MAP ESTIMATION FOR ON-LINE NOISE COMPENSATION OF TIME TRAJECTORIES OF SPECTRAL COEFFICIENTS

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

This paper presents a novel data driven compensation technique that modifies on-line the incoming spectral representation of degraded speech in order to approximate the features of high quality speech used to train a classifier. We apply the Bayesian inference framework to the degraded spectral coefficients based on the modeling of clean speech linear-spectrum with appropriate non-Gaussian distributions that allow maximum *a-posteriori* (MAP) closed form solution. The MAP solution leads to spectral magnitude estimation adapted to the spectral characteristics and noise variance of each spectral band. We perform extensive evaluation of our algorithm using white and coloured Gaussian noise on the task of improving the quality of speech perception as well as Automatic Speech Recognition (ASR), and demonstrate its robustness at very low SNRs. The enhancement process comes at little to no extra computational overhead for ASR systems, thus achieving real time performance.

1. Introduction

Although, ASR has been advanced to the point that it enables the launch of commercial products, the operational function of ASR machines is restricted by the influence of the acoustical environment. High quality speech used to train recognizers is deprived of various sources of variability that are common in operational use. This inconsistency is reflected on the degrading recognition score when we move from laboratory conditions towards real-world applications. ASR methods make implicit assumptions as to variability induced due to noise sources, by building speech sound models that are based on large speech corpora gathered in real conditions thus including in their construction common sources of variability. Nevertheless, it is practically impossible to incorporate sufficient speech training data in all different contexts, and to cope with unseen or poorly represented sources of degradation. One possible solution is to modify the training data by artificially adding noise segments from the operational environment and re-train the classifier. However, current attempts to solve this problem focus on the compensation of the noise effect in order to approximate the matched case of training and operational conditions. Comprehensive assessment of noise compensation methods, which are associated with three broad processing strategies can be found in [1]. They involve a transformation of the noisy waveform or feature vectors to speech/features used for the training of the recognizer, or corrections to the mean vectors and covariance matrices of the distributions of the clean HMM models to match the distribution of incoming degraded speech. In this work we try to reduce the mismatch between training and testing conditions by applying a nonlinear transformation to the feature-space of time trajectories of spectral coefficients. In the rest of this section we give an outline of the main presuppositions of successful techniques that apply transformations to the feature-space or to HMM models as well as our points of departure from these:

The distribution of log-spectral and cepstral coefficients is highly non-Gaussian and often multi-modal and asymmetric both for clean and degraded speech. HMM model based compensation of additive noise implicitly assumes that the corrupted speech distributions in log-spectral or cepstral domain are still Gaussian (see [2] and references therein). When a mixture of Gaussians is adopted to improve the approximation of the degraded distributions, time and memory costs pose serious constraints [2][3][4]. In [3] a Minimum Mean Square Error (MMSE) was used to infer the correction factors (mean vectors and covariance matrices) applied to mixture of Gaussians of the degraded speech in order to match the mixture pdf of clean speech. Unfortunately, as in [5] for the case of sub-band spectral estimation, MMSE did not permit closed form manipulation unless a series of simplifying assumptions were adopted. In [4] the corrupted cepstrum is assumed to follow the Gaussian mixture as is the case with the clean spectrum where the non-linearity imposed on the clean spectrum pdf, is approximated by a Taylor series. Extension of this work to model space is presented in [6]. Last but not least, the efficient MAP [7] and MLLR [8] techniques require rather long adaptation data from the noisy environment in order to obtain good estimates for the modification of the clean speech model set.

Our approach addresses the problem of noise compensation by casting it as a problem of Bayesian inference in the linear-spectral domain. We lay out its derivation in two steps:

a) We employ some well-known as well as some new probability distributions to account for the non-Gaussian representation of each band of a large ensemble of high quality clean speech. The pdf that best represents each band is inferred in the process of fitting an approximating distribution to the non-parametrically derived pdf (histogram method). The accuracy of the fit is assessed by the Kullback-Leibler (KL) divergence measure between each candidate distribution and the non-parametric estimation.

b) Based on the assumption that background noise is additive and Gaussian in the spectral domain, we use the same unified mathematical framework to derive the closed-form MAP estimation of the underlying spectral coefficient. Each band possesses its unique denoising formulation depending on the pdf that was selected in the previous step and on the noise variance of each band. 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 segmental signal to noise ratio (SNR) measurements, Itakura-Saito distortion measurements (allegedly correlated with subjective perception of speech quality), as well as word recognition accuracy of a speech recognition system. Subjective tests include visual comparison of spectrograms and informal listening tests.

2. Problem Formulation

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

$$x(m)=s(m)*h(m)+n(m) \qquad (1)$$

h(m) denotes the impulse response of the channel and {*} denotes convolution. Subsequently, based on the assumption of a linear, time invariant channel independent of the signal level and uncorrelated noise, we can derive a linear-spectral representation of Eq. 1 using Short Time Fourier Transform and a 2N point FFT.

$$|X(\omega_\kappa)|=H|S(\omega_\kappa)|+|N(\omega_\kappa)| \quad \kappa=0,\ldots,N \qquad (2)$$

where $|X(\omega_\kappa)|$ denotes the amplitude of the spectrum of the degraded sub-band $\omega_{\kappa,}$ {H} the constant over time channel effect and $|N(\omega_\kappa)|$, the noise spectral amplitude. The subsequent analysis is carried on in each band independently, which implicitly assumes independence among frequency components as in [11]. This is justified on the grounds of tractable mathematics that lead to exact formulations and real time performance. Moreover, the same analysis holds if we carry it in the linear spectrum prior to, or after applying the mel-scale filter-bank depending on whether we aim at improving the quality of perception of speech or at ASR. The posterior pdf of $H|S(\omega_\kappa)|$ can be expressed according to the Bayes rule

$$f(\mathrm{H}|S(\omega_\kappa)|\,||X(\omega_\kappa)|)=\frac{f(|X(\omega_\kappa)|\,|\mathrm{H}|S(\omega_\kappa)|)\,f(\mathrm{H}|S(\omega_\kappa)|)}{f(|X(\omega_\kappa)|)} \qquad (3)$$

The pdf of $f(||X(\omega_\kappa)||\,\mathrm{H}S(\omega_\kappa)|)$ ~$N(m_n,\sigma_n{}^2)$ when substituted to Eq. 3 formulates the MAP estimation for each individual spectral band leading to:

$$\overline{\mathrm{H}|S(\omega_\kappa)|_{map}}=\underset{\mathrm{H}|S(\omega_\kappa)|}{\arg\max}\left(\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left(-\left(\frac{|X(\omega_\kappa)|-\mathrm{H}|S(\omega_\kappa)|-m_n}{\sqrt{2}\sigma_n}\right)^2 \mathrm{f}(\mathrm{H}|S(\omega_\kappa)|)\right)\right) \qquad (4)$$

We proceed to the estimation of the appropriate spectral pdf of clean speech, that is, to the identification of the marginal pdf and its descriptive parameters for each spectral band.

2.1. Density Estimation of Clean Speech Spectral Bands

We examine the statistical behaviour of each spectral band over a large ensemble of clean recordings to derive the *a-priori* expectation of the underlying marginal pdf, which is needed in the Bayesian formulation of the denoising process.

The appropriate density to parametrize the distribution of the amplitude of each spectral band i=0,…,N in Eq. 2 is selected according to the smallest Kullback-Leibler (KL) divergence between the non-parametric density estimate of each band (histogram method) and a suitable, fitted parametric density. The KL-divergence measure between the non-parametric density $\{f_0\}$ and the selected representative pdf $\{f_s\}$, is always non-negative and zero in complete match.

$$KL(f_0, f_s)=\int f_0 \log\frac{f_0}{f_s}ds \qquad (5)$$

In order to account for the large variety of probability densities that characterize different bands of the ensemble, a family of flexible distributions is selected that allows for a closed-form MAP estimation to be derived under the assumption of additive Gaussian noise $N(m_n,\sigma_n{}^2)$ in each spectral band. The free parameters of each candidate density are assessed according to the Maximum Likelihood Estimation (MLE) criterion that gives the highest likelihood given the set of spectral observations of each band. After finding the parameters that describe the best the observational data for each pdf, all densities are tested against the non-parametric pdf $\{f_0\}$. The one that returns the lowest error according to KL-divergence is selected to represent the pdf of the spectral band. The family of curves that are tested actually comprise a cluster of variations:

Gamma pdf (Eq. 6) is a family of curves determined by two parameters that can take a versatile shape, effectively representing kurtotic, non-symmetric distributions. One should note that chi-square and exponential distributions can be derived form Gamma distribution by fixing one of the two free parameters, therefore MLE accounts for these cases also.

$$f(\mathrm{H}|\mathrm{S}(\omega_\kappa)|\,|\mathrm{a,b})=\frac{1}{\mathrm{b^a}\Gamma(\mathrm{a})}(\mathrm{H}|\mathrm{S}(\omega_\kappa)|)^{\mathrm{a}-1}\exp(-\mathrm{H}|\mathrm{S}(\omega_\kappa)|/\mathrm{b}) \qquad (6)$$

Gaussian-Laplacian pdf (Eq. 7) is a combination of the Gaussian and Laplacian density, suitable for representing moderately sparse distributions (sparser than Gaussian and less sparse than Laplace for the same variance).

$$f(\mathrm{H}|\mathrm{S}(\omega_\kappa)|\,|\mathrm{a,b})=\mathrm{C}\exp(-\frac{\mathrm{a}}{2}\mathrm{H}^2|\mathrm{S}(\omega_\kappa)|^2-\mathrm{bH}|\mathrm{S}(\omega_\kappa)|) \qquad (7)$$

The symmetric form of *Hyvärinen distribution* (Eq. 8) was originally presented in [8]. We adapted this distribution to enforce positivity in order to be consistent with the non-negative characteristic of the spectrogram. Hyvärinen's pdf is very sparse (sparser than any other pdf of the same variance) and proves very effective in capturing the pdf specifications of the spectral distributions of most of the bands.

$$f(\mathrm{H}|S(\omega_\kappa)|\,|a,b)=C(\sqrt{a(a+1)/2}+\mathrm{H}|S(\omega_\kappa)|/d)^{a+3} \qquad (8)$$

Gaussian pdf (Eq. 9) is suitable for modelling spectral bands that are slightly sub-Gaussian. Spectral bands with super-Gaussian pdfs are captured from the pre-mentioned pdfs.

$$f(\mathrm{H}|S(\omega_\kappa)|\,|\mathrm{m_s},\sigma_s)=\frac{1}{\sqrt{2\pi}\sigma_s}\exp\left(-\left(\frac{|X(\omega_\kappa)|-\mathrm{H}|S(\omega_\kappa)|-m_s}{\sqrt{2}\sigma_s}\right)^2\right) \qquad (9)$$

Spectral band probability model fitting which is the most time-consuming task of the algorithm is computed off-line and only once for all subsequent restorations.

2.2. Mean and Variance Noise Estimation

Background noise can have spectral density that is case-specific to the operational environmental conditions The symbols $\{m_n, \sigma_n{}^2\}$ in Eq. 4 denote that the algorithm requires a frequency dependent estimation of the mean and standard deviation of noise in each band. In the off-line denoising mode of the algorithm, noise can be estimated using a version of Martin's low-energy envelope tracking [12]. Specifically, we fit a normal pdf to the 55% of the lowest energies of each band and the mean value and standard deviation derived by Maximum Likelihood Estimation are used as the descriptive statistics for the noise pdf in every band. In order to address the frequency dependent SNR of coloured noise in an online way, we make use of Hirsch and Ehrlicher [12] adaptive estimation of the mean Eq. (10a) and standard deviation Eq. (10b) of noise:

$$|N_{k,i}(\omega_\kappa)|= \rho|N_{k,i-1}(\omega_\kappa)|+(1-\rho)|N_{k,\ i}(\omega_\kappa)| \quad (10a)$$

$$Var(|N_{k,i}(\omega_\kappa)|)=\rho Var\ (|N_{k,i-1}(\omega_\kappa)|)+(1-\rho)(|X_{k,\ i}(\omega_\kappa)|-N_{k,\ i}(\omega_\kappa)|)|^2 \quad (10b)$$

$$||X_{k,\ i}(\omega_\kappa)\ |-N_{k,\ i}(\omega_\kappa)||<\beta Var(|N_{k,i}(\omega_\kappa)|)^{1/2} \quad (10c)$$

where k=0,..,N indicates the band, {i} the current sample and 0.8<ρ<0.95. The update is applied only when Eq. (10c) holds, that is during speech-absence.

2.3. Spectral Amplitude Estimation

Inserting the density models in Eq. 5 and taking the derivative of the log-likelihood with respect to H|S(ω_κ)|, we get the MAP estimate of the underlying undistorted spectrum.
For the Gamma pdf the spectral estimator is:

$$\overline{H\,|\,S(\omega_\kappa)\,|_{map}} = \frac{b\,|\,X(\omega_\kappa)\,|\text{ -b}m_n\text{ -}\sigma_n{}^2}{2b} + \frac{\sqrt{(b\,|\,X(\omega_\kappa)\,|\text{ -b}m_n\text{ -}\sigma_n{}^2)^2+4b^2\sigma_n{}^2(a-1)}}{2b} \quad (11)$$

For the Gaussian Laplacian pdf the estimator is:

$$\overline{H\,|\,S(\omega_\kappa)\,|_{map}} = \frac{1}{1+a\sigma_n{}^2}(|\,X(\omega_\kappa)\,|-m_n-b\sigma_n{}^2) \quad (12)$$

For Hyvärinen’s pdf the estimator is:

$$\overline{H\,|\,S(\omega_\kappa)\,|_{map}} = \frac{|\,X(\omega_\kappa)\,|\text{-bd-}m_n}{2} + \frac{\sqrt{(|\,X(\omega_\kappa)\,|+bd)^2+2m_n(bd-|\,X(\omega_\kappa)\,|)-4\sigma_n{}^2(a+3)}}{2} \quad (13)$$

For the Gaussian pdf the amplitude estimator is:

$$\overline{H\,|\,S(\omega_\kappa)\,|_{map}} = \frac{\sigma_s{}^2}{\sigma_n{}^2+\sigma_s{}^2}(|\,X(\omega_\kappa)\,|-m_n) + \frac{\sigma_n{}^2}{\sigma_n{}^2+\sigma_s{}^2}m_s \quad (14)$$

One can clearly see the influence of a subtracting effect, (‘shrinking’), on $|X(\omega_\kappa)|$ where the magnitude of the shrinkage depends on the estimated noise variance and the expected distribution of the underlying clean spectral band.

There are several key-points that provide a strong impetus to the Bayesian viewpoint: Bayesian formulation allows a structured approach towards regulating the trade-off between distortion of spectral balance of the processed speech signal and noise suppression factor. Moreover, the Bayesian framework makes explicit use of the underlying probability distribution of the spectral coefficients in each band, which is essential in cases where the speech signal is completely masked by the overwhelming noise. Additionally, our technique has the advantage of not requiring an adaptation stage to the new operational environment.

Our algorithmic derivation is closely connected with [9][10] and with the wavelet shrinkage method [11]. It resembles a spectral subtraction in the spectral domain but it differentiates itself on being fully parametric and on deriving automatically its parameters by probability manipulation, thus avoiding the error prone procedure of empirically tuning the thresholds.

3. Results And Discussion

We conducted denoising experiments using artificially generated white Gaussian noise and six real coloured types of noise taken from the NOISEX-92 database. Each noise type was added to 100 clean speech files of 5 sec. mean duration so that the corrupted waveform ranges from –10 to 20 SNR_{dB}. The SNR improvement and the Itakura-Saito measurements of the enhancement obtained by our technique are shown in Fig. 1 and Fig. 2, while the word recognition results in Table 1.

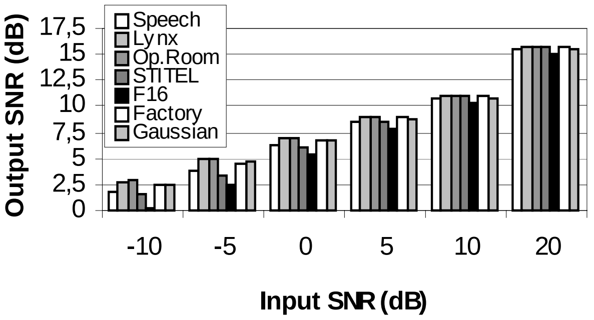

Fig. 1. Segmental SNR measurements.

In the -10 dB category the improvement for all noise types is impressive but the recognition accuracy is poor (see Fig 3). Parallel listening tests revealed that at such low SNRs musical noise is perceptible, though, noise-flooring the enhancement procedure proves very effective. Although the on-line version is able to track small noise variations, non-stationary components in certain noise types (e.g. factory noise, operations room, F16) remain intact in the enhanced version of the signal. Objective measurements demonstrate that the algorithm shows almost equal performance for both white Gaussian and coloured types of noise. We attribute the extensive denoising capability to the SNR varying estimation of noise statistics for each band independently.

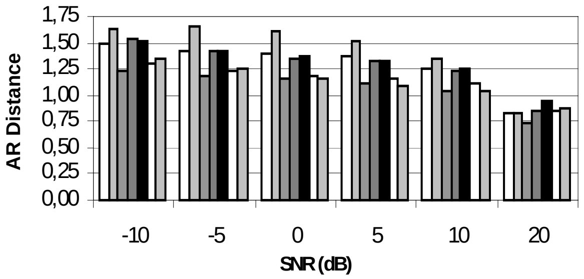

Fig. 2. Itakura-Saito AR-distance between the enhanced and the clean signals (same ordering of noises as in Fig. 1).

The Itakura-Saito distortion measure is very sensitive to speech spectrum variations but not to phase distortion and by general consensus; it is better associated with speech quality assessment. It is based on the spectral distance between AR coefficient sets of the clean and enhanced speech waveforms over synchronous frames of 15ms duration and is heavily influenced due to mismatch in formant locations. From Fig. 2 we infer that the quality of the enhanced signal degrades gracefully as SNR drops. One can observe that the optimum performance is not observed for Gaussian noise.

Our approach is formulated in the linear-spectral domain, therefore, depending on the application, it can be used for improving the quality of perception. In such a case we ought to respect the specific idiosyncrasies of human speech hearing and, therefore, reconstruct the time-domain signal. Based on the assumption that the human ear does not perceive phase distortion, we focus on the short-time amplitude of the speech signal leaving the noisy phase unprocessed. Phase is added back after the enhancement procedure has been applied and the time domain signal is subsequently reconstructed using the overlap and add method. Fig. 3 shows the spectrogram of noisy speech waveforms corrupted by Gaussian, Factory and Lynx helicopter noise types and their corresponding enhanced versions at -5 dB input SNR. The figures demonstrate extensive noise reduction.

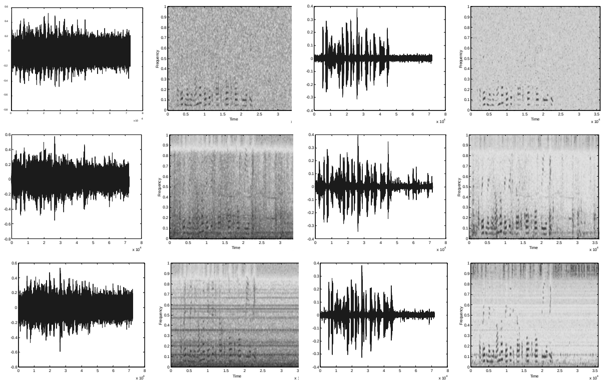

Fig. 3. Noisy and enhanced signals for: Row a) Gaussian, Row b) Factory, Row c) Lynx noise types at -5 dB input SNR

Listening tests and close lookup revealed a small amount of perceptible distortion in the form of musical noise. In Fig. 3, it can be observed that impulsively occurring components cannot be suppressed, a fact that we attribute to the violation of the stationarity assumption. Finally, enhanced signals are much more pleasant and quite perceptible even at –10 dB SNR

Word Recognition Accuracy was assessed by using a speech recognition module built with the HTK Hidden Markov Models toolkit. The basic recognition units are tied state context dependent triphones of five states each. The testing set consists of 100 files, part of the identity card corpus of the SpeechDat database (one speaker for each recording). The baseline word recognition accuracy for the clean acoustic environment was 96,07%. After applying the enhancement procedure, a 20 Mel-spaced triangular band-pass filter-bank was imposed to the spectrum. Thirteen dimensional feature vectors were formed after applying DCT to log-filter-bank outputs, which reduced the 20 output channels into 12 dimensional MFCC features plus a log-energy value. Cepstral mean normalization was applied to deal with the constant channel assumption. Deltas and double-Deltas were concatenated to form a 39-dimensional observational vector.

Noise	SNR	No Enh.	Enhanc.	Noise	SNR	No Enh.	Enhanc.
White Gaussian	-10	0	14.64	STITEL	-10	0	4.96
	-5	0	27.86		-5	0	12.59
	0	0	38.93		0	2.45	26.43
	5	5.71	52.86		5	47.14	55.36
	10	46.07	71.79		10	76.43	76.43
	20	88.93	91.07		20	93.94	92.86
Speech	-10	0	4.43	Factory	-10	0	14.64
	-5	0	24.64		-5	0	21.07
	0	0	48.93		0	0	42.50
	5	25.42	65.71		5	19.64	60.00
	10	68.7	75.36		10	63.14	73.57
	20	93.18	93.21		20	93.94	92.86
Lynx Helic.	-10	0	18.93	Operations Room	-10	0	17.86
	-5	0	35.00		-5	0	23.21
	0	0	57.14		0	1.36	39.64
	5	43.56	70.71		5	45.83	65.36
	10	80.30	86.79		10	81.82	83.21
	20	94.32	93.21		20	91.08	93.57

Table 1. (%) Word Recog. Acc. (off-line noise estimation).

4. Conclusions

We wish to emphasize the practical utility of our approach, which does not seek to revise but rather cooperates with already existing and successful front-ends that include in their construction an FFT stage (e.g. MFCC, PLP). Depending on the application, our speech enhancement method can be applied either for speech quality improvement and/or speech or speaker recognition. As our method works on a frame level basis on the spectral vectors already extracted for the recognition purpose, the restoration process comes at little to no extra computational overhead, thus achieving real time performance. We observed only a small impact on the effectiveness of our technique in noise cases, which exhibit moderate divergence from the normality assumption, and most important, the preservation of natural sound and good cooperation with the HMM framework, achieving consistently good results in very low SNRs. Moreover, the key idea of independent modeling of the spectral bands with non-Gaussian distributions selected according to the Kullback-Leibler divergence measure which leads to exact MAP formulation, can supply an efficient solution to a series of spectral estimation problems that requiring real-time response. Further work concentrates on incorporating a mixture of Gaussians into the noise model, which also results in closed-form MAP solutions, as well as on different estimation techniques of the noise variance that can be directly inserted in Eq.11-14.

5. References

[1] Gong Y., “Speech recognition in noisy environments”, *Speech. Communication,* 16, pp. 261-291, 1995.

[2] Gales M., “Predictive model-based compensation schemes for robust speech recognition”, *Speech Communication,* 25, pp. 49-74, 1998.

[3] Moreno P., Raj B., Stern R., “Data-driven environmental compensation for speech recognition: A unified approach *Speech Communication,* 24, pp. 267-285, 1998.

[4] Moreno P., Raj B., Stern R., “A vector Taylor series approach for environment-independent speech recognition”, *Proc. ICASSP,* pp. 733-736, 1996.

[5] Xie, F., Compernolle D., “Speech Enhancement by Spectral Magnitude Estimation: A unifying Approach”, *Speech Communication,* 19, pp. 89-104, 1996.

[6] Acero A., Deng L., Kristjansson T., Zhang J., “HMM adapation using vector Taylor series for noisy speech recognition”, *ICSLP Bejing,* pp. 869-872, 2000.

[7] Gauvain J., Lee C., “MAP estimation for multivariate Gaussian mixture observation of Markov Chains”, *IEEE Trans. Speech & Audio Processing,* 2, pp. 291-298, 1994.

[8] Leggetter C., Woodland P., “Maximum Likelihood Linear Regression for speaker adaptation of continuous density HMMs”, *Comp. Sp. & Lang.,* pp. 171-185, 1995.

[9] Hyvärinen A., Hoyer O., Oja E., “Sparse Code Shrinkage: Denoising of NonGaussian data by MLE”, *Tech. Rep. A51, Helsinki Univ. of Tech., Comp. Info. Sc.Lab.,* 1998.

[10] Potamitis I., Fakotakis N., Kokkinakis G., “Speech enhancement using the Sparse Code Shrinkage technique”, to appear in *Proc. of ICASSP, Utah,* 2001.

[11] Donoho D., “Denoising by soft-thresholding”, *IEEE Trans. Information Theory,* Vol. 41, pp. 613-627, 1995.

[12] Ris C., Dupont S., “Assessing local noise level estimation methods: Application to noise robust ASR”, *Speech Communication,* 34, pp. 141-158, 2001.