

# Speech Enhancement by Marginal Statistical Characterization in the Log Gabor Wavelet Domain

Suman Senapati, and Goutam Saha

**Abstract**—This work presents a fusion of Log Gabor Wavelet (LGW) and Maximum a Posteriori (MAP) estimator as a speech enhancement tool for acoustical background noise reduction. The probability density function (pdf) of the speech spectral amplitude is approximated by a Generalized Laplacian Distribution (GLD). Compared to earlier estimators the proposed method estimates the underlying statistical model more accurately by appropriately choosing the model parameters of GLD. Experimental results show that the proposed estimator yields a higher improvement in Segmental Signal-to-Noise Ratio (S-SNR) and lower Log-Spectral Distortion (LSD) in two different noisy environments compared to other estimators.

**Keywords**—Speech Enhancement, Generalized Laplacian Distribution, Log Gabor Wavelet, Bayesian MAP Marginal Estimator.

## I. INTRODUCTION

SPEECH enhancement attempts to improve one or more perceptual aspects of voice communication systems when the signal is corrupted by noise.

Among the one-channel approaches, the statistical spectral estimation methods [1] [2] are shown to be effective for the noise reduction and to produce less speech distortion [2]. 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 [2]. 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. In last decade, the number of research on non-Gaussian modeling of speech has been increased, where the approaches are carried out in different ways [3]–[7]. In [3], an implementation of Gaussian model based Ephraim-Malah filter was reported. This is achieved by spectral amplitude estimation based on the generalized gamma modeling of speech and MAP estimator. In [4], [5] 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 [5] is given just for the particular cases of the gamma modeling, where the distribution parameters are fixed, and therefore it limits the application in general cases. Alternative solutions were explored in [6]–[9]. For instance, in [6] the authors approximated the pdf of the amplitude and phase of the DFT coefficients with a parametric function to derive a joint MAP estimator. Martin *et al.* also use the super-Gaussian speech priors for MMSE Estimation of Magnitude-Squared DFT Coefficients [7]. In [8], 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 SNR and a corresponding noncausal speech enhancement algorithm is proposed in [9]. A Multiband Spectral subtraction with adjusting subtraction factor method is given in [10]. On the other hand E. Zivarehei *et al.* enhance the speech using Kalman Filtering for restoration of short time DFT trajectories [11].

It is well known that the pdf of speech samples in the time domain and DFT domain is much better modeled by a Laplacian or a Gamma density rather than a Gaussian density [4], [5]. The proposed work presents a new speech enhancement algorithm, based on the decomposition of a noisy speech signal using LGW coefficients. A method to automatically determine the appropriate shrinkage rule from the statistics of the squared amplitude response is proposed that uses Bayesian MAP Marginal Model. The pdf of the speech spectral amplitude is modeled with a simple GLD, which allows a high approximation accuracy for Laplace distributed real and imaginary parts of the speech wavelet coefficients. The pdf of the noise spectral amplitude is modeled with a zero mean Gaussian distribution. The statistical model is designed so that it fits the distribution of the speech spectral amplitudes to improve the quality of the enhanced speech signal. The proposed method is compared against four different estimators for YOHO speech corpus and POLYCOST speech corpus in two different noisy environments.

Manuscript received January 12, 2007. 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).

II. PROPOSED FRAMEWORK

A. Log Gabor Wavelet Scheme

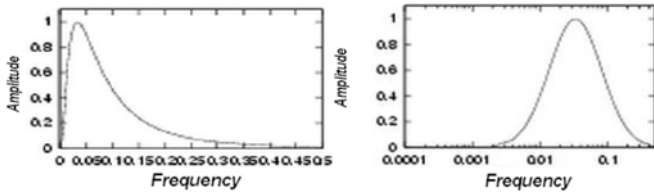


Fig. 1 Log Gabor Transfer functions viewed on linear and logarithmic frequency scales

Gabor showed [12] 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 non-orthogonal wavelets that are in symmetric/antisymmetric quadrature pairs. Here we follow the approach of Morlet et al. [13], but, rather than using Gabor filters, we prefer to use Logarithmic Gabor functions as suggested by Field [14]. These are filters having a Gaussian transfer function when viewed on the linear and logarithmic frequency scale (Fig. 1). 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:

$$G(f) = \exp \left[ \frac{-(\log(f/f_0))^2}{2(\log(k/f_0))^2} \right] \quad (1)$$

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  $f_0$ . 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.

B. Importance of Log Gabor Wavelet

The LGW is used to obtain localized frequency information. The use of the Wavelet Transform for frequency analysis was developed by Morlet et al. [13]. The basic idea behind wavelet analysis is to use a bank of filters to analyze the signal. The filters are all created from rescaling of the one wave shape, each scaling designed to pick out a particular band of frequencies from the signal being analyzed. An important point is that the scales of the filters vary geometrically, giving rise to a logarithmic frequency scale. Since we are interested in calculating local frequency in signals, we follow the approach of Morlet, that is, using wavelets based on complex valued Gabor functions - sine and cosine waves, each modulated by a Gaussian. Using two filters in quadrature enables one to calculate the squared amplitude of the signal for a particular frequency.

The important aspect of Log Gabor function is that, unlike

the Gabor function, the frequency response of the Log Gabor is symmetric on a log axis. Indeed, the log axis is the standard method for representing the frequency response. Gabor functions miss in this fit primarily because they fail to capture the relative symmetry on a log axis. The advantage of the Log Gabor is its use in which the bandwidths increase with frequency. With the constant bandwidths, the Gabor functions over-represent the low frequencies. In contrast, mapping the information into the Log Gabor spreads the information equally across the scales. At bandwidths of > 1 octave, the redundancy at the low frequencies becomes apparent. Since all the Gabor filters receive a significant and redundant input from the low frequencies, the responses represent a smaller fraction of the total energy. The frequency responses of the Log Gabor filter permits a more compact representation than the Gabor filter when the bandwidths are > 1 octave.

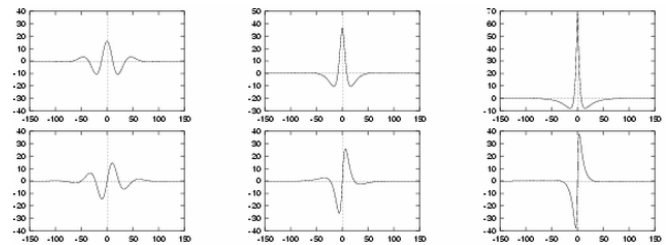


Fig. 2 Three quadrature pairs of LGWs all tuned to the same frequency, but having different bandwidths 1, 2 and 3 octaves respectively

C. Log Gabor filters in the Frequency Domain

The Fig. 2 shows three quadrature pairs of log Gabor filters of different bandwidths all tuned to the same centre frequency. It can be noted that as bandwidth increases, the sharpness of the filter also increases. Therefore, one constraint might be imposed by the maximum sharpness of the filter that we can effectively represent and a useful objective might be to minimize the width of filters in order to get maximal localization of signal's frequency information.

In the frequency domain the even symmetric filter is represented by two real-valued log-Gaussian functions symmetrically placed on each side of the origin (Fig. 3(a)) and the odd-symmetric filter by two imaginary valued log-Gaussian functions anti-symmetrically placed on each side of the origin (Fig. 3(b)). Exploiting the linearity of the Fourier Transform where  $FFT(A + B) = FFT(A) + FFT(B)$  we can do the following: Multiply the FFT of the odd-symmetric filter by  $i = \sqrt{-1}$  (to make it real valued) and add it to the FFT of the even symmetric filter. The anti-symmetric function from the odd-symmetric filter will cancel out the corresponding symmetric function from the even-symmetric filter. This leaves a single function (multiplied by 2) on the positive side of the frequency spectrum. Thus, if we construct a filter in the frequency domain with a single log-Gabor function on the positive side of the frequency spectrum we can consider this filter to be the sum of the FFT of the even and odd symmetric

filters (with the odd symmetric filter multiplied by  $i$ ). If we perform the convolution by multiplying this frequency domain filter by the FFT of the signal, we end up with the even-symmetric convolution residing in the real part of the result and the odd-symmetric convolution residing in the imaginary part.

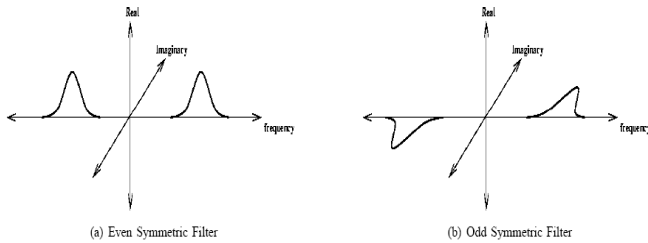


Fig. 3 Log Gabor filter Transfer Functions

**D. Design & Specifications of Log Gabor Filter Bank**

The aim is to produce a filter bank that provides even coverage of the spectrum of the represented signal. This can be achieved by making the overlap of the filter transfer functions sufficiently large so that when one sums up the individual transfer functions the net result is an even coverage of the spectrum. Thus, every point in the spectrum ends up being represented equally in the final result. For computational reasons, this even coverage of the spectrum has to be achieved with a minimal number of filters.

The second aim is to ensure that the outputs of the individual filters in the bank are as independent as possible. The whole aim of applying the filter bank is to obtain information about speech signal; if a filter's outputs are highly correlated with those of its neighbors then we have an inefficient arrangement of filters that do not provide as much information as they should. To achieve independence of output, the filters should have minimal overlap of their transfer functions. Thus the transfer functions of our filters should have the minimal overlap necessary to achieve fairly even spectral coverage. The parameters are given below:

- 1) The Maximum frequency: The maximum frequency is set by the wavelength of the smallest scale filter.
- 2) The Minimum frequency: The minimum frequency is set by the wavelength of the largest scale filter.
- 3) The filter bandwidth: The filter bandwidth is set by specifying the ratio of the standard deviation of the Gaussian.
- 4) The scaling between centre frequencies of successive filters: Having set a filter bandwidth decides the scaling between centre frequencies of successive filters.

The specifications of the Log Gabor Filter Bank are given in Table I.

**E. 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  $n(l)$ :

$$u(l) = x(l) + n(l) \tag{2}$$

TABLE I  
SPECIFICATIONS OF LOG GABOR FILTER BANK

|  |   |   |
|--|---|---|
| 1. Maximum Frequency : MaxFreq                                     | → | 3                                       |
| 2. Maximum Frequency : MinFreq                                     | → | $\frac{1}{MaxFreq * Mult^{(TotFil-1)}}$ |
| 3. Filter Bandwidth : sigmaOnf                                     | → | 0.55                                    |
| 4. Scaling between center frequencies of successive filters : Mult | → | 3                                       |
| 5. No. of Filters : TotFil   | → | 4                                       |

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

$$U(K) = X(K) + N(K) \tag{3}$$

with  $X = X_{Re} + jX_{Im}$  and  $N = N_{Re} + jN_{Im}$ , where,  $X_{Re} = Re\{X\}$  and  $X_{Im} = Im\{X\}$ .

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, we can think of the responses of each quadrature pair of filters as forming a response vector:

$$\begin{aligned} [f_1^e, f_1^o] &= [U \times M_1^e, U \times M_1^o] \tag{4} \\ &= [(X + N) \times M_1^e, (X + N) \times M_1^o] \\ &= X \times M_1^e + N \times M_1^e, X \times M_1^o + N \times M_1^o \\ &= X_1^e + N_1^e, X_1^o + N_1^o \end{aligned}$$

The values  $f_1^e$  and  $f_1^o$  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:

$$\begin{aligned} |A_1|^2 &= (f_1^e)^2 + (f_1^o)^2 \tag{5} \\ &= [X_1^e + N_1^e]^2 + [X_1^o + N_1^o]^2 \\ &\leq (X_1^e)^2 + (N_1^e)^2 + (X_1^o)^2 + (N_1^o)^2; E(XN) = 0 \\ &= (X_1^e)^2 + (X_1^o)^2 + (N_1^e)^2 + (N_1^o)^2 \\ &= (X_1)^2 + (N_1)^2 \\ &= X_c + N_c; [let, X_c = (X_1)^2, N_c = (N_1)^2] \\ &= U_c \end{aligned}$$

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. We can see that an estimate of  $F^e$  can be formed by summing the even filter convolutions. Similarly,  $F^o$  can be estimated from the odd filter convolutions.

$$\begin{aligned} F^e &= \sum_c f_c^e; F^o = \sum_c f_c^o \tag{6} \\ \sum_c |A_c|^2 &= \sum_c (X_c + N_c) = \sum_c U_c \end{aligned}$$

F. Bayesian MAP Marginal Model

The equon (6) can be written in the following form:

$$U_w = X_w + N_w \tag{7}$$

where,  $U_w = \sum_c U_c$  is the noisy LGW coefficient,  $X_w = \sum_c X_c$  is the true speech LGW coefficient and  $N_w = \sum_c N_c$  is the noise

LGW coefficient which is independent Gaussian.

The classical MAP estimator for (7) is

$$\hat{X}_w(U_w) = \arg \max_{X_w} p(X_w | U_w) \tag{8}$$

Using Bayes rule, one gets

$$\begin{aligned} \hat{X}_w(U_w) &= \arg \max_{X_w} [p(U_w | X_w) \cdot p(X_w)] \tag{9} \\ &= \arg \max_{X_w} [p(U_w - X_w) \cdot p(X_w)] \end{aligned}$$

Therefore, these equations allow us to write this estimation in terms of the pdf of the noise ( $p(N_w)$ ) and the pdf of the signal coefficient ( $p(X_w)$ ). From the assumption on the noise,  $p(N_w)$  is zero mean Gaussian with variance  $\sigma_{N_w}$ , i.e.,

$$p(N_w) = \frac{1}{\sigma_{N_w} \sqrt{2\pi}} \exp\left(-\frac{N_w^2}{2\sigma_{N_w}^2}\right) \tag{10}$$

The pdf for speech wavelet coefficients have highly non-Gaussian statistics and are modeled as a two parameter generalized Laplacian distribution (heavy tailed)

$$p(X_w) \propto \exp\left(-\left|\frac{X_w}{s}\right|^\rho\right) \tag{11}$$

where,  $\{s, \rho\}$  are the model parameters. The distribution is zero-mean and symmetric and the parameter  $\{s, \rho\}$  is directly related to the second and fourth moments. Specifically (after consultation with an integral table) one obtains

$$\sigma_{X_w}^2 = \frac{s^2 \Gamma\left(\frac{3}{\rho}\right)}{\Gamma\left(\frac{1}{\rho}\right)}; \kappa = \frac{\Gamma\left(\frac{1}{\rho}\right) \Gamma\left(\frac{5}{\rho}\right)}{\Gamma^2\left(\frac{3}{\rho}\right)} \tag{12}$$

where,  $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$  the well known ‘gamma’ function. Given the sample variance and kurtosis of a histogram, we can solve for the two parameters of our model pdf. Typical values for  $\rho$  are in the range [0.5;1]. This method of model density estimation is simple and direct, but clearly suboptimal. The equon (11) can be written as

$$p(X_w) = \frac{\exp\left(-\left|\frac{X_w}{s}\right|^\rho\right)}{Z(s, \rho)} \tag{13}$$

where,  $Z(s, \rho)$  is the parameter-dependent normalization constant. The  $Z(s, \rho)$  can be written as

$$Z(s, \rho) = 2 \cdot \frac{s}{\rho} \cdot \Gamma\left(\frac{1}{\rho}\right) \tag{14}$$

The equon (13) can be rewritten as

$$p(X_w) = \frac{\exp\left(-\left|\frac{X_w}{s}\right|^\rho\right)}{2 \cdot \frac{s}{\rho} \cdot \Gamma\left(\frac{1}{\rho}\right)} \tag{15}$$

TABLE II  
SUMMARY OF PROPOSED ALGORITHM

---

Step 1 : Take the noisy speech signal  $u_m(n)$  and calculate total length  $L = \text{length}(u_m(n))$ .

Step 2 : Take the filter’s specifications from Table I.

Step 3 : Assume  $wavelength = MaxFreq$ .

Step 4 : Take frequency values in between 0 and 0.5 to  $radius$ .

Step 5 : To stop log function complaining at origin take  $radius(1) = 1$ .

Step 6 : Calculate the FFT of noisy signal  $signalfft = \text{fft}(u_m(n))$ .

Step 7 : for all number of TotFil

Estimate Centre frequency of filter:  $f_0 = \frac{1}{wavelength}$ .

Construct filter:  $\logGabor = \exp\left[\frac{-(\log(radius/f_0))^2}{2(\log(\sigma_{onf}))^2}\right]$ .

Set value at zero frequency to 0 i.e.  $\logGabor(1) = 0$ .

Convolve and take it to EO =  $(signalfft * \logGabor)$ .

Estimate Noise and Signal variances.

for all number of bands

Take the values of  $\rho$ .

Estimate the value of  $s$ .

Estimate  $Z$ .

Estimate  $p(X_w) = \frac{e^{(-|X_w|^\rho)}}{2 \cdot \frac{s}{\rho} \cdot \Gamma(\frac{1}{\rho})}$ .

Estimate  $p(N_w) = \frac{1}{\sigma_{N_w} \sqrt{2\pi}} \cdot \exp\left(-\frac{N_w^2}{2\sigma_{N_w}^2}\right)$ .

Estimate  $\hat{X}_w(U_w)$ .

end

Increment wavelength:  $wavelength = wavelength * Mult$ .

end

Step 8 : Take the Inverse FFT and Estimate noise free signal  $\hat{x}$ .

---

The model parameter  $\{s\}$  can be formulated as

$$s = \sigma_{X_w} \cdot \sqrt{\frac{\Gamma\left(\frac{1}{\rho}\right)}{\Gamma\left(\frac{3}{\rho}\right)}} \tag{16}$$

Now, we come back to the development of MAP estimator for GLD. Equation (9) is also equivalent to

$$\hat{X}_w(U_w) = \arg \max_{X_w} [\log\{p(U_w | X_w)\} + \log\{p(X_w)\}] \tag{17}$$

Let us define,  $f(X_w) = \log(p(X_w))$ . By using equon (10), equon (17) becomes

$$\hat{X}_w(U_w) = \arg \max_{X_w} \left[-\frac{(U_w - X_w)^2}{2\sigma_{N_w}^2} + f(X_w)\right] \tag{18}$$

This is equivalent to solving the following equation if  $p(X_w)$  is assumed to be strictly convex and differentiable.

$$\left[-\frac{(U_w - \hat{X}_w)}{\sigma_{N_w}^2} + f'(\hat{X}_w)\right] = 0 \tag{19}$$

From the above equon. (19) we can write

$$\frac{(U_w - \hat{X}_w)}{\sigma_{N_w}^2} - \frac{\rho \cdot |X_w|^{\rho-1}}{\left(\sigma_{X_w} \cdot \sqrt{\frac{\Gamma\left(\frac{1}{\rho}\right)}{\Gamma\left(\frac{3}{\rho}\right)}}\right)^\rho} \cdot \text{sgn}(X_w) = 0 \tag{20}$$

For different values of model parameter  $\{\rho\}$  starting from 0.5 to 1, we can fit our model density and estimates the true speech coefficients. The Table II summarizes the proposed algorithm.

III. 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 III  
YOHO CORPUS DESCRIPTION

|                      |                                 |
|----------------------|---------------------------------|
| no. of speakers      | 138 (106 M / 32 F)              |
| no. sessions/speaker | 4 enrollments, 10 verifications |
| Interession 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 IV  
POLYCOST CORPUS DESCRIPTION

|                      |  |
|----------------------|--|
| no. of speakers      | 133 (74 M /59 F)   |
| no. sessions/speaker | > 5  |
| Interession 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  |

IV. PERFORMANCE EVALUATION

The performance of the proposed Bayesian MAP marginal estimator is evaluated under two different noise conditions by computing the average improvement in the Segmental SNR (S-SNR) and Log Spectral Deviation (LSD) after enhancing noisy speech signals. Fig. 4 shows the coefficient histograms plotted in log domain for three different values of  $\{\rho\}$  (i.e., for

$\rho = 0.5, 0.6, 0.7$ ) and fitted model densities (dashed lines). Fig. 5 shows (numerically computed) Bayesian estimators for the model of equation (20), with three different values of the exponent  $\rho$ . As with the MAP estimators, smaller values of  $\rho$  produce a nonlinear shrinkage operator. In particular, for  $\rho = 0.5$  (which is well-matched to wavelet marginal such as those shown in Fig. 4). The performance results are averaged out using 10 different utterances of 50 different speakers, drawn from the YOHO speech database. Half of the utterances are from male speakers and half are from female speakers. The noise signals include stationary White Gaussian Noise (WGN) and non-stationary Speech Babble Noise (SBN), taken from the Noisex-92 database. The speech signals are sampled at 8 kHz and degraded by the various noise types in the range [-5,10] dB. The proposed speech enhancement algorithm is applied to the noisy speech signals. For comparison, the S-SNR and LSD are calculated by Ephraim & Malah (referred as M-1), Martin’s Laplacian Prior (referred as M-2), Martin’s Gamma Prior (referred as M-3) and Lotter’s Super Gaussian Model (referred as M-4). Table V presents the results of the S-SNR improvement and Table VI presents the results of the LSD calculation using the various estimators where incase of proposed estimator it is shown for three different values of  $\{\rho\}$ . The proposed estimator yields a higher improvement in the S-SNR and lower LSD scores than the earlier estimators under all tested environmental conditions thereby giving improved performance.

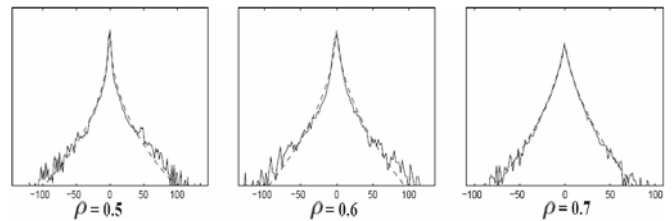


Fig. 4 Coefficient histograms for a single wavelet subband plotted in the log domain. Also shown (dashed lines) are fitted model densities respectively

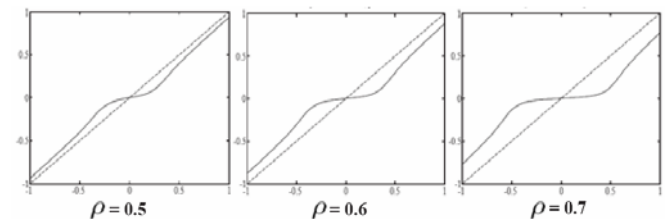


Fig. 5 MAP estimators for the model given in equation (20), with three different exponents. Dashed line indicates the identity function

From the Table V and VI it is clearly shown that our proposed systems outperform the other estimators. For WGN, the Proposed-1 model gives 1.01dB S-SNR improvement (averaged over all SNR levels), Proposed-2 model gives 0.92dB improvement and Proposed-3 model gives 0.86dB improvement than Ephraim & Malah’s MMSE estimator (i.e.

M-1) which is considered as the baseline [6] of speech enhancement field whereas the Proposed-1 model gives 0.41dB and Proposed-2 model gives 0.32dB improvement Proposed-3 model gives 0.23dB improvement than 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 SBN, the Proposed-1 model gives 0.67dB, Proposed-2 model gives 0.61dB and Proposed-3 model gives 0.57dB improvement than M-1 whereas the Proposed-1 model gives 0.18dB, Proposed-2 model gives 0.12dB and Proposed-3 model gives 0.08dB improvement than M-4. For LSD score, the proposed models (i.e. Proposed-1, Proposed-2 & Proposed-3) give 0.45, 0.38, 0.35 of lower LSD score than M-1 and 0.25, 0.18, 0.15 of lower LSD score than M-4 for SBN contamination whereas for WGN contamination the proposed models give 0.50, 0.46 and 0.41 of lower LSD score than M-1 and 0.23, 0.19, 0.14 of lower LSD score than M-4 in -5dB of SNR. The proposed models give better S-SNR improvement and lower LSD for every dB level as well as average S-SNR

improvement over all dB levels than competing estimators.

The Table VII and VIII presents Segmental SNR (S-SNR) improvement and Log Spectral Distortion (LSD) score obtained from earlier as well as proposed estimator in case of POLYCOST Corpus. The performance results are again averaged out using 10 different utterances of 50 different speakers, half of the utterances are from male speakers and half are from female speakers. From the Table VII and VIII it is noted that our proposed systems outperform the other estimators in this speech corpus also. The proposed models give better S-SNR improvement and lower LSD for every dB level as well as for average S-SNR improvement over all dB levels. For WGN, the Proposed-1 model gives 0.55dB S-SNR improvement (averaged over all SNR level), Proposed-2 model gives 0.50dB improvement and Proposed-3 model gives 0.44dB improvement than M-1 whereas the Proposed-1 model gives 0.20dB, Proposed-2 model gives 0.15dB and Proposed-3 model gives 0.12dB improvement than M-4. In case of SBN also the Proposed-1 model gives 0.44dB, Proposed-2 model gives 0.42dB improvement and Proposed-3 model gives 0.40dB improvement than M-1 whereas the Proposed-1 model gives 0.17dB, Proposed-2 model gives 0.14dB and Proposed-3 model gives 0.09dB improvement than M-4. For LSD score estimation, the proposed models (i.e. Proposed-1, Proposed-2 & Proposed-3) give 0.70, 0.66, 0.65 of lower LSD score than M-1 and 0.22, 0.18, 0.17 of lower LSD score than M-4 for WGN contamination whereas for SBN contamination the proposed models give 0.63, 0.58 and 0.56 of lower LSD score than M-1 and 0.25, 0.20, 0.18 of lower LSD score than M-4 in -5dB of SNR. The proposed systems showed that the use of our first model (i.e.  $\rho=0.5$ ) gives better results with small improvement on performance over Proposed-2 (i.e.  $\rho=0.6$ ) and Proposed-3 (i.e.  $\rho=0.7$ ) Model for both S-SNR and LSD. It is also noted that the results for S-SNR and LSD scores 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.

TABLE V  
SEGMENTAL SNR IMPROVEMENT FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED BY PROPOSED AND EARLIER ESTIMATORS FOR YOHO SPEECH CORPUS

| Type of Noises              | Used Methods                | Input Seg. SNR(dB) |             |             |             |
|-----------------------------|-----------------------------|--------------------|-------------|-------------|-------------|
|                             |                             | -5                 | 0           | 5           | 10          |
| WGN                         | M - 1                       | 8.34               | 6.12        | 5.13        | 3.07        |
|                             | M - 2                       | 8.36               | 6.55        | 5.26        | 3.21        |
|                             | M - 3                       | 8.55               | 6.71        | 5.33        | 3.65        |
|                             | M - 4                       | 8.86               | 6.81        | 5.54        | 3.82        |
|                             | Proposed - 1 ( $\rho=0.5$ ) | <b>9.11</b>        | <b>7.35</b> | <b>5.80</b> | <b>4.42</b> |
|                             | Proposed - 2 ( $\rho=0.6$ ) | <b>9.05</b>        | <b>7.23</b> | <b>5.71</b> | <b>4.35</b> |
| Proposed - 3 ( $\rho=0.7$ ) | <b>8.97</b>                 | <b>7.11</b>        | <b>5.67</b> | <b>4.21</b> |             |
| SBN                         | M - 1                       | 6.92               | 5.02        | 3.89        | 2.72        |
|                             | M - 2                       | 7.07               | 5.33        | 3.97        | 2.87        |
|                             | M - 3                       | 7.13               | 5.54        | 4.08        | 3.07        |
|                             | M - 4                       | 7.33               | 5.82        | 4.14        | 3.21        |
|                             | Proposed - 1 ( $\rho=0.5$ ) | <b>7.51</b>        | <b>5.97</b> | <b>4.34</b> | <b>3.43</b> |
|                             | Proposed - 2 ( $\rho=0.6$ ) | <b>7.46</b>        | <b>5.90</b> | <b>4.28</b> | <b>3.37</b> |
| Proposed - 3 ( $\rho=0.7$ ) | <b>7.42</b>                 | <b>5.88</b>        | <b>4.23</b> | <b>3.31</b> |             |

TABLE VI  
LOG SPECTRAL DISTORTION FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED BY PROPOSED AND EARLIER ESTIMATORS FOR YOHO SPEECH CORPUS

| Type of Noises              | Used Methods                | Input Seg. SNR(dB) |             |             |             |
|-----------------------------|-----------------------------|--------------------|-------------|-------------|-------------|
|                             |                             | -5                 | 0           | 5           | 10          |
| WGN                         | M - 1                       | 4.86               | 3.88        | 2.96        | 2.23        |
|                             | M - 2                       | 4.80               | 3.75        | 2.87        | 2.13        |
|                             | M - 3                       | 4.71               | 3.67        | 2.80        | 2.06        |
|                             | M - 4                       | 4.59               | 3.55        | 2.66        | 1.99        |
|                             | Proposed - 1 ( $\rho=0.5$ ) | <b>4.36</b>        | <b>3.27</b> | <b>2.48</b> | <b>1.79</b> |
|                             | Proposed - 2 ( $\rho=0.6$ ) | <b>4.40</b>        | <b>3.32</b> | <b>2.54</b> | <b>1.84</b> |
| Proposed - 3 ( $\rho=0.7$ ) | <b>4.45</b>                 | <b>3.39</b>        | <b>2.58</b> | <b>1.88</b> |             |
| SBN                         | M - 1                       | 4.97               | 3.92        | 3.05        | 2.44        |
|                             | M - 2                       | 4.91               | 3.87        | 2.91        | 2.39        |
|                             | M - 3                       | 4.85               | 3.80        | 2.82        | 2.33        |
|                             | M - 4                       | 4.77               | 3.75        | 2.78        | 2.25        |
|                             | Proposed - 1 ( $\rho=0.5$ ) | <b>4.52</b>        | <b>3.55</b> | <b>2.57</b> | <b>2.08</b> |
|                             | Proposed - 2 ( $\rho=0.6$ ) | <b>4.59</b>        | <b>3.60</b> | <b>2.62</b> | <b>2.12</b> |
| Proposed - 3 ( $\rho=0.7$ ) | <b>4.62</b>                 | <b>3.65</b>        | <b>2.67</b> | <b>2.17</b> |             |

TABLE VII  
SEGMENTAL SNR IMPROVEMENT FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED BY PROPOSED AND EARLIER ESTIMATORS FOR POLYCOST SPEECH CORPUS

| Type of Noises              | Used Methods                | Input Seg. SNR(dB) |             |             |             |
|-----------------------------|-----------------------------|--------------------|-------------|-------------|-------------|
|                             |                             | -5                 | 0           | 5           | 10          |
| WGN                         | M - 1                       | 7.24               | 5.32        | 4.43        | 2.97        |
|                             | M - 2                       | 7.35               | 5.65        | 4.55        | 3.11        |
|                             | M - 3                       | 7.45               | 5.70        | 4.63        | 3.19        |
|                             | M - 4                       | 7.57               | 5.85        | 4.69        | 3.25        |
|                             | Proposed - 1 ( $\rho=0.5$ ) | <b>7.99</b>        | <b>6.00</b> | <b>4.77</b> | <b>3.40</b> |
|                             | Proposed - 2 ( $\rho=0.6$ ) | <b>7.87</b>        | <b>5.95</b> | <b>4.76</b> | <b>3.38</b> |
| Proposed - 3 ( $\rho=0.7$ ) | <b>7.76</b>                 | <b>5.90</b>        | <b>4.73</b> | <b>3.36</b> |             |
| SBN                         | M - 1                       | 6.02               | 4.72        | 3.21        | 2.35        |
|                             | M - 2                       | 6.09               | 4.77        | 3.29        | 2.40        |
|                             | M - 3                       | 6.15               | 4.84        | 3.35        | 2.49        |
|                             | M - 4                       | 6.30               | 4.99        | 3.50        | 2.61        |
|                             | Proposed - 1 ( $\rho=0.5$ ) | <b>6.46</b>        | <b>5.15</b> | <b>3.68</b> | <b>2.80</b> |
|                             | Proposed - 2 ( $\rho=0.6$ ) | <b>6.43</b>        | <b>5.13</b> | <b>3.65</b> | <b>2.78</b> |
| Proposed - 3 ( $\rho=0.7$ ) | <b>6.41</b>                 | <b>5.10</b>        | <b>3.63</b> | <b>2.77</b> |             |

TABLE VIII  
LOG SPECTRAL DISTORTION FOR VARIOUS NOISE TYPES AND LEVELS,  
OBTAINED BY PROPOSED AND EARLIER ESTIMATORS FOR POLYCOST  
SPEECH CORPUS

| Type of Noises              | Used Methods                | Input Seg. SNR(dB) |             |             |             |
|-----------------------------|-----------------------------|--------------------|-------------|-------------|-------------|
|                             |                             | -5                 | 0           | 5           | 10          |
| WGN                         | M - 1                       | 5.18               | 4.08        | 3.17        | 2.35        |
|                             | M - 2                       | 4.98               | 3.89        | 2.95        | 2.26        |
|                             | M - 3                       | 4.77               | 3.71        | 2.87        | 2.19        |
|                             | M - 4                       | 4.70               | 3.62        | 2.80        | 2.07        |
|                             | Proposed - 1 ( $\rho=0.5$ ) | <b>4.48</b>        | <b>3.40</b> | <b>2.61</b> | <b>1.89</b> |
|                             | Proposed - 2 ( $\rho=0.6$ ) | <b>4.52</b>        | <b>3.45</b> | <b>2.66</b> | <b>1.91</b> |
| SBN                         | Proposed - 3 ( $\rho=0.7$ ) | <b>4.53</b>        | <b>3.47</b> | <b>2.67</b> | <b>1.92</b> |
|                             | M - 1                       | 5.23               | 4.11        | 3.25        | 2.40        |
|                             | M - 2                       | 5.11               | 4.02        | 3.16        | 2.31        |
|                             | M - 3                       | 4.97               | 3.89        | 2.91        | 2.22        |
|                             | M - 4                       | 4.85               | 3.79        | 2.81        | 2.10        |
|                             | Proposed - 1 ( $\rho=0.5$ ) | <b>4.60</b>        | <b>3.55</b> | <b>2.63</b> | <b>1.91</b> |
| Proposed - 2 ( $\rho=0.6$ ) | <b>4.65</b>                 | <b>3.60</b>        | <b>2.70</b> | <b>1.99</b> |             |
| Proposed - 3 ( $\rho=0.7$ ) | <b>4.67</b>                 | <b>3.62</b>        | <b>2.72</b> | <b>2.01</b> |             |

## V. CONCLUSION

The proposed work presents a GLD to model the statistics of Log Gabor wavelet coefficients and a simple estimator is derived from the pdfs using Bayesian MAP Marginal Estimation. The statistical model is designed and adopted to fit the distribution of the speech spectral amplitudes to improve the quality of the enhanced speech signal. The underlying statistical model can be adjusted to the demands of the specific noise reduction system. The automatic determination of threshold overcomes a problem that has plagued wavelet denoising schemes in the past. The results show superiority of the proposed method over a broad range of noise contaminations

## ACKNOWLEDGMENT

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

## REFERENCES

- [1] Boll, S. F., "Suppression of Acoustic Noise in Speech using Spectral Subtraction", IEEE ASSP, 27(2):113-120, 1979
- [2] 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. ASSP32, no. 6, pp. 1109-1121, Dec. 1984.
- [3] 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.
- [4] 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. 8790, Kyoto, Japan, Sep. 2003.
- [5] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors", IEEE ICASSP'02, Orlando, Florida, May 2002.
- [6] 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, Pages 11101126.
- [7] 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.

- [8] 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.
- [9] I. Cohen, "Speech Enhancement Using a Noncausal A Priori SNR Estimator", IEEE Signal Processing Letters, Vol. 11, No. 9, Sep. 2004, pp. 725-728.
- [10] 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.
- [11] E. Zavarzheh, 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, Page(s):219 -224.
- [12] D. Gabor, "Theory of communication", J. Inst. Electr. Eng. 93, pp. 429457, 1946.
- [13] 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, February 1982.
- [14] 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.



**Suman Senapati** received the B. Tech. in Electronics & Telecommunication Engg. in 2001 from University Of 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.