Signal Processing 102 (2014) 226–239

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Musical-noise-free blind speech extraction integrating microphone array and iterative spectral subtraction

Ryoichi Miyazaki [a], Hiroshi Saruwatari [a], Satoshi Nakamura [a], Kiyohiro Shikano [a], Kazunobu Kondo [b], Jonathan Blanchette [c], Martin Bouchard [c]

[a] Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[b] Corporate Research & Development Center, Yamaha Corporation, Japan
[c] School of Information Technology and Engineering, University of Ottawa, Canada

ARTICLE INFO

Article history:
Received 7 August 2013
Received in revised form
29 January 2014
Accepted 10 March 2014
Available online 18 March 2014

Keywords:
Blind speech extraction
Higher-order statistics
Iterative spectral subtraction
Microphone array

ABSTRACT

In this paper, we propose a musical-noise-free blind speech extraction method using a microphone array for application to nonstationary noise. In our previous study, it was found that optimized iterative spectral subtraction (SS) results in speech enhancement with almost no musical noise generation, but this method is valid only for stationary noise. The proposed method consists of iterative blind dynamic noise estimation by, e.g., independent component analysis (ICA) or multichannel Wiener filtering, and musical-noise-free speech extraction by modified iterative SS, where multiple iterative SS is applied to each channel while maintaining the multichannel property reused for the dynamic noise estimators. Also, in relation to the proposed method, we discuss the justification of applying ICA to signals nonlinearly distorted by SS. From objective and subjective evaluations simulating a real-world hands-free speech communication system, we reveal that the proposed method outperforms the conventional methods.

1. Introduction

In the past few decades, many applications of speech communication systems have been investigated, but it is well known that these systems always suffer from the deterioration of speech quality under adverse noise conditions. Spectral subtraction (SS) is a commonly used noise reduction method that has high noise reduction performance with low computational complexity [1–5]. However, in this method, artificial distortion, referred to as musical noise, arises owing to nonlinear signal processing, leading to a serious deterioration of sound quality. To achieve high-quality noise reduction with low musical noise, an iterative SS method has been proposed [6–8]. Also, some of the authors have reported the very interesting phenomenon that this method with appropriate parameters gives equilibrium behavior in the growth of higher-order statistics with increasing number of iterations [9]. This means that almost no musical noise is generated even with high noise reduction, which is one of the most desirable properties of single-channel nonlinear noise reduction methods. Following this finding, the authors have derived the optimal parameters satisfying the no-musical-noise-generation condition by analysis based on higher-order statistics. We have defined this method as musical-noise-free speech enhancement, where no musical noise is generated even for a high signal-to-noise ratio (SNR) in iterative SS [10].

In conventional iterative SS, however, it is assumed that the input noise signal is stationary, meaning that we can estimate the expectation of noise power spectral density from a time-frequency period of a signal that contains only noise. In contrast, under real-world acoustical environments, such as a nonstationary noise field, although it is

http://dx.doi.org/10.1016/j.sigpro.2014.03.010

necessary to dynamically estimate noise, this is very difficult. Therefore, in this paper, firstly, we propose a new iterative signal extraction method using a microphone array that can be applied to nonstationary noise. Our proposed method consists of iterative blind dynamic noise estimation by independent component analysis (ICA) [11,12] and musical-noise-free speech extraction by modified iterative SS.

Secondly, in relation to the proposed method, we discuss the justification of applying ICA to signals nonlinearly distorted by SS. We theoretically clarify that the degradation in ICA-based noise estimation obeys an amplitude variation in room transfer functions between the target user and microphones. Next, to reduce speech distortion, we introduce a channel selection strategy into ICA, where we automatically choose less varied inputs to maintain the high accuracy of noise estimation. Furthermore, we introduce a time-variant noise power spectral density (PSD) estimator [13] instead of ICA to improve the noise estimation accuracy. From objective and subjective evaluations, we reveal that the proposed method outperforms the conventional methods.

The rest of the paper is organized as follows. In Section 2, we describe related works on SS and the musical noise metric. In Section 3, the new musical-noise-free blind speech extraction method is proposed. In Section 4, an improvement scheme for poor noise estimation is presented. In Section 5, objective and subjective evaluations are described. Following a discussion on the results of the experiments, we present our conclusions in Section 6.

2. Related works

2.1. Conventional non-iterative spectral subtraction [2]

We apply a short-time discrete Fourier transform (DFT) to the observed signal, which is a mixture of target speech and noise, to obtain the time-frequency signal. We formulate conventional *non-iterative SS* [2] in the time-frequency domain as follows:

$$Y(f,\tau)=\begin{cases}\sqrt{|X(f,\tau)|^2-\beta \mathrm{E}[|N|^2]}\exp(j\arg(X(f,\tau)))\\ (\text{if } |X(f,\tau)|^2>\beta\ \mathrm{E}[|N|^2]),\\ \eta X(f,\tau)\ (\text{otherwise}),\end{cases} \tag{1}$$

where $Y(f,\tau)$ is the enhanced target speech signal, $X(f,\tau)$ is the observed signal, f denotes the frequency subband, τ is the frame index, β is the oversubtraction parameter, and η is the flooring parameter. Here, $\mathrm{E}[|N|^2]$ is the expectation of the random variable $|N|^2$ corresponding to the noise power spectra. In practice, we can approximate $\mathrm{E}[|N|^2]$ by averaging the observed noise power spectra $|N(f,\tau)|^2$ in the first K-sample frames, where we assume the absence of speech in this period and noise stationarity. However, this often requires high-accuracy voice activity detection.

2.2. Iterative spectral subtraction [6–8]

In an attempt to achieve high-quality noise reduction with low musical noise, an improved method based on iterative SS was proposed in the previous studies [6–8]. This method is performed through signal processing, in which the following *weak* SS processes are recursively applied to the noise signal (see Fig. 1). (I) The average power spectrum of the input noise is estimated, (II) The estimated noise prototype is then subtracted from the input with the parameters specifically set for weak subtraction, e.g., a large flooring parameter η and a small subtraction parameter β and (III) we then return to step (I) and substitute the resultant output (partially noise reduced signal) for the input signal.

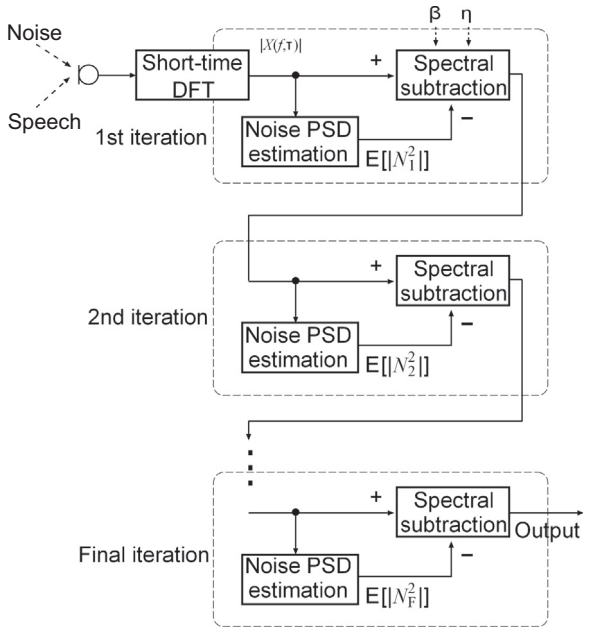

Fig. 1. Block diagram of iterative SS.

2.3. Modeling of input signal

In this paper, we assume that the input signal X in the power spectral domain is modeled using the gamma distribution as

$$P(x)=\frac{x^{\alpha-1}}{\Gamma(\alpha)\theta^{\alpha}}\exp(-x/\theta), \tag{2}$$

where $x\geq 0, \alpha>0$, and $\theta>0$. Here, α is the shape parameter, θ is the scale parameter, and $\Gamma(\alpha)$ is the gamma function, defined as $\Gamma(\alpha)=\int_0^{\infty} t^{\alpha-1}\exp(-t)\,dt$.

2.4. Mathematical metric of musical noise generation via higher-order statistics for non-iterative spectral subtraction [14]

In this study, we apply the *kurtosis ratio* to a *noise-only time-frequency period* of the subject signal for the assessment of musical noise [14]. This measure is defined as

$$\text{kurtosis ratio}=\mathrm{kurt}_{\mathrm{proc}}/\mathrm{kurt}_{\mathrm{org}}, \tag{3}$$

where $\mathrm{kurt}_{\mathrm{proc}}$ is the kurtosis of the processed signal and $\mathrm{kurt}_{\mathrm{org}}$ is the kurtosis of the observed signal. Kurtosis is defined as

$$\mathrm{kurt}=\mu_4/\mu_2^2, \tag{4}$$