Two-Channel Post-filtering Based on Adaptive Smoothing and Noise Properties

Chengshi Zheng, Yi Zhou, Xiaohu Hu, and Xiaodong Li
Key Laboratory of Noise and Vibration Research, Institute of Acoustics
Chinese Academy of Sciences, Beijing, China
Email: {cpszeng, yizhou, hxx, lxdl}@mail.ioa.ac.cn

Abstract—This paper studies the statistical properties of the gain functions, which are often used for two-channel post-filtering (TC-PF) algorithms. We reveal that the smoothing factor has a significant impact on both noise reduction and musical noise. When the smoothing factor increases, noise reduction can be improved and musical noise can be reduced simultaneously. However, the smoothing factor could not be too close to one because the system can only be assumed to be time-invariant for short durations. To solve this problem, this paper proposes an adaptive smoothing scheme by detecting the sudden change of the system. Moreover, the residual noise floor is adaptively chosen based on the structure of the noise power spectral density (NPSD) to further suppress the tonal noise components. Experimental results show the better performance of the proposed algorithm in terms of the segmental signal-to-noise-ratio (SNR) and the PESQ improvements.

I. INTRODUCTION

There are two main drawbacks of single-channel speech enhancement (SC-SE) algorithms. First, it is difficult to estimate the noise power spectral density (NPSD) accurately in extremely nonstationary environments [1]-[7]. Second, it is found that the speech intelligibility is not significantly improved for SC-SE algorithms in most noise conditions [1],[8]. Multi-channel speech enhancement (MC-SE) can solve the two problems in most cases, because a microphone-array system can get more information such as the spatial information when recording synchronously [9],[10].

Postfiltering is often needed after the spatial filtering for the MC-SE algorithms. The main reason is that the noise reduction of the spatial filtering is decreased as the reverberation time increases [9, Ch.4]. Numerous postfiltering algorithms have already been proposed in the last three decades, such as the coherence-based [11]-[15], the phase-based [16],[17], and the power-based algorithms [18],[19]. All of these methods have been shown to outperform the SC-SE algorithms in nonstationary environments.

Among the MC-SE algorithms, the two-channel adaptive wiener post-filtering (TC-PF) is a simple and effective way for both noise reduction and dereverberation, especially when the noise is uncorrelated between the two sensors [11]-[14]. The theoretical limits of the TC-PF can be found partially in [20], which is based on the assumption that all the parameters could be estimated accurately. However, it is well-known that both the speech and the noise are stochastic, only the approximate parameters could be obtained due to the limited number of available data. Based on this fact, this paper studies the statistical properties of the gain functions used for the TC-PF algorithms and reveals that the smoothing factor has a significant impact on both noise reduction and musical noise. To improve the performance of the TC-PF algorithms, we propose an adaptive smoothing scheme by detecting the sudden change of the system. Furthermore, to suppress more tonal noise components, this paper proposes an adaptive residual

This work was supported in part by National Natural Science Fund of China under Grant no. 61072123

noise floor selection approach based on the structure of the noise power spectral density (NPSD). Experimental results show the better performance of the proposed algorithm.

The rest of this paper is organized as follows. In section II, the TC-PF algorithms are reviewed in brief. Section III presents the statistical properties of the gain functions used for the TC-PF algorithms. We present an adaptive smoothing scheme and an adaptive residual noise floor selection approach in section IV and section V, respectively. Experimental results and some conclusions are given in section VI and VII, respectively.

II. REVIEW OF TWO-CHANNEL ADAPTIVE WIENER POST-FILTERING ALGORITHMS

The observed signal at the $i$th sensor is given by

$$ \hat{x}_i(n) = s_i(n) + d_i(n), \quad i = 1, 2. $$

(1)

where $s_i(n)$ and $d_i(n)$, with $i = 1, 2$, are the desired speech and the noise at the $i$th sensor, respectively. After compensating the mismatch for the two sensors and the time alignment for the desired speech, (1) can be rewritten as

$$ x_i(n) = s(n) + d_i(n), \quad i = 1, 2. $$

(2)

where $d_i(n)$ is the noise signal $\hat{d}_i(n)$ after compensating the mismatch for the two sensors and the time alignment for the desired speech. In the time frequency domain, we have

$$ X_i(k,l) = S(k,l) + D_i(k,l) $$

(3)

where $X_i(k,l)$, $S(k,l)$, and $D_i(k,l)$ are the short-time Fourier transform (STFT) of $x_i(n)$, $s(n)$, and $d_i(n)$ at the $k$th bin of the $l$th frame, respectively.

In [11], [14], the assumptions that $E\{S(k,l)D_i^*(k,l)\} = 0$ and $E\{D_i(k,l)D_j^*(k,l)\} = 0$, with $i \neq j$, were used, where the noise was assumed to be uncorrelated between sensors. For arrays with closely spaced sensors, McCowan and Bourlard used the assumption of a diffuse noise field to improve the noise reduction performance at low frequencies [15]. In this paper, we still assume the noise between sensors to be uncorrelated for two reasons. First, the assumption is valid if the sensor distance is larger than 0.4m [9, Ch. 12]. Second, the assumption simplifies the analysis and could give a deep insight into the statistical properties of the gain functions used in the TC-PF algorithms, where the gain functions are given by

$$ W_1(k,l) = \frac{2|\Phi_{12}(k,l)|}{(\Phi_{11}(k,l) + \Phi_{22}(k,l))}, $$

$$ W_2(k,l) = \frac{1}{2} \left( \frac{|\Phi_{12}(k,l)|}{\Phi_{11}(k,l)} + \frac{|\Phi_{12}(k,l)|}{\Phi_{22}(k,l)} \right), $$

$$ W_3(k,l) = \frac{4|\Phi_{12}(k,l)|^2}{(\Phi_{11}(k,l) + \Phi_{22}(k,l))^2} $$

(4)
where $W_i(k, l)$, with $i = 1, 2, 3$, could be found in [11]-[15], [20].

$\Phi_{ij}(k, l)$, with $i, j = 1, 2$, can be estimated by

$$
\Phi_{ij}(k, l) = \alpha \Phi_{ij}(k, l - 1) + (1 - \alpha) X_i(k, l) X_j^*(k, l),
$$

where $\alpha \in (0, 1)$ is a constant smoothing factor.

It is well-known that $\Phi_{ij}(k, l)$ is an asymptotically consistent estimator of $E\{X_i(k, l) X_j^*(k, l)\}$ as $\alpha$ approaches one. However, $\alpha$ should not be too large because the system is only time-invariant for short time segments. In most traditional TC-PF algorithms, a constant value of $\alpha$, i.e., $\alpha = 0.9$, was used to make a trade-off between the variance of $\Phi_{ij}(k, l)$ and the tracking performance of the system variation. By studying the statistical properties of the gain functions in (4), the following section shows that using a constant smoothing factor $\alpha$ limits the performance of the TC-PF algorithms.

III. STATISTICAL PROPERTIES OF THE GAIN FUNCTIONS

A. Distribution of the cross periodogram

The real and imaginary parts of $X_i(k, l)$ with $i = 1, 2$ are, respectively, assumed to be statistically independent Gaussian distributed. Under these assumptions, the periodograms $I_{X_iX_i}(k, l) = X_i(k, l) X_i^*(k, l)$, with $i = 1, 2$, have the $\chi^2$-distributions with 2 degrees of freedom, given by [3]

$$
I_{X_iX_i}(k, l) = \frac{U(y)}{\sigma^2_{X_iX_i}(k, l)} \exp \left( - \frac{y}{\sigma^2_{X_iX_i}(k, l)} \right),
$$

where $\sigma^2_{X_iX_i}(k, l) = E\{I_{X_iX_i}(k, l)\}$.

Define the cross periodogram $I_{X_iX_j}(k, l) = X_i(k, l) X_j^*(k, l)$, with $i \neq j$, if the real and imaginary parts of $I_{X_iX_j}$ are both Laplace distributed [21]. If $E\{I_{X_iX_j}\} = 0$ the variances of the real and the imaginary parts of $I_{X_iX_j}$ are, respectively, given by [21, (10.10)]

$$
\var\{I_{X_iX_j}(k, l)\} = \frac{1}{2} \sigma^2_{X_iX_i}(k, l) \sigma^2_{X_jX_j}(k, l),
$$

and

$$
\var\{I_{X_iX_j}(k, l)\} = \frac{1}{2} \sigma^2_{X_iX_i}(k, l) \sigma^2_{X_jX_j}(k, l),
$$

where $\Re\{\bullet\}$ and $\Im\{\bullet\}$ extract the real part and the imaginary part of a complex variable, respectively.

It is difficult to analyze the statistical properties of the gain functions in (4) by using the Laplace distributions of $R\{I_{X_iX_i}(k, l)\}$ and $R\{I_{X_iX_j}(k, l)\}$ directly. We use the Gaussian distribution to approximate the Laplace distribution for simplicity, then the probability density functions (PDF) of $R\{I_{X_iX_i}(k, l)\}$ and $R\{I_{X_iX_j}(k, l)\}$ are, respectively, given by

$$
\var\{I_{X_iX_j}(k, l)\} \propto N(0, \var\{R\{I_{X_iX_j}(k, l)\}\})
$$

and

$$
\var\{I_{X_iX_j}(k, l)\} \propto N(0, \var\{R\{I_{X_iX_j}(k, l)\}\})
$$

where $N(0, \tilde{\rho})$ indicates a Gaussian distribution with zero mean and the variance of $\tilde{\rho}$. Then the PDF of $\{I_{X_iX_j}(k, l)\}$ is the $\chi$-distribution.

B. Distribution of the estimated cross spectrum

Without loss of generality, we assume that $L$ independent frames are used to calculate $\Phi_{ij}(k, l)$, with $i, j = 1, 2$. The relationship between $L$ and $\alpha$ can be derived from [22, (28)], which is given by

$$
L = \frac{1 + \alpha}{1 - \alpha} \left( 1 + 2 \sum_{l=1}^{\infty} \alpha^l \rho(l) \right),
$$

where $\rho(l)$ can be calculated by [22, (20)]. It is well-known that the estimated spectra of $\Phi_{ij}(k, l)$, with $i = 1, 2$, have the $\chi^2$-distributions with $2L$ degrees of freedom, given by

$$
f_{\Phi_{ij}(k, l)}(y) = \frac{U(y)}{\sigma^2_{X_iX_i}(k, l) \sigma^2_{X_jX_j}(k, l)} \exp \left( - \frac{y}{\sigma^2_{X_iX_i}(k, l) \sigma^2_{X_jX_j}(k, l)} \right)
$$

where $\sigma^2_{X_iX_i}(k, l) = \frac{1}{2} \sigma^2_{X_iX_i}(k, l) \sigma^2_{X_jX_j}(k, l)$.

C. Statistical Properties of the gain functions

Assume that $\Phi_{11}(k, l)$, $\Phi_{22}(k, l)$, and $\Phi_{12}(k, l)$ are statistically independent. We have to emphasize that this assumption is only valid for noise only segments. Since this paper only studies the impact of $\alpha$ on both noise reduction and musical noise, this assumption could be used in this analysis.

The PDFs of $W_1(k, l)$ and $W_2(k, l)$ are relatively easy to derive using [23, (6-39) and (6-43)]. In (4), we have $W_3(k, l) = W_2^*(k, l)$, so the PDF of $W_3(k, l), f_{W_3}(k, l)(y)$, can be derived from $f_{W_1}(k, l)(y)$ by using [23, (5-8)].

Fig. 1 plots the PDFs of $W_i(k, l)$, with $i = 1, 2, 3$, versus different values of $\alpha$, where $\alpha = 0.8, 0.9, 0.95$ are shown in the first, the second and the third rows, respectively. In this example, we assume $\sigma^2_{X_iX_i}(k, l) = \sigma^2_{X_1X_1}(k, l) = \sigma^2_{X_2X_2}(k, l)$. For the sample rate $f_s = 16$kHz, we use a 512-point Hamming window and 75% overlap throughout this paper. The empirical PDFs of $W_i(k, l)$ with $i = 1, 2, 3$, which are obtained by Monte Carlo simulation, are compared with the theoretical PDFs. Obviously, the empirical results are well fit with the theoretical results. The amounts of noise reduction can be obtained from the PDFs of $W_i(k, l)$, given by

$$
NR_i = -20 \log \left( \int_0^\infty y f_{W_i}(k, l)(y) dy \right) \text{dB}.
$$

Fig. 2 shows the noise reduction performances for $W_i(k, l)$, with $i = 1, 2, 3$, versus different values of $\alpha$. The theoretical results fit quite well with the empirical results, which further verify the validation of the Gaussian approximation used in (9) and (10).
to obtain the adaptive smoothing factor, given by

\begin{equation}
\sigma^2_{d,z}(k,l) = \sigma_{d,z}(k,l) + 1 \sum_{i=K_f}^{K_f+1} \frac{1}{\sigma^2_{d,z}(k-i,l)}
\end{equation}

where \(K_f+1\) adjacent frequency bins are used. The ratio of the estimated NPSD to its envelop is:

\begin{equation}
\lambda(k,l) = \begin{cases} 
1 & \text{if } \frac{\sigma^2_{d,z}(k,l)}{\sigma^2_{d,z}(k,L)} \leq \lambda_{\text{th}} \\
\frac{\sigma^2_{d,z}(k,l)}{\sigma^2_{d,z}(k,L)} & \text{otherwise}
\end{cases}
\end{equation}

where \(\lambda_{\text{th}} \geq 1\) is a constant threshold. The proposed gain function is computed by

\begin{equation}
G_i(k,l) = \max \{ [1 - U(\lambda(k,l) - 1)]W_i(k,l) + U(\lambda(k,l) - 1)G_{0i}(k,l), G_{\min}(k,l) \}
\end{equation}

where \(G_{0i}(k,l) = \xi(k,l)/(1 + \xi(k,l))\). \(G_{\min}(k,l)\) is given by

\begin{equation}
G_{\min}(k,l) = G_{\max}/\sqrt{\lambda(k,l)}
\end{equation}

In (21), \(G_{0i}(k,l)\) could not be used directly, the main reason is that the NPSD is often over-estimated by using (17) [15]. Therefore, \(G_{0i}(k,l)\) is used only at the peaks of the estimated NPSD to increase noise reduction without introducing audible speech distortion. The enhanced speech can be computed by

\begin{equation}
\hat{s}_i(n) = \text{IFFT} \left\{ G_i(k,l) Z(k,l) \right\}.
\end{equation}

VI. EXPERIMENTAL RESULTS

This section compares the proposed algorithm with the traditional TC-PF (TTC-PF) algorithms using two different nonstationary processes as the noise, where their spectral properties are the same as the narrowband ARMA(2,2) example of [24] and the AR(24) example of [25], respectively. The NPSD is increased by four times from the third second for the two noise signals. Ten clean speech samples sampled at 16kHz are taken from the TIMIT database [26], where each sample has about ten seconds’ duration.

Two omnidirectional microphones are placed in an office room with the distance about 0.5m, where the reverberation distance of the room is about 1m. The desired speech speaker is located in front of the center microphone at a distance of 0.5m, and the noise speaker is located about 4m away from the center of the two microphones. The clean speech and the noise signals are recorded separately. All the clean speech samples are degraded by the two nonstationary noise signals with segmental SNR (SegSNR) in the range [0 15]dB. All the parameters for the proposed algorithm are shown in Table I. In this paper, only the comparison results of \(G_0(k,l)\) and \(W_3(k,l)\) are presented for the space limit. Moreover, \(W_3(k,l)\) is much better than \(W_1(k,l)\) and \(W_2(k,l)\) in terms of noise reduction [14].

Table II shows the comparison results of the SegSNR improvement and the PESQ improvement. With the proposed algorithm, consistent
TABLE II

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>ARMA(2,2) Nonstationary Noise</th>
<th>TTC-PF: Traditional TC-PF Algorithm Using $W_3(k, l)$ with $\alpha = 0.9$; Proposed: Proposed Algorithm.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Input SegSNR</td>
<td>PESQ Improvement</td>
</tr>
<tr>
<td>TTC-PF</td>
<td>6.01 4.92 2.12</td>
<td>2.21 0.84 0.54</td>
</tr>
<tr>
<td>Proposed</td>
<td>6.51 4.81 2.40</td>
<td>1.16 0.54 0.70</td>
</tr>
</tbody>
</table>

Fig. 3. Speech spectrograms of noisy speech at 0dB segmental SNR (a), of the signal enhanced by the traditional TC-PF algorithm (b), of the signal enhanced by the proposed algorithm (c).

improvements of both the SegSNR and the PESQ are obtained, where the better performances are more obvious for the AR(24) nonstationary noise case due to that the proposed algorithm could suppress more tonal noise components.

Speech spectrograms are shown in Fig. 3 to give indications about the structure of the residual noise and the speech distortion, where the clean speech is corrupted by the AR(24) nonstationary noise at 0dB segmental SNR. The proposed algorithm could suppress more noise components at the speech offsets and the noise-only frequency bins. What is more, the tonal noise component at about 1500Hz is entirely suppressed by the proposed algorithm. The proposed algorithm increases noise reduction without introducing audible speech distortion. This could be further confirmed by informal listening tests.

VII. CONCLUSIONS

This paper studies the statistical properties of the gain functions used in traditional TC-PF algorithms in detail and points out that the smoothing factor has a significant impact on both noise reduction and musical noise. To improve the performance of the TC-PF algorithm, we propose an adaptive smoothing scheme by detecting the sudden change of the system. Based on the noise properties, this paper proposes an adaptive residual noise floor selection approach to further suppress the tonal noise components. Experimental results verify that the proposed algorithm could suppress more noise components without introducing audible speech distortion.

We have to mention that the adaptive smoothing scheme could be directly applied to any multi-channel speech enhancement algorithms that need computing the auto/cross spectrum. Further work should concentrate on studying the correlated noise case.

REFERENCES