Perceptually Motivated Bayesian Estimators With Generalized Gamma Distribution Under Speech Presence Probability

Atanu Saha, and Tetsuya Shimamura

***Abstract*—** In this paper, we propose a family of Bayesian estimators for single channel speech enhancement. The Bayesian estimators, which utilize the cost function of the log-spectral amplitude (LSA) estimator, are based on generalized Gamma distribution under speech presence probability. The cost function obtained from the LSA estimator is weighted by psychoacoustically motivated speech distortion measure to take advantage of the perceptual interpretation. The experimental results show that the proposed estimators provide less residual noise and better speech quality compared to the traditional state-of-the-art estimators.

***Keywords*—**Speech enhancement, Bayesian estimator, Distortion measure, Generalized Gamma distribution.

I. Introduction

THERE are a wide variety of scenarios in which it is desired to enhance the signals of speech corrupted by additive background noise. Voice communication, for instance, over cellular telephone system typically suffers from background noise. Speech enhancement algorithms improve the perceptual quality of speech through extracting the desired signal from its corrupted observations [1].

Significant progress has been made in developing the speech enhancement algorithms in the last few decades [2-11]. Among the various existing algorithms, the nonlinear estimators have shown to be effective in single channel speech enhancement. Various approaches exist in the estimation theory literature [12] for deriving these nonlinear estimators. Bayesian approach is particularly attractive due to its superior performance among them. In Bayesian approach, an estimator of the clean speech is derived by minimizing the conditional expectation of a cost function that penalizes errors in the clean speech estimate.

Several Bayesian estimators of the short-time spectral amplitude (STSA), in place of the short-time Fourier transform (STFT) complex coefficients, have been proposed. The main feature of the Bayesian STSA estimators is to produce a residual background noise that is whiter than the residual musical noise produced by the STFT estimators such as Wiener filter [13]. A well-known Bayesian estimator of the STSA is the minimum mean square error (MMSE) estimator that minimizes the conditional expectation of a squared-error cost function [7]. The squared-error cost function in logarithmic domain, resulting in log-spectral amplitude (LSA) estimator [8], is more effective in reducing musical noise. A further modification of these cost functions has also been conducted by incorporating speech presence probability (SPP) in [14]-[16]. More psychoacoustically motivated Bayesian estimators that use variants of speech distortion measures as the cost function, in place of the squared-error cost function, were proposed in [17], [18].

The aforementioned approaches for speech enhancement in discrete Fourier transform (DFT) domain assume that the clean speech and noise DFT coefficients are Gaussian distributed. Although this assumption might hold for the noise DFT coefficients, it does not hold for the speech DFT coefficients. For this reason, several researchers [19], [20] have proposed the use of super-Gaussian such as Laplacian or Gamma distribution for modeling the speech DFT coefficients.

This paper is devoted to derive perceptually motivated Bayesian estimators for single channel speech enhancement. Two different kinds of estimators are derived by exploiting the generalized Gamma distribution (GGD) assumption for the speech DFT coefficients under SPP. The cost function of the proposed estimators, which is obtained from the LSA estimator, is weighted by Euclidean distortion measure so that it takes into account the loudness and masking properties of the human auditory system. The incorporation of these properties into the gain function under SPP makes the proposed method to perform well by removing a certain amount of noise while keeping the speech components as undistorted as possible. Parts of this paper have been previously reported in [21] and in [22]. The present work constitutes a substantial extension of these.

The paper is structured as follows. In Section II, we introduce the signal model with basic assumptions. Section III derives the perceptually motivated Bayesian estimators, while Section IV shows the experimental results. Finally, Section V concludes the paper with discussions.

II. Signal Model

Let the observed noisy speech signal at frame λ be assumed as

A. Saha is with the Graduate School of Science and Engineering, Saitama University, Saitama, Japan, 338-8570 (phone: 81-48-858-3496; fax: 81-48-858-3716; e-mail: saha@sie.ics.saitama-u.ac.jp).

T. Shimamura is with the Graduate School of Science and Engineering, Saitama University, Saitama, Japan, 338-8570 (phone: 81-48-858-3496; fax: 81-48-858-3716; e-mail: shima@sie.ics.saitama-u.ac.jp).

$$y_\lambda(n) = s_\lambda(n) + d_\lambda(n),\ 0 \le n \le N-1 \quad (1)$$

where n is the sampling index, $s_\lambda(n)$ is the clean speech, $d_\lambda(n)$ is the additive noise and N is the frame length. The k^{th} DFT coefficient of the noisy speech signal can be expressed as

$$Y_{\lambda,k} \cong \sum_{n=0}^{N-1} y_\lambda(n) h(n) e^{-j\frac{2\pi}{N}kn} \quad (2)$$

where $h(n)$ is the analysis window and $k \in \{0,1,\ldots,N-1\}$ is the frequency index. By considering the DFT coefficients of the clean speech and noise, denoted as $S_{\lambda,k}$ and $D_{\lambda,k}$ respectively, which are assumed to be statistically independent, (2) becomes

$$Y_{\lambda,k} = S_{\lambda,k} + D_{\lambda,k}. \quad (3)$$

The preceding equation can also be expressed by dropping the frame index for notational convenience in polar form as:

$$R_k e^{j\phi} = A_k e^{j\psi} + N_k e^{j\theta} \quad (4)$$

where $\{R_k, A_k, N_k\}$ and $\{\phi,\psi,\theta\}$ denote the amplitudes and phases of the noisy speech, clean speech and noise, respectively. The DFT coefficients of noise are assumed to obey a Gaussian distribution. The Gaussian assumption that corresponds to a Rayleigh distribution, however, is not necessarily the best model for estimation of the speech DFT amplitudes [19], [20]. A GGD assumption for speech amplitude can perform much better than the Rayleigh distribution assumption. The GGD is given by

$$f(A_k) = \frac{\delta \eta^{\upsilon}}{\Gamma(\upsilon)} A_k^{\delta\upsilon-1} \exp\left(-\eta A_k^{\delta}\right),\ \delta,\eta,\upsilon > 0 \quad (5)$$

where $\Gamma(.)$ the Gamma function, δ and υ denote the shaping parameters, and η is called the scaling parameter. The special cases of generalized priors in (5) for different estimators depend on choosing the value of δ [23]. For different cases of δ, the scaling parameter η is related to the second moment of the distribution as

$$\eta = \begin{cases} \sqrt{\frac{\upsilon(\upsilon+1)}{\varphi_S(k)}} & \text{if } \delta=1 \\ \frac{\upsilon}{\varphi_S(k)} & \text{if } \delta=2 \end{cases} \quad (6)$$

where $\varphi_s(k)$ is the variance of speech. In this paper, we consider $\delta = 1$ and $\delta = 2$ for deriving the perceptually motivated Bayesian estimators.

III. Perceptually Motivated Estimators

In this section, we derive the proposed Bayesian estimators under GGD with SPP.

The Bayesian spectral amplitude estimator minimizes the conditional expectation of a cost function, $E[C(A_k, \hat{A}_k)]$, where $\hat{A}_k$ denotes the estimated spectral amplitude of A_k. The estimator is, then, combined with the phase of the noisy speech to derive the estimator of the complex spectral component of the clean speech as $\hat{S}_k = \hat{A}_k e^{j\phi}$. Finally, the enhanced time signal is obtained by taking inverse DFT of $\hat{S}_k$. The motivation is thus to compute the gain function G_k so that it satisfies the estimator $\hat{A}_k = G_k R_k$.

In the logarithmic domain, which was proposed in [8], the cost function of the Bayesian estimator is chosen as

$$C(A_k, \hat{A}_k) = \left(\log A_k - \log \hat{A}_k\right)^2. \quad (7)$$

The LSA estimator shown in [8] can be derived by exploiting the moment generating function of $\log A_k \,|\, Y_k$ for complex Gaussian distributed clean speech and noise DFT coefficients as

$$\hat{A}_k = \exp\left(\frac{d}{d\rho} E\left[A_k^\rho \,|\, Y_k\right] \right)\Bigg|_{\rho=0} \quad (8)$$

where ρ is a real parameter. Equation (8) is equivalent to

$$\hat{A}_k = \lim_{\rho\to 0} \exp\left(\frac{\frac{d}{d\rho} E[A_k^\rho \,|\, Y_k]}{E[A_k^\rho \,|\, Y_k]} \right). \quad (9)$$

By applying L'Hopital's rule, (9) can be expressed as

$$\hat{A}_k = \lim_{\rho\to 0} \exp\left(\frac{\frac{d}{d\rho} \log E[A_k^\rho \,|\, Y_k]}{\frac{d}{d\rho}} \right). \quad (10)$$

For a small value of ρ, (10) can be simplified as

$$\hat{A}_k = E\left[A_k^\rho \,|\, Y_k\right]^{1/\rho} \quad (11)$$

where ρ is approximated as $0 < \rho << 1$. Note that (11) is a special case of the approach proposed in [24].

The spectral amplitude estimator in (11) is now considered under SPP. Given two hypotheses, $H_0(k): Y_k = D_k$ and $H_1(k): Y_k = S_k + D_k$, which indicate speech absence and presence, respectively, assuming a complex Gaussian distribution of the DFT coefficients for both speech and noise [7], the conditional SPP, $\zeta_k = P(H_1(k) \,|\, Y_k)$, is given by

$$\zeta_k = \left\{ 1 + \frac{q_k}{1-q_k} (1+\xi_k) \exp(-\upsilon_k) \right\}^{-1} \quad (12)$$

where $q_k = P(H_0(k))$ is the *a priori* probability of speech absence, $\xi_k = \varphi_s(k)/\varphi_d(k)$ is the *a priori* SNR in which $\varphi_d(k)$ denotes the variance of noise, $\gamma_k = R_k^2/\varphi_d(k)$ is called the *a posteriori* SNR, and $\upsilon_k = \xi_k \gamma_k / (1+\xi_k)$. By taking into account the SPP ζ_k, the estimator in (11) is obtained as

$$\hat{A}_k^o = \left[E\left[A_k^\rho \,|\, Y_k, H_1(k)\right] \zeta_k + E\left[A_k^\rho \,|\, Y_k, H_0(k)\right] (1-\zeta_k) \right]^{1/\rho} \quad (13)$$

where $\hat{A}_k^0$ denotes the optimal spectral amplitude estimator under consideration of SPP. It is interesting to mention that the estimator $\hat{A}_k^0$ in (13) is a special case of the method proposed in [15]. Since the gain is constrained to be larger than a threshold $G_{\min}$ during speech absence, we consider

$$E\left[A_k^\rho \,|\, Y_k, H_0(k)\right] = (G_{\min} R_k)^\rho. \quad (14)$$

Accordingly, the conditional gain function during speech

presence is defined by

$$E\left[A_k^{\rho} \mid Y_k, H_1(k)\right] = \left(G_k^w R_k\right)^{\rho} \tag{15}$$

where G_k^w is the weighted gain function considered with GGD.

The proposed method is based on deriving G_k^w with generalized Gamma distributed speech priors. As can be seen from (7), the chosen cost function of the LSA estimator is the squared-error between the estimated and actual clean speech. This type of squared-error cost function, however, is not necessarily subjectively meaningful [17]. A more perceptually significant cost function is used in [17] based on a weighted error criterion that exploits the masking properties of the human auditory system. The chosen cost function is given by

$$C\left(A_k, \hat{A}_k\right) = A_k^{\tau}\left(A_k - \hat{A}_k\right)^2 \tag{16}$$

where τ is a real parameter. To obtain the gain function G_k^w corresponding to the above cost function in (16), we simplify (11) as

$$\hat{A}_k^w = \left(\frac{E[A_k^{\rho-\tau} \mid Y_k]}{E[A_k^{-\tau} \mid Y_k]}\right)^{1/\rho}. \tag{17}$$

We now consider tow cases for GGD priors with $\delta = 1$ and $\delta = 2$.

A. The Case for GGD Priors with $\delta = 1$

Since the noise is assumed to be Gaussian distributed, the conditional probability of Y_k can be written as [25]

$$f\left(Y_k \mid A_k\right) = \frac{2R_k}{\varphi_d(k)} \exp\left(-\frac{R_k^2 + A_k^2}{\varphi_d(k)}\right) I_0\left(\frac{2A_k R_k}{\varphi_d(k)}\right) \tag{18}$$

where I_0 is the 0^{th}-order modified Bessel function of the first kind. Applying the large-value approximation of the Bessel function I_0 in (18) and by specifying the GGD prior of A_k in (5), the ρ^{th} conditional moment can be simplified as

$$E\left[A_k^{\rho} \mid Y_k\right] = \frac{\int_0^{\infty} A_k^{\upsilon+\rho-3/2} \exp\left(-\frac{A_k^2}{\varphi_d(k)} - \mu_k A_k\right) dA_k}{\int_0^{\infty} A_k^{\upsilon-3/2} \exp\left(-\frac{A_k^2}{\varphi_d(k)} - \mu_k A_k\right) dA_k} \tag{19}$$

where μ_k is defined as

$$\mu_k = \frac{2\sqrt{\gamma_k \xi_k} - \sqrt{\upsilon(\upsilon+1)}}{\sqrt{2\xi_k}}. \tag{20}$$

In terms of confluent hypergeometric function [26], the conditional moment in (19) can be determined by (21), shown in the next page. In (21), $\Phi(.)$ is called the confluent hypergeometric function. Substituting (21) in (17), the weighted gain function G_k^w via $\hat{A}_k^w = G_k^w R_k$ is determined by (22), shown in the next page. In (22), we simplify $p = -\upsilon - \rho + \tau + 0.5$ and $q = -\upsilon + \tau + 0.5$. From (13), (14), (15) and (22), we obtain

$$\hat{A}_k^{(1)} = G_k^{(1)} R_k \tag{23}$$

where the gain function $G_k^{(1)}$ is obtained in (24). In (23), the superscript is used for denoting the GGD priors considered with $\delta = 1$.

B. The Case for GGD Priors with $\delta = 2$

By specifying the GGD prior of A_k in (5), the ρ^{th} conditional moment can be simplified as

$$E\left[A_k^{\rho} \mid Y_k\right] = \left(\frac{R_k\sqrt{\mu_k}}{\gamma_k}\right)^{\rho} \frac{\Gamma\left(\upsilon + \frac{\rho}{2}\right)\Phi\left(-\upsilon + 1 - \frac{\rho}{2}; 1; -\mu_k\right)}{\Gamma(\upsilon)\Phi(-\upsilon + 1; 1; -\mu_k)} \tag{25}$$

where $\mu_k = \xi_k\gamma_k / (\upsilon + \xi_k)$. Substituting (25) in (17), the weighted gain function G_k^w via $\hat{A}_k^w = G_k^w R_k$ is determined by (26), shown in the next page. From (13), (14), (15) and (26), we obtain

$$\hat{A}_k^{(2)} = G_k^{(2)} R_k \tag{27}$$

where the gain function $G_k^{(2)}$ is obtained in (28). In (27), the superscript is used for denoting the GGD priors considered with $\delta = 2$.

It is interesting to show that the gain function $G_k^{(2)}$ converges to the Wiener gain function for $\gamma_k >> 1$ and consequently for $\mu_k >> 1$. Using the following approximation of the confluent hypergeometric function

$$\Phi(a; 1; -\mu_k) \approx \frac{\mu_k^{-a}}{\Gamma(1-a)}, \quad \mu_k >> 1 \tag{29}$$

(where a corresponds to real value) the gain function $G_k^{(2)}$, for $\zeta_k = 1$, converges to

$$G_k^{(2)} \approx \frac{\xi_k}{\upsilon + \xi_k} \tag{30}$$

which is the gain function of the Wiener filter for $\upsilon = 1$. The consideration of $\zeta_k = 1$ for large values of γ_k is not contradictory, since speech is almost present for large values of γ_k. This provides the validity of the incorporation of SPP into the gain function.

The asymptotic behavior of the gain functions for large values of γ_k in all cases is obtained as

$$G_k \approx 1. \tag{31}$$

The above two gain functions obtained in (24) and (28) are functions of both the *a priori* SNR ξ_k and *posteriori* SNR γ_k. Figs. 1 and 2 plot the gain functions as a function of the instantaneous SNR, $\gamma_k - 1$, for a fixed value of ξ_k ($\xi_k = -5$ dB in left panel and $\xi_k = 5$ dB in right panel) for several values ρ and τ. As can be observed, the shape of the gain functions is similar to all cases. The parameters ρ and τ are found to control the trade-off between the amount of noise reduction and speech distortion. Small values of ρ provide higher attenuation, while small values of τ provide lower attenuation. As a good compromise between the amount of attenuation and speech distortion, we use $\rho = -3$ dB and $\tau = -10$ dB in the experiment.

$$E\left[A_k^{\rho} \mid Y_k\right]=\left(\frac{R_k}{2\sqrt{\gamma_k}}\right)^{\rho}\frac{\Gamma(\upsilon+\rho-0.5)}{\Gamma(\upsilon-0.5)}\left\{\frac{\dfrac{\Phi\left(\dfrac{\upsilon+\rho-0.5}{2},\dfrac{1}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(\dfrac{\upsilon+\rho+0.5}{2}\right)}+\dfrac{\sqrt{2}\mu_k\Phi\left(\dfrac{\upsilon+\rho+0.5}{2},\dfrac{3}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(\dfrac{\upsilon+\rho-0.5}{2}\right)}}{\dfrac{\Phi\left(\dfrac{\upsilon-0.5}{2},\dfrac{1}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(\dfrac{\upsilon+0.5}{2}\right)}+\dfrac{\sqrt{2}\mu_k\Phi\left(\dfrac{\upsilon+0.5}{2},\dfrac{3}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(\dfrac{\upsilon-0.5}{2}\right)}}\right\} \tag{21}$$

$$G_k^{w}=\frac{1}{2\sqrt{\gamma_k}}\left[\frac{\Gamma(-p)\left\{\dfrac{\Phi\left(-\dfrac{p}{2},\dfrac{1}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(\dfrac{1-p}{2}\right)}+\dfrac{\sqrt{2}\mu_k\Phi\left(\dfrac{1-p}{2},\dfrac{3}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(-\dfrac{p}{2}\right)}\right\}}{\Gamma(-q)\left\{\dfrac{\Phi\left(-\dfrac{q}{2},\dfrac{1}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(\dfrac{1-q}{2}\right)}+\dfrac{\sqrt{2}\mu_k\Phi\left(\dfrac{1-q}{2},\dfrac{3}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(-\dfrac{q}{2}\right)}\right\}}\right]^{1/\rho} \tag{22}$$

$$G_k^{(1)}=\left[\left\{\frac{1}{2\sqrt{\gamma_k}}\left[\frac{\Gamma(-p)\left\{\dfrac{\Phi\left(-\dfrac{p}{2},\dfrac{1}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(\dfrac{1-p}{2}\right)}+\dfrac{\sqrt{2}\mu_k\Phi\left(\dfrac{1-p}{2},\dfrac{3}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(-\dfrac{p}{2}\right)}\right\}}{\Gamma(-q)\left\{\dfrac{\Phi\left(-\dfrac{q}{2},\dfrac{1}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(\dfrac{1-q}{2}\right)}+\dfrac{\sqrt{2}\mu_k\Phi\left(\dfrac{1-q}{2},\dfrac{3}{2};\dfrac{\mu_k^2}{2}\right)}{\Gamma\left(-\dfrac{q}{2}\right)}\right\}}\right]^{1/\rho}\right\}^{\rho}\zeta_k+G_{\min}^{\rho}(1-\zeta_k)\right]^{1/\rho} \tag{24}$$

$$G_k^{w}=\frac{\sqrt{\mu_k}}{\gamma_k}\left(\frac{\Gamma\left(\upsilon+\dfrac{\rho-\tau}{2}\right)\Phi\left(-\upsilon+1-\dfrac{\rho-\tau}{2};1;-\mu_k\right)}{\Gamma\left(\upsilon-\dfrac{\tau}{2}\right)\Phi\left(-\upsilon+1+\dfrac{\tau}{2};1;-\mu_k\right)}\right)^{1/\rho} \tag{26}$$

$$G_k^{(2)}=\left\{\left(\frac{\sqrt{\mu_k}}{\gamma_k}\right)^{\rho}\left(\frac{\Gamma\left(\upsilon+\dfrac{\rho-\tau}{2}\right)\Phi\left(-\upsilon+1-\dfrac{\rho-\tau}{2};1;-\mu_k\right)}{\Gamma\left(\upsilon-\dfrac{\tau}{2}\right)\Phi\left(-\upsilon+1+\dfrac{\tau}{2};1;-\mu_k\right)}\right)\zeta_k+G_{\min}^{\rho}(1-\zeta_k)\right\}^{1/\rho} \tag{28}$$

IV. Experimental Results

In this section, we investigate the performance of the proposed estimators.

The NOIZEUS speech corpus [27], which is comprised of phonetically balanced utterances, is used for investigation. The corpus comes with non-stationary noises at different SNRs. Two kinds of noises taken from the corpus are used in the experiments. These are babble noise and train noise. In addition to these, white Gaussian noise which is added by ourselves to the clean part of the corpus is also used for investigation. All utterances (30 utterances) of the corpus are used in experiments. The sampling frequency of the test utterances is 8 kHz. A 20-msec analysis Hamming window is used with 50% overlap between frames. The lower bound threshold $G_{\min}$ is set to –40 dB. The shaping parameter υ is set to –2 dB. The *a priori*

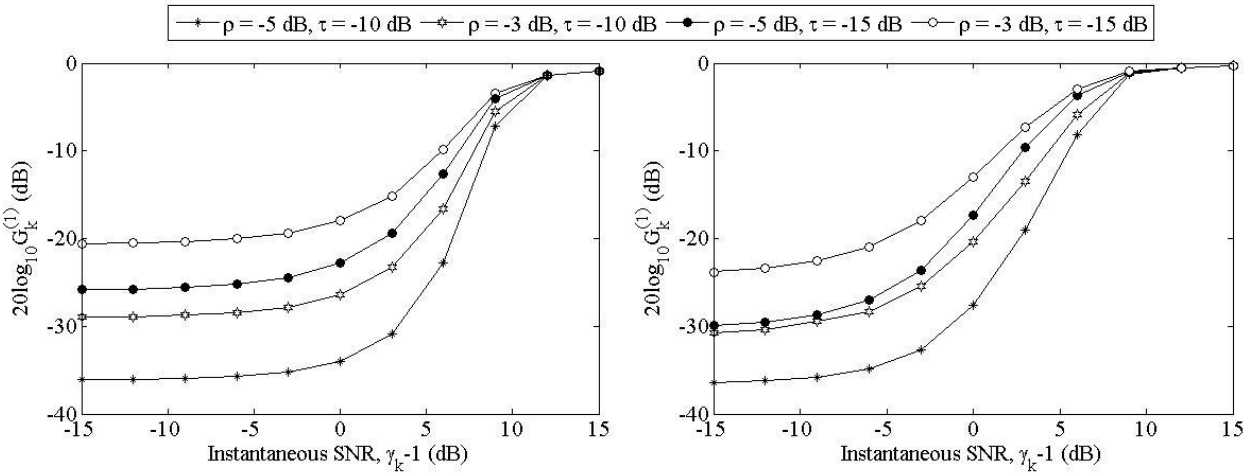

Fig. 1 Gain functions of the proposed estimator $\hat{A}_k^{(1)}$ (23) for several values of ρ and τ as a function of the instantaneous SNR, $\gamma_k - 1$, for $\upsilon = -2$ dB. The left panel plots the gain functions for $\xi_k = -5$ dB, whereas the right panel plots the gain functions for $\xi_k = 5$.

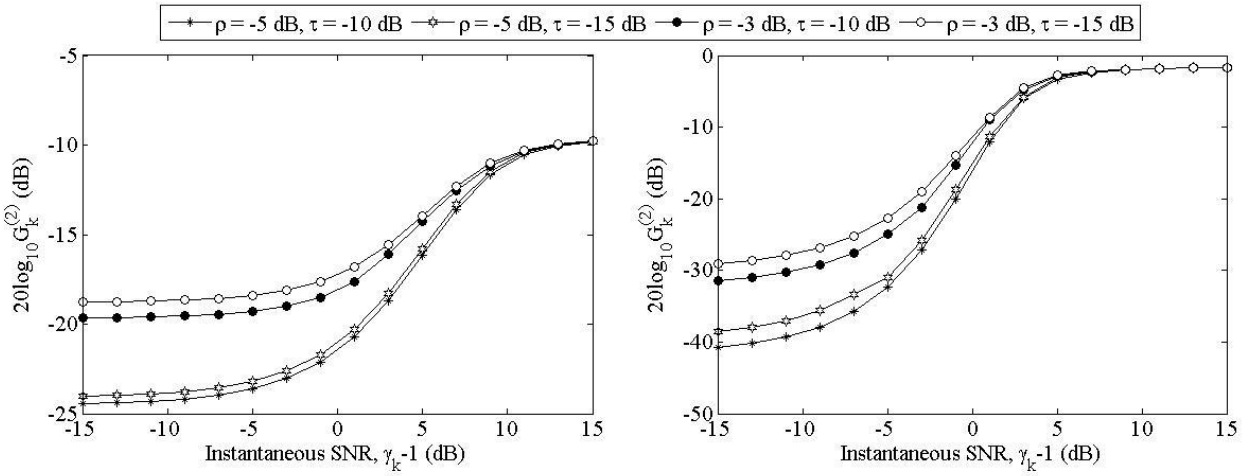

Fig. 2 Gain functions of the proposed estimator $\hat{A}_k^{(2)}$ (27) for several values of ρ and τ as a function of the instantaneous SNR, $\gamma_k - 1$, for $\upsilon = -2$ dB. The left panel plots the gain functions for $\xi_k = -5$ dB, whereas the right panel plots the gain functions for $\xi_k = 5$.

SNR is estimated by the approach proposed in [7], and the *a priori* probability of speech absence for computing the SPP is estimated according to [14]. For estimation of noise variance, we use the approach proposed in [28].

The performance of the proposed estimators $\hat{A}_k^{(1)}$ and $\hat{A}_k^{(2)}$ is compared to the weighted-Euclidean distance estimator [17] and β-order MMSE estimator [24], referred to here as $\hat{A}_k^{(E)}$ and $\hat{A}_k^{(\beta)}$, respectively. The typical parameter selection is the same as that in those approaches. The performance of all estimators is evaluated using both objective measures and subjective listening test.

A. Objective Evaluations

Many objective measures can be used to assess speech enhancement algorithms. The segmental SNR (SSNR) is generally used to measure the amount of noise reduction of the speech signals. However, the SSNR is not strongly correlated with the subjective measures such as the mean opinion score (MOS), since it does not closely emulate the signal processing involved at the auditory periphery. A study of the correlation between MOS and objective measures was conducted in [27]. One of the objective measures that was found to have the best correlation with MOS was the perceptual evaluation of speech quality (PESQ) measure with correlation coefficients of 0.89. The PESQ was found to have well correlated with both signal and background distortions. We will use the SSNR and PESQ measures for evaluation in the following.

The SSNR in each frame is calculated as follows

$$SSNR = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\frac{\sum_{n=N*j}^{N*j+N-1} s^2(n)}{\sum_{n=N*j}^{N*j+N-1}\left[s(n)-\hat{s}(n)\right]^2}\right] \quad (32)$$

where $s(n)$ is the original signal, $\hat{s}(n)$ is the enhanced signal, M is the number of frames averaged, and N is the frame length. Fig. 3 shows the averaged SSNR for various types of noise and levels. As can be seen, the proposed estimators achieve larger SSNR values than the other estimators do under all tested conditions.

The PESQ measure was not generally intended to assess speech enhancement algorithms. However, it has been used in the past years in several speech enhancement algorithms. It converts the disturbance parameters in speech to a MOS-like listening quality score in a very wide range of conditions that may include codec distortions, errors, filtering, and variable signal delay. The higher score means better perceptual speech quality.

The results of the experiments, in terms of the mean PESQ scores for different types of noise and levels, are presented in Fig. 4. As can be observed, the proposed estimators give better results compared to the conventional estimators. A competitive result between the estimators $\hat{A}_k^{(1)}$ and $\hat{A}_k^{(E)}$ is obtained only in babble noise. The PESQ, however, is deaf to residual musical noise. A listening test should be, thus, conducted to investigate the subjective speech quality.

B. Subjective Evaluations

To evaluate the quality of speech produced by the four estimators, an informal listening test is conducted. Ten utterances of the corpus (produced by five male and five female speakers) corrupted by train noise at 10 dB SNR are used in the listening test. Listeners participated in the experiments are ten persons who have normal hearing ability. The listening test is conducted using a paired-preference paradigm. The listeners are presented with pairs of sentences: one enhanced with our proposed estimators and the other enhanced with either $\hat{A}_k^{(\beta)}$ or $\hat{A}_k^{(E)}$. The listeners are asked to choose the sentence which is (1) more natural (2) easier to listen, and (3) less distorted in terms of having less residual noise. The overall preference is assessed for speech enhanced by the proposed estimators compared to the speech enhanced by the conventional estimators. A higher percentage of preference, for a particular speech type processed by a method, indicates that the speech type is more preferred. On the other hand, a smaller percentage of preference indicates that the speech type is less preferred.

The listening test results are shown in Table I. As is evident, the proposed estimators have a higher preference than that of the conventional estimators. This is mostly due to the fact that the perceptually meaningful cost function with GGD speech priors is utilized in the proposed estimators.

V. Discussion and Conclusions

In this paper, we have proposed two Bayesian estimators for single channel speech enhancement. The experimental results show that the proposed estimators, particularly the estimator considered with $\delta = 2$, are effective for speech enhancement.

The conventional approaches generally increase the quality of speech by reducing the amount of noise, while they decrease the intelligibility in terms of perceptuality of speech. In contrast to the conventional approaches, the proposed estimators are devoted to increase the amount of noise reduction as well as perceptual speech quality. This can be observed by the SSNR and PESQ results.

As can be observed from the results, the proposed estimators maximize the amount of noise reduction, while they increase or keep the same perceptuality of speech compared to that the conventional approaches tested here do. This is also consistent to the results of the listening test in which the proposed estimators are more preferred than the conventional approaches. This is mainly due to the use of the weighted criterion that takes the advantage of the perceived loudness of the LSA estimator and masking properties of the Euclidean measure. The weighted estimators have further considered with generalized Gamma distributed speech priors under SPP that makes the proposed estimators to be superior.

References

[1] J. Benesty, S. Makino and J. Chen, *Speech Enhancement*. Berlin: Springer-Verlag, 2005, ch. 1.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, 1979.

[3] S. Stankovic, J. Tilp and R. Stojanovic, "Enhancement of speech signals disturbed by noise using time-varying filtering," in *Proc. 4th WSEAS Multiconference Circuits, Systems, Communications and Computers*, vol. 2, 2000, pp. 325-329.

[4] G. Costantini and D. Casali, "Speech noise reduction using adaptive spline neural networks," *WSEAS Trans. Circuits and Systems*, vol. 3, pp. 155-158, 2004.

[5] T. Shimamura and J. Yamauchi, "Spectral subtraction with non-stationary noise estimation utilizing harmonic structure," in *Proc. 4th WSEAS Int. Conf. Electronics, Control and Signal Processing*, 2005, pp. 47-52.

[6] H. Ding, I.Y. Soon, C.K. Yeo and S.N. Koh, "2D spectrogram filter for single channel speech enhancement," in *Proc. 7th WSEAS Int. Conf. Signal, Speech and Image Processing*, 2007, pp. 89-93.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, 1984.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-23, no. 2, pp. 443-445, 1985.

[9] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 2, pp. 126-137, 1999.

[10] Y. Ephraim and H. L.V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251-266, 1995.

[11] K. C. Wang and C. L. Chin, "A time-frequency adaptation based on quantum neural networks for speech enhancement," *WSEAS Trans. Information Science and Applications*, vol. 7, pp. 11-15, 2010.

[12] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. NJ: Prentice Hall, 1993.

[13] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 345-349, 1994.

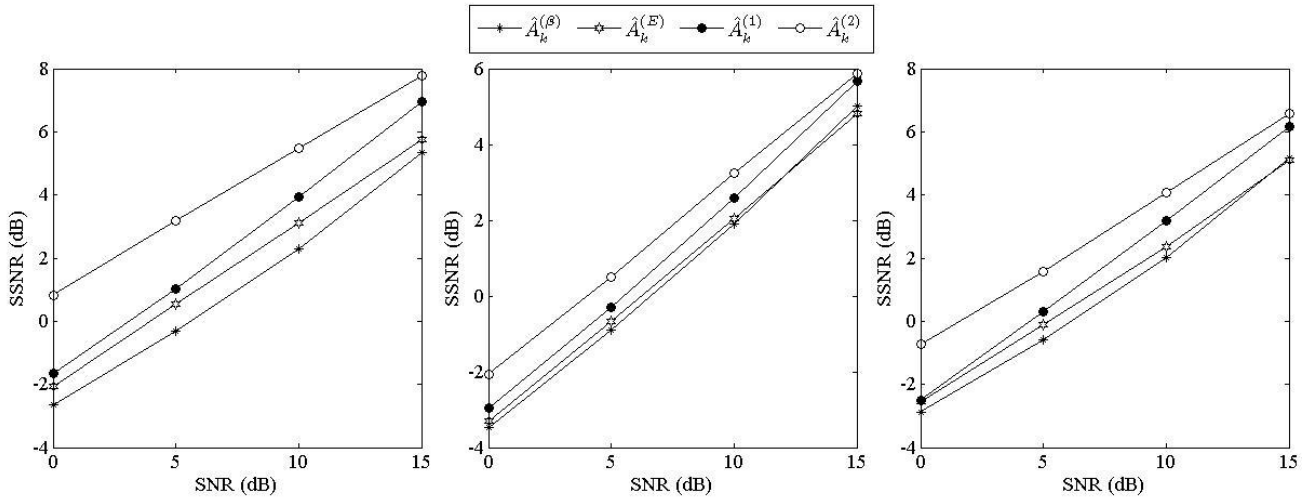

Fig. 3 Speech quality, in terms of SSNR, for various kinds of noise and levels. The left, middle and right panels show the SSNR for white Gaussian noise, babble noise and train noise, respectively.

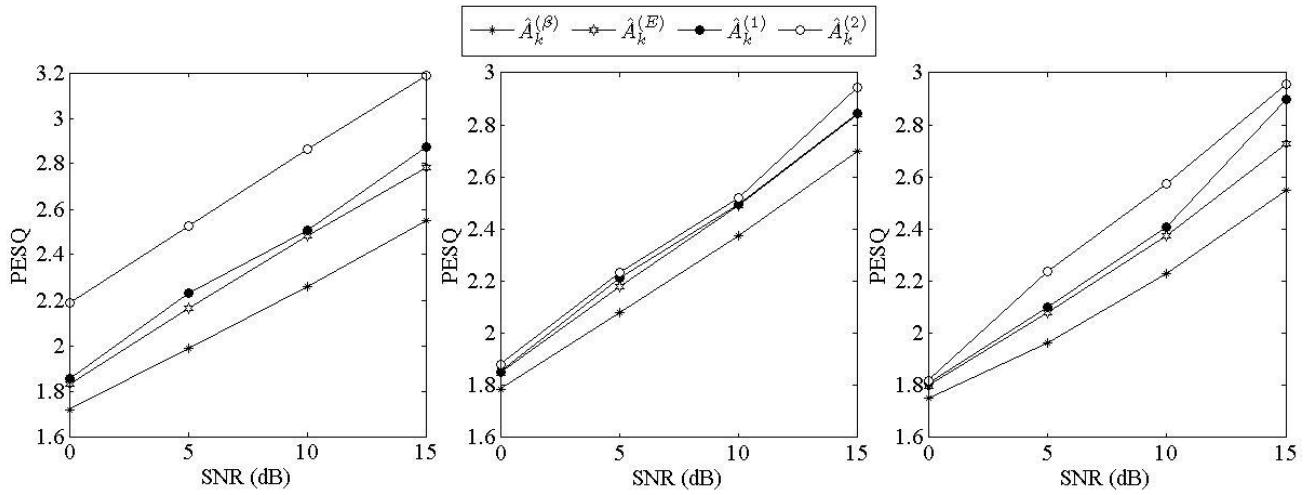

Fig. 4 Speech quality, in terms of PESQ, for various kinds of noise and levels. The left, middle and right panels show the PESQ values for white Gaussian noise, babble noise and train noise, respectively.

Table I Preference percentage for the proposed method compared to other methods

Method	Preference
$\hat{A}_k^{(1)}$ over $\hat{A}_k^{(\beta)}$	71%
$\hat{A}_k^{(1)}$ over $\hat{A}_k^{(E)}$	59%
$\hat{A}_k^{(2)}$ over $\hat{A}_k^{(\beta)}$	76%
$\hat{A}_k^{(2)}$ over $\hat{A}_k^{(E)}$	63%

[14] D. Malah, R.V. Cox and A.J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1999, pp. 789-792.

[15] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, pp. 2403-2418, 2001.

[16] B. Dashtbozorg and H.R. Abutalebi, "Adaptive MMSE speech spectral amplitude estimator under signal presence uncertainty," in *Proc. 17th European Signal Processing Conf.*, 2009, pp. 209-212.

[17] P.C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 475-486, 2005.

[18] E. Plourde and B. Champagne, "Generalized Bayesian estimators of the spectral amplitude for speech enhancement," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 485-488, 2009.

[19] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Processing*, vol. 7, pp. 1110-1126, 2005.

[20] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 845.856, 2005.
[21] A. Saha and T. Shimamura, "Weighted log-spectral amplitude estimation with generalized Gamma distribution under speech presence probability," in *Proc. Conf. Signal Processing and Applied Mathematics for Electronics and Communications*, 2011, pp. 37-40.
[22] A. Saha and T. Shimamura, "Generalized Gamma distributed Bayesian estimator under speech presence probability," in *Proc. 11th WSEAS Int. Conf. Applied Computer Science*, 2011, pp. 118-123.
[23] J.S. Erkelens, R.C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741-1752, 2007.
[24] C.H. You, S.N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 4, pp. 475-486, 2005.
[25] R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-28, pp. 137-145, 1980.
[26] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*. NY: Academic, 2000.
[27] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229-238, 2008.
[28] A. Saha and T. Shimamura, "Noise spectrum estimation based on optimum smoothing for robust speech enhancement," in *Proc. Int. Workshop on Nonlinear Circuits, Communication and Signal Processing*, 2010, pp. 293-296.
[29] ITU, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recommendation* 2000, p. 862.