Complex laplacian probability density function for noisy speech enhancement

Joon-Hyuk Chang

School of Electronic and Electrical Engineering,

Inha University, Incheon, 402–751, Korea

changjh@inha.ac.kr

Abstract: In this letter, we propose a new approach to speech enhancement based on a complex Laplacian probability density function (pdf). With a use of a goodness-of-fit (GOF) test, we discover that the complex Laplacian pdf is more desirable to describe noisy speech distribution than the conventional Gaussian pdf for speech enhancement. The likelihood ratio (LR) is computed and then applied to computation of the speech absence probability in the speech enhancement algorithm. Our evaluation shows that the proposed complex Laplacian pdf scheme provides further speech quality enhancement compared with the conventional method through a number of objective tests.

Keywords: GOF, likelihood ratio, complex laplacian, speech enhancement

Classification: Science and engineering for electronics

References

- [1] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Lett.*, vol. 7, no. 5, pp. 108–110, May 2000.
- [2] J.-H. Chang and N. S. Kim, "Speech enhancement : new approaches to soft decision," *IEICE Trans. Inf. & Syst.*, vol. 27, E84-D, pp. 1231–1240, Sept. 2001.
- [3] J.-H. Chang and N. S. Kim, "Voice activity detection based on complex Laplacian model," *Electron. Lett.*, vol. 39, no. 7, pp. 632–634, April 2003.
- [4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
- [5] A. G. Glen, L. M. Leemis, and D. R. Barr, "Order statistics in goodness-of-fit testing," *IEEE Trans. Rel.*, vol. 50, no. 2, pp. 209–213, June 2001.
- [6] R. C. Reininger and J. D. Gibson, "Distributions of the two dimensional DCT coefficients for images," *IEEE Trans. Commun.*, vol. Com-31, no. 6, pp. 835–839, June 1983.
- [7] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*,

vol. 12, no. 3, pp. 247–251, July 1993.

- [8] N. Ma, M. Bouchard, and R. Goubran, “Perceptual Kalman filtering for speech enhancement in colored noise,” in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, vol. 1, pp. 717–720, Montreal, May 2004.

1 Introduction

Noisy speech enhancement has become a crucial part of the variable-rate speech coding and automatic speech recognition [1]. Recently, a number of efforts to improve the performance of speech enhancement by adopting soft decision estimation have been made since soft decision gain modification have better performances compared with earlier methods employing hard decisions in which each frame is classified into either speech or non-speech employing a voice activity detector (VAD) [2]. The most popular algorithms for the enhancement of noisy speech are based on the discrete Fourier transform (DFT) domain, it is assumed that the distributions of the clean speech and noise spectra are characterized by the complex Gaussian probability density function (pdf) [1, 2]. Recently, it has been reported that the DFT coefficients of clean and noise are better modeled by the Gamma and Laplacian pdf, respectively. Furthermore, complex Laplacian pdf was introduced for the purpose of VAD with the DFT coefficients of noisy speech under various noise environments.

In this paper, we present the complex Laplacian pdf for speech enhancement in the DFT domain. At first, we compare the Laplacian pdf with the conventional pdf by applying the goodness-of-fit (GOF) test under various noisy signal conditions. The GOF test indicates that the empirical noisy speech distribution is more close to approximated by the Laplacian pdf than the Gaussian pdf. Based on the statistical analysis, we devise a novel method to compute the speech absence probability (SAP) based on the complex Laplacian pdf and apply it not only to spectral gain modification, but also to update noise spectrum estimates under a unified framework. The performances of the proposed algorithm are evaluated by the ITU-T P.862 perceptual evaluation of speech quality (PESQ) tests and they are shown better than those of the method used in [1].

2 Statistical Analysis

We assume that a noise signal n is added to a speech signal s , with their sum being denoted by x . Given two hypotheses, H_0 and H_1 , which respectively indicate speech absence and presence, it is assumed that

$$H_0 : \text{speech absent} : \mathbf{X}(t) = \mathbf{N}(t) \quad (1)$$

$$H_1 : \text{speech present} : \mathbf{X}(t) = \mathbf{N}(t) + \mathbf{S}(t) \quad (2)$$

in which $\mathbf{X}(t) = [X_0(t), X_1(t), \dots, X_{M-1}(t)]^T$, $\mathbf{N}(t) = [N_0(t), N_1(t), \dots, N_{M-1}(t)]^T$, and $\mathbf{S}(t) = [S_0(t), S_1(t), \dots, S_{M-1}(t)]^T$ are the discrete Fourier

transform (DFT) coefficients of the noisy speech, noise and clean speech, respectively. The above statistical model is completed with a suitable specification of the DFT coefficients' distribution. In this paper, we consider two different probabilistic density functions (pdf's) for the candidate distribution. The first one is the conventional complex Gaussian pdf which is most widely applied to characterize the DFT coefficients distribution in speech analysis [1].

For a successful statistical model-based speech enhancement, we must select a model that fits better to the given noisy speech spectra. For this, we carry out a statistical fitting test for the noisy spectral components conditioned on both H_0 and H_1 under various noise conditions. For the pdf selection, we apply the *Kolmogorov-Smirnov* (*KS*) test which is well-known as the goodness-of-fit (GOF) test [5, 6]. Due to the incorporation of *KS* test, a reliable survey of each statistical assumption is guaranteed. From the *KS* test, we can conclude that the Laplacian model is more adequate for characterizing the DFT coefficients of noisy speech than the Gaussian model.

3 Improved Speech Enhancement Employing Complex Laplacian pdf

With the Gaussian pdf assumption, the distributions of the noisy spectral components conditioned on both hypotheses are given by

$$p_G(X_k|H_0) = \frac{1}{\pi\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}\right\} \quad (3)$$

$$p_G(X_k|H_1) = \frac{1}{\pi[\lambda_{n,k} + \lambda_{s,k}]} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k} + \lambda_{s,k}}\right\} \quad (4)$$

where $\lambda_{n,k}$ and $\lambda_{s,k}$ denote the variances of N_k and S_k , respectively [1]. In this paper, we introduce the complex Laplacian pdf. Let $X_{k(R)}$ and $X_{k(I)}$ denote the real and imaginary parts of the DFT coefficient X_k , respectively. Then, according to the complex Laplacian pdf, $X_{k(R)}$ and $X_{k(I)}$ are assumed to be distributed as follows:

$$p(X_{k(R)}) = \frac{1}{\sigma_x} \exp\left\{-\frac{2|X_{k(R)}|}{\sigma_x}\right\} \quad (5)$$

$$p(X_{k(I)}) = \frac{1}{\sigma_x} \exp\left\{-\frac{2|X_{k(I)}|}{\sigma_x}\right\} \quad (6)$$

where σ_x^2 denotes the variance of X_k . If the real and imaginary parts of X_k are assumed to be independent, we can obtain $p(X_k)$ such that

$$p(X_k) = p(X_{k(R)}) \cdot p(X_{k(I)}) \quad (7)$$

$$= \frac{1}{\sigma_x^2} \exp\left\{-\frac{2(|X_{k(R)}| + |X_{k(I)}|)}{\sigma_x}\right\}. \quad (8)$$

To justify the assumption of independence real and imaginary parts, we evaluated scatter plots of the real and imaginary parts of DFT coefficients. From these plots, it is concluded that the real and imaginary parts are only weakly dependent and the error is negligible.

Using (8), the distributions of the noisy DFT coefficients given two hypotheses are described as follows:

$$p_L(X_k|H_0) = \frac{1}{\lambda_{n,k}} \exp\left\{-\frac{2(|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{\lambda_{n,k}}}\right\} \quad (9)$$

$$p_L(X_k|H_1) = \frac{1}{\lambda_{n,k} + \lambda_{s,k}} \cdot \exp\left\{-\frac{2(|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{\lambda_{n,k} + \lambda_{s,k}}}\right\}. \quad (10)$$

Conditioned on current observation $\mathbf{X}(t)$, the global speech absence probability (GSAP) is given by

$$p(H_0|\mathbf{X}(t)) = \frac{p(\mathbf{X}(t)|H_0)p(H_0)}{p(\mathbf{X}(t)|H_0)p(H_0) + p(\mathbf{X}(t)|H_1)p(H_1)} \quad (11)$$

where $p(H_0)$ ($= 1 - p(H_1)$) is the *a priori* probability for speech absence. Since the spectral component in each frequency bin is assumed to be statistically independent, (11) can be rewritten as

$$\begin{aligned} p_L(H_0|\mathbf{X}(t)) &= \frac{p(H_0) \prod_{k=1}^M p_L(X_k(t)|H_0)}{p(H_0) \prod_{k=1}^M p_L(X_k(t)|H_0) + p(H_1) \prod_{k=1}^M p_L(X_k(t)|H_1)} \\ &= \frac{1}{1 + q \prod_{k=1}^M \Lambda_k(X_k(t))} \end{aligned} \quad (12)$$

in which q is the ratio defined by

$$q = \frac{p(H_1)}{p(H_0)} \quad (13)$$

and $\Lambda_k(X_k(t))$ is the likelihood ratio computed in the k th frequency bin as follows:

$$\Lambda_k(X_k(t)) = \frac{p_L(X_k|H_1)}{p_L(X_k|H_0)} = \frac{1}{1 + \xi_k} \exp\left\{2(|X_{k(R)}| + |X_{k(I)}|)\right\} \quad (14)$$

$$= \left(\frac{|X_k| - \sqrt{\lambda_{n,k}}}{|X_k| \sqrt{\lambda_{n,k}}}\right)\}. \quad (15)$$

where

$$\xi_k(t) \equiv \frac{\lambda_{s,k}(t)}{\lambda_{n,k}(t)}. \quad (16)$$

4 Speech Enhancement

The success or failure of the global soft decision approach adopting (12) depends mostly on the reliable estimation of $\{\lambda_{s,k}(t)\}$ and $\{\lambda_{n,k}(t)\}$ [1]. Under a general stationarity assumption of $\mathbf{N}(t)$ and $\mathbf{X}(t)$, we use the long-term smoothed power spectra of the background noise and clean speech as the estimates for $\{\lambda_{n,k}(t)\}$ and $\{\lambda_{s,k}(t)\}$, respectively. Let $\hat{\lambda}_{n,k}(t)$ and $\hat{\lambda}_{s,k}(t)$ be the estimates for $\lambda_{n,k}(t)$ and $\lambda_{s,k}(t)$. Then,

$$\begin{aligned} \hat{\lambda}_{n,k}(t+1) &= \zeta_n \hat{\lambda}_{n,k}(t) + (1 - \zeta_n) E[|N_k(t)|^2 | \mathbf{X}(t)] \\ \hat{\lambda}_{s,k}(t+1) &= \zeta_s \hat{\lambda}_{s,k}(t) + (1 - \zeta_s) E[|X_k(t)|^2 | \mathbf{X}(t)] \end{aligned} \quad (17)$$

where $0 < \zeta_n < 1$ and $0 < \zeta_s < 1$ are the smoothing parameters. It is not difficult to see from (19) that $\hat{\lambda}_{n,k}(t)$ and $\hat{\lambda}_{s,k}(t)$ do not rely on the current observation, $\mathbf{X}(t)$, which implies that they are some kind of predicted estimates of relevant parameters.

Let $\hat{\mathbf{S}}(t) = [\hat{S}_1(t), \hat{S}_2(t), \dots, \hat{S}_M(t)]$ denote the estimated clean speech spectrum at t th frame. In most of the conventional spectral enhancement techniques, $\hat{X}(t)$ is obtained by applying a specific gain to each spectral component of the noisy speech signal. Among a number of ways to compute the spectral gains, we choose the noise suppression rule proposed by Ephraim and Malah [4] (which we will call the EMSR method hereafter), because of its superiority in reducing musical noise phenomena after enhancement.

The EMSR method leads us to

$$\hat{S}_k(t) = G(\eta_k(t), \gamma_k(t)) X_k(t) \quad (18)$$

in which $\eta_k(t)$ is referred to as the *a priori* signal-to-noise-ratio (SNR) while $\gamma_k(t)$ is the *a posteriori* SNR described in [1]. The gain function $G(\cdot, \cdot)$ in (18) is given by

$$G(\eta, \gamma) = \frac{\sqrt{\pi}}{2} \sqrt{\frac{\eta}{\gamma(1+\eta)}} \times M \left[\frac{\gamma\eta}{1+\eta} \right] \quad (19)$$

where

$$M[\theta] = \exp\left(-\frac{\theta}{2}\right) \left[(1+\theta)I_0\left(\frac{\theta}{2}\right) + \theta I_1\left(\frac{\theta}{2}\right) \right] \quad (20)$$

with I_0 and I_1 being the modified Bessel functions of zero and first order, respectively. It is well-known that the decision directed approach for estimating the *a priori* SNR makes it possible for the EMSR method to reduce the musical noise artifacts. If $\hat{\gamma}_k(t)$ and $\hat{\eta}_k(t)$ are the estimates for $\gamma_k(t)$ and $\eta_k(t)$, then with the help of the decision directed approach we have

$$\begin{aligned} \hat{\gamma}_k(t) &= \frac{|X_k(t)|^2}{\hat{\lambda}_{n,k}(t)} \\ \hat{\eta}_k(t) &= \alpha \frac{|\hat{S}_k(t-1)|^2}{\hat{\lambda}_{n,k}(t-1)} + (1-\alpha)P[\hat{\gamma}_k(t)-1] \end{aligned} \quad (21)$$

where $\alpha \in [0, 1]$ and $P[x] = x$ if $x \geq 0$, and $P[x] = 0$ otherwise.

$\hat{S}_k(t)$ shown in (18) represents an estimate for the short-time spectral amplitude conditioned on the observed noisy spectrum, and it is considered desirable to modify the suppression rule if we are provided with the knowledge on speech absence or presence. Based on the soft decision scheme, we can derive

$$\begin{aligned} \hat{S}_k(t) &= E[S_k(t)|\mathbf{X}(t)] \\ &= E[S_k(t)|\mathbf{X}(t), H_0]p_L(H_0|\mathbf{X}(t)) + E[S_k(t)|\mathbf{X}(t), H_1]p_L(H_1|\mathbf{X}(t)) \\ &= E[S_k(t)|\mathbf{X}(t), H_1]p_L(H_1|\mathbf{X}(t)) . \end{aligned} \quad (22)$$

Table I. PESQ results for the proposed enhancement algorithm (CL-SEGSD) and conventional SEGSD technique

noise	white			vehicular			babble		
SNR (dB)	5	10	15	5	10	15	5	10	15
none	1.50	1.80	2.61	2.37	2.50	2.64	1.81	2.03	2.21
SEGSD	1.71	1.85	2.81	2.60	2.65	2.81	1.97	2.22	2.36
CL-SEGSD	1.81	1.90	2.87	2.68	2.71	2.84	2.09	2.30	2.44

From the result of (22), one of the most natural ways to modify the suppression rule is to replace $G(\cdot, \cdot)$ in the right hand side of (19) by $\tilde{G}(\cdot, \cdot)$ where

$$\tilde{G}(\eta_k(t), \gamma_k(t)) = p_L(H_1|\mathbf{X}(t))G(\eta_k(t), \gamma_k(t)) . \quad (23)$$

5 Experimental Results

In order to evaluate the performance of the proposed enhancement algorithm, which we denote by complex Laplacian pdf-based speech enhancement by global soft decision (CL-SEGSD), we conducted objective quality test experiments under various noisy conditions. Ten test sentences in which five were spoken by a male speaker and the others were generated by a female speaker were used for evaluation. Three types of noise sources, the white, babble and vehicular noises from the NOISEX-92 database were added to the clean speech waveforms by varying SNR [7]. Objective scores were decided by the PESQ results of the ITU-T P.862 [8]. The PESQ scores are shown in Table I. From the PESQ results, we can see that in most noisy conditions the proposed method yielded higher scores than the previous our work which is the SEGSD enhancement algorithm [1].

Acknowledgments

This work was supported in part by IITA through IT Leading R&D Support Project.