SPEECH KURTOSIS ESTIMATION FROM OBSERVED NOISY SIGNAL BASED ON GENERALIZED GAUSSIAN DISTRIBUTION PRIOR AND ADDITIVITY OF CUMULANTS

†Ryo Wakisaka, †Hiroshi Saruwatari, †Kiyohiro Shikano, and ‡Tomoya Takatani

†Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara, 630-0192 Japan
‡Toyota Motor Corporation, 543 Kirigabora Nishihirose-cho, Toyota, Aichi, 470-0309 Japan

ABSTRACT

In this paper, we propose a new method for stable estimation of the kurtosis of a speech power spectrum. Speech kurtosis can be used for the prediction of speech recognition accuracy as reported in recent studies. However, the conventional estimation method is very unstable owing to the high sensitivity of higher-order statistics. To overcome this problem, we introduce the generalized Gaussian distribution prior in order to avoid the calculation of higher-order statistics, and construct a kurtosis table that directly represents the relationship among the kurtosis of speech, noise, and their mixture in the power spectrum domain. Speech kurtosis can be estimated stably from observable data by looking up values in the table. An experimental evaluation confirms the efficacy of the proposed method.

***Index Terms*—** Speech kurtosis estimation, Generalized Gaussian distribution, Kurtosis table, Moment-cumulant transformation.

1. INTRODUCTION

In recent years, many applications of speech communication systems have been developed, resulting in real-world human interfaces. In such applications, the essential requirement is robustness against environmental noise. Therefore, many nonlinear noise reduction methods, such as spectral subtraction and MMSE short-time spectral amplitude estimation, have been actively studied [1].

Several metrics have been proposed as measures of evaluation for these methods, namely, the noise reduction rate (or SNR) [2], cepstral distortion [3], and kurtosis ratio [4], which correspond to the amounts of noise reduction, speech distortion, and musical noise generation, respectively. Since speech distortion affects speech recognition accuracy, a measure of speech distortion is particularly indispensable as an indicator for optimizing speech recognition systems. The calculation of cepstral distortion, which is a commonly used measure of speech distortion, requires a reference (clean) speech signal. However, in actual situations, the speech component is always overlapped with noise, and we cannot obtain a clean speech signal. Consequently, the optimal parameters in the noise reduction method cannot be estimated. To overcome this problem, as an unsupervised measure of speech distortion estimated in a reference-free manner, *the kurtosis of the speech power spectrum* has been proposed by the authors, which is effective for optimizing parameters in the noise reduction method and predicting speech recognition accuracy [5, 6].

The main problem in our previous method [5] is the low robustness in the estimation of higher-order statistics. In this method, it is necessary to calculate up to eighth-order statistics in the observed signal waveform domain. Since sixth- and eighth-order statistics are very sensitive to outliers, we cannot estimate them stably from observable finite samples, causing considerable degradation of the estimated kurtosis of the speech power spectrum. To solve this problem, in this paper we propose a new method using a statistical prior of waveform signals, where the waveform signals of speech and noise obey the generalized Gaussian distribution [7]. We can construct a *kurtosis table*, which represents the direct relationship among the kurtosis of speech, noise, and their mixture in the power spectrum domain, using the prior, the additivity of cumulants, and a moment-cumulant transformation. Then, the kurtosis of the speech power spectrum is estimated from observable signals without any references by looking up values in the table. We conduct an evaluation experiment and confirm that the accuracy of speech kurtosis estimation is markedly improved by the proposed method even if the snapshot data length is only 1 s.

This work was partly supported by JST Core Research of Evolutional Science and Technology (CREST), Japan.

2. PREVIOUS WORKS

2.1. Problem and strategy

In this section, we describe the conventional method of speech kurtosis estimation for evaluating *pure* distortion that arises only in the speech component. We consider an acoustic mixing model, where the observed signal consists of a target speech signal and an additive noise signal. Hereafter, the observed signal in the time-frequency domain, $x(f,\tau)$, is given by

$$x(f,\tau) = s(f,\tau) + n(f,\tau), \tag{1}$$

where f is the frequency bin number, τ is the time-frame index number, $s(f,\tau)$ is the target speech signal component, and $n(f,\tau)$ is the additive noise signal. Since the speech component is always contaminated with noise at every time-frequency grid, it is difficult to estimate the speech kurtosis via theoretical analysis. Therefore, we inversely calculate the kurtosis of the speech power spectrum in a data-driven manner, utilizing two observable statistics of the noisy speech signal $x(f,\tau)$ and noise signal $n(f,\tau)$ (we assume that the noise statistics can be measured in a speech-absent time period by voice-activity detection or BSS-based noise estimation [5]). Note that the proposed speech kurtosis estimation is an unsupervised method because it requires no reference (clean) speech signals, unlike cepstral distortion.

To cope with the mathematical problem that the mixing of speech and noise is additive but generally their higher-order moments are not additive, we introduce the *cumulant*, which retains the additivity for additive variables. Meanwhile, in the transformation from a waveform to its power spectrum, the exponentiation operation is conducted. However, the cumulant does not have a straightforward relationship. In this case, we use the moment instead of the cumulant. Thus, we previously proposed the use of a *moment-cumulant transformation* [5].

2.2. Moment-cumulant transformation

In this section, we give some formulas regarding the moment-cumulant transformation. They explicitly represent the relations

between the moment and cumulant in each order, which are useful for estimating the kurtosis of a speech power spectrum. The mth-order moment $\mu_m(y)$ can be written as

$$\mu_m(y) = \sum_{\pi(m)} \prod_{B\in\pi(m)} \kappa_{|B|}(y), \tag{2}$$

where $\pi(m)$ runs through the list of all partitions of a set of size m, $B \in \pi(m)$ means that B is one of the blocks into which the set is partitioned, and $|B|$ is the size of set B. In the same manner, the mth-order cumulant $\kappa_m(y)$ is given by

$$\kappa_m(y) = \sum_{\pi(m)} (-1)^{|\pi(m)|-1}(|\pi(m)|-1)! \prod_{B\in\pi(m)} \mu_{|B|}(y). \tag{3}$$

2.3. Estimation of speech kurtosis from observations [5]

Hereafter, to deal with time-frequency-domain signals, we define complex-valued variables of the observed (noisy speech) signal, the original speech signal, and the noise signal as $(x_{\mathrm{R}}+ix_{\mathrm{I}})$, $(s_{\mathrm{R}}+is_{\mathrm{I}})$, and $(n_{\mathrm{R}}+in_{\mathrm{I}})$, respectively, where $x_{\mathrm{R}} = s_{\mathrm{R}}+n_{\mathrm{R}}$ and $x_{\mathrm{I}} = s_{\mathrm{I}}+n_{\mathrm{I}}$ hold. Only the statistics of $(x_{\mathrm{R}} + ix_{\mathrm{I}})$ and $(n_{\mathrm{R}} + in_{\mathrm{I}})$ are observable, but that of $(s_{\mathrm{R}} + is_{\mathrm{I}})$ is a hidden value to be estimated. First, we measure the following mth-order moments from the data:

$$\mu_m(x_{\mathrm{R}}) = \mu_m(x_{\mathrm{I}}) = \mathrm{E}[x_{\mathrm{R}}^m], \tag{4}$$
$$\mu_m(n_{\mathrm{R}}) = \mu_m(n_{\mathrm{I}}) = \mathrm{E}[n_{\mathrm{R}}^m], \tag{5}$$

where we assume that x_{R} and x_{I} are i.i.d., and this also holds for the noise and observed signals.

In [5], the kurtosis of the speech power spectrum is estimated from the following equation using (4), (5), and the additivity of cumulants:

$$\mathrm{kurt}_{\mathrm{speech}} = \frac{\mu_4(s_{\mathrm{R}}^2+s_{\mathrm{I}}^2)}{\mu_2^2(s_{\mathrm{R}}^2+s_{\mathrm{I}}^2)} = \frac{\mathcal{N}\left(\mu_m(x_{\mathrm{R}}),\mu_m(n_{\mathrm{R}})\right)}{\mathcal{D}\left(\mu_m(x_{\mathrm{R}}),\mu_m(n_{\mathrm{R}})\right)}, \tag{6}$$

where

$$\begin{aligned}
&\mathcal{N}\left(\mu_m(x_{\mathrm{R}}),\mu_m(n_{\mathrm{R}})\right)\\
&= \mu_8(x_{\mathrm{R}}) - \mu_8(n_{\mathrm{R}})\\
&\quad + \left[4\mu_2(x_{\mathrm{R}}) - 32\mu_2(n_{\mathrm{R}})\right]\mu_6(x_{\mathrm{R}})\\
&\quad + \left[-32\mu_2(x_{\mathrm{R}}) + 60\mu_2(n_{\mathrm{R}})\right]\mu_6(n_{\mathrm{R}})\\
&\quad + \left[-76\mu_4(n_{\mathrm{R}}) - 96\mu_2(x_{\mathrm{R}})\mu_2(n_{\mathrm{R}}) + 516\mu_2^2(n_{\mathrm{R}})\right]\mu_4(x_{\mathrm{R}})\\
&\quad + \left[-60\mu_2^2(x_{\mathrm{R}}) + 1056\mu_2(x_{\mathrm{R}})\mu_2(n_{\mathrm{R}}) - 1416\mu_2^2(n_{\mathrm{R}})\right]\mu_4(n_{\mathrm{R}})\\
&\quad + 3\mu_4^2(x_{\mathrm{R}}) + 73\mu_4^2(n_{\mathrm{R}})\\
&\quad + 468\mu_2^2(x_{\mathrm{R}})\mu_2^2(n_{\mathrm{R}}) - 3456\mu_2(x_{\mathrm{R}})\mu_2^3(n_{\mathrm{R}}) + 2988\mu_2^4(n_{\mathrm{R}}),
\end{aligned} \tag{7}$$

$$\begin{aligned}
&\mathcal{D}\left(\mu_m(x_{\mathrm{R}}),\mu_m(n_{\mathrm{R}})\right)\\
&= 2\left(\mu_4(x_{\mathrm{R}}) - \mu_4(n_{\mathrm{R}}) + \mu_2^2(x_{\mathrm{R}}) - 8\mu_2(x_{\mathrm{R}})\mu_2(n_{\mathrm{R}}) + 7\mu_2^2(n_{\mathrm{R}})\right)^2.
\end{aligned} \tag{8}$$

3. PROPOSED METHOD

3.1. Problem of conventional method and motivation

The conventional method can estimate the kurtosis of a speech power spectrum without a clean speech signal to some extent. However, the accuracy of speech kurtosis estimation using the conventional method is often very unstable because of the instability in estimating very high order (sixth- and eighth-order) statistics obtained from finite samples. To avoid this, we propose to estimate speech kurtosis directly in the power spectrum domain instead of estimating the kurtosis of the speech power spectrum using sixth- and eighth-order statistics obtained from waveform signals in the time-frequency domain. More specifically, we calculate the kurtosis of the speech power spectrum by looking up values in a *kurtosis table*, which represents the direct relation among the kurtosis of speech, noise, and observed (noisy speech) signals in the power spectrum domain.

To construct the kurtosis table, we have to determine the mathematical relationship among the kurtosis of signals. However, there exist infinite patterns of signals that have an equivalent kurtosis value. Therefore, it is quite difficult to uniquely determine each signal. To avoid this problem, we apply a statistical assumption to the waveform signals of speech and noise. In the following section, we describe the statistical assumption in detail.

3.2. Parametric model

In the proposed method, we introduce the generalized Gaussian distribution for modeling the waveform signals of speech and noise. The probability density function (p.d.f.) of the generalized Gaussian distribution is defined as

$$p(y) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|y|/\alpha)^\beta}, \tag{9}$$

where α is a scale parameter, β is a shape parameter, and $\Gamma(\cdot)$ denotes the gamma function. Next, the mth-order moment of the generalized Gaussian distribution is given by

$$\mu_m(y) = \int_{-\infty}^{\infty} y^m p(y)dy = \alpha^m \frac{\Gamma((m+1)/\beta)}{\Gamma(1/\beta)}. \tag{10}$$

The kurtosis table is constructed using (10) and the moment-cumulant transformation in the next subsection.

3.3. Speech kurtosis estimation based on generalized Gaussian distribution prior

First, the mth-order moments of the waveform (time-frequency gird) signals of speech and noise are calculated as

$$\mu_m(s_{\mathrm{R}}) = \alpha_{\mathrm{s}}^m \frac{\Gamma((m+1)/\beta_{\mathrm{s}})}{\Gamma(1/\beta_{\mathrm{s}})}, \tag{11}$$

$$\mu_m(n_{\mathrm{R}}) = \alpha_{\mathrm{n}}^m \frac{\Gamma((m+1)/\beta_{\mathrm{n}})}{\Gamma(1/\beta_{\mathrm{n}})}, \tag{12}$$

where α_{s} and α_{n} are the scale parameters in the distributions for speech and noise signals, and β_{s} and β_{n} are the shape parameters in these distributions, respectively.

Next, the moment of the square of s_{R} is given by

$$\mu_m(s_{\mathrm{R}}^2) = \mu_{2m}(s_{\mathrm{R}}) = \alpha_{\mathrm{s}}^{2m} \frac{\Gamma((2m+1)/\beta_{\mathrm{s}})}{\Gamma(1/\beta_{\mathrm{s}})}. \tag{13}$$

Then we can calculate the cumulant of the power spectrum $s_{\mathrm{R}}^2 + s_{\mathrm{I}}^2$ as

$$\begin{aligned}
\kappa_m(s_{\mathrm{R}}^2+s_{\mathrm{I}}^2) &= 2\kappa_m(s_{\mathrm{R}}^2)\\
&= 2\sum_{\pi(m)} (-1)^{|\pi(m)|-1}(|\pi(m)|-1)! \prod_{B\in\pi(m)} \mu_{|B|}(s_{\mathrm{R}}^2),
\end{aligned} \tag{14}$$

and the mth-order moment of the power spectrum is given by

$$\mu_m(s_{\mathrm{R}}^2+s_{\mathrm{I}}^2) = \sum_{\pi(m)} \prod_{B\in\pi(m)} \kappa_{|B|}(s_{\mathrm{R}}^2+s_{\mathrm{I}}^2). \tag{15}$$

Finally, using (11) and (13)–(15), the kurtosis of the speech power spectrum is derived as a function of the shape parameter β_{s},

$$\mathrm{kurt}_{\mathrm{speech}} = \frac{\mu_4(s_{\mathrm{R}}^2 + s_{\mathrm{I}}^2)}{\mu_2(s_{\mathrm{R}}^2 + s_{\mathrm{I}}^2)^2} = \frac{\mathcal{N}_{\mathrm{s}}(\beta_{\mathrm{s}})}{\mathcal{D}_{\mathrm{s}}(\beta_{\mathrm{s}})}, \tag{16}$$

where

$$\begin{aligned}&\mathcal{N}_{\mathrm{s}}(\beta_{\mathrm{s}})\\ &= \Gamma\left(\frac{9}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)^3 + 4\Gamma\left(\frac{7}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)^2 + 3\Gamma\left(\frac{5}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)^2,\end{aligned} \tag{17}$$

$$\begin{aligned}&\mathcal{D}_{\mathrm{s}}(\beta_{\mathrm{s}})\\ &= 2\Gamma\left(\frac{5}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)^3 + 4\Gamma\left(\frac{5}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{s}}}\right)^2\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right) + 2\Gamma\left(\frac{3}{\beta_{\mathrm{s}}}\right)^4.\end{aligned} \tag{18}$$

Next, in the same manner, the kurtosis of the noise power spectrum is defined as

$$\mathrm{kurt}_{\mathrm{noise}} = \frac{\mu_4(n_{\mathrm{R}}^2 + n_{\mathrm{I}}^2)}{\mu_2(n_{\mathrm{R}}^2 + n_{\mathrm{I}}^2)^2} = \frac{\mathcal{N}_{\mathrm{n}}(\beta_{\mathrm{n}})}{\mathcal{D}_{\mathrm{n}}(\beta_{\mathrm{n}})}, \tag{19}$$

where

$$\begin{aligned}&\mathcal{N}_{\mathrm{n}}(\beta_{\mathrm{n}})\\ &= \Gamma\left(\frac{9}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)^3 + 4\Gamma\left(\frac{7}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)^2 + 3\Gamma\left(\frac{5}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)^2,\end{aligned} \tag{20}$$

$$\begin{aligned}&\mathcal{D}_{\mathrm{n}}(\beta_{\mathrm{n}})\\ &= 2\Gamma\left(\frac{5}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)^3 + 4\Gamma\left(\frac{5}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{n}}}\right)^2\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right) + 2\Gamma\left(\frac{3}{\beta_{\mathrm{n}}}\right)^4.\end{aligned} \tag{21}$$

Next, we calculate the kurtosis of the observed (speech-noise mixture) signal. Generally, the cumulant has additivity for additive independent valuables, i.e., $\kappa_m(a+b) = \kappa_m(a) + \kappa_m(b)$. Using this relation and (3), we can estimate the cumulant of the observed signal as

$$\begin{aligned}\kappa_m(x_{\mathrm{R}}) =& \kappa_m(s_{\mathrm{R}}) + \kappa_m(n_{\mathrm{R}})\\ =& \sum_{\pi(m)} (-1)^{|\pi(m)|-1}(|\pi(m)|-1)! \prod_{B\in\pi(m)} \mu_{|B|}(s_{\mathrm{R}})\\ &+ \sum_{\pi(m)} (-1)^{|\pi(m)|-1}(|\pi(m)|-1)! \prod_{B\in\pi(m)} \mu_{|B|}(n_{\mathrm{R}}).\end{aligned} \tag{22}$$

The moment of the square of x_{R} is given by

$$\mu_m(x_{\mathrm{R}}^2) = \mu_{2m}(x_{\mathrm{R}}) = \sum_{\pi(2m)} \prod_{B\in\pi(2m)} \kappa_{|B|}(x_{\mathrm{R}}). \tag{23}$$

Then, we can estimate the kurtosis of the observed signal in the power spectrum domain in a similar way to (14)–(16). The kurtosis of the observed signal in the power spectrum domain is calculated as

$$\mathrm{kurt}_{\mathrm{observed}} = \frac{\mu_4(x_{\mathrm{R}}^2 + x_{\mathrm{I}}^2)}{\mu_2(x_{\mathrm{R}}^2 + x_{\mathrm{I}}^2)^2} = \frac{\mathcal{N}_{\mathrm{x}}(\beta_{\mathrm{s}}, \alpha_{\mathrm{s}}, \beta_{\mathrm{n}}, \alpha_{\mathrm{n}})}{\mathcal{D}_{\mathrm{x}}(\beta_{\mathrm{s}}, \alpha_{\mathrm{s}}, \beta_{\mathrm{n}}, \alpha_{\mathrm{n}})}, \tag{24}$$

where

$$\begin{aligned}&\mathcal{N}_{\mathrm{x}}(\beta_{\mathrm{s}}, \alpha_{\mathrm{s}}, \beta_{\mathrm{n}}, \alpha_{\mathrm{n}})\\ &= \alpha_{\mathrm{s}}^8\left\{\Gamma\left(\frac{9}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right) + 4\Gamma\left(\frac{7}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{s}}}\right) + 3\Gamma\left(\frac{5}{\beta_{\mathrm{s}}}\right)^2\right\}\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)^2\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)^4\\ &+ \alpha_{\mathrm{s}}^6\alpha_{\mathrm{n}}^2\left\{32\Gamma\left(\frac{7}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right) + 96\Gamma\left(\frac{5}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{s}}}\right)\right\}\Gamma\left(\frac{3}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)^2\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)^3\\ &+ \alpha_{\mathrm{s}}^4\alpha_{\mathrm{n}}^4\left\{76\Gamma\left(\frac{5}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{5}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right) + 60\Gamma\left(\frac{5}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{n}}}\right)^2\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)\right.\\ &\quad \left. + 60\Gamma\left(\frac{5}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{s}}}\right)^3\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right) + 108\Gamma\left(\frac{3}{\beta_{\mathrm{s}}}\right)^2\Gamma\left(\frac{3}{\beta_{\mathrm{n}}}\right)^2\right\}\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)^2\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)^2\\ &+ \alpha_{\mathrm{s}}^2\alpha_{\mathrm{n}}^6\left\{32\Gamma\left(\frac{7}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right) + 96\Gamma\left(\frac{5}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{n}}}\right)\right\}\Gamma\left(\frac{3}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)^3\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)^2\\ &+ \alpha_{\mathrm{n}}^8\left\{\Gamma\left(\frac{9}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right) + 4\Gamma\left(\frac{7}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{n}}}\right) + 3\Gamma\left(\frac{5}{\beta_{\mathrm{s}}}\right)^2\right\}\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)^4\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)^2,\end{aligned} \tag{25}$$

$$\begin{aligned}&\mathcal{D}_{\mathrm{x}}(\beta_{\mathrm{s}}, \alpha_{\mathrm{s}}, \beta_{\mathrm{n}}, \alpha_{\mathrm{n}})\\ &= 2\left[\alpha_{\mathrm{s}}^4\left\{\Gamma\left(\frac{5}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right) + \Gamma\left(\frac{3}{\beta_{\mathrm{s}}}\right)^2\right\}\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)^2\right.\\ &+ \alpha_{\mathrm{s}}^2\alpha_{\mathrm{n}}^2\left\{\Gamma\left(\frac{3}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{3}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right)\right\}\\ &\left. + \alpha_{\mathrm{n}}^4\left\{\Gamma\left(\frac{5}{\beta_{\mathrm{n}}}\right)\Gamma\left(\frac{1}{\beta_{\mathrm{n}}}\right) + \Gamma\left(\frac{3}{\beta_{\mathrm{n}}}\right)^2\right\}\Gamma\left(\frac{1}{\beta_{\mathrm{s}}}\right)^2\right]^2.\end{aligned} \tag{26}$$

From the above-mentioned results, we can confirm that (a) the speech kurtosis, $\mathrm{kurt}_{\mathrm{speech}}$, given by (16), is a function of β_{s} only, and noise the kurtosis, $\mathrm{kurt}_{\mathrm{noise}}$, given by (19), is a function of β_{n} only. Thus, they are independent variables. (b) The kurtosis of the observed signal, $\mathrm{kurt}_{\mathrm{observed}}$, is a complex function of β_{s}, β_{n}, α_{s}, and α_{n}. Thus, given α_{s} and α_{n}, we can plot the value of $\mathrm{kurt}_{\mathrm{observed}}$ on the two-dimensional independent axes of $\mathrm{kurt}_{\mathrm{speech}}$ (of β_{s}) and $\mathrm{kurt}_{\mathrm{noise}}$ (of β_{n}), yielding the desired kurtosis table. The scale parameter is defined using the variance and shape parameter of each signal as

$$\alpha = \sqrt{\sigma^2\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}}, \tag{27}$$

where σ^2 is the variance of the signal. We determine the scale parameter from the input SNR, which expresses the variance ratio of the speech to noise signals. The variance of noise, σ_{n}^2, is measured in a noise-only time period, and that of speech, σ_{s}^2, can be estimated as $\sigma_{\mathrm{x}}^2 - \sigma_{\mathrm{n}}^2$. Thus, we can construct the kurtosis table at each input SNR by varying the shape and scale parameters. Then, the kurtosis of the speech power spectrum can be estimated from the kurtosis of the power spectra of the noise and observed signals by looking up values in the table. An example of such a kurtosis table is shown in Fig. 1. In this figure, the variance ratio of speech to noise signals is fixed to unity. Thus, this table should be used when the input SNR between the speech and noise signals is 0 dB.

4. EXPERIMENTS

4.1. Experimental setup

To confirm the effectiveness of the proposed method, we conducted an experiment on kurtosis estimation of the speech power spectrum. In this experiment, the conventional method and the proposed

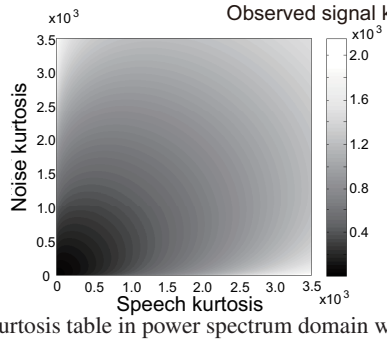

Fig. 1. Kurtosis table in power spectrum domain when input SNR is 0 dB.

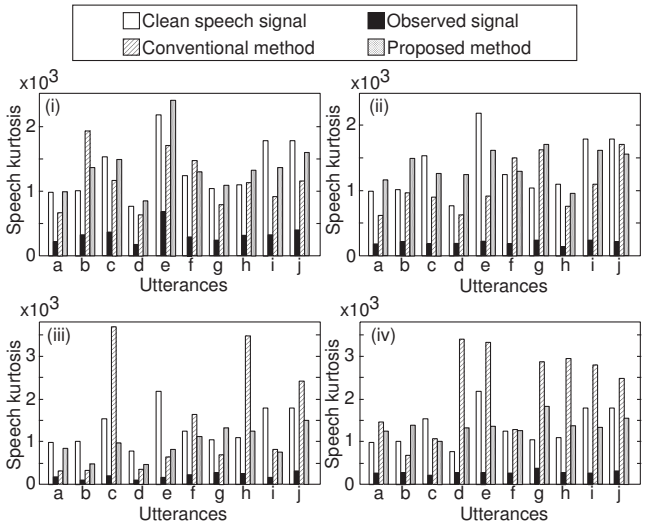

Fig. 2. Examples of speech kurtosis estimates for (i) white Gaussian noise, (ii) railway station noise, (iii) museum noise, and (iv) babble noise. Whole sentences are used in estimation.

method based on the generalized Gaussian distribution prior were compared.

We used 200 utterances (100 males and 100 females from the Japanese newspaper dictation database) as the target speech signals and four types of noise signals, namely white Gaussian noise, railway station noise, museum noise, and babble noise. Furthermore, we used three different types of data length: (a) the full length of each utterance of 2–15 s **(whole sentence)**, (b) the first half of each utterance **(half of sentence)**, and (c) the first second of each utterance **(one second)**. The test data were obtained by combining the target speech signals and noise signals. All signals used in this experiment were 16-kHz-sampled signals. The input SNR of the test data was set to 0, -5, or -10 dB. The speech kurtosis in the power spectrum domain of the test data was estimated using the conventional and proposed methods. In the proposed method, we constructed the kurtosis table for each SNR. In these kurtosis tables, $\mathrm{kurt}_{\mathrm{speech}}$ and $\mathrm{kurt}_{\mathrm{noise}}$ were changed from 0 to 3500 by three. We calculated the normalized error of the estimates in the conventional and proposed methods and compared the accuracy of speech kurtosis estimation. The normalized error is defined as $e_{\mathrm{norm}} = |\mathrm{kurt}_{\mathrm{oracle}} - \mathrm{kurt}_{\mathrm{speech}}|/\mathrm{kurt}_{\mathrm{oracle}}$, where $\mathrm{kurt}_{\mathrm{oracle}}$ is the power spectral kurtosis of the clean speech signal and $\mathrm{kurt}_{\mathrm{speech}}$ is the estimate of the speech power spectral kurtosis.

4.2. Experimental results

Figure 2 shows examples of speech kurtosis estimation results for 10 utterances using the conventional and proposed methods when the input SNR is -10 dB. In Fig. 2, although the kurtosis of the clean speech signal has quite a high value, the kurtosis of the observed (noisy speech) signal is lower than that of the clean speech signal. Estimates close to original values can be obtained by using the conventional and proposed methods. However, sometimes the estimates of speech kurtosis have a large error in the conventional method; in contrast, the proposed method's results are very stable and accurate.

Table 1 shows the average normalized error of estimates of speech kurtosis in the power spectrum domain using the conventional and proposed methods. In Table 1, estimates using the conventional method for the museum noise and babble noise have quite a large error, especially in the case of 1 s data length. However, these errors are reduced by using the proposed method. Also, in estimates for the white Gaussian noise and railway station noise, errors are lower for the proposed method. From these results, we can confirm that accuracy of speech kurtosis estimation is markedly improved using the proposed method.

Table 1. Average normalized error of estimates of speech kurtosis in power spectrum domain using conventional and proposed methods for (i) white Gaussian noise, (ii) railway station noise, (iii) museum noise, and (iv) babble noise

Data length	Method	(i)	(ii)	(iii)	(iv)
Whole sentence	Conventional	0.27	0.35	0.45	0.49
	Proposed	**0.11**	**0.22**	**0.33**	**0.28**
Half of sentence	Conventional	0.33	0.37	0.55	0.88
	Proposed	**0.18**	**0.25**	**0.31**	**0.35**
One second	Conventional	0.30	0.60	1.28	16.50
	Proposed	**0.16**	**0.32**	**0.39**	**0.40**

5. CONCLUSION

In this paper, we proposed a new method for the stable estimation of speech kurtosis in the power spectrum domain based on the generalized Gaussian distribution prior in order to avoid the calculation of higher-order statistics. Using this prior and the additivity of cumulants, we can construct a kurtosis table that directly represents the relationship among the kurtosis of speech, noise, and their mixture in the power spectrum domain, and speech kurtosis can be estimated stably from the observable data. An experimental evaluation confirmed the efficacy of the proposed method.

6. REFERENCES

[1] P. C. Loizou, *Speech Enhancement Theory and Practice*, CRC Press, Taylor & Francis Group FL, 2007.

[2] H. Saruwatari et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing.*, vol.2003, no.11, pp.1134–1146, 2003.

[3] L. Rabiner et al., *Fundamentals of speech recognition,* Upper Saddle River, NJ: Prentice Hall PTR, 1993.

[4] Y. Uemura et al., "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," *Proc. IWAENC2008*, 2008.

[5] R. Miyazaki et al., "Theoretical analysis of parametric blind spatial subtraction array and its application to speech recognition performance prediction," *Proc. HSCMA*, pp.19–24, 2011.

[6] R. Wakisaka et al., "Blind speech prior estimation for generalized minimum mean-square error short-time spectral amplitude estimator," *Proc. INTERSPEECH*, pp.361–364, 2011.

[7] R. Prasad et al., "Probability distribution of time-series of speech spectral components," *IEICE Trans. Fundamentals*, vol.E87-A, no.3, pp.584–597, 2004.