THEORETICAL ANALYSIS OF PARAMETRIC BLIND SPATIAL SUBTRACTION ARRAY AND ITS APPLICATION TO SPEECH RECOGNITION PERFORMANCE PREDICTION

† Ryoichi Miyazaki, † Hiroshi Saruwatari, † Ryo Wakisaka, † Kiyohiro Shikano, and † Tomoya Takatani
† Nara Institute of Science and Technology, Nara 630-0192, Japan
† Toyota Motor Corporation, Aichi, 470-0309, Japan

ABSTRACT

In this paper, an improved parametric postfiltering is introduced in our previously proposed blind spatial subtraction array (BSSA), and its theoretical analysis of the amounts of musical noise and noise reduction is conducted via higher-order statistics. Compared with the conventional BSSA, it is clarified that parametric BSSA can improve speech recognition performance. Next, we propose an unsupervised speech-recognition-performance prediction metric based on higher-order statistics in BSSA. We successfully reveal that the noise and speech extraction methods based on independent component analysis (ICA) have been proposed (see, e.g., [1, 2, 3]). We previously proposed blind spatial subtraction array (BSSA) [4] that consists of accurate noise estimation by ICA and the following speech extraction procedure based on nonlinear noise reduction such as spectral subtraction (SS) [5]; in this paper, we call the latter nonlinear speech extraction part postfiltering. The postfiltering in BSSA markedly improves the noise reduction performance, particularly in the presence of diffuse noise. However, BSSA always suffers from artificial distortion, so-called musical noise, due to nonlinear signal processing. This leads to a serious tradeoff between the noise reduction performance and amount of signal distortion in speech recognition.

In this paper, first, we propose a new improved BSSA method introducing two types of flexible postfiltering, generalized SS (GSS) [6] and quasi-parametric Wiener filter (QPWF) [7]. This modification aims to achieve a less-musical-noise speech enhancement and high speech recognition performance, but causes an additional problem of complex optimization of many internal parameters. Therefore, next, we propose a novel unsupervised (reference-less) speech-recognition-performance prediction metric based on the theoretical analysis of higher-order statistics in BSSA. This idea is motivated by our previous findings that the kurtosis, the 4th-order-moment-based descriptive statistics, of power spectra is strongly correlated to the generation of musical noise [9].

This paper successfully reveals that two explanatory variables, i.e., the kurtosis and signal-to-noise ratio (SNR), can predict speech recognition performance without using any reference signals. Moreover, this reference-less property of the proposed metric can contribute to the easy optimization of the parameter settings in BSSA, unlike the traditionally used cepstral distortion (CD), because CD needs clean speech signals before being added with noise but we cannot measure them in advance in real applications.

2. BLIND SPATIAL SUBTRACTION ARRAY [4]

2.1. Overview of BSSA

BSSA consists of a delay-and-sum array (DS)-based primary path and a reference path for the ICA-based noise estimation (see Fig. 1). The noise component estimated by ICA is efficiently subtracted from the primary path in the power spectrum domain without phase information. The detailed signal processing is shown below.

2.2. Partial speech enhancement in primary path

The observed signal vector of the J-channel array in the time-frequency domain, $x(f, \tau) = [x_1(f, \tau), \ldots, x_J(f, \tau)]^T$, is given by

$$x(f, \tau) = h(f)x(f, \tau) + n(f, \tau), \quad (1)$$

where $f$ is the frequency bin, $\tau$ is the frame number, $h(f) = [h_1(f), \ldots, h_J(f)]^T$ is a column vector of transfer function from the target signal component to each microphone, $x(f, \tau)$ is a target speech signal component, and $n(f, \tau) = [n_1(f, \tau), \ldots, n_J(f, \tau)]^T$ is a column vector of the additive noise signal. The target speech signal is partially enhanced in advance by DS. This procedure is given by

$$y_{DS} = w_{DS}^T x(f, \tau), \quad (2)$$

$$w_{DS} = [w_{DS1}, \ldots, w_{DSJ}]^T, \quad (3)$$

$$w_{DSj} = \frac{1}{J} \exp(-2i(f/M)\delta_d \sin \theta_i / c), \quad (4)$$

where $y_{DS}$ is the primary-path output that is a slightly enhanced target speech signal, $w_{DS}(f)$ is the filter coefficient vector of DS. $M$ is
2.3 ICA-based noise estimation in reference path

ICA-based noise estimation can be represented as:

\[ a(f, t) = W_{ICA} f_x(f, t) = [a_1(f, t), \ldots, a_J(f, t)]^T. \]  

\[ W_{ICA}^H f_x(f, t) = \rho I - \phi(a(f, t))^\delta f_x(f, t) \]  

\[ W_{ICA}^H f_x(f, t) = W_{ICA} f_x(f, t) + W_{ICA} f_x(f). \]

where \( \rho \) is the step-size parameter, \( I \) is used to express the value of the \( p \)th step in iterations, and \( L \) is the identity matrix. Besides, \( (\cdot) \) denotes a time-averaging operator, \( M^H \) denotes conjugate transpose of matrix \( M \), and \( \phi(\cdot) \) is an appropriate nonlinear vector function.

Next, in reference path, the estimated target speech signal is discarded at it is not required because we want to estimate only the noise component. Instead, we construct a noise-only vector \( \sigma_{noi} \) for \( f_x(f, t) \) from the output signal obtained by ICA using (5), where \( a(f, t) \) is assumed to be the speech component. Following this, we apply the projection back operation to remove the ambiguity of amplitude and construct the estimated noise signal \( c(f, t) \) by applying DS, as

\[ z(f, t) = W_{ICA}^H f_x(f, t) a_{noi}(f, t). \]

2.4 Speech extraction processing

It is known that ICA is proficient in noise estimation under the non-point-source noise condition [4]. Hence, in BSSA, noise reduction is carried out by subtracting the noise power spectrum \( z(f, t) \) from the partially enhanced target speech signal power spectrum \( y_{BSSA}(f, t) \), as

\[ y_{BSSA}(f, t) = \begin{cases} \sqrt{\beta} \Delta(f, t)^2 - \beta \cdot [\Delta(f, t)]^2 & \text{if } [\Delta(f, t)^2 > 0] \\ 0 & \text{otherwise} \end{cases} \]

where \( y_{BSSA}(f, t) \) is the output of BSSA, \( \beta \) is an oversubtraction parameter, and \( [\Delta(f, t)^2] \) is the smoothed noise power spectrum within a certain time frame window.

3. IMPROVED BSSA WITH PARAMETRIC POSTFILTERING: PARAMETRIC BSSA

In the conventional BSSA, SS is used to extract the target speech as described in (8). However, nonlinear noise reduction such as SS always generates a large musical noise and speech distortion. Therefore, to reduce the generation of distortion, we introduce GSS [6] and QPWF [7] as improved noise reduction processing (hereafter, we call them postfiltering). These are given by

\[ y_{GSS}(f, t) = \begin{cases} \sqrt{\beta} \Delta(f, t)^2 - \beta \cdot [\Delta(f, t)]^2 & \text{if } [\Delta(f, t)^2 > 0] \\ 0 & \text{otherwise} \end{cases} \]

\[ y_{QPWF}(f, t) = \begin{cases} \sqrt{\beta} \Delta(f, t)^2 - \beta \cdot [\Delta(f, t)]^2 & \text{if } [\Delta(f, t)^2 > 0] \\ 0 & \text{otherwise} \end{cases} \]

\[ y_{BSSA}(f, t) = \begin{cases} \sqrt{\beta} \Delta(f, t)^2 - \beta \cdot [\Delta(f, t)]^2 & \text{if } [\Delta(f, t)^2 > 0] \\ 0 & \text{otherwise} \end{cases} \]

where \( y_{GSS}(f, t) \) is the output of BSSA with GSS-based postfiltering, \( y_{QPWF}(f, t) \) is the output of BSSA with QPWF-based postfiltering, and \( n \) is an exponent parameter. We call this improved BSSA structure parametric BSSA in this paper. Parametric BSSA enables us to achieve both high noise reduction and low musical noise generation because of the flexibility of its parameter settings. However, one expected problem is the complex parameter optimization and prediction of the best parameters for speech recognition.

4. THEORETICAL ANALYSIS OF PARAMETRIC BSSA

4.1 Analysis strategy

In this section, we analyze the amount of musical noise generated and noise reduction via parametric BSSA under two typical noise conditions: spatially uncorrelated noise and point-source noise with strong spatial correlation.

Note that the two above-mentioned noise conditions were carefully selected because they are good representatives of real-world acoustical environments, e.g., a diffuse noise field is well approximated by spatially uncorrelated noise in the high-frequency range, and by spatially correlated noise in the low-frequency range. In our proposed analysis, we evaluate the amounts of musical noise and noise reduction independently in two frequency subbands, namely, high-frequency and low-frequency subbands.

In this paper, we assume that the input noise signal \( x \) in the power spectral domain can be modeled using the gamma distribution as

\[ P_G(\lambda) = \Gamma(\alpha)^{-1} \lambda^{\alpha-1} e^{-\lambda / \theta}, \]

where \( \lambda \geq 0, \alpha > 0, \) and \( \theta > 0 \). Here, \( \alpha \) is the shape parameter, \( \theta \) is the scale parameter, and \( \Gamma(\alpha) \) is the gamma function.

4.2 Analysis of amount of musical noise

4.2.1 Metric of musical noise generation: kurtosis ratio

We speculate that the amount of musical noise is highly correlated with the number of isolated power spectral components and their level of isolation. In this paper, we call these isolated components tonal components. Since such tonal components have relatively high power, they are strongly related to the weight of the skin of their probability density function (p.d.f.). Therefore, quantifying the skin of the p.d.f. makes it possible to measure the number of tonal components. Thus, we adopt kurtosis, one of the most commonly used higher-order statistics, to evaluate the percentage of tonal components among all components. A larger kurtosis value indicates a signal with a heavy skin, denoting that the signal has many tonal components. Kurtosis is defined as

\[ \text{kurt} = \mu_4 / \mu_2^2, \]

where "kurt" is the kurtosis and \( \mu_n \) is the \( n \)-th order moment, as

\[ \mu_n = \int x^n p(x) dx, \]

where \( p(x) \) is the p.d.f. of a signal \( x \). Note that \( \mu_n \) is not a central moment but a raw moment. Thus, (12) is not kurtosis in the mathematically strict definition but a modified version; however, we still refer to (12) as kurtosis in this paper.

In this study, we apply such kurtosis-based analysis to a noise-only time-frequency period of subject signals for the assessment.
of musical noise, even though these signals contain target-speech-
dominant periods. Thus, this analysis should be conducted during,
for example, periods of silence in speech. This is because we aim
to quantify the tonal components arising in the noise-only part,
which is the main cause of musical noise perception, and not in
the target-speech-dominant part.

Although kurtosis can be used to measure the number of tonal
components, note that the kurtosis itself is not sufficient to measure
the amount of musical noise. This is obvious because the kurtosis of
some unprocessed noise signals, such as an interfering speech
signal, is also high, but we do not recognize speech as musical noise.

Hence, we turn to the change in kurtosis between before
and after signal processing to identify only the musical-noise
components. Thus, we adopt the kurtosis ratio as a measure to
assess musical noise [8]. This measure is defined as

\[
kurtosis = \frac{kurt_{\text{proc}}}{kurt_{\text{orig}}}, \tag{14}\]

where \(kurt_{\text{proc}}\) is the kurtosis of the processed signal and \(kurt_{\text{orig}}\) is
the kurtosis of the observed signal. This measure increases as the
amount of generated musical noise increases. In Ref. [8], it was
reported that the kurtosis ratio is strongly correlated with the human
perception of musical noise.

### 4.2.2. Analysis in the case of spatially uncorrelated noise

First, we show the amount of kurtosis change after DS (primary path
of parametric BSSA). Using the shape parameter \(\alpha_m\), we can express
the kurtosis of a gamma distribution, \(\alpha_m\) as

\[
kurt_{\text{ds}} = \frac{2J_{\alpha_m} - (\alpha_m + 2)(\alpha_m + 3)}{\alpha_m (\alpha_m + 1)}. \tag{15}\]

The power spectral domain kurtosis after DS is given by [9]

\[
kurt_{\text{ds}} = J_{\alpha_m} - 6, \tag{16}\]

where \(\alpha_m\) is the kurtosis of the input signal power spectrum. Using
(15) and (16), we can derive a shape parameter \(\alpha_{\text{os}}\) after DS as

\[
\alpha_{\text{os}} = \left[2J^{\alpha_m} - (\alpha_m + 2)(\alpha_m + 3)\right]^{-1} \tag{17}\]

Next, we calculate the amount of kurtosis change after postfiltering.
With the shape parameter after DS, \(\alpha_{\text{os}}\), the \(m\)-th order moment
of the p.d.f after GSS or QPWF is given by [7, 10]

\[
\mu_m = \theta_m M_{\alpha_m}(\alpha_{\text{os}}, \beta, m, n), \tag{18}\]

where \(\theta_m\) is the scale parameter after DS, and \(*\) = GSS for
generalized SS and \(*\) = QPWF for quasi-parametric WF. Furthermore,
\(M_{\alpha_m}(\alpha_m, \beta, m, n)\) [10] and \(M_{\alpha_m}(\alpha_m, \beta, m, n)\) [7] can be expressed as

\[
M_{\alpha_m}(\alpha_m, \beta, m, n) = \frac{1}{\Gamma(\alpha_m)} \sum_{l=0}^{\infty} \beta^l \Gamma(\alpha_m + l) \Gamma(\alpha) \frac{\Gamma(m/n + 1)}{\Gamma(m/n - l + 1)} \left(\frac{\Gamma(\alpha_m + l)}{\Gamma(\alpha)} \right)^{\frac{1}{n}}, \tag{19}\]

where \(\Gamma(\alpha, z)\) is the upper incomplete gamma function defined as

\[
\Gamma(\alpha, z) = \int_{z}^{\infty} t^{\alpha-1} \exp(-t) dt. \tag{20}\]

Using (14), (18), (19), and (20), we can obtain the resultant kurtosis
ratio through parametric BSSA as

\[
kurtosis \text{ ratio}_{\text{BSSA}} = \frac{M_{\alpha_m}(\alpha_{\text{os}}, \beta, 0, n)}{M_{\alpha_m}(\alpha_m, 0, n)} \tag{21}\]

### 4.2.3. Analysis in the case of point-source noise

The kurtosis in the DS part does not change because point-source
noise has a strong spatial correlation [9]. Therefore, the resultant
kurtosis ratio depends on only the kurtosis of the postfiltering, i.e.,

\[
kurtosis \text{ ratio}_{\text{BSSA}} = \frac{M_{\alpha_m}(\alpha_{\text{os}}, \beta, 0, n)}{M_{\alpha_m}(\alpha_m, 0, n)} \tag{22}\]

### 4.3. Analysis of amount of noise reduction

#### 4.3.1. Metric of noise reduction performance

We define the noise reduction rate (NRR) as a measure of the noise
reduction performance, which is defined as the output SNR in dB
minus the input SNR in dB [4]. The NRR is given by

\[
NRR = 10 \log_{10} \frac{E[s_n^2]}{E[n_m^2]} = NRR_{\text{os}} + NRR_{\text{ds}} \tag{23}\]

where \(s_n\) and \(s_m\) are the input and output speech signals, and \(n_n\)
and \(n_m\) are the input and output noise signals, respectively. Also,
we assume that the amount of noise reduction is much larger than
that of speech distortion in processing, i.e., \(E[s_n^2] = E[s_m^2]\).

#### 4.3.2. Analysis in the case of spatially uncorrelated noise

First, we estimate the NRR of the DS part, \(NRR_{\text{ds}}\). For spatially
uncorrelated noise, it is well known that the \(NRR_{\text{ds}}\) is proportional
to the number of microphones, \(J\). The \(NRR_{\text{ds}}\) is given by

\[
NRR_{\text{ds}} = 10 \log_{10} J. \tag{24}\]

Next, we calculate the NRR of GSS or QPWF. In (24), since \(E[n_m^2] = \mu_i\)
when \(\beta = 0\) in (19) or (20) and \(E[s_n^2] = \mu_i\) for a specific
\((\text{nonzero})\beta\),

\[
NRR_{\text{os}} = 10 \log_{10} \frac{M_{\alpha_m}(\alpha_{\text{os}}, 0, 1, n)}{M_{\alpha_m}(\alpha_{\text{os}}, \beta, 1, n)}, \tag{25}\]

Finally, we obtain the amount of noise reduction through the
parametric BSSA procedure, \(NRR_{\text{BSSA}}\). Using (25) and (26),

\[
NRR_{\text{BSSA}} = NRR_{\text{ds}} + NRR_{\text{os}} = 10 \log_{10} \frac{J \cdot M_{\alpha_m}(\alpha_m, 0, 1, n)}{M_{\alpha_m}(\alpha_{\text{os}}, \beta, 1, n)}. \tag{26}\]

#### 4.3.3. Analysis in the case of point-source noise

In the case of point-source noise, the resultant NRR is only in the
postfiltering part given by (26), thus

\[
NRR_{\text{BSSA}} = 10 \log_{10} \frac{M_{\alpha_m}(\alpha_m, 0, 1, n)}{M_{\alpha_m}(\alpha_m, \beta, 1, n)}. \tag{27}\]

In summary, we can derive theoretical estimates for the amounts
of musical noise and noise reduction using (22) and (27) for the high-
frequency subband, and (23) and (28) for the low-frequency sub-
band, respectively.
FIG. 2. Theoretical behaviors of NRR and kurtosis ratio in parametric BSSA with GSS-based postfiltering and QPW F-based postfiltering for (a) white Gaussian noise and (b) speech noise.

4.4. Comparison of amount of musical noise under same NRR condition

According to the previous analysis, we can compare the amount of musical noise among parametric BSSA with GSS-based postfiltering and QPW F-based postfiltering under the same amount of noise reduction. Figure 2 shows the theoretical behaviors of the kurtosis ratio and NRR for various parameter values. In this figure, the shape parameter of input noise, $\alpha_{in}$, is set to 0.95 and 0.83 corresponding to almost white Gaussian noise and multiple speech noise that simulates a crowded place. Here, we assume a two-element array with an interelement spacing of 2.15 cm, denoting that noise is almost spatially correlated in 0-4 kHz and almost spatially uncorrelated in 4-8 kHz. The NRR is varied from 2 to 10 dB, and the oversubtraction parameter $\beta$ is adjusted so that the target speech NRR is achieved. In GSS and QPW F, the signal exponent parameter $2n$ is set to 2.0, 1.0, 0.5, and 0.1. Note that the kurtosis ratio and NRR in Fig. 2 are averages for all subbands. Figure 2 indicates that less amount of musical noise is generated when a lower exponent parameter is used, regardless of the type of noise and NRR as well as the postfiltering method. This result theoretically proves that parametric BSSA outperforms the conventional BSSA.

5. EXPERIMENT

5.1. Experimental setup

We conduct a speech recognition experiment with the proposed parametric BSSA, where we mainly evaluate the relation among the NRR, kurtosis ratio, and speech recognition performance. Moreover, we investigate the feasibility of predicting speech recognition performance by using theoretically derived metric, i.e., the kurtosis ratio and NRR in BSSA.

We used a two-element microphone array with an interelement spacing of 2.15 cm, and the direction of the target speech is set to be normal to the array. The size of the experimental room is 4.2 x 3.5 x 3.0 m, and the reverberation time is about 200 ms. All the signals used in this experiment are 16-kHz-sampled signals with 16-bit accuracy. The observed signal consists of the target speech signal of 200 speakers (100 males and 100 females) and two types of diffuse noise (white Gaussian noise and multiple speech noise) emitted from eight surrounding loudspeakers. The input SNR of test data is set to 10 dB for white Gaussian noise and 5 dB for speech noise.

Fig. 3. Results of word accuracy in various methods, unprocessed, conventional ICA, conventional BSSA, and parametric BSSA with GSS- or QPW F-based postfiltering for (a) white Gaussian noise and (b) speech noise.

Fig. 4. Relation among CD, NRR, and speech recognition performance, and among kurtosis ratio, NRR, and speech recognition performance in parametric BSSA with GSS-based postfiltering. (a) and (b) are for white Gaussian noise, and (c) and (d) are for speech noise.

The FFT size is 1024, and the frame shift length is 256 in BSSA. The speech recognition task is a 20-k-word Japanese newspaper dictation, where we used Julius 3.4.2 [11] as the speech decoder. The acoustic model is a phonetic tied mixture [11], and we use 260 speakers (150 sentences/speaker) for training the acoustic model.

In this experiment, NRR is varied from 2 to 10 dB, the exponent parameter $2n$ is set from 0.1 to 2.0, and the oversubtraction parameter $\beta$ is adjusted so that the target speech NRR is achieved.

5.2. Speech recognition performance

Figure 3 shows the results of word accuracy in several methods with their best parameter settings, where Conventional ICA directly uses the separated speech component (5) in ICA, Conventional BSSA corresponds to (8), Parametric BSSA corresponds to (9) and (10). This result reveals that the word accuracy of the parametric BSSA with GSS or QPW F-based postfiltering is remarkably superior to those of the conventional methods.

Next, we show the relation among NRR, CD, kurtosis ratio, and word accuracy in parametric BSSA. Figures 4(a) and 4(c), and Figs. 5(a) and 5(c) indicate that the better speech recognition performance can be obtained when the CD is smaller under the same amount of noise reduction. Also, in Figs 4(b) and 4(d), and Figs. 5(b) and 5(d), we can confirm the same tendency that the better...
speech recognition performance is well related with the smaller kurtosis ratio under the same amount of noise reduction. Therefore, we speculate that a pair of the kurtosis ratio and NRR is an alternative good indicator of speech recognition performance instead of CD.

Next, we evaluate the multiple correlation coefficient with speech recognition performance (word accuracy) as response variable and explanatory variables (NRR, CD, and kurtosis ratio) for white Gaussian noise and speech noise, where we merge the results of GSS- and QPWF-based postfiltering. The multiple correlation coefficients with the kurtosis ratio and NRR are 0.78 for white Gaussian noise and 0.69 for speech noise; the corresponding scores of using (superimposed) CD and NRR are 0.98 and 0.96. This result means that the kurtosis ratio can be used for predicting speech distortion but noise distortion only. Therefore, we will discuss an issue of estimating speech distortion in the next section.

6. SPEECH KURTOSIS ESTIMATION FOR SPEECH RECOGNITION PERFORMANCE PREDICTION

6.1. Problem and strategy

In this section, a new method of speech kurtosis estimation is proposed for evaluating pure distortion that arises only in the speech component. Since the speech component is always contaminated with noise 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 the data-driven manner, utilizing two observable statistics of the noisy speech signal and noise signal estimated by ICA or in the speech-absent part. Note that the proposed speech kurtosis estimation is still an unsupervised method because this method requires no reference (clean) speech signals, unlike CD.

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 holds the additivity for additive variables. Meanwhile, in transformation from a waveform to its power spectrum, the exponentiation operation is conducted but the cumulant does not have a straightforward relationship. In this case, we use the moment instead of the cumulant. Thus, we propose to use moment-cumulant transformation.

6.2. Moment-cumulant transformation

In this section, we derive some formulae regarding moment-cumulant transformation. They explicitly represent the relations between the moment and cumulant in each order, which are useful for estimating the kurtosis of the speech power spectrum.

First, the characteristic function \( \phi_\tau(t) \) of the random variable \( x \) is defined as

\[
\phi_\tau(t) = \mathbb{E}[e^{itx}] = \int e^{itx} f(x) dx.
\]

Then we can define the \( m \)-th moment \( \mu_m(x) \) and the \( m \)-th cumulant \( \kappa_m(x) \) of \( x \), as follows:

\[
\mu_m(x) = \frac{\partial^m \log \phi_\tau(t)}{\partial t^m} \bigg|_{t=0}, \quad \kappa_m(x) = \frac{\partial^m \log \phi_\tau(t)}{\partial t^m} \bigg|_{t=0}.
\]

Next, polynomial forms of interrelations between the moment and cumulant are derived below. From (30), the \( m \)-th moment \( \mu_m(x) \) can be rewritten as

\[
\mu_m(x) = \sum_{n=0}^{m} \frac{\mu_n \mu_{m-n}}{n!} = \sum_{n=0}^{m} \mu_n \kappa_{m-n}.
\]

where we use a **combinational form of Faà di Bruno's formula**,

\[
\frac{\partial^m f(g(x))}{\partial x^m} = \sum_{|B|=m} \frac{m!}{|B|!} \prod_{B \in \binom{m}{B}} \mu_m(g(x)^B),
\]

where \( m \) runs through the list of all partitions of a set of size \( m \), \( B \in \binom{m}{B} \) means that \( B \) is one of the blocks into which the set is partitioned, and \( |B| \) is the size of the set \( B \).

In the same manner, from (31), the \( m \)-th cumulant \( \kappa_m(x) \) is given by

\[
\kappa_m(x) = \sum_{n=0}^{m} \frac{\mu_n \kappa_{m-n}}{n!} = \sum_{n=0}^{m} \kappa_n \mu_{m-n}.
\]

6.3. Estimation of speech kurtosis from observations

Hereafter we define complex-valued variables of the observed (noisy speech) signal, the original speech signal, and the noise signal, as \((x_k + is_k), (s_k + is_k), \) and \((n_k + in_k)\), respectively, where \( x_k = s_k + n_k \) and \( s_k = s_k + n_k \) hold. Only the statistics of \( (s_k + is_k) \) and \( (n_k + in_k) \) are observable, but that of \( (s_k + is_k) \) is a hidden value to be estimated. First, we measure the following \( m \)-th moments from data:

\[
\mu_m(x_k) = \mathbb{E}[x_k^m], \quad \mu_m(s_k) = \mathbb{E}[s_k^m], \quad \mu_m(n_k) = \mathbb{E}[n_k^m].
\]
Generally, the cumulant has the additivity for the additive independent variables, i.e., \( \kappa_a(a + b) = \kappa_a(a) + \kappa_a(b) \). Using this relation and (34), we can estimate the kurtosis of the speech signal as

\[
\kappa_n(a_k) = \kappa_n(b_k) = \sum_{m} (-1)^{|m|-1} E(m) - 1! \prod_{\ell \neq m} \mu_\ell(b_k).
\]

Next, the statistics of the squared variable of \( \nu_k \) is given by

\[
\mu_n(a_k^2) = \mu_n(b_k^2) = \sum_{m} \prod_{\ell \neq m} \kappa_n(a_k^2) + \kappa_n(b_k^2).
\]

Finally, using (35)-(40), we can estimate the resultant kurtosis of the speech power spectrum as

\[
\text{kurt}_{\text{speech}} = \frac{\mu_4(x_k^2) + \mu_4(n_k^2)}{\mu_2(x_k^2) + \mu_2(n_k^2)} = \frac{N(\mu_4(x_k), \mu_4(n_k))}{D(\mu_2(x_k), \mu_2(n_k))},
\]

where

\[
N(\mu_4(x_k), \mu_4(n_k)) = \mu_4(x_k) + \mu_4(n_k) + 4\mu_2(x_k)\mu_2(n_k) + 6\mu_2(x_k)^2 + 6\mu_2(n_k)^2,
\]

\[
D(\mu_2(x_k), \mu_2(n_k)) = 2(\mu_2(x_k) + \mu_2(n_k))^2 - 8\mu_4(x_k)\mu_2(n_k) + 7\mu_4(n_k)^2.
\]

6.4. Experiment of speech recognition performance prediction

We again evaluate the multiple correlation coefficient with speech recognition performance as response variable and the speech kurtosis ratio as a new explanatory variable added to the NRR and noise kurtosis ratio. Speech kurtosis estimation by (41) is applied to the microphone-array input and BSSA output signals, and then the ratio of both kurtosis of speech power spectra is calculated.

Figure 6 shows that the multiple correlation coefficient with NRR, noise kurtosis ratio, and speech kurtosis ratio as explanatory variables is better than that with NRR and noise kurtosis ratio. Moreover, this result is almost close to that with the conventional supervised explanatory variables such as NRR and CD. Therefore, we conclude that the speech kurtosis ratio is appropriate to a new unsupervised measurement for evaluating speech recognition performances.

7. CONCLUSION

In this paper, first, we introduced GSS and QPWF in the nonlinear speech extraction part of BSSA for improving the sound quality, and performed its theoretical analysis based on higher-order statistics. Compared with the conventional BSSA, it was clarified that parametric BSSA can improve speech recognition performance. Next, we proposed the unsupervised speech-recognition-performance prediction method based on higher-order statistics in BSSA. We successfully reveal that the noise and speech kurtosis ratios can predict speech recognition performance without using any reference signals.

8. REFERENCES


