ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 2012. No. 6(122)
ELEKTRONIKA IR ELEKTROTECHNIKA

SYSTEM ENGINEERING, COMPUTER TECHNOLOGY
T 120
SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS

# A Novel Magnitude-Squared Spectrum Cost Function for Speech Enhancement

**Huan Zhao, Zhiqiang Lu**
*School of Information Science and Engineering, Hunan University,*
*Lushan South Rood, Changsha 410082, P. R. China, phone: +86 731 88821563, e-mail: hzhao@hun.edu.cn*
**Fei Yu**
*Jiangsu Provincial Key Lab of Image Processing & Image Communications Nanjing University of Posts and Telecommunicationsy,*
*Nanjing 310002, P. R. China, phone: +86 18229764458, e-mail: hunanyufei@126.com*
**Cheng Xu**
*School of Information Science and Engineering, Hunan University,*
*Lushan South Rood, Changsha 410082, P. R. China,*
*Jiangsu Provincial Key Lab of Image Processing & Image Communications Nanjing University of Posts and Telecommunicationsy,*
*Nanjing 310002, P. R. China*

crossref http://dx.doi.org/10.5755/j01.eee.122.6.1812

## Introduction

The problem of improving the quality and intelligibility of speech in noisy environments has attracted a great deal of interest in a long time. The existence of noise is inevitable in real-world application of speech processing. In particular, speech coders and speech recognition systems might be rendered useless in the presence of background noise.

Numerous techniques have been developed, and conventional speech enhancement algorithms basically consist of four classes of algorithms, including spectral subtraction [1], subspace [2], statistical model based [3] and Wiener filter based algorithms [4]. The well-known Ephraim-Malah algorithm which base on statistical model is an MMSE [3] estimator for the speech DFT amplitude. In this study, we also choose the Byes risk as the basis since it is the most fundamental statistical model approach, and many algorithms are closely connected to this technique [5]. Minimizing the Byes risk for a given cost function results in a variety of estimators. In fact, the maximum a posteriori (MAP) [6] estimator, minimum mean square error (MMSE) and maximum likelihood (ML) [7] estimators can be derived from the different Bayes risk cost functions. Also it is not difficult to notice that the Bayesian estimators based on perceptually motivated cost functions in place of traditional cost function are tightly related to the Byes risk [8-10]. In summary, different Bayesian estimators can be derived depending on the choice of the cost function. In recent years, Yang Lu and Loizou et al [11] propose a new speech enhancement algorithm which assumes that the magnitude-square spectrum of the noisy speech signal can be computed as the sum of the (clean) signal and noise magnitude-squared spectra, and finally, they derive a MMSE-MSS estimator which uses the squared-error cost function. Motivated by the previously mentioned assumption, we derive a novel speech enhancement by using other distortion measure in this paper. Results show that the proposed estimator yielded lower residual noise and lower speech distortion than the conventional MMSE-MSS estimator, in terms of yielding better speech quality.

This paper is organized as follows. In section 2, the background information of Bayes risk is given. In section 3, the proposed algorithm is presented. The experimental results of comparing the algorithm proposed in this paper with other algorithms are also presented in section 4. Finally, our work of this paper is summarized in the last section.

## Bayes risk

Supposing the observed noisy speech $y(n)=x(n)+d(n)$, is assumed to be clean speech signals $x(n)$ perturbed by statistically independent additive noise signals $d(n)$. The short-time Fourier transform of $y(n)$ can be expressed as $Y(\omega_k)=X(\omega_k)+D(\omega_k)$. For $\omega_k=2\pi k/N$ and $k=0,1,2,\ldots,N-1$, the equation is equivalent to the polar form as

$$Y_k \exp(j\theta_y(k)) = X_k \exp(j\theta_x(k)) + D_k \exp(j\theta_d(k)), \quad (1)$$

where $\theta_y(k)$, $\theta_x(k)$ and $\theta_d(k)$ denote the phases and $Y_k$, $X_k$, $D_k$ are spectral magnitudes at frequency bin $k$ of the noisy speech, clean signal, and noise respectively. The short-time spectral magnitude estimation of $X_k$ can be expressed in a form of $\hat{X}_k = G(k) \cdot Y_k$, where the gain function $G(k) = G(\xi_k, \gamma_k) = \hat{X}_k / Y_k$ are interpreted as the a priori and a posteriori SNRs, respectively. $\sigma_x^2(k) \equiv E\{X_k^2\}$ and $\sigma_d^2(k) \equiv E\{D_k^2\}$ denote clean speech and noise speech variances, respectively.

It is noted that when we measure the speech quality, the spectral magnitude is more important than its phase. So we focus on the estimation of the spectral magnitude, $X_k$ from the noisy spectral magnitude $Y_k$. Let $d(\varepsilon) = d(X_k, \hat{X}_k)$ represents a nonnegative function of $\varepsilon$. Where $\varepsilon$ denotes the error in estimating the magnitude $X_k$ at frequency bin $k$. The well-known Bayes risk can be given by the following

$$R_B = E\left\{d(X_k, \hat{X}_k)\right\} =$$
$$= \iint d(X_k, \hat{X}_k) p(X_k, Y(\omega_k)) \mathrm{d}X_k \mathrm{d}Y(\omega_k) =$$
$$= \int [\int d(X_k, \hat{X}_k) p(X_k \mid Y(\omega_k)) \mathrm{d}X_k] p(Y(\omega_k)) \mathrm{d}Y(\omega_k), \quad (2)$$

where $E(\cdot)$, $p(\cdot)$, $p(\cdot|\cdot)$ denote expectation function, probability density function and conditional probability density function, respectively. It is of great interest in using different distortion measures to derive the variety of estimators.

**Proposed algorithm**

Minimizing $R_B$ is just minimizing the following

$$R(X_k \mid Y(\omega_k)) = \int_0^\infty d(X_k, \hat{X}_k) p(X_k \left| Y(\omega_k)\right.) \mathrm{d}X_k, \quad (3)$$

where $R(X_k \left| Y(\omega_k)\right.)$ is corresponding to the inner integral in (2), we refer to this equation as conditional average cost function. A variety of traditional cost functions have been developed, when we substitute the squared-error cost function $d(X_k, \hat{X}_k) = (X_k - \hat{X}_k)^2$, into (3), and take the derivative $R$ with respect to $X_k$ and set it equal to zero, then we can get the well-known MMSE estimator [3]. while we can get the MAP estimator [6] with the given function

$$d(X_k, \hat{X}_k) = \begin{cases} 1, & \left|X_k - \hat{X}_k\right| \geq \Delta/2, \\ 0, & \left|X_k - \hat{X}_k\right| < \Delta/2, \end{cases} \quad (4)$$

where $\Delta$ denotes minimum and positive parameter. It should be noticed that the ML estimator is a special case of the MAP estimator, and it assumes that the density of $X_k$ obey uniform distribution.

*1) Minimum Mean Squared-Error Estimator.* Generally, the above analysis is in contrary to magnitude spectrum but not to magnitude-squared spectrum. However, some special cost functions also can be appropriate for magnitude-squared spectrum. To do that, we must derive the corresponding conditional average cost function. In order to get this, we take the $X_k^2$ and $Y_k^2$ as a whole, respectively. And then replace (3) as following

$$R(X_k^2 \mid Y_k^2) = \int_0^{Y_k^2} d(X_k^2, \hat{X}_k^2) p(X_k^2 \left| Y_k^2\right.) \mathrm{d}X_k^2, \quad (5)$$

where $X_k^2 \in [0, Y_k^2]$. Depending on the above equation, we can derive the MMSE-MSS estimator. Yang Lu, et al [11] propose a new solution with this distortion measure $d(X_k^2, \hat{X}_k^2) = (X_k^2 - \hat{X}_k^2)$. Finally, we can obtain

$$G_{\text{MMSE-MSS}}(\xi_k, \gamma_k) = \begin{cases} \left(\frac{1}{\upsilon_k} - \frac{1}{\exp(\upsilon_k) - 1}\right)^{1/2}, & \sigma_x^2 \neq \sigma_d^2, \\ (0.5)^{1/2}, & \sigma_x^2 = \sigma_d^2. \end{cases} \quad (6)$$

Additionally, where $\upsilon$ is defined as $\upsilon_k \equiv (1 - \xi_k)\, \gamma_k / \xi_k$.

*2) Conditional Median Estimator.* In this section, we investigate other cost functions which rely on the magnitude-squared spectrum assumption. The distortion measure is defined as magnitude-squared absolute error function $d(\varepsilon) = d(X_k^2, \hat{X}_k^2) = \left|X_k^2 - \hat{X}_k^2\right|$. Taking the magnitude-squared absolute error function into the function of $R(X_k^2 \left| Y_k^2\right.)$ with respect to $X_k^2$ and setting it to zero, it can be shown as

$$\int_0^{\hat{X}_k^2} p(X_k^2 \left| Y_k^2\right.) \mathrm{d}X_k^2 = \int_{\hat{X}_k^2}^{Y_k^2} p(X_k^2 \left| Y_k^2\right.) \mathrm{d}X_k^2. \quad (7)$$

The above equation is defined as conditional median, and utilizing the conditional median to estimate $X_k^2$, owing to:

$$p\left(X_k^2 \left| Y_k^2\right.\right) = \begin{cases} \psi_k \exp\left[-\frac{X_k^2}{\lambda(k)}\right], & \sigma_x^2 \neq \sigma_d^2, \\ 1/Y_k^2, & \sigma_x^2 = \sigma_d^2, \end{cases} \quad (8)$$

$$\psi_k \equiv \frac{1}{\lambda(k)\left\{1 - \exp\left[-\frac{Y_k^2}{\lambda(k)}\right]\right\}}, \quad (9)$$

where $\psi_k$ is a positive parameter. Substituting (8), (9), into (7), and using $1/\lambda(k) = 1/\sigma_x^2 - 1/\sigma_d^2$, then yielding

$$\hat{X}_k^2 = \begin{cases} -\lambda(k) \ln \frac{1}{2}\left[1 + \exp(-\frac{Y_k^2}{\lambda(k)})\right], & \sigma_x^2 \neq \sigma_d^2, \\ Y_k^2/2, & \sigma_x^2 = \sigma_d^2. \end{cases} \quad (10)$$

The simplification of the above equation using $\lambda(k) = Y_k^2 / \upsilon_k$, we can get the conditional median estimator of MMS (CM-MMS)

$$G_{\text{CM-MSS}}(\xi_k, \gamma_k) = \begin{cases} \frac{1}{\upsilon_k}\left|\ln 2 - \ln(1 + \exp(-\upsilon_k))\right|^{1/2}, & \sigma_x^2 \neq \sigma_d^2, \\ (0.5)^{1/2}, & \sigma_x^2 = \sigma_d^2. \end{cases} \quad (11)$$

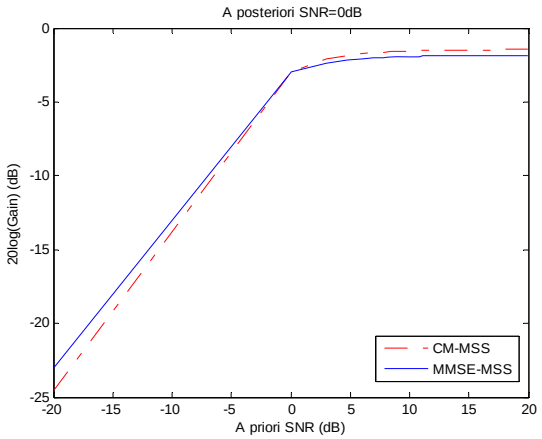


**Fig. 1.** Gain function of the proposed estimator and MMSE-MSS estimator

It is clearly that the MMSE-MMS estimator and CM-MSS estimator are dependent on both the $\xi_k$, $\gamma_k$ values. Fig. 1 plots the gain functions of the two estimators as a function of $\xi_k$ (for fixed values of $\gamma_k$). As shown from the figure, the CM-MSS estimator provides more suppression then the MMSE-MMS estimator for low a priori SNRs. For this reason, we expect that the CM-MSS estimator can perform better than the MMSE-MMS estimator.

In order to carry out the comparison between the estimators, we need to know the a priori SNR $\xi_k$. Thus, we use the"decision-directed" [3] approach:

$$\xi_k(l) = \alpha \frac{\hat{X}_k^2(l-1)}{\hat{\sigma}_d^2(k,l-1)} + (1-\alpha)\max[\gamma_k(l)-1,0], \quad (12)$$

$$\xi_k(l) = \max(\xi_k(l), \xi_{\min}), \quad (13)$$

where $l$ denotes the frame index and $\alpha$ denotes tunable coefficient. $\hat{\sigma}_d^2(k,l)$ denotes the estimate of the noise variance. Particularly, $\xi_{\min}$=-20dB.

**Table 1.** Performance comparison, in terms of SNRseg, between the various estimations

| Noise | Method | 0dB | 5dB | 10dB | 15dB |
|---|---|---|---|---|---|
| Car | MMSE-MSS | -1.851 | 0.602 | 3.393 | 6.414 |
| | CM-MSS | -1.219 | 1.124 | 3.759 | 6.639 |
| Street | MMSE-MSS | -1.918 | 0.493 | 3.587 | 6.347 |
| | CM-MSS | -1.541 | 0.764 | 3.755 | 6.375 |
| Babble | MMSE-MSS | -2.647 | 0.054 | 3.024 | 6.144 |
| | CM-MSS | -2.288 | 0.314 | 3.210 | 6.259 |
| White | MMSE-MSS | -1.214 | 1.383 | 4.231 | 7.190 |
| | CM-MSS | -0.344 | 2.102 | 4.766 | 7.485 |

**Table 2.** Performance comparison, in terms of PESQ, between the various estimations

| Noise | Method | 0dB | 5dB | 10dB | 15dB |
|---|---|---|---|---|---|
| Car | MMSE-MSS | 1.875 | 2.181 | 2.563 | 2.890 |
| | CM-MSS | 1.906 | 2.248 | 2.624 | 2.933 |
| Street | MMSE-MSS | 1.821 | 2.165 | 2.514 | 2.822 |
| | CM-MSS | 1.845 | 2.200 | 2.541 | 2.847 |
| Babble | MMSE-MSS | 1.837 | 2.186 | 2.531 | 2.904 |
| | CM-MSS | 1.838 | 2.190 | 2.548 | 2.923 |
| White | MMSE-MSS | 1.847 | 2.208 | 2.566 | 2.896 |
| | CM-MSS | 1.938 | 2.337 | 2.672 | 2.974 |

## Experiments and results

To evaluate the performance of the proposed and derived estimators, a total of 30 sentences taken from the publicly-available NOIZEUS database were used. The sentences were corrupted by car, babble, white and street noise at 0, 5, 10, and15 dB SNRs. Speech was segmented into 20 ms frames and han-windowed with 50% overlap. The overlap-add method was used to obtain the enhanced signal. The estimation of the noisespectra was using the algorithm of minimum controlled recursive average (MCRA) [12]. In (12), the value of $\alpha$ was set to 0.97. In order to assess the performance, two objective measurements, namely, average segmental signal to noise radio(SNRseg) and Perceptual Evaluation of Speech Quality (PESQ) [13], were utilized. The PESQ measure which has been found to yield a high correlation with the speech quality [14], is the best measure for overall speech quality prediction both of the speech quality and noise distortion. Higher PESQ values indicate better performance, i.e., better speech quality. In terms of background noise distortion, SNRseg is the best measure. Like the PESQ, higher SNRseg values indicate that the enhanced signal is more similar to clean speech. The SNRseg measure is defined by

$$\text{SNRseg} = \frac{10}{M}\sum_{l=0}^{M-1}\log_{10}\frac{\sum_{k=0}^{L-1}x_k^2}{\sum_{k=0}^{L-1}[\hat{x}_k(l)-x_k(l)]^2}. \quad (14)$$

In the above equation, in which $x_k$ , $\hat{x}_k$ denote clean speech and estimated speech, respectively, here, $M$ and $L$ denote total number of frames and the length of frames, respectively.

Table 1 and Table 2 show the performance comparison in terms of SNRseg and PESQ between the various estimations. In terms of SNRseg, which is easy to implement and is better correlated with Mean opinion score (MOS) than SNR, it has been widely used to qualify the enhanced speech. It is not difficult to see from the

Table 1, as for the four types of noise conditions at all SNR levels, the CM-MSS estimator yields significantly higher SNRseg values than the MMSE-MSS estimator.

PESQ is more reliable and correlated better with MOS than the traditional measures in most situations. In terms of PESQ, the overall results were shown in Table 2**.** As well as the performance of SNRseg, under different types of noise conditions at all SNR levels, the CM-MSS estimator yielded significantly higher PESQ scores than the MMSE-MSS estimator, either. In summary, the proposed estimator offers better speech quality and lower speech distortion than the MMSE-MSS estimator.

## Conclusions

In this paper, we report several existing Bayesian short-time spectral amplitude cost functions for speech enhancement. There are no prior studies in the condtional median, therefore,we propsoe a new MMS estimator where the distortion measure is the absolute error function. The derived estimator, which markedly reduces the background noise without introducing speech distortion, it is superior to the MMSE-MSS estimator in terms of both SNRseg and PESQ. Our future work is to calculate the condtional median estimato of the magnitude spectrum.

## Acknowledgements

This work was supported by National Science Foundation of China (Grant No. 61173106), the Key Program of Hunan Provincial Natural Science Foundation of China (Grant No.10JJ2046), and the Planned Science and Technology Key Project of Hunan Province, China (Grant No.2010GK2002).

## References

1. **Boll S. F.** Suppression of Acoustic Noise in Speech Using Spectral Subtraction // IEEE Trans. Acoust., Speech, Signal Processing, 1979. – Vol. 27. – No. 2. – P. 113–120.
2. **Hu Y., Loizou P. C.** A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise // IEEE Trans Speech and Audio Processing, 2003. – Vol. 11. – No. 4. – P. 334–341.
3. **Ephraim Y., Malah D.** Speech Enhancement Using a Minimum Mean-Square Error Short-time Spectral Amplitude estimator // IEEE Trans. Acoust., Speech, Signal Processing, 1984. – Vol. 32. – No. 6. – P. 1109–1121.
4. **Hu Y., Loizou P. C.** Speech Enhancement Based on Wavelet Thresholding the Multitaper Spectrum // IEEE Transactions on Speech and Audio Processing, 2004. – Vol. 12. – No. 1. – P. 59–67.
5. **Wolfe P. J., Godsill S. J.** Efficient Slternatives to Ephraim and Malah Suppression Rule for Audio Signal Enhancement // Journal of EURASIP Journal on Applied Signal Processing, 2003. – Vol. 2003. – No. 10. – P. 1043–1051.
6. **Kay S.** Fundamentals of Statistical Signal Processing: Estimation Theory . – Upper Saddle River, NJ: Prentice-Hall, 1993. – 303p.
7. **McAulay R. J., Malpass M. L.** Speech Enhancement Using a Soft-decision Noise Suppression Filter // IEEE Trans. Acoustics Speech, and Signal Processing, 1980. – Vol. 28. – No. 2. – P. 137–145.
8. **Loizou P. C.** Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum // IEEE Transactions on Audio, Speech, and Signal Processing, 2005. – Vol. 13. – No. 5. – P. 857–869.
9. **Plourde E., Champagne B.** Generalized Bayesian Estimators of the Spectral Amplitude for Speech Enhancement // IEEE Signal procession letters, 2009. – Vol. 16. – No. 6. – P. 485–488.
10. **Nguyen A. D., Naoe K., Takefuji Y.** A New Log-Spectral Amplitude Estimator Using the Weighted Euclidean Distortion Measure For Speech Enhancement // 2010 IEEE 26-th Convention of Electrical and Electronics Engineers. – Israel, 2010. – P. 675–679.
11. **Yang Lu, Loizou P. C.** Estimators of the magnitude-squared Apectrum and Methods for Incorporating SNR Uncertainty // IEEE Transactions on Audio, Speech, and Signal Processing, 2011. – Vol. 19. – No. 5. – P. 1123–1136.
12. **Cohen I., Berdugo B.** Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement // IEEE Signal Process. Lett., 2002. – Vol. 9. – No. 1. – P. 12–15.
13. **ITU-T Rec. P.862.** Perceptual Evaluation of Speech Quality (PESQ), and Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, 2000.
14. **Hu Y., Loizou P. C.** Evaluation of Objective Quality Measures for Speech Enhancement // IEEE Transactions on Audio, Speech, Languge Process, 2008. – Vol. 16. – No. 1. – P. 229–238.

Received 2011 04 04
Accepted after revision 2012 04 27

**Huan Zhao, Zhiqiang Lu, Fei Yu, Cheng Xu. A Novel Magnitude-Squared Spectrum Cost Function for Speech Enhancement // Electronics and Electrical Engineering. – Kaunas: Technologija, 2012. – No. 6(122). – P. 11–14.**

In Bayesian approaches for speech enhancement, the enhanced speech is estimated by minimizing the Bayes risk. In detail, an estimate of the clean speech is derived by minimizing the expectation of a cost function. Various estimators have been derived by the classic cost function, squared-error cost function, and "hit-or-miss" function. However, absolute error function was paid less attention. In this paper, we consider a magnitude-squared spectrum (MSS) motivated estimator for speech enhancement based on statistics and Bayesian cost function in the frequency domain. Specifically, we derive a novel estimator of which the cost function is the absolute error distortion measure of the MSS. By studying experimental results with NOIZEUS database, we find that the performance of the proposed scheme can achieve a significant noise reduction and a better speech quality as compared to minimum mean-squared error (MMSE) estimator of the MSS. Ill. 1, bibl. 14, tabl. 2 (in English; abstracts in English and Lithuanian).

**Huan Zhao, Zhiqiang Lu, Fei Yu, Cheng Xu. Nauja kvadratūrinio spektro nustatymo funkcija kalbai gerinti // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2012. – Nr. 6(122). – P. 11–14.**

Kuriant Bajeso kalbos gerinimo metodus, pagerinta kalba yra apskaičiuota minimizuojant Bajeso riziką. Įvairūs įvertiniai buvo išvesti iš klasikinės kaštų funkcijos, kvadratinės paklaidos kaštų funkcijos ir iš atsitiktinio pasirinkimo funkcijos. Tačiau į absoliutinę klaidos funkciją nebuvo kreipiama dėmesio. Pateikiamas kvadratūriniu spektru pagrįstas kalbos gerinimo įvertinys, kuris remiasi statistine ir Bajeso funkcija dažnių srityje. Analizuojant eksperimentinius rezultatus su NOIZEUS duomenų baze, buvo nustatyta, kad siūlomoji schema gali gerokai sumažinti triukšmą ir užtikrinti geresnę kalbos kokybę, palyginti su vidutinės kvadratinės klaidos įvertiniu. Il. 1, bibl. 14, lent. 2 (anglų kalba; santraukos anglų ir lietuvių k.).