A Dual-microphone Sub-band Post-filter Using Simplified TBRR for Speech Enhancement

Haiping Wang[1], Yi Zhou[1], Yongbao Ma[2], and Hongqing Liu[1]
[1]*School of Communication and Information Engineering,*
Chongqing University of Posts and Telecommunications, Chongqing, China
[2]*Suresense Technology, Chongqing, China*
s180131236@stu.cqupt.edu.cn, zhouy@cqupt.edu.cn

2020 15th IEEE International Conference on Signal Processing (ICSP) | 978-1-7281-4480-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICSP48669.2020.9320979

Abstract—**Post-filtering is an effective method for further reducing noise components at a beamformer output. Existing techniques either are inefficient at suppressing highly non-stationary noise, or possess complex calculations. This paper proposes a novel dual-microphone sub-band post-filtering algorithm using a simplified transient beam to reference ratio (TBRR), applicable to adaptive beamformer, which is named SS-TBRR (simplified sub-band-TBRR). The sub-band signal processing approach is extended to the post-filtering for reducing computational complexity and smoothing the spectrum of the processed signal to eliminate musical noise. The relation between the observed signal, beamformer primary output, and the reference noise signal is exploited to differentiate non-stationary noise components from speech components. Based on speech presence probability and combined with an appropriate spectral enhancement approach, non-stationary noise is reduced significantly without canceling desired signal. Experimental results verify the effectiveness and robustness of the proposed algorithm.**

Keywords—post-filtering, simplified transient beam to reference ratio, sub-band, speech presence probability

I. Introduction

Microphone array systems are usually utilized for spatially filtering interfering signals coming from undesired directions. The generalized sidelobe canceller (GSC) is an effective technique for adaptive microphone array, which is widely used in speech enhancement applications [1]. The GSC can suppress the coherent noise components well, but it lacks the ability to reduce the incoherent noise components like diffuse noise. Post-filter is designed to suppress the residual noise components at a beamformer output. The existing post-filtering technologies are mainly divided into two categories: multi-channel post-filtering and single-channel post-filtering. A single-channel post-filtering approach does not provide sufficient attenuation for highly non-stationary noise components because such components are indistinguishable from the speech components. Multi-channel post-filtering is preferred to improve noise reduction performance, and many research outputs have been proposed.

Multi-channel post-filtering based on Wiener filtering and the auto and cross spectral densities of the observation signals was introduced by Zelinski [2]. In [3], Simmer and Wasiljeff proposed a modified version of Zelinski post-filtering, which employs the power spectral density (PSD) of the beamformer output rather than the average of the power spectral densities of individual observation signals. To take into account the presence of coherent noise components, McCowan and Bourlard proposed a post-filter based on noise field coherence [4]. A Wiener post-filter constructed using the signal-to-noise ratio (SNR) values computed by the coherence function between the input signals was suggested by Yousefian and Loizou [5]. These post-filters belong to the category of multi-channel Wiener filtering. Other approaches to multi-channel post-filtering are, for example: The *Multi-Channel Speech Presence Probability* (MC-SPP) [6], which estimates the noise PSD matrix utilizing the *a priori* speech absence probability (SAP) and the *Minima Controlled Recursive Averaging* (MCRA) algorithm generalized from the single-channel to the multi-channel case. The *Transient Beam to Reference Ratio* (TBRR) [7] is an effective approach to suppressing non-stationary noise, which depends on the transient power ratio of the beamformer primary output to the blocking matrix output. A detailed analysis about the TBRR can be found in [8]. However, MCRA algorithm has been used multiple times in the TBRR to estimate noise floor in both the beamformer primary output and the blocking matrix outputs, which leads to a large amount of computation for the TBRR. And there is strong musical noise in the signal processed by the TBRR. More recently, a model-based post filter using neural network was proposed by Xiong *et al.* [9].

Although various post-filtering approaches have been proposed, noise reduction performance in the highly non-stationary noise and low SNR conditions is still very challenging. The dual-microphone array has the advantages of low cost, small size and ultra-low power consumption and has been widely used in wearable devices like hearing aids, earphones, and smart glasses. Therefore, the paper focuses on the dual-microphone array framework and proposes a dual-microphone post-filtering algorithm using a simplified TBRR. Mel filter bank is utilized to decompose the speech into sub-band signals, and the process of obtaining the post-filtering gain is performed in the Mel domain, which aims to reduce calculation and musical noise. The relation between the observed signal of the first microphone, GSC primary output and the reference noise signal is explored to distinguish whether a transient is desired or interfering. The speech presence probability (SPP) controls the rate of recursive averaging to obtain a noise PSD estimate. Wiener filtering is utilized as the spectral enhancement approach. Experimental results confirm the good performance of the proposed post-filtering algorithm.

The rest of the paper is organized as follows. Section II introduces the conventional TBRR post-filtering algorithm. Section III elaborates on the details of the proposed SS-TBRR algorithm. Simulation experiments and results are presented and discussed in section IV. Conclusions are given in section V.

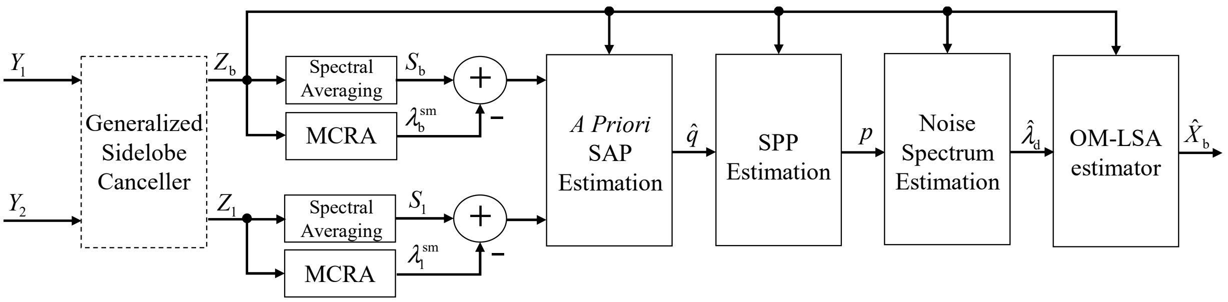

Fig. 1. Block diagram of the TBRR post-filtering algorithm proposed in [7].

II. Conventional TBRR Post-filtering

This section presents a brief review of the TBRR post-filtering [7] on which the proposed technique is based. The structure of the TBRR post-filtering is shown in Fig. 1, for a dual-microphone array. Let k and ℓ denote the frequency bin and the frame indices respectively. $\{Y_i\}_{i=1}^{2}$ represents the noisy observed signal. Z_b and Z_1 represent the GSC primary output and the reference noise signal respectively. $S_b(k,\ell)$ and $S_1(k,\ell)$ denote the smoothed spectrograms of the GSC output signals. The MCRA approach [10] is used to obtain the noise spectrum estimates $\lambda_b^{sm}(k,\ell)$ and $\lambda_1^{sm}(k,\ell)$. Then, the TBRR, which is the ratio between the transient power at the GSC primary output and the transient power at the reference noise signal, is used for indicating whether such a transient is desired or interfering, which is given by

$$\psi(k,\ell) = \frac{\max\left\{S_b(k,\ell) - \lambda_b^{sm}(k,\ell), 0\right\}}{\max\{S_1(k,\ell) - \lambda_1^{sm}(k,\ell), \varepsilon\lambda_b^{sm}(k,\ell)\}} \tag{1}$$

where ε is a constant for preventing the denominator from decreasing to 0. The *a priori* SAP is obtained by

$$\hat{q}(k,\ell) = \begin{cases} 1, \quad \text{if } \gamma_{\min}(k,\ell) \le 1 \text{ or } \psi(k,\ell) < \psi_{\text{low}} \\ \max\left\{\dfrac{\gamma_0 - \gamma_{\min}(k,\ell)}{\gamma_0 - 1}, \dfrac{\psi_{\text{high}} - \psi(k,\ell)}{\psi_{\text{high}} - \psi_{\text{low}}}, 0\right\}, \text{otherwise,} \end{cases} \tag{2}$$

where ψ_{low} , ψ_{high} and γ_0 are constants. $\gamma_{\min}(k,\ell)$ is obtained in the process of estimating $\lambda_b^{sm}(k,\ell)$. Then the SPP is estimated by

$$p(k,\ell) = \left\{1 + \frac{\hat{q}(k,\ell)}{1-\hat{q}(k,\ell)}(1+\xi(k,\ell))\exp(-\upsilon(k,\ell))\right\}^{-1} \tag{3}$$

where $\xi(k,\ell)$ is the *a priori* SNR estimated by a method proposed in [11], $\upsilon(k,\ell) = \gamma(k,\ell)\xi(k,\ell)/(1+\xi(k,\ell))$, and $\gamma(k,\ell) = |Z_b(k,\ell)|^2 / \lambda_d(k,\ell)$ is the *a posteriori* SNR. The noise spectrum estimation $\lambda_d(k,\ell)$ at the GSC primary output is updated with the SPP-based noise estimation algorithm as follows:

$$\hat{\lambda}_d(k,\ell+1) = \tilde{\alpha}_d(k,\ell)\hat{\lambda}_d(k,\ell) + \beta[1-\tilde{\alpha}_d(k,\ell)]|Z_b(k,\ell)|^2 \tag{4}$$

where β is a factor which compensates the bias, and the time-frequency dependent smoothing factor

$$\tilde{\alpha}_d(k,\ell) = \alpha_d + (1-\alpha_d)p(k,\ell) \tag{5}$$

where α_d is a constant satisfying $0 < \alpha_d < 1$. Then the *Optimally-Modified Log-Spectral Amplitude* (OM-LSA) algorithm [11] is performed to estimate the desired speech components, $\hat{X}_b(k,\ell)$.

The above TBRR post-filtering technique needs to utilize MCRA algorithm M times (where M is the number of microphones in the array), which results in a large amount of calculational complexity. Therefore, it can not meet the real-time requirements of some devices. And lots of musical noise remains in the signal processed by the above TBRR post-filtering technique. To cope with these issues, this paper proposed a simplified TBRR-based post-filtering algorithm for reducing computational complexity and improving noise reduction performance.

III. Proposed Post-filter

A. System Overview

The structure of the proposed dual-microphone post-filter is plotted in Fig. 2. In this paper, we extend the sub-band signal processing approach utilizing Mel filter bank to microphone array post-filtering to reduce the amount of calculation and smooth the spectrum after noise reduction to eliminate musical noise. The relation between the observed signal of the first microphone, GSC primary output and the reference noise signal is exploited to get a simplified TBRR to distinguish non-stationary noise and non-stationary speech, which further reduces the amount of calculation. Then an estimate $\hat{q}(k,\ell)$ for the *a priori* SAP is obtained. Based on a Gaussian statistical model and the decision-directed approach for the *a priori* SNR, an estimate $p(k,\ell)$ for the SPP is produced. Then the noise PSD at the GSC primary output is estimated using the SPP-based noise estimation algorithm. Finally, spectral enhancement of the GSC primary output is achieved by applying a Wiener gain function.

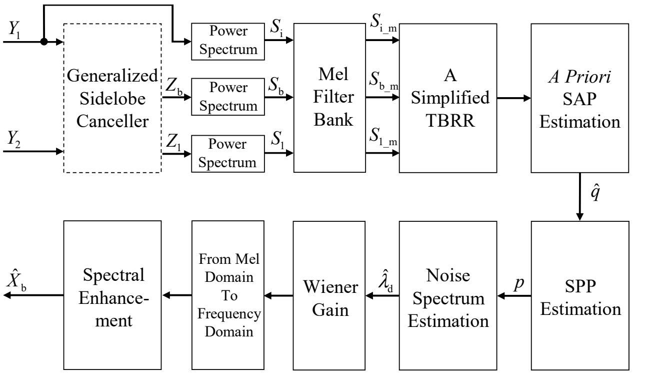

Fig. 2. Block diagram of the proposed post-filtering algorithm.

B. Mel Filter Bank

The primary step of the proposed algorithm is to decompose the speech into sub-band signals using the Mel filter bank [12], [13]. The Mel-scale aims to mimic the non-linear human ear perception of sound, which is more discriminative at lower frequencies and less discriminative at higher frequencies. The transform between the actual frequency f in Hz and the perceived frequency f_{mel} in Mel can be determined by

$$f = \text{F}_{\text{mel}}^{-1}(f_{\text{mel}}) = 700(e^{f_{\text{mel}}/1125} - 1), \tag{6}$$

$$f_{\text{mel}} = \text{F}_{\text{mel}}(f) = 1125\ln(1 + f/700). \tag{7}$$

Several band-pass filters with transfer function $H_m(f)$ ($0 \le m < N$, where N is the number of filters) are set within the frequency range of speech. Each filter has a triangular filtering characteristic with a center frequency of $f(m)$. These filters are of equal bandwidth in the Mel frequency range. The transfer function of each band-pass filter is

$$H_m(f) = \begin{cases} 0, & \text{if } f < f(m-1) \\ \dfrac{f - f(m-1)}{f(m) - f(m-1)}, & \text{if } f(m-1) \le f \le f(m) \\ \dfrac{f(m+1) - f}{f(m+1) - f(m)}, & \text{if } f(m) < f \le f(m+1) \\ 0, & \text{if } f > f(m+1) \end{cases} \tag{8}$$

where $f(m)$ can be defined as

$$f(m) = \left(\frac{L}{f_s}\right)\text{F}_{\text{mel}}^{-1}\left(\text{F}_{\text{mel}}(f_l) + m\frac{\text{F}_{\text{mel}}(f_h) - \text{F}_{\text{mel}}(f_l)}{N+1}\right). \tag{9}$$

f_l and f_h are the lowest and the highest frequencies in the filter frequency range, respectively. L is the length of short-time Fourier transform (STFT), and f_s is the sampling frequency.

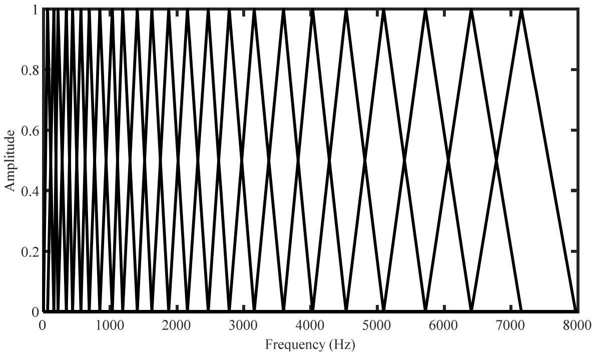

Fig. 3. The response curve of the Mel filter bank.

In this paper, above parameters are set as follows: $f_s = 16\text{kHz}$, $f_l = 0\text{Hz}$, $f_h = f_s/2 = 8000\text{Hz}$, $N = 24$, $L = 512$. Then the response curve of the Mel filter bank is shown in Fig. 3. Note that these triangular filters are overlapped in frequency domain.

After applying the filter bank to the power spectra (periodogram) of the Y_1, Z_b, and Z_1 respectively, corresponding power spectra in the Mel domain $S_{\text{i\_m}}$, $S_{\text{b\_m}}$ and $S_{\text{1\_m}}$ can be obtained. Then these power spectra are used to achieve the gain of the post-filtering. Due to the conjugate symmetry of the Fourier transform, only 257 frequency bins are used per frame to calculate the gain of the post-filtering, and only 24 values per frame are involved in the calculation in the Mel domain, which greatly reduces the amount of calculation.

C. Simplified TBRR Post-filtering

After obtaining the power spectra in the Mel domain, a simplified TBRR is defined by

$$\psi(m,\ell) = \frac{\tilde{S}_{\text{b\_m}}(m,\ell)}{\tilde{\mu}(m,\ell)}. \tag{10}$$

The smoothed power spectra $\tilde{S}_{\text{b\_m}}(m,\ell)$ and $\tilde{\mu}(m,\ell)$ are obtained by

$$\tilde{S}_{\text{b\_m}}(m,\ell) = \alpha_{\text{s}}\tilde{S}_{\text{b\_m}}(m,\ell-1) + (1-\alpha_{\text{s}})S_{\text{b\_m}}(m,\ell), \tag{11}$$

$$\tilde{\mu}(m,\ell) = \alpha_{\text{s}}\tilde{\mu}(m,\ell-1) + (1-\alpha_{\text{s}})\mu(m,\ell) \tag{12}$$

where α_{s} is a smoothing factor, and $\mu(m,\ell)$ is calculated by

$$\mu(m,\ell) = \frac{S_{\text{1\_m}}(m,\ell)S_{\text{b\_m}}(m,\ell)}{\max\left\{S_{\text{i\_m}}(m,\ell),\sigma\right\}} \tag{13}$$

where σ is a very small constant for preventing the denominator from 0. The *a priori* SAP is estimated by

$$\hat{q}(m,\ell)=\begin{cases}1, & \text{if } \psi(\mathrm{m},\ell)<\psi_{\mathrm{low}}\\ \dfrac{\psi_{\mathrm{high}}-\psi(m,\ell)}{\psi_{\mathrm{high}}-\psi_{\mathrm{low}}}, & \text{if } \psi_{\mathrm{low}}\le\psi(m,\ell)\le\psi_{\mathrm{high}}\\ 0, & \text{if } \psi(m,\ell)>\psi_{high}\end{cases} \tag{14}$$

where ψ_{low} and ψ_{high} are constants that denote the uncertainty in $\psi(m,\ell)$ during weak speech activity.

Assuming a Gaussian statistical model, the SPP is produced by

$$p(m,\ell)=\left\{1+\frac{\hat{q}(m,\ell)}{1-\hat{q}(m,\ell)}(1+\xi(m,\ell))\exp(-\upsilon(m,\ell))\right\}^{-1} \tag{15}$$

where $\gamma(m,\ell)$ and $\xi(m,\ell)$ are respectively the *a posteriori* and *a priori* SNRs, and $\upsilon=\gamma\xi/(1+\xi)$.

The *a prior* SNR is calculated by

$$\hat{\xi}(m,\ell)=\alpha G_{H_1}^2(m,\ell-1)\gamma(m,\ell-1)+(1-\alpha)\max\{\gamma(m,\ell)-1,0\} \tag{16}$$

where α is a constant for controlling the trade-off between noise suppression and desired signal distortion, and

$$G_{H_1}(m,\ell)=\frac{\xi(m,\ell)}{1+\xi(m,\ell)} \tag{17}$$

is the spectral gain function of the Wiener estimator.

Computing a smoothing parameter $\tilde{\alpha}_{\mathrm{d}}(m,\ell)$ by

$$\tilde{\alpha}_{\mathrm{d}}(m,\ell)=\alpha_{\mathrm{d}}+(1-\alpha_{\mathrm{d}})p(m,\ell) \tag{18}$$

where α_{d} is a constant satisfying $0<\alpha_{\mathrm{d}}<1$, we obtain the following estimate for the noise PSD at the GSC primary output:

$$\hat{\lambda}_{\mathrm{d}}(m,\ell)=\begin{cases}(\tilde{\alpha}_{\mathrm{d}}(m,\ell)-\eta)\hat{\lambda}_{\mathrm{d}}(m,\ell-1)+(1-\tilde{\alpha}_{\mathrm{d}}(m,\ell)+\eta)\mu(m,\ell), & \\ \qquad\qquad \text{if } \tilde{\alpha}_{\mathrm{d}}(m,\ell)>\tau & \\ \tilde{\alpha}_{\mathrm{d}}(m,\ell)\hat{\lambda}_{\mathrm{d}}(m,\ell-1)+(1-\tilde{\alpha}_{\mathrm{d}}(m,\ell))S_{\mathrm{b\_m}}(m,\ell), & \\ \qquad\qquad \text{if } \tilde{\alpha}_{\mathrm{d}}(m,\ell)\le\tau & \end{cases} \tag{19}$$

where τ is a constant satisfying $0.9<\tau<1$ for slowing down the noise PSD update at strong speech components, and η is a very small constant for compensating the bias.

The above described noise PSD estimate takes into account the transient, as well as stationary, noise components. Then an appropriate spectral enhancement algorithm (e.g., the Wiener filtering technique) is utilized to achieve an estimate of the desired signal. Note that the post-filtering gain in the Mel domain is obtained first. The necessary step is to decompose the gain in the Mel domain into the gain in the frequency domain using the inverse transformation of (8).

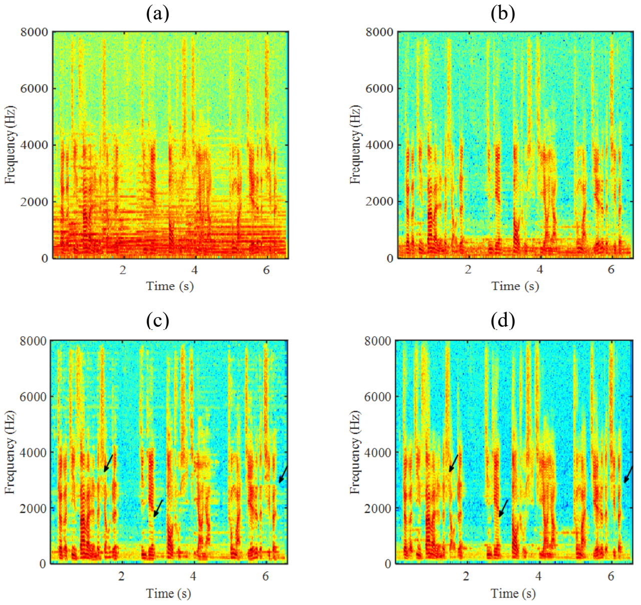

Fig. 4. Speech spectrograms. (a) Noisy signal at a single microphone; (b) GSC output; (c) GSC output enhanced by the conventional TBRR algorithm; (d) GSC output enhanced by the proposed SS-TBRR algorithm.

IV. Experimental Results

We compare the proposed SS-TBRR algorithm with the conventional TBRR algorithm and show improvements can be achieved. To simulate a wearable device application, the distance between the two microphones was set to 3 cm. A root Hanning window of 512 samples was used with 50% overlap between two consecutive frames. Other parameters used in the algorithm were as follows: $\alpha_{\mathrm{s}}=0.8$, $\sigma=10^{-10}$, $\psi_{\mathrm{low}}=2$, $\psi_{\mathrm{high}}=20$, $\alpha=0.95$, $\alpha_{\mathrm{d}}=0.8$, $\tau=0.97$, and $\eta=10^{-3}$.

The test speech signals were obtained by corrupting the clean speech with noise under various SNR (-5 dB, 0 dB, 5 dB and 10 dB). To demonstrate the robustness of the proposed SS-TBRR algorithm, four types of noise including directional noise and diffuse noise are used. The clean speech signals at 0° were recorded in the absence of background noise. Fig. 4 depicts the spectrograms of the noisy signal, GSC output, GSC output enhanced by the conventional TBRR algorithm, and GSC output enhanced by the proposed SS-TBRR algorithm in the case of the speech signal with 0 dB SNR and music noise at 90°. It can be seen that desired signal cancellation exists in Fig. 4 (c) where the black arrows point, while the desired signal is preserved in Fig. 4 (d) where the black arrows point. Much of weak noise remains in the GSC output enhanced by the conventional TBRR algorithm, especially in the high frequency portion. However, those noise can be eliminated by the proposed SS-TBRR algorithm. In short, the SS-TBRR algorithm proposed in this paper not only suppresses noise better, but also prevents the desired signal distortion.

To further verify the advantage of the proposed algorithm, the objective test perceptual evaluation of speech quality

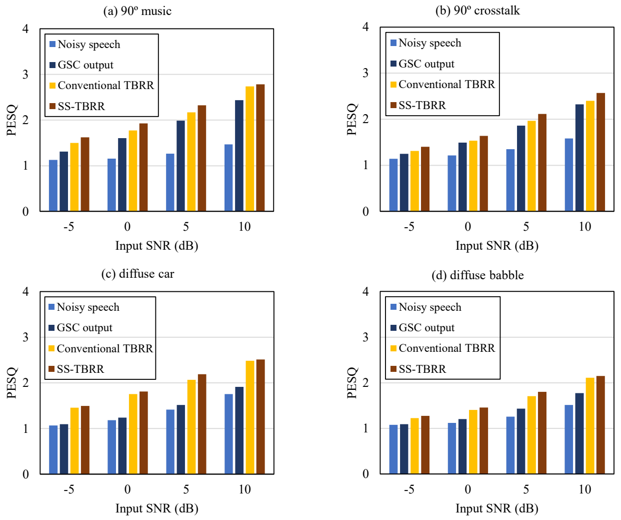

Fig. 5. PESQ values at different background noise levels.

(PESQ) [14], which is highly correlated with the subjective listening test, was conducted. Fig. 5 shows the PESQ values at four background noise levels. It can be seen that both the conventional TBRR algorithm and the proposed SS-TBRR algorithm have achieved improvements on PESQ scores in various noise environments, but the SS-TBRR algorithm improves PESQ scores even more, especially in the case of directional noise. More importantly, the SS-TBRR algorithm is much less computationally intensive than the conventional TBRR algorithm.

V. Conclusions

In this paper, an improved robust dual-microphone post-filtering algorithm for generalized sidelobe canceller is proposed, which is particularly advantageous in non-stationary noise environments. Sub-band signal processing approach was extended to the proposed post-filtering to reduce computational complexity and smooth the spectrum of the processed signal to eliminate musical noise. Lots of effective simplifying operations based on the conventional TBRR algorithm have been performed to further reduce calculation, like the simplified methods for obtaining TBRR and the *a prior* SAP. The proposed SS-TBRR algorithm enjoys robustness under various background noise conditions, and provides good noise suppression while protecting the desired signal. Compared with the conventional TBRR algorithm, the proposed SS-TBRR algorithm not only reduces the computational complexity significantly, but also improves noise reduction performance.

References

[1] M. Brandstein and D. Ward, Microphone Arrays, Springer Verlag, 2001.

[2] R. Zelinski, “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms,” in Proc. 13th IEEE Internat. Conf. Acoust. Speech Signal Process., New York, pp. 2578–2581, Apr. 11–14, 1988.

[3] K. U. Simmer and A. Wasiljeff, “Adaptive microphone arrays for noise suppression in the frequency domain,” in Proc. 2nd Cost-229 Workshop on Adaptive Algorithms in Communications, Bordeaux, France, pp. 185–194, October 30, 1992.

[4] I. A. McCowan and H. Bourlard, “Microphone array post-filter based on noise field coherence, ” in IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pp. 709-716, Nov. 2003.

[5] N. Yousefian and P. C. Loizou, “A Dual-Microphone Algorithm That Can Cope With Competing-Talker Scenarios,” in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 1, pp. 145-155, Jan. 2013.

[6] Mehrez Souden, Jingdong Chen, Jacob Benesty and Sofiene Affes, “An integrated solution for online multichannel noise tracking and reduction,”IEEE Transactions on Audio Speech and Language Processing, vol. 19, no. 7, Sept. 2011.

[7] I. Cohen and B. Berdugo, “Microphone array post-filtering for non-stationary noise suppression,” in Proc. 27th IEEE Int. Conf. Acoust. Speech Signal Process., Orlando, FL, pp. 901–904, May 13–17, 2002.

[8] I. Cohen, "Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering," in IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pp. 684-699, Nov. 2003.

[9] Y. Xiong *et al*., “Model-Based Post Filter for Microphone Array Speech Enhancement,” 2018 7th International Conference on Digital Home (ICDH), Guilin, China, pp. 82-88, 2018.

[10] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," in IEEE Transactions on Speech and Audio Processing, vol. 11, no. 5, pp. 466-475, Sept. 2003.

[11] I. Cohen and B. Berdugo, “Speech Enhancement for Non-Stationary Noise Environments,” Singnal Processing, vol. 81, no. 11, pp. 2403-2418, October 2001.

[12] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. on Acoustics Speech and Signal Processing, vol. 28, no. 4, pp. 357-366, Agust 1980.

[13] J. Dai, Y. Zhang, J. Hou, X. Wang, L. Tan and J. Jiang, “Sparse Wavelet Decomposition and Filter Banks with CNN Deep Learning for Speech Recognition,” 2019 IEEE International Conference on Electro Information Technology (EIT), Brookings, SD, USA, pp. 098-103, 2019.

[14] Rix A W, Beerends J G, Hollier M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221). pp. 749-752, 2001.