Generalized Gamma Distributed Bayesian Estimator under Speech Presence Probability

ATANU SAHA† and TETSUYA SHIMAMURA‡
Graduate School of Science and Engineering
Saitama University
255 Shimo-okubo, Sakura-ku, Saitama, 338-8570
JAPAN
{saha†, shima‡}@sie.ics.saitama-u.ac.jp

Abstract: This paper presents an approach for speech enhancement based on the Bayesian estimator. The cost function in logarithmic domain of the Bayesian estimator is weighted by psychoacoustically motivated speech distortion measure. This weighted cost function exploits the generalized Gamma distributed speech priors under speech presence probability. The experimental results show that the proposed method provides better perception in speech quality compared to state-of-the-art speech enhancement approaches.

Key Words: Speech enhancement, Gamma distribution, Bayesian estimator, Speech presence probability

1 Introduction

The rationale behind the use of speech enhancement algorithms is to enhance the perceptual quality of speech corrupted by additive background noise. To accomplish this, a large number of researches on speech enhancement has been going on for a long time with various degrees of success.

Recently, the trend of deriving nonlinear estimators of the amplitude rather than the complex spectrum of the signal has shown itself to be effective in speech enhancement. Various techniques exist in the estimation theory literature [1] for deriving these nonlinear estimators. Bayesian estimators of the amplitude spectrum have received a lot of attention among these. A well-known Bayesian estimator is the minimum mean square error (MMSE) estimator that minimizes the conditional expectation of a squared-error cost function [2]. The squared-error cost function in logarithmic domain, called log-spectral amplitude (LSA) estimator [3], is more effective in reducing musical noise. The generalization of these cost functions under speech presence probability (SPP) was also proposed in [4,5]. More perceptually 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 [6,7].

The aforementioned approaches for speech enhancement in discrete Fourier transform (DFT) domain assume that the clean speech and noise DFT coefficients are complex 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 [8-10] have proposed the use of super-Gaussian such as Laplacian or Gamma distribution for modeling the speech DFT coefficients.

In this paper, we propose a speech enhancement approach that utilizes the generalized Gamma distribution (GGD) under SPP. The GGD is incorporated into a weighted cost function derived from the LSA estimator. The cost function of the LSA estimator is weighted by Euclidean distance measure to take advantage of perceptual interpretation that makes the proposed method more perceptual in speech quality compared to state-of-the-art speech enhancement approaches.

The organization of this paper is as follows. Section 2 provides the preliminary definitions of the proposed method. Section 3 describes the proposed Gamma distributed Bayesian estimator. In Section 4, the experimental results are presented. Finally, we conclude the paper in Section 5.

2 Basic Definitions

Let the observed noisy speech signal at frame $\lambda$ be assumed as

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

where $n$ is the sampling time 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

ISBN: 978-1-61804-039-8
signal can be expressed as
\[ Y_{\lambda,k} = \sum_{n=0}^{N-1} y_{\lambda}(n)h(n)e^{-j\frac{2\pi}{N}kn} \] (2)

where \( h(n) \) is the analysis window and \( k \in \{0, 1, ..., N - 1\} \) 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 and assumed to be statistically independent, (2) becomes

\[ Y_{\lambda,k} = S_{\lambda,k} + D_{\lambda,k}. \] (3)

The preceding equation can also be expressed by dropping the frame index for notational convenience in polar form as:
\[ R_k e^{j\phi} = A_k e^{j\psi} + N_k e^{j\omega} \] (4)

where \( \{R_k, A_k, N_k\} \) and \( \{\phi, \psi, \omega\} \) 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 [8-10]. A GGD assumption for speech amplitude can perform much better than the Rayleigh distribution assumption. The GGD is given by
\[ f(A_k) = \frac{\delta^\nu}{\Gamma(\nu)} A_k^{\nu-1} e^{-\eta A_k^\delta}, \quad \delta, \eta, \nu > 0 \] (5)

where \( \Gamma(.) \) is the Gamma function, \( \delta \) and \( \nu \) denote the shaping parameters, and \( \eta \) is called the scaling parameter. The special cases of generalized priors in (5) for different estimators depends on choosing the value of \( \delta \) [10]. The value of \( \delta = 2 \) has been considered in the approach proposed in [11]. In this study, we use \( \delta = 1 \) for which \( \nu \) is related to \( \nu \) and the variance of speech, \( \lambda_s(k) \), as \( \nu = \sqrt{\nu(\nu + 1)/\lambda_s(k)} \).

3 Proposed Gamma Distributed Enhancement Method under SPP

In this section, we derive the gain function of the proposed method with GGD under 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 \( \hat{S}_k = \hat{A}_k e^{j\psi} \). Finally, the enhanced time signal is obtained by taking inverse DFT of \( \hat{S}_k \).

In the logarithmic domain, which was proposed in [3], the cost function of the Bayesian estimator is chosen as
\[ C(A_k, \hat{A}_k) = (\log A_k - \log \hat{A}_k)^2. \] (6)

The LSA estimator shown in [3] 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
\[ \hat{A}_k = \exp \left( \frac{d}{d\rho} E[A_k^\rho | Y_k] \right) |_{\rho=0}. \] (7)

Equation (7) is equivalent to
\[ \hat{A}_k = \lim_{\rho \to 0} \exp \left( \frac{d}{d\rho} \log E[A_k^\rho | Y_k] \right). \] (8)

By applying L'Hopital's rule, (8) can be expressed as
\[ \hat{A}_k = \lim_{\rho \to 0} \exp \left( \frac{d}{d\rho} \log E[A_k^\rho | Y_k] \right). \] (9)

For a small value of \( \rho \), (9) can be simplified as
\[ \hat{A}_k = E[A_k^\rho | Y_k]^{\frac{1}{\rho}} \] (10)

where \( \rho \) is approximated as \( 0 < \rho \ll 1 \). Equation (10) is a special case of the approach proposed in [12].

The spectral amplitude estimator in (10) 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 respectively speech absence and presence, and assuming a complex Gaussian distribution of the DFT coefficients for both speech and noise [2], the conditional SPP, \( \zeta_k \) is given by
\[ \zeta_k = P(H_1(k) | Y_k), \] (11)

where \( q_k \) is the \( a \ priori \) probability of speech absence, \( \xi_k = \lambda_s(k)/\lambda_d(k) \) is the \( a \ priori \) SNR in which \( \lambda_d(k) \) denotes the variance of noise, \( \gamma_k = R_k^2/\lambda_d(k) \) is called the \( a \ posteriori \) SNR, and \( \nu_k = \xi_k \gamma_k/(1 + \xi_k) \). By taking into account the SPP \( \zeta_k \), the estimator in (10) is obtained as
\[ \hat{A}_k^\rho = \left[ E[A_k^\rho | Y_k, H_1(k)] \zeta_k + E[A_k^\rho | Y_k, H_0(k)] (1 - \zeta_k) \right]^{\frac{1}{\rho}}. \] (12)

where \( \hat{A}_k^\rho \) denotes the optimal spectral amplitude estimator under consideration of SPP. It is interesting
to mention that the estimator \( \hat{A}_k^p \) in (12) is a special case of the method proposed in [13]. Since the gain is constrained to be larger than a threshold \( G_{\text{min}} \) during speech absence, we consider

\[
E \left[ A_k^p | Y_k, H_0(k) \right] = (G_{\text{min}} \cdot R_k)^{\rho}.
\]  

Accordingly, the conditional gain function during speech presence is defined by

\[
E \left[ A_k^p | Y_k, H_1(k) \right] = (G_k \cdot R_k)^{\rho} \tag{14}
\]

where \( G_k \) is a gain function considered with GGD.

The proposed method is based on deriving \( G_k \) with generalized Gamma distributed speech priors. As can be seen from (6), 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 [6]. A more perceptually significant cost function is used in [6] based on a weighted error criterion that exploits the masking properties of the human auditory system. The chosen cost function is given by

\[
C \left( A_k, \hat{A}_k \right) = A_k^2 \left( A_k - \hat{A}_k \right)^2 \tag{15}
\]

where \( \tau \) is a real parameter. To obtain the gain function \( G_k \) corresponding to the above cost function in (15), we simplify (10) as

\[
\hat{A}_k = \left( \frac{E \left[ A_k^{\rho-\tau} | Y_k \right]}{E \left[ A_k^{-\tau} | Y_k \right]} \right)^{\frac{1}{\rho}}. \tag{16}
\]

Since the noise is assumed to be Gaussian distributed, the conditional probability of \( Y_k \) can be written as [14]

\[
f \left( Y_k | A_k \right) = \frac{2R_k}{\lambda_d(k)} \exp \left( -\frac{R_k^2 + A_k^2}{\lambda_d} \right) I_0 \left( \frac{2A_k R_k}{\lambda_d} \right) \tag{17}
\]

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 (17) and by specifying the GGD prior of \( A_k \) in (5), the \( \rho^{th} \) conditional moment can be simplified as

\[
E[A_k^p | Y_k] = \int_0^\infty A_k^{\rho + \frac{3}{2}} \exp \left( -\frac{A_k^2}{\lambda_d(k)} - \mu_k A_k \right) dA_k \tag{18}
\]

where \( \mu_k \) is defined as

\[
\mu_k = \frac{2\sqrt{\gamma_k \xi_k} - \sqrt{\nu(\nu + 1)}}{\sqrt{2\xi_k}} \tag{19}
\]

In terms of confluent hypergeometric function [15], the conditional moment in (18) can be determined by (20), shown in the next page. In (20), \( \Phi(\cdot) \) is called the confluent hypergeometric function. Substituting (20) in (16), we obtain

\[
\hat{A}_k = G_k R_k \tag{21}
\]

where \( G_k \) is determined by (22), shown in the next page. In (22), we simplify \( \rho = -\nu - \rho + \tau + 0.5 \) and \( q = -\nu + \tau + 0.5 \). From (12), (13), (14) and (22), the gain function via \( \hat{A}_k^p = G_k^p R_k \) is determined by (23), shown in the next page. This gain function \( G_k^o \) is a function of both the a priori SNR \( \xi_k \) and posterior SNR \( \gamma_k \). Figure 1 plots the gain function \( G_k^o \) as a function of the instantaneous SNR, \( (\gamma_k - 1) \), for a fixed value of \( \xi_k = 5 \) dB in left panel and \( \xi_k = 5 \) dB in right panel) for several values of \( \rho \) and \( \tau \). As can be seen, the shape of the gain function \( G_k^o \) is similar to both cases. The trade-off between the amount of attenuation and speech distortion seems to be dependent on the value of the parameters \( \rho \) and \( \tau \). Small values of \( \rho \) provide higher attenuation, while small values of \( \tau \) provide lower attenuation. As a compromise between the amount of noise reduction and speech distortion, we use \( \rho = -3 \) dB and \( \tau = -10 \) dB in the experiment.

4 Experimental Results

In this section, the performance of the proposed estimator, referred to as Gamma distributed Bayesian (GDB) estimator, is investigated. The NOIZEUS speech corpus [16] is used for evaluation in the experiments. 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 exhibition noise and street noise. All utterances (30 utterances) of the corpus are used in our evaluation. Half of the utterances come from male speakers and half are from female speakers. The sampling frequency of the utterances is 8 kHz. An analysis Hamming window of 20-msec is used with 50% overlap between frames. The lower bound threshold \( G_{\text{min}} \) is set to \(-40 \) dB. The shaping parameter \( \nu \) is set to \(-2 \) dB. The a priori SNR \( \xi_k \) is estimated as the approach proposed in [2] and the a priori probability of speech absence for computing the SPP is estimated according to [4]. The noise spectrum is estimated as the approach proposed in [17].

The performance of the GDB estimator is investigated by comparing that of the weighted Euclidean (WE) estimator proposed in [6]. The parameter selection for WE is the same as the approach proposed in
Figure 1: Gain curves of the proposed estimator for several values of $\rho$ and $\tau$ as a function of the instantaneous SNR ($\gamma_k - 1$) for $\nu = -2$ dB. The left panel plots the gain curves for $\xi_k = -5$ dB, whereas the right panel plots the gain curves for $\xi_k = 5$ dB.

[6]. For evaluation the speech quality, objective evaluation and subjective listening test are conducted.

Objective measure, in terms of PESQ (perceptual evaluation of speech quality) [18], is used to evaluate the performance of two estimators. The PESQ measure is employed to predict high correlation with subjective speech quality measure. The PESQ converts the disturbance parameters in speech to a MOS (mean opinion score)-like listening quality score. The higher the score, the better the perceptual quality.

The results, in terms of average PESQ scores for all utterances, are listed in Table I for exhibition noise and street noise. The input SNR of both noise conditions are set to 0, 5, 10 and 15 dB. As is evident from Table 1, the GDB estimator outperforms the WE estimator for both exhibition noise and street noise.
Table 1: Performance, in terms of PESQ, of the GDB estimator for different types of noise and levels. The performance of the WE estimator is also shown for comparison.

<table>
<thead>
<tr>
<th>Noise Method</th>
<th>0 dB</th>
<th>5 dB</th>
<th>10 dB</th>
<th>15 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhibition WE</td>
<td>1.65</td>
<td>2.01</td>
<td>2.42</td>
<td>2.87</td>
</tr>
<tr>
<td>GDB</td>
<td>1.78</td>
<td>2.19</td>
<td>2.51</td>
<td>2.99</td>
</tr>
<tr>
<td>Street WE</td>
<td>1.71</td>
<td>2.11</td>
<td>2.39</td>
<td>2.65</td>
</tr>
<tr>
<td>GDB</td>
<td>1.82</td>
<td>2.25</td>
<td>2.47</td>
<td>2.82</td>
</tr>
</tbody>
</table>

Table 2: Preference percentage for the GDB estimator compared to WE estimator.

<table>
<thead>
<tr>
<th>Method</th>
<th>Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDB over WE</td>
<td>69%</td>
</tr>
</tbody>
</table>

noise. The noticeable improvement in performance is found at low SNR. The PESQ measure, however, does not sound to musical noise which is often introduced in the processed speech by the exaggerated suppression of the enhancement algorithms, especially at low SNR. A listening test is, thus, conducted to investigate the subjective speech quality.

Ten sentences of the corpus (produced by five male and five female speakers) corrupted by street noise at 5 dB SNR are used in the listening test. The listening test is conducted using a paired-preference paradigm. Five listeners are participated in the listening test. The listeners are presented with pairs of sentences: one enhanced with the GDB estimator and the other enhanced with the WE estimator. The listeners are asked to choose the sentence which is more easier to listen and less distorted in terms of having less musical noise. The overall preference is assessed for speech enhanced by the GDB estimator compared to the speech enhanced by the WE estimator.

The results, in terms of preference percentage, are presented in Table 2. As is evident, the GDB estimator has the higher preference than that of the WE estimator. We suppose that this is mostly due to the fact that the perceptually meaningful cost function with the GGD is used in the proposed estimator.

5 Conclusion

In this paper, a speech enhancement approach has been proposed. The experimental results show that the proposed method yields more improvement in perceptual speech quality than that of the conventional method. This is mainly happened due to the fact that a mixture model of Gaussian and Gamma distribution under SPP is exploited into a cost function that is perceptually weighted by Euclidean distance measure. In the weighted cost function, the exponent is set to a very small value that preserves the characteristics of LSA estimator. Furthermore, the utilization of the Euclidean distance measure makes the processed speech produced by the proposed method perceptual.

References:


