WEIGHTED LOG-SPECTRAL AMPLITUDE ESTIMATION WITH GENERALIZED GAMMA DISTRIBUTION UNDER SPEECH PRESENCE PROBABILITY

Atanu SAHA and Tetsuya SHIMAMURA

Graduate School of Science and Engineering, Saitama University, Saitama, Japan
{saha,shima}@sie.ics.saitama-u.ac.jp

ABSTRACT

In this paper, we propose a speech enhancement approach. The approach is based on deriving weighted log-spectral amplitude estimator that exploits the generalized Gamma distributed speech priors under speech presence probability. The log-spectral amplitude estimator is weighted by psychoacoustically motivated speech distortion measure to take advantage of the perceptual interpretation. The experimental results show that the proposed approach provides less residual noise and better speech quality compared to state-of-the-art speech enhancement approaches.

Keywords: *Speech enhancement, weighted log-spectral amplitude estimator, generalized Gamma distribution, speech presence probability.*

1. INTRODUCTION

Significant progress has been made in developing speech enhancement algorithms over the last few decades [1]. The enhancement algorithms are generally concerned with improving the perceptual aspects of speech quality while extracting the desired signal from its corrupted observations.

Enhancement algorithms based on Bayesian estimators of the amplitude spectrum have received a lot of attention. 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, resulting in log-spectral amplitude (LSA) estimator [3], has shown itself to be more effective in reducing musical noise. The generalization of these cost functions under speech presence probability (SPP) was also reported in [4-6]. More perceptually motivated Bayesian estimators that use variants of speech distortion measures as the cost functions in place of the squared-error cost function were proposed in [7,8].

The aforementioned approaches estimate the clean speech discrete Fourier Transform (DFT) coefficients based on the assumption that the clean speech and noise DFT coefficients are complex Gaussian distributed. The recent studies [9,10], however, show that the clean speech DFT coefficients have a super-Gaussian behavior. Based on this observation, the super-Gaussian distribution assumptions such as Laplacian or Gamma distributions have been derived in [10].

This paper is devoted to deriving a weighted LSA, referred in this paper to as WLGSP, estimator that exploits the generalized Gamma distribution (GGD) assumption for the speech DFT coefficients under SPP. The LSA estimator is weighted by a perceptually meaningful cost function that uses Euclidean distance 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 component as undistorted as possible.

The organization of the paper is as follows. Section 2 provides basic assumptions and some preliminary notations that we will use through out this paper. In Section 3, we describe the proposed method. Section 4 shows the experimental results. Finally, Section 5 concludes the paper.

2. PRELIMINARY DEFINITIONS

Let us assume that the noisy speech at sampling time index n, $x(n)$, consists of clean speech $s(n)$ and additive noise $d(n)$. That is $x(n) = s(n) + d(n)$. The noisy signal in the time domain is transformed into the frequency domain by the application of a window function and DFT. In the frequency domain, we have

$$X(\lambda,k) = S(\lambda,k) + D(\lambda,k) \tag{1}$$

where $X(\lambda,k)$, $S(\lambda,k)$ and $D(\lambda,k)$ are the DFT coefficients obtained at frequency index k in frame λ from the noisy speech, clean speech and noise, respectively. The DFT coefficients $S(\lambda,k)$ and $D(\lambda,k)$ are assumed to be statistically independent.

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

$$R_k e^{j\phi} = A_k e^{j\psi} + N_k e^{j\omega} \tag{2}$$

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 [9,10]. A GGD assumption for speech amplitude can perform much better than the Rayleigh distribution assumption. The GGD is given by

$$f(A_k) = \frac{\delta \eta^{\nu}}{\Gamma(\nu)} A_k^{\delta\nu-1} \exp(-\eta A_k^{\delta}), \quad \delta, \eta, \nu > 0 \tag{3}$$

where $\Gamma(.)$ is the Gamma function, δ and ν denote shaping parameters, and η is a parameter specified by ν. The special cases of generalized priors in (3) for different estimators depends on choosing the value of δ [11]. In this study, we use $\delta = 2$ for which η is related to ν and the variance of speech, $\lambda_s(k)$, as $\eta = \nu/\lambda_s(k)$.

3. WEIGHTED LOG-SPECTRAL AMPLITUDE ESTIMATION WITH GGD UNDER SPP

In this section, we derive the WLGSP estimator under SPP with generalized Gamma distributed speech priors.

Bayesian spectral amplitude estimator minimizes the conditional expectation of a cost function, $E\left[C(A_k,\hat{A}_k)\right]$, 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 $\hat{S}_k = \hat{A}_k e^{j\phi}$. 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 is chosen as

$$C\left(A_k,\hat{A}_k\right) = \left(\log\ A_k - \log\ \hat{A}_k\right)^2. \tag{4}$$

The LSA estimator shown in [3] can be derived by exploiting the moment generating function of $\log\ A_k|X_k$ for complex Gaussian distributed clean speech and noise DFT coefficients as

$$\hat{A}_k = \exp\left(\frac{d}{d\rho} E[A_k^\rho|X_k]\right)|_{\rho=0}\,. \tag{5}$$

The equation (5) is equivalent to

$$\hat{A}_k = \lim_{\rho\to 0} \exp\left(\frac{\frac{d}{d\rho}E[A_k^\rho|X_k]}{E[A_k^\rho|X_k]}\right). \tag{6}$$

By applying L'Hopital's rule, (6) can be expressed as

$$\hat{A}_k = \lim_{\rho\to 0} \exp\left(\frac{\frac{d}{d\rho}\log\ E[A_k^\rho|X_k]}{\frac{d}{d\rho}\rho}\right). \tag{7}$$

For a small value of ρ, (7) can be expressed as

$$\hat{A}_k = E\left[A_k^\rho|X_k\right]^{\frac{1}{\rho}} \tag{8}$$

where ρ is approximated as $0 < \rho \ll 1$. Equation (8) is a special case of the approach proposed in [12].

The spectral amplitude estimator in (8) is now considered under SPP. Given two hypotheses, $H_0(k) : X_k = D_k$ and $H_1(k) : X_k = S_k + D_k$, 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, $p_k \triangleq P(H_1(k)|X_k)$, is given by

$$p_k = \left\{1 + \frac{q_k}{1-q_k}(1+\xi_k)\exp(-\nu_k)\right\}^{-1} \tag{9}$$

where $q_k \triangleq P(H_0(k))$ 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 p_k, the estimator in (8) is obtained as

$$\hat{A}_k^o = \left[E\left[A_k^\rho|X_k,H_1(k)\right]p_k + E\left[A_k^\rho|X_k,H_0(k)\right](1-p_k)\right]^{\frac{1}{\rho}} \tag{10}$$

where $\hat{A}_k^o$ denotes the optimal spectral amplitude estimator under consideration of SPP. It is interesting to mention that the estimator $\hat{A}_k^o$ in (10) is a special case of the method proposed in [5]. Since the gain is constrained to be larger than a threshold $G_{\min}$ during speech absence, we consider

$$E\left[A_k^\rho|X_k,H_0(k)\right] = (G_{\min}.R_k)^\rho\,. \tag{11}$$

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

$$E\left[A_k^\rho|X_k,H_1(k)\right] = \left(G_k^{wlg}.R_k\right)^\rho \tag{12}$$

where G_k^{wlg} is a gain function considered with GGD.

The proposed method is based on deriving G_k^{wlg} with generalized Gamma distributed speech priors. As can be seen from (4), 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 [7]. A more perceptually significant cost function is used in [7] based on a weighted error criterion that exploits the masking properties of the human auditory system. The chosen cost function is given by

$$C\left(A_k,\hat{A}_k\right) = A_k^\tau\left(A_k - \hat{A}_k\right)^2 \tag{13}$$

where τ is a real parameter. To obtain the gain function G_k^{wlg} corresponding to the above cost function in (13), we simplify (8) as

$$\hat{A}_k = \left(\frac{E\left[A_k^{\rho-\tau}|X_k\right]}{E\left[A_k^{-\tau}|X_k\right]}\right)^{\frac{1}{\rho}}. \tag{14}$$

By specifying the GGD prior of A_k in (3), the ρ^{th} conditional moment can be simplified as

$$E\left[A_k^\rho|X_k\right] = \left(\frac{\sqrt{\mu}_k}{\gamma_k}\right)^\rho \frac{\Gamma(\nu+\frac{\rho}{2})\Phi(-\nu+1-\frac{\rho}{2};1;-\mu_k)}{\Gamma(\nu)\Phi(-\nu+1;1;-\mu_k)} R_k^\rho \tag{15}$$

where $\mu_k = \xi_k\gamma_k/(\nu+\xi_k)$ and $\Phi(.)$ denotes the confluent hypergeometric function [13]. Substituting (15) in (14), we obtain

$$\hat{A}_k = G_k^{wlg} R_k \tag{16}$$

where

$$G_k^{wlg} = \frac{\sqrt{\mu}_k}{\gamma_k}\left(\frac{\Gamma\left(\nu+\frac{\rho-\tau}{2}\right)\Phi(-\nu+1-\frac{\rho-\tau}{2};1;-\mu_k)}{\Gamma\left(\nu-\frac{\tau}{2}\right)\Phi(-\nu+1+\frac{\tau}{2};1;-\mu_k)}\right)^{\frac{1}{\rho}} \tag{17}$$

is a gain function considered with generalized Gamma distributed speech priors. From (10), (11), (12) and (17), the gain function via $\hat{A}_k^o = G_k^{wlgsp} R_k$ is determined by (18),

$$G_k^{wlgsp} = \left\{ \left(\frac{\sqrt{\mu_k}}{\gamma_k} \right)^{\rho} \left(\frac{\Gamma(\nu + \frac{\rho-\tau}{2})\Phi(-\nu+1-\frac{\rho-\tau}{2};1;-\mu_k)}{\Gamma(\nu-\frac{\tau}{2})\Phi(-\nu+1+\frac{\tau}{2};1;-\mu_k)} \right) p_k + G_{\min}^{\rho}(1-p_k) \right\}^{\frac{1}{\rho}} \quad (18)$$

shown in top of this page. In (18), G_k^{wlgsp} is the gain function for the proposed WLGSP estimator with generalized Gamma distributed speech priors under SPP. In (18), the parameters ρ and τ are found to control the trade-off between speech distortion and noise reduction. Figure 1 presents gain curves for several values of ρ and τ as a function of the instantaneous SNR, $\gamma_k - 1$, for a fixed value of $\xi_k = 0$ dB. As can be seen, lower attenuation is obtained for lower value of ρ, whereas lower attenuation is obtained for higher value of τ. This interpretation helps us to select the value of ρ and τ used in the experiment. Consequently, we select $\rho = -5$ dB and $\tau = -3$ dB as a good compromise between the desired noise reduction performed by the estimator and the speech distortion introduced.

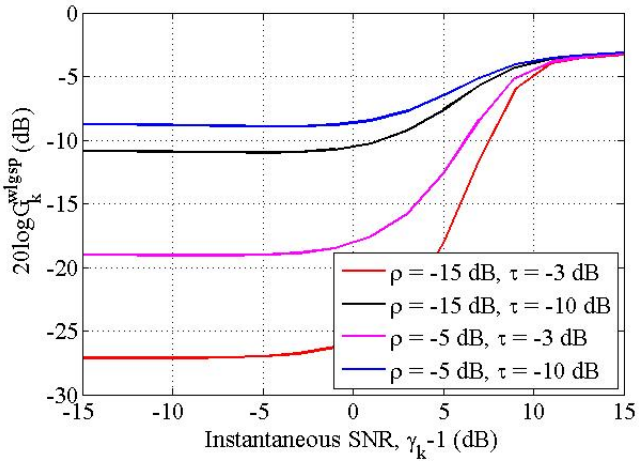

Figure 1*: Gain curves for several values of ρ and τ as a function of $(\gamma_k - 1)$ for $\xi_k = 0$ dB and $\nu = -2$ dB.*

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

$$\Phi(a;1;-\mu_k)\,|_{\mu_k \gg 1} \approx \frac{\mu_k^{-a}}{\Gamma(1-a)}, \quad (19)$$

the gain function G_k^{wlgsp}, for $p_k = 1$, converges to

$$G_k^{wlgsp} \approx \frac{\xi_k}{\nu + \xi_k} \quad (20)$$

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

Table 1: Performance, in terms of segmental SNR, of the WLGSP estimator for different types of noise and levels. The performance of the WE estimator is also shown for comparison.

Noise	Method	0 dB	5 dB	10 dB	15 dB
Exhibition	WE	-0.78	1.15	3.62	6.72
	WLGSP	1.54	3.10	5.50	8.65
Train	WE	-1.61	1.55	3.88	6.31
	WLGSP	1.56	3.89	5.92	8.30

4. EXPERIMENTAL RESULTS

In this section, we investigate the performance of the proposed WLGSP estimator by simulation experiments. The NOIZEUS speech corpus [14] 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 train noise. A total of four utterances from the NOIZEUS corpus are used in our evaluation. Half of them are from male speakers and half are from female speakers. The test utterances are sampled at 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 ν is set to -2 dB. The decision directed approach [2] is used for estimating the *a priori* SNR and the *a priori* probability of speech absence for computing the SPP is estimated according to [4]. For estimation of noise variance, we use the method proposed in [15].

Speech quality is evaluated by the segmental SNR (SSNR) in speech segments and perceptual evaluation of speech quality (PESQ) [16]. The SSNR is calculated as follows:

$$SSNR = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\frac{\sum_{n=N*j}^{N*j+N-1} s(n)^2}{\sum_{n=N*j}^{N*j+N-1}[s(n)-\hat{s}(n)]^2} \right] \quad (21)$$

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.

The performance of the WLGSP estimator is investigated by comparing that of the weighted Euclidean (WE) estimator proposed in [6]. The performance, in terms of SSNR, of the estimators is shown in Table 1 for different types of noise and levels. As can be observed, the WLGSP estimator achieves a larger SSNR for all cases tested in the experiments. Table 2 shows the performance, in terms of PESQ measure, of the estimators for different kinds of noise and levels. It is worth noting that a higher PESQ score results in more perceptual speech quality. As can be seen from Table 2, the proposed WLGSP estimator achieves a higher

Table 2: Performance, in terms of PESQ, of the WLGSP estimator for different types of noise and levels. The performance of the WE estimator is also shown for comparison.

Noise	Method	0 dB	5 dB	10 dB	15 dB
Exhibition	WE	1.76	1.97	2.32	2.81
	WLGSP	1.69	2.05	2.47	2.89
Train	WE	1.73	2.05	2.33	2.68
	WLGSP	1.75	2.11	2.44	2.77

PESQ score than the WE estimator in most of the cases. An informal listening test also confirms that the WLGSP estimator performs significantly better by providing less residual noise than the WE estimator. A further improvement of the proposed estimator may be still possible by considering time-and frequency dependent parameters ρ and τ, since they control a trade-off between noise reduction and speech distortion.

5. CONCLUSIONS

In this paper, we have proposed a speech enhancement approach based on the WLGSP estimator. The experimental results show that the proposed estimator yields more improvements over other tested estimators in terms of speech quality measures. 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 estimator has further considered with generalized Gamma distributed speech priors under SPP that makes the proposed estimator to be superior.

REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, Signals and Communication Technology, Springer, 2005.

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

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-33, pp. 443-445, 1985.

[4] 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. Acoustics, Speech, and Signal Process., pp. 789-792, 1999.

[5] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator", IEEE Signal Processing Letters, vol. 9, no. 4, pp. 113-116, 2002.

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

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

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

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

[10] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors", IEEE Trans. Speech Audio Process., vol. 13, no. 5, pp. 845.856, Sep. 2005.

[11] 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 Process., Vol. 15, no. 6, 2007.

[12] C.H. You, S.N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement", IEEE Trans. Speech Audio Process., vol. 13, no. 4, pp. 475-486, 2005.

[13] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. New York: Academic, 2000.

[14] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement", IEEE Trans. Audio, Speech, and Language Process., vol. 16, no. 1, pp. 229-238, 2008.

[15] 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 Process., pp. 293-296, 2010.

[16] 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 P. 862, 2000.