International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:1, No:9, 2007

Speaker Identification by Joint Statistical Characterization in the Log Gabor Wavelet Domain

Suman Senapati, Goutam Saha

***Abstract*—**Real world Speaker Identification (SI) application differs from ideal or laboratory conditions causing perturbations that leads to a mismatch between the training and testing environment and degrade the performance drastically. Many strategies have been adopted to cope with acoustical degradation; wavelet based Bayesian marginal model is one of them. But Bayesian marginal models cannot model the inter-scale statistical dependencies of different wavelet scales. Simple nonlinear estimators for wavelet based denoising assume that the wavelet coefficients in different scales are independent in nature. However wavelet coefficients have significant inter-scale dependency. This paper enhances this inter-scale dependency property by a Circularly Symmetric Probability Density Function (CS-PDF) related to the family of Spherically Invariant Random Processes (SIRPs) in Log Gabor Wavelet (LGW) domain and corresponding joint shrinkage estimator is derived by Maximum a Posteriori (MAP) estimator. A framework is proposed based on these to denoise speech signal for automatic speaker identification problems. The robustness of the proposed framework is tested for Text Independent Speaker Identification application on 100 speakers of POLYCOST and 100 speakers of YOHO speech database in three different noise environments. Experimental results show that the proposed estimator yields a higher improvement in identification accuracy compared to other estimators on popular Gaussian Mixture Model (GMM) based speaker model and Mel-Frequency Cepstral Coefficient (MFCC) features.

***Keywords*—**Speaker Identification, Log Gabor Wavelet, Bayesian Bivariate Estimator, Circularly Symmetric Probability Density Function, SIRP.

I. Introduction

AUTOMATIC Speaker Identification (ASI) is becoming a task of high relevance in many fields, specially for security purposes as biometric authentication tool. These systems, usually developed under laboratory conditions, severely degrade their performance level when an acoustical mismatch appears among training and testing phases. Such a problem has limited the development of real-world nonspecific applications, as testing conditions are highly variant or even unpredictable during the training process. This has guided researchers to design robust speaker identifiers by enhancing the speech coming from noisy environment. The process of providing robustness to the identifiers can be accomplished in the acoustical stage, giving rise to speech enhancement techniques that may improve the input signal with different types of noise at different Signal to Noise Ratio (SNR). Development of speech enhancement techniques remains an area of active interest towards development of a robust speaker identification system.

Among the one-channel approaches, the statistical spectral estimation methods [1]–[4] are shown to be effective for the noise reduction and to produce less speech distortion [3], [4]. The Gaussian modeling of speech and noise spectral components have been reported in the literatures and it was successfully combined with the Minimum Mean Square Error (MMSE) estimator in speech enhancement systems [3], [4]. The Gaussian assumption is indeed true in the asymptotic case of large Discrete Fourier Transform (DFT) frames when the span of correlation of the signal under consideration is much shorter than the DFT frame size. But the pdf of speech samples in the time domain as well as in DFT domain is much better modeled by a Laplacian or a Gamma density rather than a Gaussian density [5]–[9]. In last decade, the number of works reported on non-Gaussian modeling of speech has been increased [5]–[7], [10], [11]. In [5], an implementation of Gaussian model based Ephraim-Malah filter was presented. This is achieved by spectral amplitude estimation based on the generalized Gamma modeling of speech and MAP estimator. In [6], [7] authors proposed MMSE spectral components estimation approaches using Laplacian or a special case of the Gamma modeling of speech and noise spectrum. However the estimation presented in [7] are given just for a particular cases of the Gamma modeling, where the distribution parameters are fixed, and therefore it limits the application in general cases. The use of Gamma or Laplacian distributions, however, complicates the derivation of the MMSE estimate of the magnitude spectrum. This is partly because there is no analytical expression for the pdf of the magnitude of the DFT coefficients when the real and imaginary parts of the DFT coefficients are modeled by a Laplacian (or Gamma) distribution. Alternate solutions were explored in [10]–[13]. In [10], the authors approximated the pdf of the amplitude and phase of the DFT coefficients with a parametric function to

This work is partly supported by Indian Space Research Organization (ISRO), Govt. of India.

*S. Senapati is with Department of E & ECE, Indian Institute of Technology, Kharagpur 721302, India (e-mail: suman@ece.iitkgp.ernet.in).

G. Saha is with Department of E & ECE, Indian Institute of Technology, Kharagpur 721302, India (e-mail: gsaha@ece.iitkgp.ernet.in).

derive a joint MAP estimator. Martin *et al.* also use the super-Gaussian speech priors for MMSE Estimation of Magnitude-Squared DFT Coefficients [11]. In [12], a new algorithm for statistical speech feature enhancement in the cepstral domain is presented. The algorithm exploits joint prior distributions (in the form of Gaussian mixture) in the clean speech model, which incorporate both the static and frame-differential dynamic cepstral parameters. A noncausal estimator for the a priori Signal-to-Noise Ratio (SNR) and a corresponding noncausal speech enhancement algorithm is proposed in [13]. A Multi-band Spectral subtraction with adjusting subtraction factor method is given in [14]. On the other hand E. Zavarehei *et al.* enhance the speech using Kalman Filtering through restoration of short time DFT trajectories [15]. In [16], 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 [17]. The efficient MAP [18] and Maximum Log-Likelihood ratio (MLLR) [19] 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.

The wavelet transform has proved to be very successful in making signal and noise components of a noisy signal distinct. As wavelets have compact support the wavelet coefficients resulting from a signal are localized, whereas the coefficients resulting from noise in the signal are distributed. Thus the energy from the signal is directed into a limited number of coefficients which stand out from the noise. Wavelet shrinkage denoising consists of identifying the magnitude of wavelet coefficients from the noise (the threshold), and then shrinking the magnitudes of all the coefficients by this amount. The shrunk remains of the coefficients should represent valid signal data, and the transform can then be inverted to reconstruct an estimate of the signal [20]–[22]. A problem with wavelet shrinkage denoising is that the discrete wavelet transform is not translation invariant. If the signal is displaced by one data point the wavelet coefficients do not simply move by the same amount. They are completely different because there is no redundancy in the wavelet representation. Thus, the shape of the reconstructed signal after wavelet shrinkage and transform inversion will depend on the translation of the signal - clearly this is not very satisfactory.

The proposed work presents a new speech enhancement algorithm based on the decomposition of a noisy speech signal using complex-valued LGW coefficients. The underlying speech statistical model is representative for a class of stationary stochastic processes i.e. Spherically Invariant Random Processes (SIRPs) that has been introduced to model speech signals by [8], [23]. These processes are characterized by multivariate pdf depending on a quadratic form $x^T M^{-1} x$ of their arguments built up by the inverse covariance matrix M^{-1} where x is data vector. The Gaussian process represents SIRP in a better way, but there exists a variety of other SIRPs too, including various speech-model densities. For each case, the corresponding bivariate pdf exhibit ellipsoidal or circular contour lines, leading to the nomenclature 'spherical invariance'. The major advantage arises from the SIRP-model is that multivariate pdf can be derived analytically from the univariate one. From the theory of SIRPs it is known that these processes may be interpreted as a random mixture of Gaussian ones, which is equivalent to multiplying a Gaussian process with a randomly chosen constant. This work thus proposes a new joint non-Gaussian model to characterize the dependency between a coefficient and its parent and derives the corresponding bivariate MAP estimators based on noisy wavelet coefficients. The performance evaluations of the proposed estimators are done in ASI application context. The proposed method is compared against four different estimators. The Mel Frequency Cepstral Coefficient (MFCC) [24]–[26] is used to extract the features. Perhaps the most compelling reason for using the Mel warped cepstrum is that it has been demonstrated to work well in ASI systems. The robustness of the proposed model is presented with 100 speakers each for POLYCOST and YOHO speech database against three different noise contamination with SNR varying from 30dB to 0dB. The speaker model used is GMM based, considered most suitable for text-independent speaker recognition applications [27].

II. Problem Formulation

A. Wavelet based denoising & Sparseness Property

The speech signal corrupted by White Gaussian Noise (WGN) will be modeled as:

$$u = s + n \tag{1}$$

We observe u (a noisy signal) and wish to estimate the desired speech signal s as accurately as possible. The quantity n is the additive WGN that contains independent samples of a zero-mean Gaussian variable of variance σ^2_n. The goal of denoising signal is to recover s from the observed u. In wavelet-based denoising, it is applied to the noisy data yielding the noisy wavelet coefficients w; these are described by an analogous model:

$$w \equiv Wu = Ws + Wn = \theta + n' \tag{2}$$

where $\theta = Ws$ and $n' = Wn$. The standard Bayesian approach of the three step wavelet-based denoising is:

1) Compute the wavelet transform of the data $w = Wu$.
2) Obtain a Bayes estimate by given w.
3) Reconstruct the estimated speech signal s.

The wavelet transforms of most real-world signals tend to be dominated by a few large coefficients [28]. This is the so-called sparseness property which, in probabilistic terms, corresponds to a wavelet coefficient density function with a marked peak at zero and heavy tails; that is, a strongly non-Gaussian density (also called super-Gaussian). On the other hand, the DWT of WGN produces Gaussian distributed coefficients; these are bounded in magnitude by a suitable threshold proportional to their standard deviation. Therefore, a

natural denoising criterion results from this statistical difference between the coefficients of the signal and the noise: if the magnitude of an observed wavelet coefficient is large, its signal component is probably much larger than the noise and it should be kept; conversely, if a coefficient has small absolute value, it is probably due to noise and it should be attenuated or even removed.

B. Joint Subband Statistics

The coefficients of wavelet subbands are approximately decorrelated. Nevertheless, it is clear that wavelet coefficients are not statistically independent. Large magnitude coefficients tend to occur at neighboring locations, and also at the same relative locations of subbands at adjacent scales and orientations. We wish to explicitly examine and utilize the statistical relationship between wavelet coefficient magnitudes. Consider, two coefficients representing information at adjacent scales, but the same orientation (e.g., horizontal). Fig. 1 shows the joint histogram of the "child" coefficient conditioned on the coarser-scale of "parent" coefficient for speech signal taken from 50 speakers of POLYCOST database. The histogram illustrates several important aspects of the relationship between the two coefficients. First, they are (approximately) second-order decorrelated, since the expected value of "child" is roughly zero for all values of "parent". Second, the variance of "child" exhibits a strong dependence on the value of "parent". Thus,

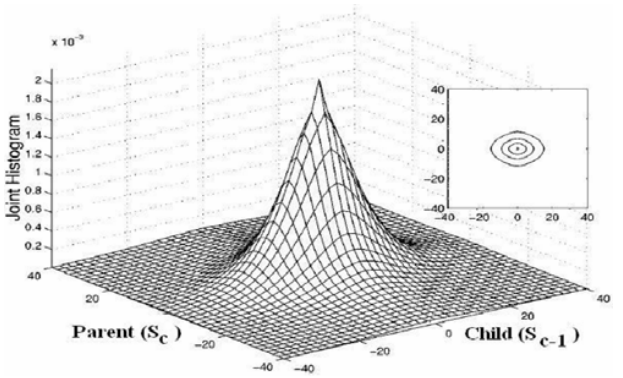

Fig. 1 Empirical joint parent-child histogram of Log Gabor wavelet coefficients

although "child" and "parent"' are uncorrelated, they are still statistically dependent.

III. Proposed Framework

A. LGW scheme and its performance

Gabor showed [29] how to represent time varying signals in terms of functions that are localized in both time and frequency. The LGW transform is used to obtain localized frequency information in a signal. To preserve such frequency information we must use nonorthogonal wavelets that are in symmetric/antisymmetric quadrature pairs. Here we follow the approach of Morlet et al. [30], but, rather than using Gabor filters, we prefer to use Logarithmic Gabor functions as suggested by Field [31]. These are filters having a Gaussian transfer function when viewed on the linear and logarithmic frequency scale. Log Gabor filters allow arbitrarily large bandwidth filters to be constructed while still maintaining a zero DC component in the even-symmetric filter. A zero DC value cannot be maintained in Gabor functions for bandwidths over one octave. It has a frequency response described by

$$G(f) = \exp\left[\frac{-(\log(f / f_0))^2}{2(\log(k / f_0))^2}\right] \quad (3)$$

where f_0 is the filter's centre frequency. To obtain constant-shape ratio filters (i.e. filters that are all geometric scaling of some reference filter) the term k/f_0 must also be held constant for varying $f0$. For example, a k/f_0 value of 0.75 will result in a filter bandwidth of approximately one octave and a value of 0.55 will result in a two-octave bandwidth. Details of Log Gabor Wavelet and specifications are found in [32].

B. Signal Analysis by LGW

The noisy time signal $u(l)$ sampled at regular time intervals $l \cdot T$ is composed of clean speech $x(l)$ and additive noise $n(l)$:

$$u(l) = s(l) + n(l) \quad (4)$$

where, $u(l)$ is the observed noisy speech signal, $s(l)$ is the original speech signal, and $n(l)$ is additive noise, uncorrelated with the original speech signal s(l). Taking the Fast Fourier Transform (FFT) the noisy coefficient $U(k)$ of frequency bin k consists of speech part $S(k)$ and noise $N(k)$:

$$U(K) = S(K) + N(K) \quad (5)$$

with $S = S_{Re} + jS_{Im}$ and $N = N_{Re} + jN_{Im}$, where, $S_{Re} = Re\{S\}$ and $S_{Im} = Im\{S\}$.

Analysis of noisy speech signal is done by multiplying the signal with each of the quadrature pairs of wavelets. If we let M^e and M^o denote the even-symmetric (cosine) and odd-symmetric (sine) wavelets at a scale c, we can think of the responses of each quadrature pair of filters as forming a response vector:

$$\begin{aligned}[f_c^e, f_c^o] &= [U \times M_c^e, U \times M_c^o] \quad (6)\\ &= [(S+N) \times M_c^e, (S+N) \times M_c^o]\\ &= S \times M_c^e + N \times M_c^e, S \times M_c^o + N \times M_c^o\\ &= S_c^e + N_c^e, S_c^o + N_c^o\end{aligned}$$

The values $f^e{}_c$ and $f^o{}_c$ can be thought of as real and imaginary parts of complex valued frequency component. The squared amplitude of the transform at a given wavelet scale is given by:

$$\begin{aligned}|A_c|^2 &= (f_c^e)^2 + (f_c^o)^2 \quad (7)\\ &= [S_c^e + N_c^e]^2 + [S_c^o + N_c^o]^2\\ &\le (S_c^e)^2 + (N_c^e)^2 + (S_c^o)^2 + (N_c^o)^2; E(SN) = 0\\ &= (S_c^e)^2 + (S_c^o)^2 + (N_c^e)^2 + (N_c^o)^2\\ &= (S_c)^2 + (N_c)^2\\ &= S_c + N_c; [let, S_c = (S_c)^2, N_c = (N_c)^2]\\ &= U_c\end{aligned}$$

We will have an array of these response vectors, one response vector for each scale of filter. These response vectors form the basis of our localized representation of the signal. The design of the LGW filter bank needs to be such that the transfer function of each filter overlaps sufficiently with its neighbors so that the sum of all the transfer functions forms a relatively uniform coverage of the spectrum.

C. Bayesian Bivariate Model

Marginal models from Laplacian, Gaussian, and Gamma etc. cannot model the statistical dependencies between LGW coefficients. However, there are strong dependencies between neighbor coefficients such as between a coefficient, its parent (adjacent coarser scale locations), and their siblings (adjacent scale locations). Let S_c represent the parent of S_{c-1}. (S_c is the LGW coefficient at the same position as S_{c-1}, but at the next coarser scale.) Then from (7)

$$U_{c-1} = S_{c-1} + N_{c-1} \tag{8}$$

$$U_c = S_c + N_c \tag{9}$$

where, U_{c-1} and U_c are noisy observations of S_{c-1} and S_c. N_{c-1}, N_c are noise samples. We can write

$$U^w = S^w + N^w \tag{10}$$

where, $S^w = (S_{c-1}, S_c)$, $U^w = (U_{c-1}, U_c)$ and $N^w = (N_{c-1}, N_c)$. The standard MAP estimator for S^w given the corrupted observation U^w is

$$\hat{S}^w(U^w) = \arg \max_{S^w} \; p(S^w \mid U^w) \tag{11}$$

Using Bayes rule, one gets

$$\hat{S}^w(U^w) = \arg \max_{S^w} [\, p(U^w \mid S^w) \cdot p(S^w)] \tag{12}$$

$$= \arg \max_{S^w} [\, p(U^w - S^w) \cdot p(S^w)]$$

From this equation, the Bayes rule allows us to write this estimation in terms of the probability densities of noise and the prior density of the LGW coefficients. In order to use this equation to estimate the original signal, we must know both pdf. We assume the noise is i.i.d. Gaussian, and we write the noise pdf as:

$$p(N^w) = \frac{1}{2\pi\sigma_{N^w}^2} \exp\left(- \frac{N_{c-1}^2 + N_c^2}{2\sigma_{N^w}^2} \right) \tag{13}$$

1) With Constant Inter-Scale Variance Model: It is hard to find a model for the empirical histogram in Fig. 1, so we propose the following pdf:

$$p(S^w) = \frac{\sqrt{3}}{2\pi\sigma_{S^w}^2} \exp\left(- \frac{\sqrt{3}\,(\sqrt{S_{c-1}^2 + S_c^2})}{\sigma_{S^w}} \right) \tag{14}$$

With this pdf, S_{c-1}, S_c is uncorrelated but not independent. This is a CS-PDF and is related to the family of SIRPs. Before going further with this new model, let us consider the case where S_{c-1}, S_c are assumed to be independent Laplacian (equn. (15)) and independent Gaussian (equn. (16)); then, the joint pdf can be written as

$$p(S^w) = \frac{1}{2\sigma_{S^w}^2} \exp\left(- \frac{\sqrt{2}\,(|S_{c-1}| + |S_c|)}{\sigma_{S^w}} \right) \tag{15}$$

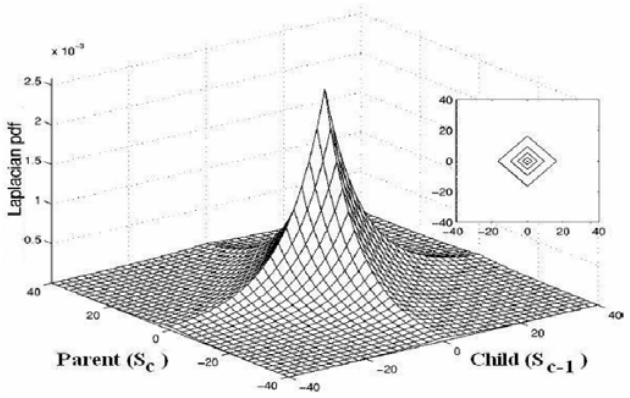

Fig. 2 Independent Laplacian model (15) for joint pdf of parent-child Log Gabor wavelet coefficient pairs

$$p(S^w) = \frac{1}{2\pi\sigma_{S^w}^2} \exp\left(- \frac{(S_{c-1}^2 + S_c^2)}{2\sigma_{S^w}^2} \right) \tag{16}$$

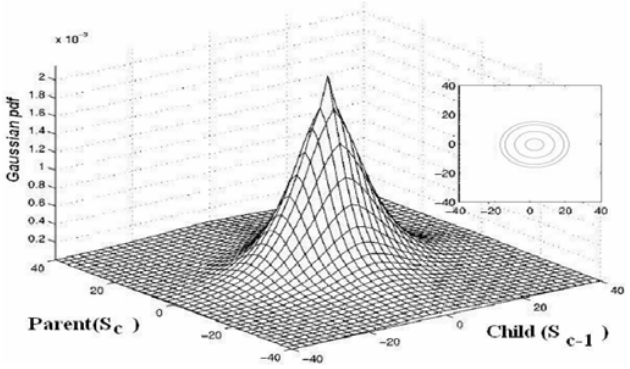

Fig. 3 Independent Gaussian model (16) for joint pdf of parent-child Log Gabor wavelet coefficient pairs

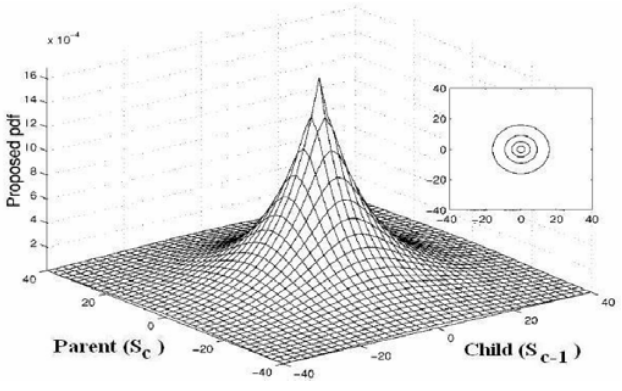

Fig. 4 New bivariate pdf (14) proposed for joint pdf of parent-child Log Gabor wavelet coefficient pairs

A plot of the independent Laplacian and Gaussian model is

illustrated in Fig. 2 and 3 respectively. If this model is compared with Fig. 1, the difference between them can be easily observed. Let us consider our proposed model, the plot of this pdf and its contour plot is illustrated in Fig. 4. As one can easily notice, this model is a much better approximation to the empirical histogram illustrated in Fig. 1.

Let us continue on developing the MAP estimator given in (12), which is equivalent to

$$\hat{S}^w(U^w) = \arg\max_{S^w}[\log[\,p(U^w - S^w)] + \log[\,p(S^w)]] \quad (17)$$

Let's define $f(S^w) = \log(p(S^w))$. By using (13), (17) becomes

$$\hat{S}^w(U^w) = \arg\max_{S^w}\left[-\frac{(U_{c-1}-S_{c-1})^2}{2\sigma^2_{N^w}} - \frac{(U_c - S_c)^2}{2\sigma^2_{N^w}} + f(S^w)\right] \quad (18)$$

This is equivalent to solving the following equations together, if $p(S^w)$ is assumed to be strictly convex and differentiable:

$$\left[\frac{(U_{c-1}-\hat{S}_{c-1})}{\sigma^2_{N^w}} + f_{c-1}(\hat{S}^w)\right] = 0 \quad (19)$$

$$\left[\frac{(U_c-\hat{S}_c)}{\sigma^2_{N^w}} + f_c(\hat{S}^w)\right] = 0$$

where, f_{c-1} and f_c represent the derivative of $f(S^w)$ w.r.t S_{c-1} and S_c respectively. The independent Laplacian model in (15) applies the soft threshold function to U_{c-1} to estimate S_{c-1} whereas the independent Gaussian model in (16) applies the wiener function to U_{c-1} to estimate S_{c-1}.

Let us find the MAP estimator corresponding to our proposed model given in (14), $f(S^w)$ can be written as

$$f(S^w) = \log\left(\frac{\sqrt{3}}{2\pi\sigma^2_{S^w}}\right) - \left(\frac{\sqrt{3}(\sqrt{S^2_{c-1}+S^2_c})}{\sigma_{S^w}}\right) \quad (20)$$

from this

$$f_{c-1}(S^w) = \left(-\frac{\sqrt{3}S_{c-1}}{\sigma_{S^w}(\sqrt{S^2_{c-1}+S^2_c})}\right) \quad (21)$$

$$f_c(S^w) = \left(-\frac{\sqrt{3}S_c}{\sigma_{S^w}(\sqrt{S^2_{c-1}+S^2_c})}\right)$$

Solving (19) by using (21), the MAP estimator (or "the joint shrinkage function") can be written as

$$\hat{S}_{c-1} = \frac{\sqrt{U^2_{c-1}+U^2_c} - \frac{\sqrt{3}\sigma^2_{N^w}}{\sigma_{S^w}}}{\sqrt{U^2_{c-1}+U^2_c}} \cdot U_{c-1} \quad (22)$$

The derivation can be found in Appendix A. Fig. 5 shows the plot of this bivariate shrinkage function. Denoising methods derived using the independence assumption disregard the parent value when estimating each coefficient. However, our results clearly show that the estimated value should depend on the parent value. The smaller the parent value, the greater the shrinkage. This result is very interesting because it illustrates the effect of taking into account the parent-child dependency. Note that when the parent value is zero, the MAP estimate of *Sc* is obtained by the soft threshold function.

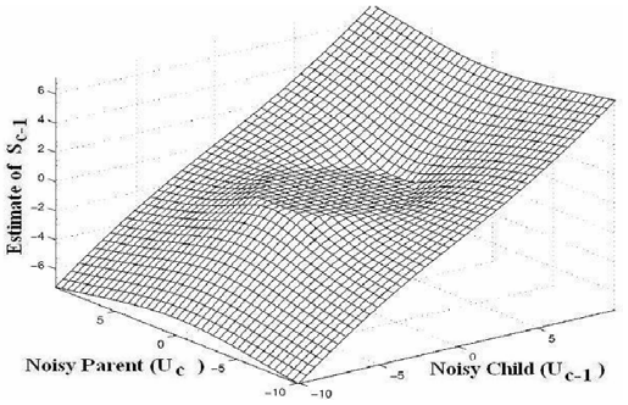

Fig. 5 New bivariate shrinkage function derived from the Constant Inter-Scale Variance Model proposed in (14) (fig. 4)

2) With Variable Inter-Scale Variance Model: Now if we consider the variance of the LGW coefficients are very much different from scale to scale, we would like to generalize the above Model. For this purpose, we propose a new Model, which has adjustable marginal variances, i.e.,

$$p(S^w) = \frac{\sqrt{3}}{2\pi\sigma_{S_{c-1}}\sigma_{S_c}}\exp\left(-\sqrt{3}\sqrt{\frac{S^2_{c-1}}{\sigma^2_{S_{c-1}}} + \frac{S^2_c}{\sigma^2_{S_c}}}\right) \quad (23)$$

Let us develop the MAP estimator for this model. From the pdf

$$f(S^w) = \left(-\sqrt{3}\sqrt{\frac{S^2_{c-1}}{\sigma^2_{S_{c-1}}} + \frac{S^2_c}{\sigma^2_{S_c}}}\right) \quad (24)$$

and this gives

$$f_{c-1}(S^w) = \left(-\frac{\sqrt{3}S_{c-1}}{\sigma^2_{S_{c-1}}\sqrt{\frac{S^2_{c-1}}{\sigma^2_{S_{c-1}}} + \frac{S^2_c}{\sigma^2_{S_c}}}}\right) \quad (25)$$

$$f_c(S^w) = \left(-\frac{\sqrt{3}S_c}{\sigma^2_{S_c}\sqrt{\frac{S^2_{c-1}}{\sigma^2_{S_{c-1}}} + \frac{S^2_c}{\sigma^2_{S_c}}}}\right)$$

Substituting (25) into the (19) gives

$$\hat{S}_{c-1}\left(1 + \frac{\sqrt{3}\sigma^2_{N^w}}{\sigma^2_{S_{c-1}}r}\right) = U_{c-1} \quad (26)$$

$$\hat{S}_c\left(1 + \frac{\sqrt{3}\sigma^2_{N^w}}{\sigma^2_{S_c}r}\right) = U_c$$

where, $$r = \left(\sqrt{\frac{\hat{S}^2_{c-1}}{\sigma^2_{S_{c-1}}} + \frac{\hat{S}^2_c}{\sigma^2_{S_c}}}\right)$$

These two equations do not have a simple closed-form solution like first Model. However, the solution can be found using iterative numerical methods. The solution using the successive substitution method is described in Appendix B.

IV. Database Description

A. YOHO Database

The YOHO database contains a large scale; high-quality speech corpus to support text-dependent speaker authentication research, such as is used in "secure access" technology. The data was collected in 1989 by ITT under a US Government contract to support Government secure access applications. A high-quality telephone handset (Shure XTH-383) was used to collect the speech; however, the speech was not passed through a telephone channel. YOHO was recorded in a fairly quiet office environment with low-level office noise, fan noise, and occasional pages over a public address system. The phrases are randomized and prompted in a text-dependent speaker verification scenario using "combination lock" phrase syntax.

TABLE I
YOHO Corpus Description

no. of speakers	138 (106 M / 32 F)
no. sessions/speaker	4 enrollments, 10 verifications
Intersession interval	Days-month (3 days nominal)
Type of speech	Prompted digit phrases
Microphones	Fixed high-quality in handset
Channels	3.8KHz/clean
Acoustic environment	Office

B. POLYCOST Database

The POLYCOST corpus was collected under the COST 250 European project. Most of the speech is non-native English with some speech in speaker's native tongue covering 13 European countries. The speech was collected digitally over international ISDN telephone lines. The different languages in this corpus allow for experimentation on the effect of language on speaker recognition performance.

TABLE II
POLYCOST Corpus Description

no. of speakers	133 (74 M /59 F)
no. sessions/speaker	> 5
Intersession interval	Days-weeks
Type of speech	Fixed and prompted digit strings, read sentences, free monologue
Microphones	Variable telephone handsets
Channels	Digital ISDN
Acoustic environment	Home/office

V. Performance Evaluation

A. Spectrogram Representation

The spectrographic representation of the proposed enhancement schemes were compared with the other enhancement schemes. Fig. 6 shows the spectrograms of a digits utterance thirty-one thirty-two thirty-three (a) in clean, (b) corrupted by SBN at 10dB SNR, (c) enhanced by Ephraim & Malah's Gaussian approach (Me-1) (d) enhanced by Martin's Laplacian approach (Me-2) (e) enhanced by Martin's Gamma approach (Me-3) (f) enhanced by Lotter & Vary's

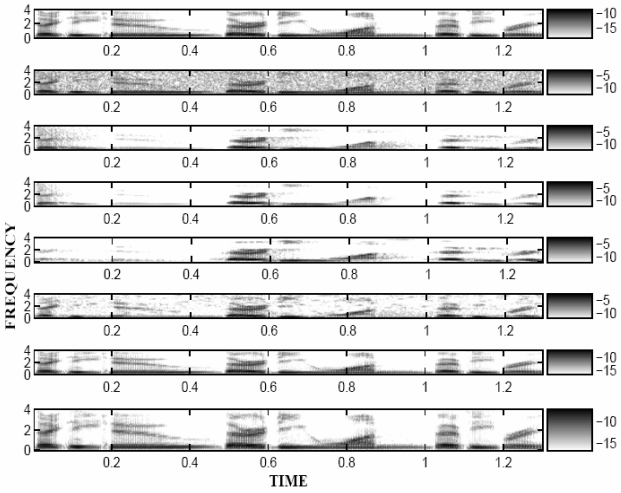

Fig. 6 Spectrogram of: (a) clean speech (b) speech corrupted by SBN at 10dB SNR (c) Me-1 enhanced speech (d) Me-2 enhanced speech (e) Me-3 enhanced speech (f) Me-4 enhanced speech (g) Proposed 1 enhanced speech (h) Proposed 2 enhanced speech

super Gaussian approach (Me-4) (g) enhanced by Proposed 1 approach and (h) enhanced speech by Proposed 2 approach. It is evident from the fig. 6 that the proposed schemes strike a better balance between the amount of noise removed and the amount of speech information retained compared to the other methods. Fig. 7 shows the spectrograms by CIN at 0dB SNR. As is obvious from the fig. 7, proposed methods lose very little of the speech information when the SNR level is 0dB. Notice that the proposed methods are able to enhance the weak higher formants while passing very little noise. It is also worth pointing out that only the proposed methods were able to process the digit corrupted by the most noise whereas the other methods pass it with little or no enhancement.

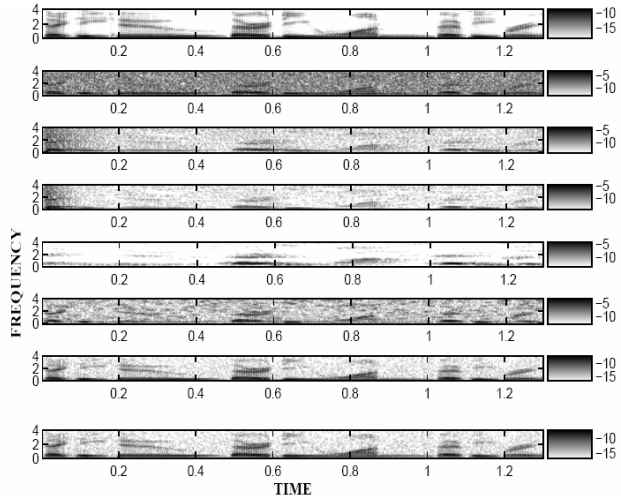

Fig. 7 Spectrogram of: (a) clean speech (b) speech corrupted by CIN at 0dB SNR (c) Me-1 enhanced speech (d) Me-2 enhanced speech (e) Me-3 enhanced speech (f) Me-4 enhanced speech (g) Proposed 1 enhanced speech (h) Proposed 2 enhanced speech

B. Identification Accuracy

TABLE III
IDENTIFICATION ACCURACY FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED BY PROPOSED AND EARLIER ESTIMATORS FOR YOHO SPEECH CORPUS

Noises	Used Method	Input SNR(dB)						
		0	5	10	15	20	25	30
WGN	M-1	72.40	80.00	87.66	92.40	95.20	96.80	97.40
	M-2	77.60	85.20	91.60	94.60	96.00	97.40	97.40
	M-3	84.80	87.60	93.60	95.60	96.80	97.40	97.40
	M-4	85.20	88.80	94.40	96.80	97.40	98.55	98.55
	Prop-1	**88.80**	**91.60**	**96.80**	**98.55**	**98.55**	**98.55**	**98.55**
	Prop-2	**89.40**	**92.00**	**97.40**	**98.55**	**98.55**	**98.55**	**98.55**
CIN	M-1	77.60	83.20	88.80	93.60	95.20	96.00	97.40
	M-2	80.00	87.60	92.00	94.60	96.00	97.40	97.40
	M-3	85.20	88.80	93.60	95.60	96.80	98.55	98.55
	M-4	86.80	91.60	94.40	96.80	97.40	98.55	98.55
	Prop-1	**90.00**	**92.40**	**96.00**	**98.55**	**98.55**	**98.55**	**98.55**
	Prop-2	**91.60**	**93.60**	**96.80**	**98.55**	**98.55**	**98.55**	**98.55**
FCN	M-1	71.60	79.20	88.80	94.00	95.20	96.80	97.40
	M-2	76.80	84.80	91.60	94.60	96.00	97.40	97.40
	M-3	83.20	85.20	94.40	95.60	96.80	97.40	98.55
	M-4	84.80	86.80	95.60	96.80	97.40	98.55	98.55
	Prop-1	**87.60**	**90.00**	**96.80**	**98.55**	**98.55**	**98.55**	**98.55**
	Prop-2	**88.80**	**91.60**	**97.40**	**98.55**	**98.55**	**98.55**	**98.55**

The performance of the proposed bivariate estimators are evaluated under three different noise conditions by computing the average improvement in the identification accuracy after enhancing noisy speech signals. The table III presents identification scores obtained from normal feature extraction technique, earlier as well as proposed estimators in case of YOHO Corpus. The performance results are averaged out using 100 different speakers, drawn from the YOHO speech database. For each speaker, there are 4 enrollment sessions of 24 utterances and 10 verification sessions of 4 utterances each. We used total $4 \times 24 = 96$ training utterances for building speaker models and total $4 \times 10 = 40$ test utterances are taken from 'VERIFY' session for testing. The noise signals include stationary White Gaussian Noise (WGN), Car Interior Noise (CIN) and F-16 Cockpit Noise (FCN), taken from the Noisex92 database. The speech signals are sampled at 8kHz and degraded by the various noise types in the range [0,30] dB. The proposed speech enhancement algorithms as well as the other methods are applied to the noisy speech signals. For comparison, the identification accuracy is also calculated by normal feature extraction method without noise removal (referred as M-1), Ephraim & Malah [3] (referred as M-2), Martin's Laplacian Prior [6] (referred as M-3) and Lotter's Super Gaussian Model [10] (referred as M-4). Table III presents the results of the identification accuracy for YOHO database using the various estimators where in case of proposed estimator it is shown for two cases of variances i.e. constant over scale (referred as Prop-1) and variable in between scales (referred as Prop-2). From the table III it is clearly shown that our proposed systems outperform than the other estimators. For WGN, the Prop-1 model gives 7.07% of identification accuracy improvement (averaged over all SNR levels) and Prop-2 model gives 7.15% improvement than normal feature extraction method without noise removal technique (i.e. M-1), 4.51% and 4.58% than Ephraim & Malah's MMSE estimator (i.e. M-2) which is considered as the baseline [10] of speech enhancement field. The Prop-1 model gives 1.67% and Prop-2 model gives 1.75% improvement over the closest Lotter & Vary's Super-Gaussian estimator (i.e. M-4). Note that method M-4 always gives closest but poorer performance than the proposed methods in all the test experiments. In case of CIN, the Prop-1 model gives 5.82% and Prop-2 model gives 5.88% improvement over M-1, 3.94% and 4% over M-2 whereas the Prop-1 model gives 1.21% and Prop-2 model gives 1.28% improvement than M-4. In case of FCN, the Prop-1 model gives 6.51% and Prop-2 model gives 6.58% improvement over M-1, 4.28% and 4.35% over M-2 whereas the Prop-1 model gives 1.44% and Prop-2 model gives 1.51% improvement than M-4. The proposed models give better identification accuracy improvement as well as average identification accuracy improvement over all dB levels than competing estimators.

The table IV presents identification scores obtained from POLYCOST Corpus. For each speaker, there are 10 enrolment sessions of 9 digit utterances out of which we used odd sessions for enrollment and even sessions for testing purposes i.e. total $5 \times 9 = 45$ training utterances for building speaker models for a speaker and total $5 \times 9 = 45$ utterances for testing. From the table IV it is noted that our proposed systems outperform the other estimators in this speech corpus also. The proposed models give better identification accuracy improvement for every dB level as well as for identification accuracy improvement over all dB levels. For WGN, the Prop-1 model gives 7.38% of identification accuracy improvement (averaged over all SNR levels) and Prop-2 model gives 7.47% improvement than M-1, 2.75% and 2.83% than M-2 and Prop-1 model gives 0.67% and Prop-2 model

TABLE IV
IDENTIFICATION ACCURACY FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED BY PROPOSED AND EARLIER ESTIMATORS FOR POLYCOST SPEECH CORPUS

Noises	Used Method	Input SNR(dB)						
		0	5	10	15	20	25	30
WGN	M-1	78.22	80.00	82.85	85.71	88.57	91.42	95.71
	M-2	82.85	87.14	90.00	91.42	92.85	94.28	96.37
	M-3	85.71	89.14	91.42	92.57	93.42	95.14	96.48
	M-4	87.42	90.00	91.85	92.85	94.85	96.00	96.51
	Prop-1	**87.74**	**90.54**	**92.42**	**93.68**	**95.71**	**96.68**	**97.42**
	Prop-2	**87.88**	**90.60**	**92.48**	**93.77**	**95.77**	**96.77**	**97.55**
CIN	M-1	79.62	82.85	83.14	86.00	89.14	91.85	96.00
	M-2	85.71	86.00	88.57	89.14	91.85	94.68	96.57
	M-3	87.42	87.74	89.14	90.54	93.42	95.85	96.77
	M-4	88.57	89.14	90.54	93.42	94.68	96.00	96.80
	Prop-1	**90.00**	**91.42**	**93.68**	**94.68**	**96.00**	**96.80**	**97.71**
	Prop-2	**90.10**	**91.48**	**93.74**	**94.77**	**96.08**	**96.89**	**97.78**
FCN	M-1	79.05	81.23	83.14	86.00	88.57	90.54	95.71
	M-2	83.14	85.71	87.74	88.57	92.66	94.28	96.48
	M-3	86.00	87.42	90.54	92.66	93.42	95.85	96.57
	M-4	87.74	89.14	91.85	93.42	94.85	96.00	96.62
	Prop-1	**88.57**	**91.85**	**93.42**	**94.28**	**95.85**	**96.77**	**97.57**
	Prop-2	**88.64**	**91.96**	**93.48**	**94.38**	**95.92**	**96.84**	**97.64**

gives 0.75% improvement than the M-4. Note that in this case also method M-4 always gives closest but poorer performance than the proposed methods in all the test experiments. In case of CIN, the Prop-1 model gives 7.38% and Prop-2 model gives 7.44% improvement over M-1, 3.96% and 4.04% over M-2 whereas the Prop-1 model gives 1.59% and Prop-2 model gives 1.67% improvement than M-4. In case of FCN, the Prop-1 model gives 7.72% and Prop-2 model gives 7.79% improvement over M-1, 4.24% and 4.31% over M-2 whereas the Prop-1 model gives 1.24% and Prop-2 model gives 1.31% improvement than M-4. The proposed systems showed that the use of our second model (i.e. variable inter-scale variances) resulted in small improvement on performance over Prop-1 bivariate Model (with constant inter-scale variance). It is also noted that the results are poorer in case of POLYCOST speech corpus than YOHO may be due to the fact that the first one (i.e. POLYCOST) is telephone based and second one (i.e. YOHO) is microphone based speech corpus.

VI. Conclusion

In this paper, two new bivariate distributions are proposed based on the class of Spherically Invariant Random Processes for Log Gabor Wavelet coefficients of speech signals to characterize the dependencies between a coefficient and its parent in speaker identification application. This is followed by derivation of corresponding bivariate shrinkage functions using Bayesian MAP estimation. The Proposed 1 model maintains the simplicity, efficiency, intuition of the classical soft thresholding approach whereas Proposed 2 model is proposed in order to characterize larger group of distributions. To show the effectiveness of these new estimators, three noise contaminations (WGN, CIN & FCN) over two public speech databases (YOHO & POLYCOST), one collected from microphone speech and the other from telephone speech, are presented for automatic speaker identification problem. It is shown that both methods yield higher identification accuracy than others. Of the proposed two methods, the second model (i.e. variable inter-scale variances) always perform better than the first model (with constant inter-scale variance) but requires numerical techniques (e.g. successive substitution) for solution while a closed form relation exists for the first. We obtained these results by observing the dependencies between coefficients and their parents. It is expected that the results can be further improved if the other dependencies between a coefficient and its other neighbors are exploited.

Acknowledgment

The work is partly supported by Indian Space Research Organization (ISRO), Government of India.

Appendix

A. Derivation of the Shrinkage Function With Constant Inter-Scale Variance Model

Solving (19) by using (21), we get

$$\hat{S}_{c-1}\left(1+\frac{\sqrt{3}\sigma^2_{N^w}}{\sigma^2_{S^w} r}\right) = U_{c-1} \tag{27}$$

$$\hat{S}_{c}\left(1+\frac{\sqrt{3}\sigma^2_{N^w}}{\sigma^2_{S^w} r}\right) = U_{c}$$

where,

$$r = \left(\sqrt{\hat{S}^2_{c-1} + \hat{S}^2_{c}}\right)$$

Using (27)

$$r^2 = \frac{(U_{c-1})^2}{\left(1+\frac{\sqrt{3}\sigma^2_{Nw}}{\sigma_{Sw}\cdot r}\right)^2} + \frac{(U_c)^2}{\left(1+\frac{\sqrt{3}\sigma^2_{Nw}}{\sigma_{Sw}\cdot r}\right)^2}$$

$$or, \left(r+\frac{\sqrt{3}\sigma^2_{Nw}}{\sigma_{Sw}}\right)^2 = (U_{c-1})^2 + (U_c)^2$$

$$or, r = \left(\sqrt{(U_{c-1})^2 + (U_c)^2} - \frac{\sqrt{3}\sigma^2_{Nw}}{\sigma_{Sw}}\right) \tag{28}$$

Substituting the value of r into (27) the MAP estimator (or "the joint shrinkage function") can be written as

$$\hat{S}_{c-1} = \frac{\sqrt{U^2_{c-1}+U^2_c} - \frac{\sqrt{3}\sigma^2_{N^w}}{\sigma_{S^w}}}{\sqrt{U^2_{c-1}+U^2_c}} \cdot U_{c-1}$$

B. Derivation of the Shrinkage Function With Variable Inter-Scale Variance Model

Step 1: $U_{c-1} = S_{c-1} + N_{c-1}$ and $U_c = S_c + N_c$,
where, $U_{c-1}, S_{c-1}, N_{c-1} \in c-1$ and $U_c, S_c, N_c \in c$.

Step 2: Since we consider variable inter-scale variances
therefore, $\sigma^2_{U_{c-1}} = \sigma^2_{S_{c-1}} + \sigma^2_{N^w}$ and $\sigma^2_{U_c} = \sigma^2_{S_c} + \sigma^2_{N^w}$.

Step 3: $\sigma^2_{U_{c-1}}$ and $\sigma^2_{U_c}$ can be found by, $\hat{\sigma}^2_{U_{c-1}} = \frac{1}{L^2_{c-1}} \sum_{U_{(c-1)i} \in c-1} U^2_{(c-1)i}$
and $\hat{\sigma}^2_{U_c} = \frac{1}{L^2_c} \sum_{U_{(c)i} \in c} U^2_{(c)i}$, where, L_{c-1}
and L_c are the size of the scales

Step 4: Estimate the noise variance $\hat{\sigma}^2_{Nw}$.

Step 5: Estimate the $\sigma_{S_{c-1}}$ and σ_{S_c} as, $\hat{\sigma}_{S_{c-1}} = \sqrt{\hat{\sigma}^2_{U_{c-1}} - \hat{\sigma}^2_{N^w}}$
and $\hat{\sigma}_{S_c} = \sqrt{\hat{\sigma}^2_{U_c} - \hat{\sigma}^2_{Nw}}$.

Step 6: Initialize $\hat{S}^{[0]}_{c-1}$, $\hat{S}^{[0]}_{c}$ and k.

Step 7: Calculate r using, $r = \sqrt{\left(\frac{\hat{S}^{[k]}_{c-1}}{\sigma_{c-1}}\right)^2 + \left(\frac{\hat{S}^{[k]}_{c}}{\sigma_c}\right)^2}$.

REFERENCES

[1] Boll, S. F., "Suppression of Acoustic Noise in Speech using Spectral Subtraction", *IEEE ASSP*, 27(2):113-120, 1979.
[2] Berouti M., Schwartz R., and Makhoul J., "Enhancement of speech corrupted by acoustic noise", *IEEE ICASSP*, 1979, vol. 1, pp. 208-211.
[3] Y. Ephraim and D. Malah, "Speech Enhancement using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.
[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator", *IEEE Trans. on Acoust., Speech ,Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.
[5] T. H. Dat, K. Takeda and F. Itakura, "Generalized Gamma Modeling of Speech and its Online Estimation for Speech Enhancement", *Proceedings of ICASSP-2005*, 2005.
[6] R. Martin and C. Breithaupt, "Speech Enhancement in the DFT Domain using Laplacian Speech Priors", *in Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC 03)*, pp. 87-90, Kyoto, Japan, Sep. 2003.
[7] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors", *IEEE ICASSP'02*, Orlando, Florida, May 2002.
[8] H. Brehm, E.W. J¨ungst and D. Wolf, "Simulation von Sprachsignalen", *AE¨U*, Vol. 28, 1974, pp. 445-450.
[9] W. B. Davenport, "An experimental study of speech wave probability distributions", *J. Acoust. Soc. Amer.*, Vol. 24, July 1952, pp. 390-399.
[10] Thomas Lotter and Peter Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model", *EURASIP Journal on Applied Signal Processing* , vol. 2005, Issue 7, pp. 1110-1126.
[11] C. Breithaupt and R. Martin, "MMSE Estimation of Magnitude-Squared DFT Coefficients with Super-Gaussian Priors", *IEEE Proc. Intern. Conf. on Acoustics, Speech and Signal Processing*, vol. I, pp. 896-899, April 2003.
[12] Deng, J. Droppo, and A. Acero. "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features", *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, May 2004, pp. 218-233.
[13] I. Cohen, "Speech Enhancement Using a Noncausal A Priori SNR Estimator", *IEEE Signal Processing Letters*, Vol. 11, No. 9, Sep. 2004, pp. 725-728.
[14] S. Kamath and P. Loizou, "A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise", *In Proceedings International Conference on Acoustics, Speech and Signal Processing*, 2002.
[15] E. Zavarehei, S. Vaseghi and Q. Yan, "Speech Enhancement using Kalman Filters for Restoration of Short-Time DFT Trajectories", *Automatic Speech Recognition and Understanding (ASRU), 2005 IEEE Workshop*, Nov. 27, 2005, pp. 219 - 224.
[16] Moreno P., Raj B., Stern R., "A vector Taylor series approach for environment-independent speech recognition", *Proc. ICASSP*, pp. 733-736, 1996.
[17] Acero A., Deng L., Kristjansson T., Zhang J., "HMM adapation using vector Taylor series for noisy speech recognition", *ICSLP Bejing*, pp. 869-872, 2000.
[18] Gauvain J., Lee C., "MAP estimation for multivariate Gaussian mixture observation of Markov Chains", *IEEE Trans. Speech & Audio Processing*, 2, pp. 291-298, 1994.
[19] Leggetter C., Woodland P., "Maximum Likelihood Linear Regression for speaker adaptation of continuous density HMMs", *Comp. Sp. & Lang.*, pp. 171-185, 1995.
[20] D. L. Donoho, "De-noising by soft-thresholding", *IEEE Transactions on Information Theory*, 41(3):613-627, 1995.
[21] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage", *Biometrika*, 81(3):425-455, 1994.
[22] R. R. Coifman and D. Donoho, "Time-invariant wavelet denoising", In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, volume 103 of *Lecture Notes in Statistics*, pages 125-150, New York, 1995. Springer-Verlag.
[23] H. Brehm, "Description of spherically invariant random processes by means of G-functions", *in: Lecture Notes in Computer Science*, Vol. 969, Springer, New York, 1982, pp. 39-73.
[24] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. On ASSP*, vol. ASSP 28, no. 4, pp. 357-365, Aug. 1980.
[25] Molla, M. K. I., and K. Hirose, "On the effectiveness of mfccs and their statistical distribution properties in speaker identification", *in Virtual Environments, Human-Computer Interfaces and Measurement Systems, VCIMS2004 IEEE Symposium*, July 12-14, 2004, pp. 136-141.
[26] R. Vergin, D. OShaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition", *IEEE Trans. On Speech and Audio Processing*, vol. 7, no. 5, pp. 525-532, Sep. 1999.
[27] Douglas A. Reynolds, Richard C. Rose, "Robust Text- Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, pp. 72-83, vol. 3, no. 1, January 1995.
[28] D. Donoho and I. Johnstone, "Ideal adaptation via wavelet shrinkage", *Biometrika*, vol. 81, pp. 425-455, 1994.
[29] D. Gabor, "Theory of communication", J. Inst. Electr. Eng. 93, pp. 429457, 1946.
[30] J. Morlet, G. Arens, E. Fourgeau and D. Giard, "Wave Propagation and Sampling Theory - Part II: Sampling theory and complex waves", *Geophysics*, 47(2):222-236, Feb. 1982.
[31] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells", *Journal of the Optical Society of America A*, 4(12):2379-2394, Dec. 1987.
[32] S. Senapati and G. Saha, "Speech Enhancement by Marginal Statistical Characterization in Log gabor Wavelet domain", *International J. of Signal Processing*, vol. 4, no. 2, pp. 107-113, 2007.

Suman Senapati received the B. Tech. in Electronics & Telecommunication Engg. in 2001 from University 0f Kalyani, India. He got his M. Tech. degree in Optics & Optoelectronics in 2003 from University of Calcutta, India. He has been with Indian Institute of Technology, Kharagpur, India since 2004, where he is research scholar of Electronics and Electrical Comm. Engg. department. During 2003-04 he was a Research Assistant at Applied Physics department, University of Calcutta, India under Prof. Asit K. Datta. His research interests are Speech Enhancement and Speech Processing.

Goutam Saha graduated in 1990 from Dept. of Electronics & Electrical Communication Engineering, Indian Institute of Technology (IIT), Kharagpur, India. The author worked in Tata Steel, India in the period 1990-1994, joined IIT Kharagpur as CSIR research Fellow in 1994 and completed Ph. D work in 1999. He worked in Institute of Engineering & Management, Salt Lake, Kolkata as a faculty member during 1999-2002 and since 2002 serving IIT Kharagpur as Assistant Professor till date. An active researcher in the field of speech processing, biomedical signal processing, modeling and prediction he has published papers in reputed journals like Physical Review E, IEEE Trans. on Systems, Man & Cybernetics, IEEE Trans. on Biomedical Engineering etc.