BLIND SPEECH EXTRACTION COMBINING ICA-BASED NOISE ESTIMATION AND LESS-MUSICAL-NOISE NONLINEAR POST PROCESSING

Hiroshi Saruwatari, Yu Takahashi, Kiyorhiro 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 propose a new blind speech extraction microphone array combining an independent component analysis (ICA)-based noise estimator and nonlinear signal processing for achieving high-quality speech enhancement. The proposed method consists of three parts, namely, the ICA-based noise estimator for a robust target cancellation, channel-wise spectral subtraction (chSS), and post-beamforming to sum up the chSS outputs. We provide a detailed proof of the less musical noise generation property in the proposed method via higher-order statistics analysis, compared with the conventional multichannel speech enhancement methods. The superiority of the proposed method is assessed in the experimental subjective evaluation.

1. INTRODUCTION
In recent years, integration methods of microphone array signal processing and nonlinear signal processing have been studied for better noise reduction, e.g., [1, 2]. It is reported that such an integration method can achieve higher noise reduction performance rather than a conventional adaptive microphone array [3], e.g., Griffith-Jim array. However, in such methods, artificial distortion (so-called musical noise) due to nonlinear signal processing arises, causing user's discomfort. The strength of nonlinear signal processing part in the integration methods is heuristically determined to mitigate musical noise generation.

Recently, it has been reported that the amount of generated musical noise is strongly related with the difference between higher-order statistics before/after nonlinear signal processing [4]. Therefore, based on higher-order statistics, it would be possible to design and optimize integration methods of microphone array signal processing and nonlinear signal processing from the viewpoint of not only noise reduction performance but also the sound quality.

In this paper, first, we analyze two integration methods of microphone array signal processing and nonlinear signal processing based on higher-order statistics. Particularly, we focus on spectral subtraction (SS)[5] method, i.e., the most popular and simplest nonlinear signal processing, as a nonlinear signal processing. Figure 1 shows a typical architecture...
beamforming technique, e.g., null beamformer [6] or adaptive beamforming. Finally, we obtain the speech enhanced signal based on SS. The detailed signal processing is shown below.

We consider the following J-channel observed signal in time-frequency domain as

\[ x(f, \tau) = h(f)x(f, \tau) + m(f, \tau), \]

where \( x(f, \tau) = [x_1(f, \tau), \ldots, x_J(f, \tau)]^T \) is the observed signal vector, \( h(f) = [h_1(f), \ldots, h_J(f)]^T \) is the transfer function vector, \( n(f, \tau) \) is the target speech, and \( m(f, \tau) = [n_1(f, \tau), \ldots, n_J(f, \tau)]^T \) is the noise vector. For enhancing the target speech, DS is applied to the observed signal. This can be represented by

\[ y_{\text{DS}}(f, \theta) = g_{\text{DS}}(f, \theta_\parallel)^T x(f, \tau), \]

where \( g_{\text{DS}}(f, \theta) = [g_{1, \text{DS}}(f, \theta), \ldots, g_{J, \text{DS}}(f, \theta)]^T \), \( g_j(f, \theta_\parallel) = J^{-1} \cdot \exp(-i2\pi(f/M)d_j \sin \theta/c) \), and \( j \) is the sampling frequency and \( c \) is the sound velocity. Finally, we obtain the enhanced target speech by channel-wise SS. This can be designated as

\[ y_{\text{chSS}}(f, \tau) = \left\{ \begin{array}{ll}
\sqrt{|y_{\text{DS}}(f, \tau)|^2 - \beta \cdot E[|\hat{n}_j(f, \tau)|^2]} & \text{if } |y_{\text{DS}}(f, \tau)|^2 - \beta \cdot E[|\hat{n}_j(f, \tau)|^2] > 0,
0 & \text{otherwise},
\end{array} \right. \]

where \( y_{\text{chSS}}(f, \tau) \) is the target speech enhanced signal by SS at channel, and \( \hat{n}_j(f, \tau) \) is the estimated noise signal in \( j \) channel.

Finally, we obtain the target speech enhanced signal by applying DS to \( y_{\text{chSS}}(f, \tau) \). This procedure can be represented by

\[ y(f, \tau) = g_{\text{DS}}(f, \theta_\parallel)y_{\text{chSS}}(f, \tau), \]

where \( y(f, \tau) \) is the final output of the proposed method.

3.3. Kurtosis-based analysis

3.3.1. Analysis strategy

It has been reported by the authors that the amount of generated musical noise is strongly related with the difference between the before-and-after kurtosis of a signal in nonlinear signal processing [4]. Thus, in this section, we analyze the amount of generated musical noise through the proposed chSS+BF and BF+SS, based on kurtosis. Basically, kurtosis increases through nonlinear signal processing, and larger increment of the kurtosis by nonlinear signal processing leads to more amount of musical noise generation. Thus, it can be expected that the generated musical noise becomes smaller with lower-kurtosis-increment signal processing. In the following subsections, hence, we analyze the kurtosis of BF+SS and the proposed chSS+BF, and prove which method can reduce the resultant kurtosis. Note that our analysis has no limitation in assumption of noise model, thus any noises including Gaussian and non-Gaussian can be under consideration.

3.3.2. Kurtosis

Kurtosis is one of the popular higher-order statistics for assessment of non-Gaussianity. Kurtosis is defined as

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

where \( x \) is the probability variable, \( \text{kurt}_x \) is the kurtosis of \( x \), and \( \mu_n \) is the \( n \)-th order moment of \( x \). Although \( \text{kurt}_x \) becomes 3 if \( x \) is Gaussian signal, note that the kurtosis of Gaussian signal in power spectral domain becomes 6. This is because Gaussian signal in time domain obeys chi-square distribution with two degrees of freedom in power spectral domain. In chi-square distribution with two degrees of freedom, \( \mu_4/\mu_2^2 = 6 \).

3.3.3. Resultant kurtosis in spectral subtraction

In this section, we analyze the kurtosis after SS. For evaluating the resultant kurtosis of SS, we utilize gamma distribution as a model of input signal in power spectral domain [7]. The probability density function (p.d.f.) of the gamma distribution for probability variable \( x \) is defined as

\[ p(x) = \Gamma^{-1}(\alpha) \cdot a^{\alpha-1} \cdot x^{\alpha-1} \cdot e^{-ax}, \]

where \( x \geq 0 \), \( \alpha > 0 \), and \( a > 0 \). Here, \( \alpha \) denotes the shape parameter and \( \theta \) is the scale parameter. Besides, \( \Gamma(\cdot) \) is the gamma function. Gamma distribution with \( \alpha = 1 \) corresponds to chi-square distribution with two degrees of freedom. Moreover, it is well-known that the average of the gamma distribution is given by \( E[P(x)] = \alpha \theta \). Furthermore, the kurtosis of
Gamma distribution, \( \text{kurt}_{GM} \), can be designated as [4]

\[
\text{kurt}_{GM} = \frac{(\alpha + 2)(\alpha + 3)}{\alpha(\alpha + 1)}. \tag{11}
\]

In SS, the average of observed power spectra is utilized as an estimated noise power spectrum. So the amount of subtraction is \( \beta \cdot \alpha \). Subtraction of the estimated noise power spectrum in each frequency band can be regarded as deforming of the p.d.f., which is the lateral shift of the p.d.f. to zero power direction. As a result, the probability of the negative power component arises. To avoid this, such a negative component probability is replaced by zero (so-called flooring technique). The resultant p.d.f. after SS can be written as

\[
P(x) = \begin{cases} 
C \cdot (x + \beta \cdot \alpha)^{\alpha-1} e^{-\frac{(x + \beta \cdot \alpha)^{\alpha}}{\alpha}} & (x > 0), \\
C \int_0^x x^{\alpha-1} e^{-\frac{x^\alpha}{\alpha}} dx & (x = 0), \end{cases} \tag{12}
\]

where \( C = 1/\Gamma(\alpha \cdot \beta) \).

Thus, the resultant kurtosis after applying SS, \( \text{kurt}_{SS} \), can be given as

\[
\text{kurt}_{SS} = \frac{\alpha^\beta}{\alpha + 1} \left( \alpha + 2\alpha + 3 \right) + \frac{\alpha^\beta}{\alpha + 1} \left( \beta \alpha \right)^2 \frac{1}{2} \left( \alpha + 3\alpha - 1 \right). \tag{13}
\]

Although we cannot describe details of the derivation of (13) due to the limitation of the paper space, reference [4] helps you to understand the derivation of (13).

### 3.3.4. Resultant kurtosis after DS

In this section, we analyze the kurtosis after DS, and we reveal that DS can reduce the kurtosis of input signals.

Now let \( x_j \) (\( j = 1, \ldots, J \)) be \( J \)-channel input signal, and they are i.i.d. signal each other. Moreover, we assume that the p.d.f. of \( x_j \) is both side symmetry and its average is zero. These assumptions make odd order cumulants zero except the first order cumulant. For cumulants, it is well known that the following relation holds;

\[
\text{cum}_n(aX + bY) = a^n \text{cum}_n(X) + b^n \text{cum}_n(Y), \tag{14}
\]

where \( \text{cum}_n(X) \) expresses the \( n \)-th order cumulant of probability variable \( X \). Based on the relation (14), the resultant cumulant after DS, \( \text{cum}_{n,DS} \), can be given by,

\[
\text{cum}_{n,DS} = \text{cum}_n(J^{-1}), \tag{15}
\]

where \( \text{cum}_n \) is the \( n \)-th order cumulant of \( x_j \). Using (15) and well-known mathematical relation between cumulant and moment, the power-spectral-domain kurtosis of DS can be expressed by

\[
\text{kurt}_{DS} = \frac{K_8 + 38JK_2K_4 + 32J^2K_2^3K_4 + 192J^2K_2K_4^2 + 288J^2K_2^3K_4 + 192J^2K_2K_4^2}{2J^2K_2^2 + 16J^2K_2^2K_4 + 32J^2K_2^2K_4^2}. \tag{16}
\]

for this derivation, see [8]. Considering an actual acoustic signal and its cumulants, we can illustrate the relation between input and output kurtosis via DS as Fig. 3. This relation can be approximated as

\[
\text{kurt}_{DS} \approx J^{-1} \cdot (\text{kurt}_{in} - 6) + 6, \tag{17}
\]

where \( \text{kurt}_{in} \) is the kurtosis of the input signal power spectra. As we can see in Fig. 3, the output kurtosis decreases in reverse proportion to the number of microphones.

### 3.3.5. Resultant kurtosis: chSS+BF vs. BF+SS

In the previous subsections, we have discussed the resultant kurtosis of SS and DS. In this subsection, we formulate the resultant kurtosis of the proposed chSS+BF and BF+SS. As described in Sect. 3.3.1, it can be expected that the smaller kurtosis increment leads to the less amount of generated musical noise. Thus, we compare the resultant kurtosis increment of both methods.

In BF+SS, first, DS is applied to multi-channel observed signal. At this point, the resultant kurtosis in power spectral domain, \( \text{kurt}_{DS} \), is

\[
\text{kurt}_{DS} = J^{-1} \cdot (\text{kurt}_{obs} - 6) + 6, \tag{18}
\]

where \( \text{kurt}_{obs} \) is the kurtosis of the observed signal in power spectral domain. Using (11), we can derive a shape parameter of gamma distribution corresponding to \( \text{kurt}_{DS} \) as

\[
d = \frac{\sqrt{\text{kurt}_{DS}^2 + 14 \cdot \text{kurt}_{DS} + 1} - \text{kurt}_{DS} + 5}{2 \cdot \text{kurt}_{DS} - 2}. \tag{19}
\]

Next, SS is applied to the DS output; using (13), we can obtain the resultant kurtosis of BF+SS, \( \text{kurt}_{BF,SS} \), as

\[
\text{kurt}_{BF,SS} \geq \frac{\alpha^\beta}{\alpha + 1} \left( \alpha + 2\alpha + 3 \right) + \frac{\alpha^\beta}{\alpha + 1} \left( \beta \alpha \right)^2 \frac{1}{2} \left( \alpha + 3\alpha - 1 \right). \tag{20}
\]

In the proposed chSS+BF, first, SS is applied to each input channel. Thus, the output kurtosis of channel-wise SS, \( \text{kurt}_{SS,SS} \), can be given by,

\[
\text{kurt}_{SS,SS} \geq \frac{\alpha^\beta}{\alpha + 1} \left( \alpha + 2\alpha + 3 \right) + \frac{\alpha^\beta}{\alpha + 1} \left( \beta \alpha \right)^2 \frac{1}{2} \left( \alpha + 3\alpha - 1 \right), \tag{21}
\]

where \( \alpha \) is a shape parameter of gamma distribution for the observed signal power spectra. Here, \( \alpha \) and \( \text{kurt}_{obs} \) satisfy (11). Finally, DS is performed and its resultant kurtosis can be written as

\[
\text{kurt}_{chSS,BF} = J^{-1} \cdot (\text{kurt}_{ASS,ASS} - 6) + 6, \tag{22}
\]

where \( \text{kurt}_{ASS,ASS} \) is the resultant kurtosis of the proposed chSS+BF.

Here, we consider the following equation to compare the resultant kurtosis of chSS+BF and BF+SS.

\[
D = \text{kurt}_{BF,SS} - \text{kurt}_{ASS,ASS,BF}. \tag{23}
\]

![Fig. 3. Relation between input kurtosis and output kurtosis of DS.](image)
where $D$ expresses the difference of the output kurtosis between chSS+BF and BF+SS. The positive $D$ indicates that the proposed chSS+BF reduced the resultant kurtosis compared with BF+SS. The relation about $D$ is depicted in Fig. 4. In the figure, oversubtraction parameter $\beta$ is fixed to 2. From this figure, we can confirm that the proposed chSS+BF can reduce the resultant kurtosis rather than BF+SS for almost all the input signals with various kurtosis. When input kurtosis is smaller than 4, the proposed chSS+BF cannot reduce the resultant kurtosis rather than BF+SS. However, such an input kurtosis corresponds to sub-Gaussian signal. In a common acoustical environment, such a sub-Gaussian signal cannot be expected to exist. Therefore, the proposed chSS+BF can be considered to reduce the resultant kurtosis rather than BF+SS in acoustic signals.

4. APPLICATION TO LOW-MUSICAL-NOISE BLIND SPEECH EXTRACTION

4.1. BSSA and chBSSA

Based on the discussion described in the previous section, we propose a new less-musical-noise structure for blind speech extraction method, referred to as chBSSA.

Conventional BSSA has been proposed in our previous study, which consists of ICA-based noise estimation and the following noise reduction processing based on SS. Figure 5(a) shows the configuration of BSSA. The ICA-based noise estimator provides a robust target-speech cancellation, and post-SS processing can efficiently reduce the noise components. However, this method has an inherent drawback concerning musical noise generation, just like the same case of BF+SS.

To cope with the musical noise problem in BSSA, in this paper, we propose chBSSA for achieving high-quality speech enhancement. The proposed method consists of three parts, namely, the ICA-based noise estimator for a robust target cancellation, channel-wise spectral subtraction, and post-beamforming to sum up the channel-wise outputs (see Fig. 5(b)); this guarantees less musical noise property similar to the theory of chSS+BF. In the following sections, we give

4.2. Objective evaluation

First, we compared the conventional BSSA and the proposed chBSSA in kurtosis difference and noise reduction performance. We used the following 16 kHz sampled signals as test data; the target speech is the original speech convoluted with the impulse responses which were recorded in a room with 200 ms reverberation, and to which artificially generated spatially uncorrelated white Gaussian was added. Besides, we use 6 speakers (6 sentences) as sources of the original source. The number of microphone elements in the simulation is changed from 2 to 16. The subtraction coefficient in the SS part, $\beta$, is set to 2.0, and the flooring parameter for BSSA, $\gamma$, is set to 0.0, 0.1, 0.2, 0.4 and 0.8. Note that flooring is not performed in chBSSA.

Here, we utilize the kurtosis difference as the measure for the amount of generated musical noise. This is given by

\begin{equation}
\text{Kurtosis difference} = \text{kurt}(n_{\text{proc}}(f, \tau)) - \text{kurt}(n_{\text{org}}(f, \tau)),
\end{equation}

(24)

where $n_{\text{proc}}(f, \tau)$ is the power spectrum of the residual noise signal after processing, and $n_{\text{org}}(f, \tau)$ is the power spectrum of the noise signal before processing. This kurtosis difference indicates how kurtosis is increased with processing. Thus, it is desired that the kurtosis difference becomes smaller. Moreover noise reduction performance is measured based on the power of the residual noise; this is given by

\begin{equation}
\text{Power of residual noise [dB]} = 10 \log_{10} \left\{ \frac{\sum_{f, \tau} |n_{\text{proc}}(f, \tau)|^2}{\sum_{f, \tau} |n_{\text{org}}(f, \tau)|^2} \right\}.
\end{equation}

(25)
Figure 6 shows the simulation results. From Fig. 6(a), we can see that the kurtosis difference of chBSSA is monotonically decreasing with increasing the number of microphones. On the other hand, the kurtosis difference of BSSA is constant regardless of the number of microphones. Indeed BSSA with the specific flooring parameter can achieve the same kurtosis difference as chBSSA, e.g., the case of flooring parameter of 0.4 in 10 microphones. However, BSSA with the large flooring parameter degrades the noise reduction performance itself (see Fig. 6(b)). On the other hand, the proposed chBSSA can reduce the kurtosis difference, i.e., musical noise generation, without degradation of noise reduction performance.

4.3. Subjective evaluation

Next, we conducted a subjective evaluation to confirm that the proposed chBSSA can mitigate the musical noise. In the evaluation, we gave two processed signals by the proposed chBSSA and conventional BSSA respectively to examinees with random order, and let 7 examinees (7 males) forcedly select which signal is less amount of musical noise. In this experiment, 3 types of noises, i.e., (a) artificial spatially uncorrelated white Gaussian, (b) real-recorded railway-station noise emitted from 36 loudspeakers, and (c) real-recorded human speech emitted from 36 loudspeakers, were used. 10 pairs of signal per one kind of noise, totally 30 pairs of processed signal were displayed to each examinee.

Figure 7 shows the subjective evaluation results, and we can confirm that the output of the proposed chBSSA is preferred by more than 90% examinees compared with that of BSSA even for the real acoustic noises including non-Gaussianity and inter-channel correlation properties.

5. CONCLUSION

In this paper, first, we analyzed musical noise generation in two integrated methods of microphone array and SS, i.e., chSS+BF and BF+SS, based on higher-order statistics. We revealed that the proposed chSS+BF can reduce the kurtosis compared with BF+SS. Next, based on the above-mentioned analysis, we proposed a new less-musical-noise structure for a combination method of ICA-based noise estimator and channel-wise nonlinear signal processing. Finally, the superiority of the proposed method was assessed in the experimental subjective evaluation. These analytic and experimental results imply great potential of higher-order-statistics-based optimization for musical noise [9].

6. REFERENCES