MUSICAL NOISE GENERATION ANALYSIS FOR NOISE REDUCTION METHODS BASED ON SPECTRAL SUBTRACTION AND MMSE STSA ESTIMATION

†Yoshihisa Uemura, †Yu Takahashi,†Hiroshi Saruwatari, †Kiyohiro Shikano, and ‡Kazunobu Kondo

†Nara Institute of Science and Technology, Nara, 630-0192, JAPAN
‡SA Group, Center for Advanced Sound Technologies, Yamaha Corp., Shizuoka, 438-0192, JAPAN

ABSTRACT

In this paper, we reveal new findings about the generated musical noise in minimum mean-square error short-time spectral amplitude (MMSE STSA) processing. Recently we have proposed a objective metric of musical noise based on kurtosis change ratio on spectral subtraction (SS). Also we found an interesting relationship among the degree of generated musical noise, the shapes of signal's probability density function, the strength parameter of SS processing. This paper is aimed to automatically evaluate the sound quality of various types of noise reduction methods using kurtosis change ratio. We give a mathematical analysis based on higher-order statistics viewpoint, and lead to a valuable relation in that MMSE STSA has a weakness in speech period distortion rather than noise period, and vice versa in SS.

Index Terms— Musical noise, higher-order statistics, minimum mean-square error short-time spectral amplitude, spectral subtraction, speech enhancement

1. INTRODUCTION

Nonlinear processing, e.g., spectral subtraction (SS) [1] and minimum mean-square error short-time spectral amplitude (MMSE STSA) [2], often generates particular distortion, the so-called *musical noise*. It is one of the critical problems inherent in nonlinear processing because musical noise is perceived as harsh and artificial tone. Thus a lot of countermeasures to handle the musical noise have been proposed. However musical noise can not be evaluated by traditional metric about sound quality, e.g., cepstrum distance, and we just had to subjectively assess the sound quality [3]. To begin with, we did not know so much thing and theory about subjective evaluation of musical noise. Thus we can not evaluate the actual performance of each of countermeasures against musical noise.

We have proposed a novel mathematical metric of musical noise [4]. The metric based on change of kurtosis, 4th-order statistics, through nonlinear processing, has high correlation with the amount of perceived musical noise by human. Also we found that the degree of generated musical noise in processing is strongly related with kurtosis ratio. Therefore now we can objectively evaluate the degree of generated musical noise in nonlinear processing with our kurtosis ratio.

Recently it is widely accepted for speech-enhancement studies that MMSE STSA sets up less musical noise than SS, and we can obtain the high-quality (less degraded) output signal. However no one confirmed it from theoretical and analytical aspects because it is so difficult and unrealistic that both of MMSE STSA and SS are compared via subjective evaluation in every parameters of itself. In this paper, we realize the comparison using kurtosis ratio, and consequently we reveal a new findings about the degree of generated musical noise in MMSE STSA vs. SS.

This work was partly supported by MIC Strategic Information and Communications R&D Promotion Programme in Japan.

2. OVERVIEW OF NONLINEAR PROCESSING

2.1. SS

At first, we introduce two kinds of representative nonlinear processing, i.e., SS and MMSE STSA. Although various types of SS methods are proposed, we address single-channel SS in the power domain, which is used for any speech enhancement [5].

Let the corrupted speech signal $o(t)$ be represented as

$$o(t) = s(t) + d(t), \tag{1}$$

where $s(t)$ is a clean speech signal and $d(t)$ is a noise signal. This processing is conducted on a frame-by-frame basis. The short-time Fourier transform (STFT) is used and the previous model can be rewritten as

$$O(k,m) = S(k,m) + D(k,m), \tag{2}$$

where k denotes the frequency subband and m is the frame index. In SS, noise reduction is achieved by subtracting the power spectrum of the estimated noise from the power spectrum of the noisy observation. This procedure is given by

$$Y(k,m) = \sqrt{|O(k,m)|^2 - \beta \cdot \mathrm{E}_m\left[|D(k,m)|^2\right]} \cdot e^{j \arg(O(k,m))}, \tag{3}$$

where $Y(k,m)$ is an estimated speech signal, β is an over-subtraction coefficient (i.e., strength parameter) and $\mathrm{E}[\cdot]$ is an expectation operator of $\cdot$ with respect to m.

2.2. MMSE STSA estimator

MMSE STSA is a method for estimating the clean speech spectral amplitude from corrupted speech signal by minimization of mean-square error. It is supposed that the statistical model of noise is Gaussian model, and the model is statistically independent and has zero mean. The given spectral gain by MMSE is written by

$$G(k) = \Gamma(1.5)\frac{\sqrt{(v_k)}}{\gamma_k}\exp\left(-\frac{v_k}{2}\right)\cdot\left[(1+v_k)I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right], \tag{4}$$

where $\Gamma(\cdot)$ denotes the gamma function. $I_0$and I_1denotes the modified Bessel functions of zero and first order, respectively.

Also functions in the equation are defined by

$$v_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k. \tag{5}$$

Here ξ_k and γ_k are defined by

$$\xi_k \triangleq \frac{\lambda_s(k)}{\lambda_d(k)}, \tag{6}$$

$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)}, \tag{7}$$

where $\lambda_s(k) \triangleq \mathrm{E}[|S_k|^2]$ and $\lambda_d(k) \triangleq \mathrm{E}[|D_k|^2]$. ξ_k and γ_k are called *a*

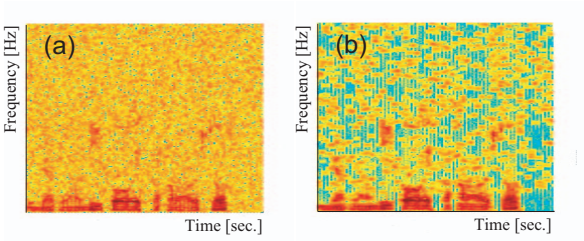

Fig. 1. (a) Observed signal spectrogram. (b) Processed signal spectrogram.

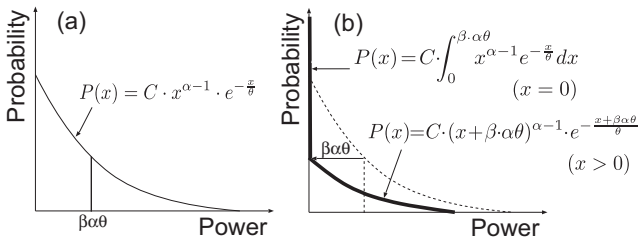

Fig. 2. Shapes of p.d.f. (a) Original signal. (b) Processed signal.

priori and *a posteriori* signal-to-noise ratios (SNR), respectively.

MMSE STSA can estimate the clean speech spectral amplitude as above, ideally. However, in actual case, we can not know *a priori* and *a posterior* SNR, and thus we estimate them by the following equation,

$$\hat{\xi}_k(m) = \eta \frac{\hat{A}_k^2(m-1)}{\lambda_d(k, m-1)} + (1-\eta) P[\gamma_k(m) - 1], \qquad 0 < \eta < 1, \quad (8)$$

where $\hat{A}_k(m-1)$ is the amplitude estimator of the kth signal spectral component in the $(m-1)$th analysis frame, and $P[\cdot]$ is an operator which is defined by

$$P[x] = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Consequently MMSE STSA is managed by estimating the mth frame spectral gain using previous frame spectral gain.

3. MUSICAL NOISE METRIC VIA KURTOSIS RATIO

In this section, we introduce a basic idea of musical noise evaluation and musical noise metric for SS [4]. Hereinafter we give an explanation of the adequacy applying the metric to general nonlinear processing.

3.1. Relationship between kurtosis ratio and musical noise

Nonlinear processing, including SS and MMSE STSA, often generates characteristic isolated power spectral components (see Fig. 1 (a) and (b)). We define the musical noise as the generated audible isolated spectral components through nonlinear processing. Thus we speculate that the amount of musical noise is highly related to the number of isolated components and the isolated level of them. Consequently we realize the evaluation of the isolated components by kurtosis.

We could say that kurtosis can evaluate the percentage of tonal components in total components. Bigger value indicates a signal with heavy skirt in its probability density function (p.d.f.); it means that a signal has a lot of tonal components. Kurtosis is defined as

$$\text{kurt} = \frac{\mu_4}{\mu_2^2}, \quad (10)$$

where kurt denotes kurtosis and μ_n is the nth order moment which is given by

$$\mu_n = \int_0^\infty x^n P(x) dx. \quad (11)$$

Here, $P(x)$ is p.d.f. of the signal. We consider the SS in power spectral domain, so the integral range is only positive.

Although we can measure the number of the tonal components by kurtosis, note that kurtosis itself is not enough to measure the musical noise. This is obvious in that kurtosis of some unprocessed signals, e.g., speech signals, is also high, but we do not recognize speech as musical noise. In order to set aside the genuine tonal components, we focus on the fact that musical noise is generated only in artificial signal processing. Hence, we turn our attention to kurtosis change ratio (kurtosis ratio) between before/after signal processing.

3.2. Kurtosis ratio in SS

We derive the relationship between kurtosis and the strength of SS. Moreover, the relationship between kurtosis of processed signal and kurtosis of unprocessed signal are revealed.

3.2.1. Gamma distribution modeling

We utilize the gamma distribution as a model of speech or noise signal [6]. The gamma distribution have a lot of useful mathematically attributes which are derived from the gamma function.

The p.d.f. of the gamma distribution is written as

$$P(x) = \frac{1}{\Gamma(\alpha)\, \theta^{\alpha}} \cdot x^{\alpha-1}\, e^{-\frac{x}{\theta}}, \quad (12)$$

where $x \geq 0$, $\alpha > 0$ and $\theta > 0$. Also α denotes the shape parameter and θ is the scale parameter. The Gamma function is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \cdot dx. \quad (13)$$

Hereafter, in this paper, let $C = 1/[\Gamma(\alpha)\, \theta^{\alpha}]$. If $\alpha = 1$, this is the exponential distribution. It is well known that the average of the gamma distribution is given by

$$\mathrm{E}[P(x)] = \alpha\theta, \quad (14)$$

where E[·] is an expectation operator. The gamma distribution modeling is the estimation of the shape and the scale parameters from the raw input signal. In this paper, we use the maximum likelihood estimation method for estimating two parameters α and θ, as follows,

$$\hat{\alpha} = \frac{3 - \gamma + \sqrt{(\gamma-3)^2 + 24\gamma}}{12\gamma}, \quad (15)$$

$$\hat{\theta} = \frac{\mathrm{E}[x]}{\hat{\alpha}}, \quad (16)$$

where $\gamma = \log(\mathrm{E}[x]) - \mathrm{E}[\log x]$ (see Refs. [7]).

3.2.2. Kurtosis of modeling signal

In this way of modeling by the gamma distribution, kurtosis is determined by the shape and the scale parameters as below. At first, we represent the nth-order moment as

$$\mu_n = \int_0^\infty x^n P(x)\, dx = C \cdot \theta^{\alpha+n} \cdot \Gamma(\alpha + n). \quad (17)$$

Using (17) and useful relation, $\Gamma(\alpha) = (\alpha-1)\cdots(\alpha-j)\Gamma(\alpha-j)$, we can obtain kurtosis of the modeled raw signal by the gamma distribution as follows [4].

$$\text{kurt}_{\text{org}} = \frac{\mu_4}{\mu_2^2} = \frac{(\alpha+2)(\alpha+3)}{\alpha(\alpha+1)}. \quad (18)$$

3.2.3. Change of modeling signal's kurtosis in processing

SS is regarded as p.d.f. deforming processing, i.e., lateral shift of p.d.f. (See Fig. 2). Thus processed signal's kurtosis depends on the strength of processing. We formulated the deformed signal's kurtosis

Table 1. Expected situation of musical noise generation

	SS	MMSE STSA estimator
In speech period	Moderate	(Unclear)
In noise period	Awful	Moderate

Table 2. Subjective evaluation conditions

Database	noise: white Gaussian noise: speech: Japan News Article Sentences
Mixing	equivalent SNR mixing
SS	strength parameter: configure from 0 to 2.5 with 0.05 increments in between flooring: negative power components are replaced by zero
MMSE STSA	strength parameter: configure from 0.5 to 0.99 with 0.01 increments in between
Evaluation norm	SNR and kurtosis ratio of clean speech and clean noise signal

and change of kurtosis in SS [4]. The resultant p.d.f. of the processed signal is written as

$$P(x) = \begin{cases} C \cdot (x+\beta\cdot\alpha\theta)^{\alpha-1} e^{-\frac{x+\beta\cdot\alpha\theta}{\theta}} & (x>0), \\ C\int_0^{\beta\cdot\alpha\theta} x^{\alpha-1}e^{-\frac{x}{\theta}}dx & (x=0). \end{cases} \quad (19)$$

Here we approximate $(x+\beta\alpha\theta)^{\alpha-1}$ in (19) by Taylor expansion, and we have

$$\mu_4 \approx Ce^{-\alpha\beta}\Big[\int_0^\infty x^{(\alpha+4)-1}e^{-\frac{x}{\theta}}dx+\beta\alpha\theta(\alpha-1)\int_0^\infty x^{(\alpha+3)-1}e^{-\frac{x}{\theta}}dx + \frac{(\beta\alpha\theta)^2}{2}(\alpha-2)(\alpha-1)\int_0^\infty x^{(\alpha+2)-1}e^{-\frac{x}{\theta}}dx\Big]. \quad (20)$$

Also the 2nd-order moment is estimated as below,

$$\mu_2 = \int_0^\infty x^2\left[C(x+\beta\cdot\alpha\theta)^{\alpha-1}e^{-\frac{x+\beta\alpha\theta}{\theta}}\right]dx \le Ce^{-\alpha\beta}\theta^{\alpha+2}\Gamma(\alpha+2). \quad (21)$$

Thus processed signal's kurtosis is estimated as

$$\text{kurt}_{\text{ss}} \ge \frac{e^{\alpha\beta}}{\alpha(\alpha+1)}\left\{(\alpha+2)(\alpha+3)+\beta\alpha(\alpha+2)(\alpha-1)+\frac{(\beta\alpha)^2}{2}(\alpha-3)(\alpha-1)\right\}. \quad (22)$$

Here, α denotes the shape parameter of modeled gamma distribution, and β is the parameter of SS processing strength. Thus kurtosis ratio in SS can be given as

$$\text{kurtosis ratio} = \frac{\text{kurt}_{\text{ss}}}{\text{kurt}_{\text{org}}} = e^{\alpha\cdot\beta}\left\{1+\frac{\beta\alpha(\alpha-1)}{(\alpha+3)}+\frac{(\beta\alpha)^2(\alpha-2)(\alpha-1)}{2\,(\alpha+2)(\alpha+3)}\right\}. \quad (23)$$

Also we found that the musical noise metric based on kurtosis ratio is highly related with the amount of perceived musical noise by human [4].

Kurtosis ratio is strongly related with the generated isolated components in nonlinear processing. Thus we can evaluate the degree of generated musical noise according to magnitude of kurtosis-ratio value.

4. THEORETICAL ANALYSIS

As indicated in (23), we can now find a relationship between the amount of generated musical noise in SS and the original signal's kurtosis. That is, SS for high kurtosis signal (α is small) results in less musical noise than SS for low kurtosis signal (α is large) even if we set the fixed subtraction parameter β. Thus in this paper, we

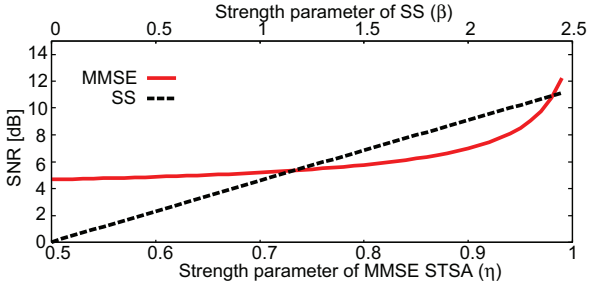

Fig. 3. Processed signal's SNR on SS and MMSE STSA.

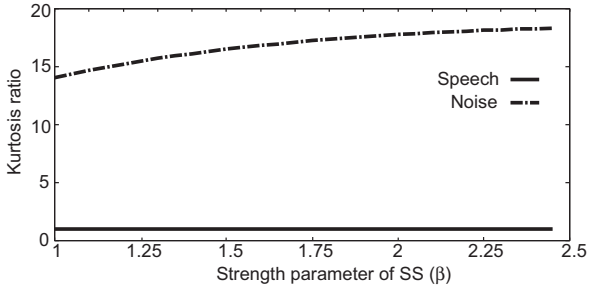

Fig. 4. Relationship between noise's and speech's kurtosis ratio and the strength parameter of SS (β).

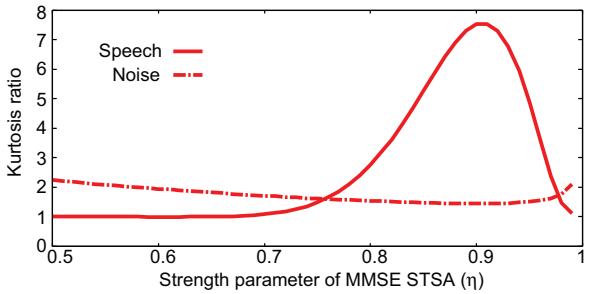

Fig. 5. Relationship between noise's and speech's kurtosis ratio and the strength parameter MMSE STSA (η).

apply the above-mentioned theory to speech-noise mixed signal, and obtain the following theoretical prediction about the generated musical noise when SS is applied to speech/noise-dominant intervals.

Speech signal has higher kurtosis than noise signal, in general. Therefore generated musical noise in speech-dominant intervals is relatively less. Conversely, a lot of musical noises generate in noise-only intervals because noise signal is low kurtosis signal, e.g., Gaussian noise.

On the other hand, we can not formulate generated musical noise amount in MMSE STSA estimator, but we can still speculate the degree of generated musical noise in only instance of noise intervals. Here, we suppose that noise signal is stationary and has low kurtosis. In this instance, the obtained spectral gain by MMSE STSA estimator is stationary and small value in noise intervals. Thus output signal's p.d.f. does not change so much from original one. Consequently, the amount of generated musical noise is less than the case of using SS. However, in speech intervals, we can not forecast because the estimation of spectral gain using *a prior* and *a posterior* SNR depends on previous frame and is extremely complicated.

Table 1 lists the summary of the points about musical noise generation. We will confirm the predictions and bring out the unclear points with experiment in the next section.

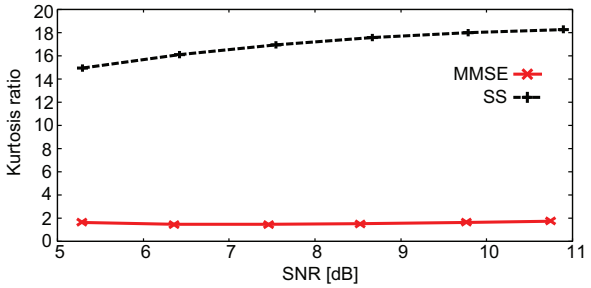

Fig. 6. Comparison of noise's kurtosis ratio on equivalent SNR.

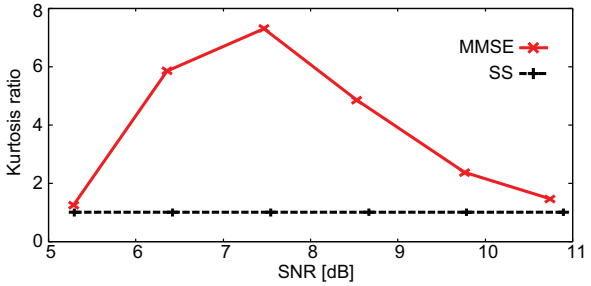

Fig. 7. Comparison of speech's kurtosis ratio on equivalent SNR.

5. EXPERIMENT

5.1. Conditions

We conduct an experiment and objective evaluation for musical noise on SS and MMSE STSA. One of our great interest is comparative merits and demerits of SS and MMSE STSA. Particularly, we are interested in the difference on the degree of generated musical noise between both method, and between both signal of speech or noise.

Conditions of experiment are listed in Table 2. The strength parameter of MMSE STSA is set to commonly-used value and the strength parameter of SS is controlled as to equal SNR performance of noise reduction.

5.2. Results

Figure 3 depicts processed signal's SNR on SS and MMSE STSA. As we can see from Fig. 3, both the upper limit performance of MMSE STSA and SS are about 11 dB. Figures 4 and 5 show the relationship between kurtosis ratio and the strength parameter of each method. In SS, as we expected, kurtosis ratio is smaller value in speech intervals than noise intervals. This is because kurtosis ratio in SS depends on unprocessed signal's original kurtosis.

On the other hand, in MMSE STSA, noise's kurtosis ratio is very small, but speech's kurtosis is very high and changes rapidly. It is particular note around commonly-used parameter. This phenomenon has already confirmed as the sensitive relationship between the strength parameter of MMSE STSA estimator and musical noise via the subjective evaluation [2]. Consequently, in SS, musical noise mainly arises in noise interval, on another front, in MMSE STSA, the problem of musical noise generation is mainly boiled up in speech interval.

We compare MMSE STSA with SS in terms of the degree of generated musical noise. Figures 6 and 7 present the kurtosis ratio on SS and MMSE STSA in condition of equivalent SNR. These results are very interesting. Figure 6 shows the familiar phenomenon of SS that the amount of generated musical noise in SS is very much and gradually increasing as bigger the strength parameter, on the other hand, it is less in MMSE STSA. Figure 7 presents the interesting result that MMSE STSA generates more musical noise in speech signal than SS. This is a new finding. We have believed the commonly accepted theory that MMSE STSA is superior to SS in terms of musical noise. It is still true in noise-only part, but, *in speech-dominant part, this is misconception*. This new finding is confirmed by spectrogram (Fig. 8). Spectrogram of SS processed signal shows the degradation of speech signal but we can not detect the isolated component. On the other hand, spectrogram of MMSE show the apparent isolated components. This is consistent with our subjective impressions. Consequently MMSE STSA generates the isolated components in speech signal and we have conscious access to musical noise.

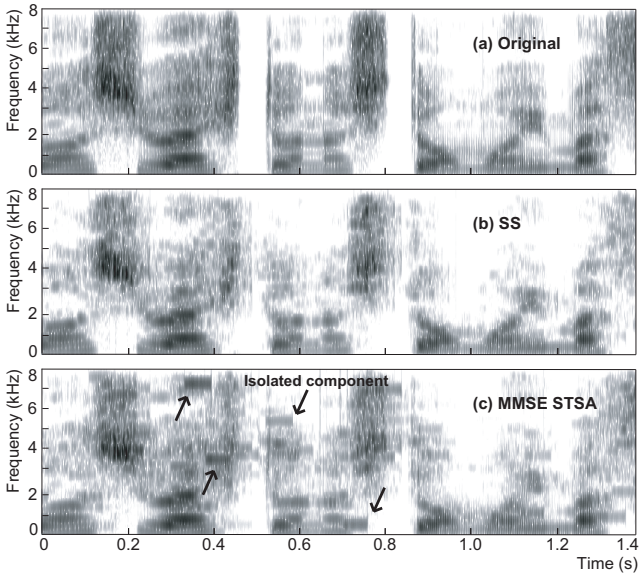

Fig. 8. Spectrogram (a) Original speech signal. (b) SS processed speech signal (SNR = 11 dB). (c) MMSE STSA processed speech signal (SNR = 11 dB).

6. CONCLUSION

We analyze the degree of generated musical noise in SS and MMSE STSA. Also we came up with the novel fact about how to characteristic generate musical noise in MMSE STSA.

7. REFERENCES

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

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol.32, no.6, pp.1109-1121, December 1984.

[3] M. Kato, et al., "Noise Suppression with High Speech Quality Based on Weighted Noise Estimation and MMSE STSA (Digital Signal Processing)" *IEICE Trans. Fundamentals*, vol.E85-A, no.7, July 2002.

[4] Y. Uemura, et al., "Automatic Optimization Scheme of Spectral Subtraction Based on Musical Noise Assessment via Higher-Order Statistics" *IWAENC*, 2008.

[5] J. Li, et al., "noise reduction based on adaptive β-order generalized spectral subtraction for speech enhancement," *INTERSPEECH*, pp,802–805, 2007.

[6] T. H. Dat, et al., "Gamma modeling of speech power and its on-line estimation for statistical speech enhancement," *IEICE Trans. INF & SYST.* vol.E89-D, no.3, 2006.

[7] M. Evans, et al., *Statistical Distributions, 2nd ed.* Wiley. 1993.