of Wiener/Kalman and HMM-based algorithms require an accurate estimate of corrupting noise statistics. This is acquired during speech pauses through the use of a Voice Activity Detector (VAD). The construction of a robust VAD at low SNRs is a task still open to research. The framework of SCS has no need of a VAD to estimate speech-presence.

Though background noise can have spectral density that is case-specific to the operational environmental conditions, in many cases it can be adequately simulated as additive Gaussian. SCS is based on the assumption that background noise is additive and Gaussian (although it can be coloured). However, we suggest in theory and demonstrate in practice that SCS is quite robust in noise types that show modest deviation from normality.

We apply Independent Component Analysis to a large ensemble of frames derived from clean, phonetically balanced recordings, revealing their underlying independent component structure based on Bells' seminal work on the higher order structure of images and sounds [10] along with Hyvärinen's MAP formulation [11] on the independent bases of images. ICA is a statistical technique that determines a linear coordinate system, whose axes are defined by all higher moments of the

data to which it is applied. Projecting the frames of a noisy recording on this basis, the time-domain observations are linearly transformed so that the resulting data are as statistically independent as possible. MAP inference leads to a soft threshold function optimally derived from the statistics of each independent component that effectively reduces white and coloured Gaussian noise. Subsequently, an inverse transformation from the ICA transformed domain back to time domain reconstructs the enhanced signal.

We support theoretical derivations by extensive experimentation using recorded speech signals and real noise sources from the NOISEX-92 database. The assessment criteria are based on total and segmental signal to noise ratio (SNR) measures, as well as Itakura-Saito distortion measurements (allegedly correlated with subjective perception of speech quality). We also include visual comparison of speech spectrograms and informal listening tests.

2. PROBLEM FORMULATION

Consider a clean, time-domain speech signal x(m) subsequently corrupted by additive Gaussian noise n(m) producing the noisy signal y(m), where m is the sample index:

$$y(m)=x(m)+n(m) \quad (1)$$

We reshape signals y(m), x(m), n(m) as matrices formed by setting as columns, consecutive non-overlapping speech windows of 10 ms duration. The resulting matrices are of size NxF, where F denotes the number of frames and N=80 the 10 ms window size in samples at a sampling rate of 8 kHz. N is selected to guarantee stationarity and must be small enough to capture rapid varying phenomena as transients and stop bursts. After size normalization of the last frame (zero padding), the signals are represented as:

$$\mathbf{Y=X+N}, \quad (2)$$

where $\mathbf{Y}=[\mathbf{y}_1,..,\mathbf{y}_F]$, $\mathbf{X}=[\mathbf{x}_1,..,\mathbf{x}_F]$, $\mathbf{N}=[\mathbf{n}_1,..,\mathbf{n}_F]$. Each $\mathbf{x}_i$, $\mathbf{y}_i$, $\mathbf{z}_i$, i=1,..,F, is an N-dimensional vector corresponding to a frame.

Let assume the existence of a suitable orthogonal matrix $\mathbf{W} \in \Re^{NxN}$ which, when applied to each vector $\mathbf{x} \in \Re^{N}$ renders its components independent. We postpone the derivation of **W** until section 3. By applying **W** from the left side to Eq. 2 we get:

$$\mathbf{Z=WY=WX+WN=U+N'} \quad (3)$$

Gausssian distributions are invariant to linear transformations, therefore $\mathbf{N'=WN}$ is also Gaussian. Any $\mathbf{u} \in \mathbf{U} : \mathbf{U=WX}$, where $\mathbf{U}=[\mathbf{u}_1,..,\mathbf{u}_F]$, corresponds to an uncorrupted speech frame in the ICA transform domain. The components composing each **u** are now independent and in terms of p.d.f. can be set as:

$$f_{\mathbf{u}}(\mathbf{u}) = \prod_{i=1}^{N} f_{ui}(u_i) \qquad \forall\, \mathbf{u} \in \mathbf{U} \quad (4)$$

Since the statistics of noise and speech share the same statistical properties from frame to frame we drop subscript i denoting that the subsequent analysis holds for every vector of the corresponding matrices **Z**, **U**, **N'**.

$$\mathbf{z=u+n'} \quad (5)$$

The vector **z** corresponds to the noisy observation **y** in the ICA transformed domain and **n'** the transformed noisy frame. The posterior p.d.f. of **u** can be expressed according to the Bayes rule

$$f_{\mathbf{u/z}}(\mathbf{u}/\mathbf{z}) = \frac{f_{\mathbf{z/u}}(\mathbf{z}/\mathbf{u}) f_{\mathbf{u}}(\mathbf{u})}{f_{\mathbf{z}}(\mathbf{z})} = \frac{1}{f_{\mathbf{z}}(\mathbf{z})} f_{\mathbf{n'}}(\mathbf{z-u}) f_{\mathbf{u}}(\mathbf{u}) \quad (6)$$

The likelihood $f_{\mathbf{n'}}(\mathbf{z-u})$ is given by:

$$f_{\mathbf{n'}}(\mathbf{z-u})=(2\pi)^{-N/2}|\Sigma_{\mathbf{n'n'}}|^{-1}\exp(-1/2(\mathbf{z-u})^{\mathbf{T}}\,\Sigma_{\mathbf{n'n'}}^{-1}(\mathbf{z-u}) \quad (7)$$

From the minimization of the uniform cost function we obtain the classical MAP estimate of **u**. That is:

$$\bar{\mathbf{u}} = \underset{u}{\operatorname{argmax}} \left\{ f_{\mathbf{u/z}}(\mathbf{u}/\mathbf{z}) \right\} = \underset{u}{\operatorname{argmax}} \left\{ f_{\mathbf{n'}}(\mathbf{z-u}) f_{\mathbf{u}}(\mathbf{u}) \right\} \quad (8)$$

Assuming that there is no correlation of noise between components, and based on the factorization of $f_{\mathbf{u}}(\mathbf{u})$, we obtain the MAP estimate of **u**:

$$\bar{u} = \underset{u}{\operatorname{argmax}} \left\{ \prod_{i=1}^{N} f_{n'}(z_i - u_i) \prod_{i=1}^{N} f_{ui}(u_i) \right\} \quad (9)$$

As $\bar{\mathbf{u}}$ is now factorized, it can be decomposed and the denoising method can be applied to each individual component leading to:

$$\bar{u}_i = \underset{ui}{\operatorname{argmax}} \left\{ f_{n'_i}(z_i - u_i) f_{ui}(u_i) \right\} \quad (10)$$

Evaluating Eq.10 for $i = 1,.., N$ allows $\bar{\mathbf{u}}_i = [\bar{u}_1,..,\bar{u}_N]^T$ to be constructed and subsequently $\bar{\mathbf{U}} = [\bar{\mathbf{u}}_1,..,\bar{\mathbf{u}}_F]$. By making use of the relation $\bar{\mathbf{U}} = \mathbf{WX}$ and the orthogonality of **W**, we return back to time domain using the transformation:

$$\mathbf{X=W^TU} \quad (11)$$

Reshaping matrix **X** to vector form by concatenating all columns reconstructs the enhanced waveform. What remains to obtain is **W** and a closed-form of the densities $f_{ui}(u_i)$ required in Eq. 10.

3. INDEPENDENT COMPONENT ANALYSIS

Let $\mathbf{x}(t)=[x_1(t),..,x_N(t)]^T$ be a zero mean vector of scalar-valued components, where t is the time index. ICA seeks to find a suitable transformation matrix $\mathbf{W} \in \Re^{NxN}$ which, when applied to $\mathbf{x}(t)$, produces a vector $\mathbf{u}(t)=\mathbf{Wx}(t) \in \Re^{N}$ composed of variables that are as mutually independent as possible over time. That is:

$$MI(\mathbf{u}(t)) = \int f(\mathbf{u}(t)) \log \frac{f(\mathbf{u}(t))}{\prod_{i=1}^{N} f_{ui}(u_i(t))} d\mathbf{u}(t) = 0 \quad (12)$$

MI(**u**(t)) denotes the concept of mutual information that measures the level of dependency between the variables composing **u**(t) at each observation instance t. MI is strictly non-negative and zero only when the components comprised in a vector are independent, that is, when $f_{\mathbf{u}}(\mathbf{u}(t))$ is factorized.

In [12], different algorithmic versions of ICA can be put under the unifying framework of information maximization leading to the same iterative learning formula.

Information maximization: Let $\mathbf{y}(t)=g(\mathbf{WX})$ where g(.) has the form of cumulative density function of the prior distribution of the data which in the case of speech is super-Gaussian. Minimization of mutual information between the $y_i(t)$ components of **y**(t) implies minimization between the $u_i(t)$ components of **u**(t), since g (.) is an invertible mapping from $u_i(t)$ to $y_i(t)$ and the MI measure is invariant to component-wise monotonic transformations. **W** is iteratively calculated by taking the gradient of MI(**y**(t)) with respect to **W** (see [10],[12]).

$$\Delta\mathbf{W} \propto \frac{\partial MI(\mathbf{y})}{\partial \mathbf{W}} = [\mathbf{I} + (\frac{\partial f(\mathbf{u})}{\partial \mathbf{u}} / f(\mathbf{u}))\mathbf{u}^{\mathbf{T}}]\mathbf{W} \quad (13)$$

In order to derive the transformation matrix **W** we make use of 1000 clean, phonetically balanced recordings uttered by an equal

number of speakers of both genders. Each recording was sampled at 8 kHz sampling rate. Since **W** must learn all intrinsic information about quickly varying phenomena such as transients, we used small windows 10 ms long (80 samples with 79 samples overlap) to capture all the necessary detail.. A batch version of the infomax algorithm has been derived where **W** is iteratively calculated through successive presentations of the frames of each recording separately until convergence in order to avoid processing at once the huge amount of total observational data. The transformation kernel **W** is derived by applying ICA to the ensemble of observational frames from all recordings, denoted by $\mathbf{X}(t) \in \Re^{NxFt}$; where the time index corresponds to the frame index, N=80 corresponding to the samples of a 10 ms window and Ft is the number of total frames.

4. ESTIMATING THE SPARSE DENSITIES

We estimated the kurtosis of each component u_i defined as:

$$K(u_i) = \frac{E\{u_i^4\}}{E\{u_i^2\}^2} - 3 \qquad (14)$$

for every testing recording and we found with no exception that all components are super-Gaussian, therefore very sparse. (Refer to part 5 for the test set. Mean values of variance-normalized kurtosis over all test recordings are depicted in Fig. 1 and it is obvious that they diverge significantly from the zero value of normalized kurtosis of a Gaussian p.d.f.).

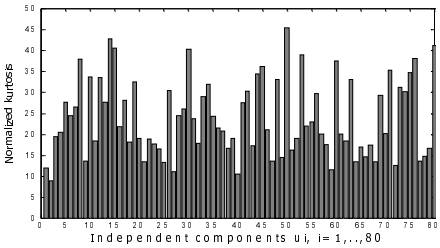

Fig. 1: Assessing the sparsity of independent components.

Therefore, the representative of a family of very sparse p.d.f.s is selected in advance. We adopted a density sparser than the Laplacian described in [12]:

$$f_{ui}(u_i) = \frac{1}{2d} \frac{(a+2)[a(a+1)/2]^{a/2+1}}{[\sqrt{a(a+1)/2} + |u_i/d|]^{a+3}} \qquad (15)$$

d denotes standard deviation and $a=1/d/(E^2\{|u_i|\}-1)$ controls the sparsity of the distribution. Substituting Eq. 15 in Eq. 10 and setting the derivative of the log-likelihood to zero results to a non-linearity applied to z_i (see [12] for details in derivation).

$$\bar{u}_i = \text{sign}(z_i)\max(0, \frac{|z_i| - bd}{2} + \frac{1}{2}\sqrt{(|z_i| + bd)^2 - 4\sigma^2(a+3)}) \qquad (16)$$

where I=1,..,N and $b = \sqrt{a(a+1)/2}$. As can be observed from the coefficient $|z_i|$-bd , the non-linearity has a thresholding, 'shrinking' effect by setting small values to zero.

5. EXPERIMENTAL RESULTS

5.1. Noise Types

From Eq. 3, it becomes obvious that every **z**(t) accumulates a number of noise components. According to the Central Limit Theorem for each **z**(t) component the density of noise will be closer to Gaussian than the distribution of each of the noise components. Therefore, we expect small impact on the effectiveness of SCS for noise cases which exhibit moderate divergence from the normality assumption. As regards the SNR of the recordings: each noise type is added to 34 clean speech files of 5 sec. mean duration so that the corrupted waveform ranges from –10 to 20 SNR_{dB}. To be more specific: let y1, y2, y be the clean/noisy/corrupted signal respectively. y=y1+H*y2, such that $10\log(E\{y1\}/E\{H*y2\})=SNR_{dB}$ where E{.} denotes energy. The objective criteria are based on the mean value over all recordings.

5.2. Objective criteria

Global SNR provides a simple error estimation over time and frequency although it has been questioned as to its suitability for speech quality assessment since it weights all time-domain errors equally while noise is known to be especially disruptive in low energy parts of the waveform. Let s(t) and $\hat{s}$(t) denote the original undegraded speech and enhanced speech signals, respectively. The total output SNR is defined:

$$SNR_{dB} = 10\log_{10} \frac{\sum_s s^2(t)}{\sum_s (s(t) - \hat{s}(t))^2} \qquad (17)$$

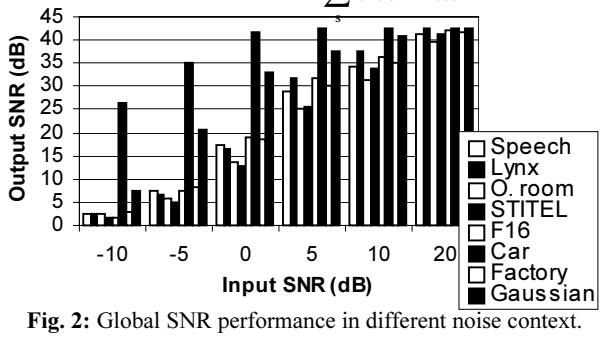

Fig. 2: Global SNR performance in different noise context.

SNR improvement may be misleading. In high input SNRs, the enhancement algorithm produces signals having retained a very small amount of residual noise which results in high SNR output values. In the -10 dB category the improvement for all noise types is impressive but the remaining noise and the distortion induced to the original signal may render it close to being inaudible. Objective measurements demonstrate that the algorithm shows its best performance for the case of car noise which is again misleading since we made use of speech files recorded over telephone lines. Telephone line inflicts a filter of 300-3400 Hz and the transformation matrix **W** which is adapted to the spectral characteristics of the recorded speech, leaves the part of the spectrum lower than 300 Hz unprocessed. Therefore, we made use of a fourth order high-pass Butterworth filter after enhancement to remove the part of the spectum that was not processed by the algorithm. Since car-noise is concentrated in low frequencies, the high-pass filter suppresses part of the noise. The second best performance is for the white Gaussian type of noise. We attribute the extensive denoising capability to the fact that white Gaussian noise complies with the assumptions of the algorithm, more than any other type of noise.

Segmental SNR is suggested to be better correlated with speech quality evaluations than global SNR and is calculated

for non-overlapping frames of 15 ms duration while the result is averaged over all waveform segments. Seg.SNR is defined:

$$\mathrm{SNR}_{\mathrm{dB}} = \frac{1}{\mathrm{M}} \sum_{j=0}^{M-1} 10 \log_{10} \frac{\sum_{\mathrm{s,M}} s^2(t)}{\sum_{\mathrm{s,M}} (s(t) - \hat{s}(t))^2} \qquad (18)$$

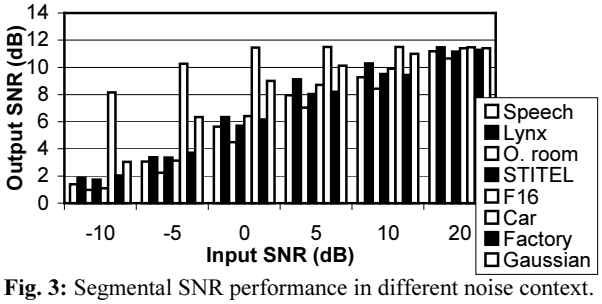

Fig. 3: Segmental SNR performance in different noise context.

The segmental SNR confirms our observation based on total SNR measure and further, it is able to order the performance of the algorithm clearly for the different noise types. Very good performance is observed for coloured types of noise, though, it seems that they form a category of their own compared to the Gaussian noise.

Itakura-Saito distortion measure is known to be closely associated with speech quality assessment since it is very sensitive to spectrum variations but not to phase distortion. It is based on the spectral distance between AR coefficient sets of the clean and enhanced speech waveforms over synchronous frames of 15ms duration. We used the median value of the distances between all frames in order to eliminate possible outliers.

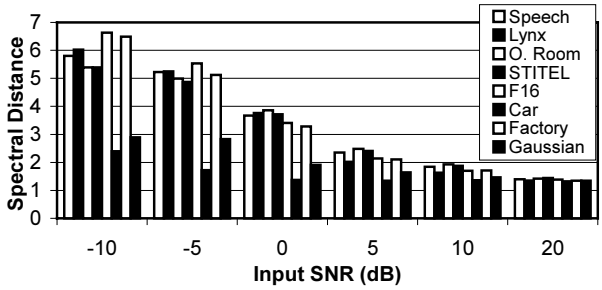

Fig. 4: Itakura-Saito performance in different noise context.

Itakura measure is heavily influenced due to mismatch in formant locations. From Fig. 4 we infer that the enhancement procedure distorts formants the same way for all coloured noises.

5.3. Subjective criteria

Fig. 5 shows the spectrogram of clean speech waveforms and the corresponding noisy versions corrupted by Gaussian, factory and speech noise types and the corresponding enhanced versions at 0 dB input SNR. The figures demonstrate extensive noise reduction. Close lookup does not reveal visually perceptible distortion of the formants due to the enhancement procedure. In row 2, (Factory noise), it can be observed that impulsively occurring components cannot be suppressed, a fact that we attribute to the violation of the stationarity assumption. Parallel listening tests confirm that these noises have non-stationary components (e.g. factory noise, operations room, F16) that remain intact in the enhanced version of the signal.

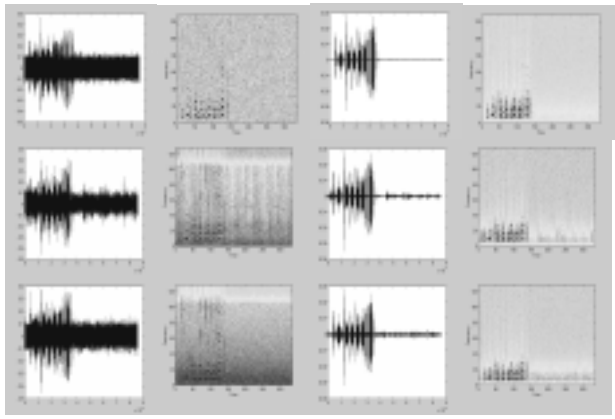

Fig. 5: Noisy and enhanced signals for: Row a) Gaussian, Row b) Factory, Row c) Speech noise types, at 0 dB input SNR.

6. SUMMARY AND CONCLUSIONS

A novel view for the enhancement of signals is applied successfully to speech. It is based on the idea of decomposing speech frames corresponding to 10 ms phonetic segments into their independent basis functions. This transformation facilitates the application of Bayesian inference to each of the resulting independent components separately. The MAP formulation leads to a shrinkage function optimally derived from the statistics of each component. Extensive experimentation with this technique gave excellent results in the case of white Gaussian noise. It proved very effective in coloured types of Gaussian noise that do not diverge significantly from the stationarity and the normality assumption. We observed considerable improvement of the enhanced signal versus the noisy one, in all objective criteria and most important, the preservation of natural sound. As regards the computation load, the most time-consuming task of the algorithm is the computation of the projection matrix **W**, which is computed off-line and only once for all subsequent restorations.

7. REFERENCES

[1] Gong Y., (1995), "Speech recognition in noisy environments: A survey," Speech Comm, 16, pp. 261-291.

[2] Boll S. (1979), "Suppression of acoustic noise using spectral subtraction," IEEE Trans. ASSP-27, pp. 113-120.

[3] Sambur M., (1978), "Adaptive noise canceling for speech signals," IEEE Trans. ASSP-26, No.5, pp. 419-423.

[4] Lim J., Oppenheim A., (1978), "All-pole modeling of degraded speech," IEEE Trans. ASSP-26, pp. 197-210.

[5] Ephraim Y., (1992), "Statistical model-based speech enhancement systems," Proc. IEEE, Vol. 80, pp. 1526-1555.

[6] Ephraim Y., (1992), "A Bayesian approach for speech enhancement using HMM," IEEE Trans. on SP, p. 725-735.

[7] Soon I., et all., (1998), "Noisy speech enhancement using DCT," Speech Communication, Vol 24, No. 3, pp. 249-259.

[8] Wang D., Lim J., (1982), "The unimportance of phase in speech enhancement," IEEE Trans. ASSP-30, pp. 679-682.

[9] Vary P., (1985), "Noise suppression by spectral magnitude estimation," Signal Processing, Vol. 8, pp. 387-400.

[10] Bell A., (1997), "The Independent Components of Natural Scenes are Edge Filters," Vision Research, pp. 3327-3338.

[11] Hyvärinen A., (2000), "Image denoising by Sparse Code Shrinkage," Intelligent Signal Processing, IEEE Press.

[12] Lee T., (1998), "Independent Component Analysis: Theory and Applications", Kluwer Academic Publishers.