ADAPTIVE FILTERING EXPLOITING BAND-LIMITED PROPERTY OF SOUND FIELD IN THE WAVE DOMAIN

Satoru Emura, Yusuke Hiwasaki, and Hitoshi Ohmuro

NTT Media Intelligence Laboratories, NTT Corporation

ABSTRACT

Massive multichannel sound reproduction is necessary for highly immersive full-duplex communication. Efficient echo cancellation between massive loudspeakers and microphones is indispensable, and wave-domain adaptive filtering reduces the computational complexity drastically. We propose to further reduce this complexity by exploiting the band-limited property that the sound field has in the space-time-frequency representation. Experimental evaluations support the applicability of the proposed method.

Index Terms— wave domain, adaptive filtering, space-time-frequency representation

1. INTRODUCTION

For full-duplex immersive communication, multichannel sound reproduction is essential. Adaptive filtering for multichannel echo cancellation is indispensable for this and has been studied extensively [1][2][3].

The Wave Field Synthesis (WFS) method [4] has recently been gaining attention since it can overcome the limitation of a narrow listening area inherent to sound reproduction techniques such as that using a 5.1-channel system. The WFS is a massive multichannel sound reproduction technique that requires many (>30) channels. When applying it for full-duplex communication, the number of echo paths between loudspeakers and microphones is proportional to the square of the number of channels, and accordingly, the computational complexity of echo cancellation increases.

Buchner et al. [5] proposed wave-domain adaptive filtering to drastically reduce this complexity, where a $P \times P$ multi-input multi-output (MIMO) system in the frequency domain is modeled as only P decoupled single-input single-output (SISO) systems in the wave domain. Schneider et al. [6] extended this decoupled wave-domain model from SISO to multi-input single-output (MISO) by including adjacent spatial frequency bins. This extension is effective especially when multiple sound sources are reproduced at the same time. Though the above wave-domain algorithms are far more efficient than the corresponding frequency-domain algorithm, a further reduction of complexity is still required to handle massive multichannel systems.

We propose to further reduce the computational complexity of wave-domain adaptive filtering based on the analysis [7] [8] that a 2D Fourier transform of the sound field has a band-limited property and a region where the signal energy is almost zero.

2. WAVE-DOMAIN ADAPTIVE FILTERING

The wave-domain adaptive filtering proposed by Buchner et al. [5] is based on the temporal and spatial decomposition of multichannel echo paths using orthogonal basis functions. Due to the spatial orthogonality, a $P \times P$ MIMO echo path in the frequency domain can be modeled as only P decoupled SISO systems in the wave domain. Figure 1 shows a diagram of the wave-domain adaptive filtering for echo cancellation between linear arrays of loudspeakers and microphones.

Let the p-th reproduced signal of a P channel be denoted as $x(p,t)$. We use the following notations hereafter.

n : *index of frame*
p : *index of channel*
f : *index of temporal frequency*
s : *index of spatial frequency*

The details of this processing are as follows. Note that the signals and transfer functions in the wave domain are hereafter denoted by a line under the respective variables.

The time-domain echo replica is obtained using the overlap-add method via wave-domain processing. First, the P channel signals $x(p,t)$ are transformed to the frequency domain, and then its P channel elements at the frequency $\omega = f\Delta\omega$ are transformed to the wave domain as

$$\begin{aligned}&\left[\ \underline{X}_f(0,n)\cdots\underline{X}_f(S-1,n)\quad \underline{X}_f(-S,n)\cdots\underline{X}_f(-1,n)\ \right]\\ &\qquad = FFT\left(\left[\ X_f(1,n)\quad \ldots\quad X_f(P,n)\ \right]\right),\end{aligned} \tag{1}$$

where $S = P/2$ and $\Delta\omega$ is the resolution of temporal frequency. The wave-domain echo replica is estimated by using the SISO model [5]

$$\underline{\hat{Y}}_f(s,n) = \underline{H}_f(s,s,n)\underline{X}_f(s,n), \tag{2}$$

where s $(-S \le s \le S-1)$ corresponds to the spatial frequency $k_x = s\Delta k_x$ and Δk_x is the resolution of spatial frequency, or the modified MISO model [6]

$$\underline{\hat{Y}}_f(s,n) = \sum_{s'=s-\sigma}^{s+\sigma} \underline{H}_f(s,s',n)\underline{X}_f(s',n), \tag{3}$$

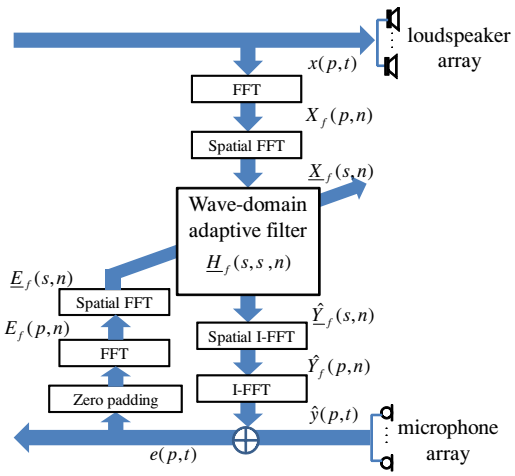

Fig. 1. Schematic diagram of conventional wave-domain adaptive filter.

where $\underline{H}_f(s,s',n)$ is the coefficient of the wave-domain adaptive filter and σ is the parameter of additional inputs. These echo replicas are transformed inversely to the time domain.

After subtracting the time-domain echo replicas from microphone signals, the zero-padded error signals are transformed to the wave domain as $\underline{E}_f(s,n)$. We can update coefficients of the wave-domain adaptive filter in the normalized least mean square (NLMS)-like manner as follows

$$\underline{H}_f(s,s',n+1) = \underline{H}_f(s,s',n) + \mu \frac{conj\left(\underline{X}_f(s',n)\right)\underline{E}_f(s,n)}{\underline{Z}_f(s,i)+\epsilon}, \quad (4)$$

where

$$\underline{Z}_f(s,n) = \beta \underline{Z}_f(s,n-1) + (1-\beta)\sum_{s'=s-\sigma}^{s+\sigma}\left|\underline{X}_f(s',n)\right|^2, \quad (5)$$

and μ is the step size, ϵ is the small constant for avoiding zero division, and β is the smoothing factor.

3. SPECTRUM OF SOUND FIELD

We review the spectral representation of the sound fields generated by plane waves [7] and a point sound source [8] when the Fourier transform is taken over time and space. First, consider a plane wave emitting all possible frequencies arriving with angle θ at the x axis corresponding to the linear microphone array. The space-time representation of the sound field along the x axis is given by $p_r(x,t) = \delta\left(t + x\cos\theta/c\right)$, where c is the speed of sound propagation. The space-time-frequency representation of this sound field, that is, the 2D Fourier transform, is given by using delta function $\delta(\,)$ as

$$P_r\left(k_x,\omega\right) = 2\pi\delta\left(k_x - \omega\cos\theta/c\right), \quad (6)$$

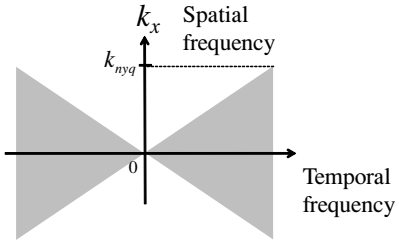

Fig. 2. 2D spectrum of sound field when plane waves emitting all possible frequencies arrive from all possible angles.

where ω is the temporal frequency in rad/s and k_x is the spatial frequency in rad/m.

When plane waves arrive from all possible angles, the temporal frequency ω and spatial frequency k_x satisfy $|k_x| \leq \omega/c$ since $-1 \leq \cos\theta \leq 1$. This 2D spectrum is shown as the triangular-shaped regions in Fig. 2.

Next, consider the sound field generated by a point sound source located at $(x,y,z) = (x_s,d,0)$. The space-time representation of its sound field along the x axis is given by

$$p(x,t) = \frac{\delta\left(t - \sqrt{(x-x_s)^2+d^2}/c\right)}{4\pi\sqrt{(x-x_s)^2+d^2}}. \quad (7)$$

Its space-time-frequency representation is given by

$$P_r(k_x,\omega) = -\frac{j}{4\pi}e^{-jk_x x_s}H_0^*\left(d\sqrt{\left(\frac{\omega}{c}\right)^2 - k_x^2}\right), \quad (8)$$

where H_0^* represents the complex conjugate of the zero-order Hankel function of the first kind.

We plot the cross sections of the magnitude of $P_r(k_x,\omega)$ in (8) along the spatial frequency in Fig. 3 for (a) $d = 5$ m and (b) $d = 0.5$ m. From this figure, we can see that for $d = 5$ m the spectrum along the spatial frequency is decaying very fast for all temporal frequencies, and that almost all energy is contained in the region $|k_x| \leq |\omega|/c$.

As for $d = 0.5$ m, we can see that the decay of the spectrum along the spatial frequency depends on the temporal frequency. It becomes slower for lower temporal frequencies. Hence, there remains energy corresponding to the evanescent mode of the waves in the region $|k_x| > |\omega|/c$.

The above analyses show that the 2D Fourier transform of the sound field has a band-limited property and that it has a region where the signal energy is almost zero.

4. COMPLEXITY REDUCTION IN 2D SPECTRUM

We propose to reduce the computational complexity of wave-domain adaptive filtering by using the band-limited property of the space-time-frequency representation of the sound field. We can reduce the computational complexity by omitting the wave-domain processing in the region where the signal energy is almost zero.

Figure 4 shows how we specify the target region of signal processing, where α is the factor for taking into account

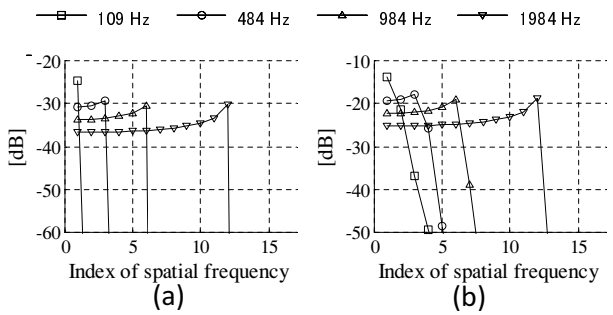

Fig. 3. Cross sections of 2D spectrum of sound field along spatial frequency axis generated by a point sound source: (a) $d = 5$ m and (b) $d = 0.5$ m.

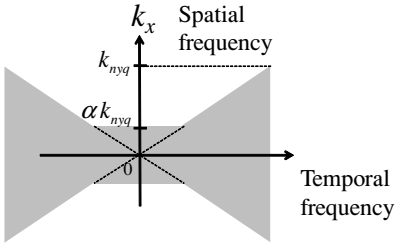

Fig. 4. Proposed target region of signal processing in 2D spectrum of sound field.

the slow decay of the spectrum along the spatial frequency for lower temporal frequencies. The target region is given as αk_{nyq}, where k_{nyq} is the spatial Nyquist frequency.

Since the linear microphone array is placed close to the linear loudspeaker array in full duplex WFS communication, the spectrum decay along the spatial frequency cannot be ignored, and the model with $\alpha > 0$ is more appropriate than the model with $\alpha = 0$ proposed in [9].

The wave-domain processing and its complexity can be reduced to $(1+\alpha^2)/2$ when the temporal and spatial Nyquist frequencies ω_{nyq} and k_{nyq} satisfy $\omega_{nyq} = k_{nyq}$. For example, $\alpha = 1/2$ reduces the wave-domain processing to 62.5% of the conventional one.

Next, we derived the overall complexity of the proposed adaptive filtering with a practical multi-delay structure [2] [10], where the processing delay is significantly reduced. It is given as a number of complex multiplications by

$$3P\frac{2F}{4}\log_2(2F) + 3F\frac{P}{2}\log_2(P) + \rho FP(2\sigma+1)I + \rho FP\left(1+3(2\sigma+1)I\right), \quad (9)$$

where $2F$-point temporal fast Fourier transform/inverse fast Fourier transform (FFT/I-FFT) and spatial P-point FFT/I-FFT are used, each wave-domain MISO model has $2\sigma+1$ input channels, the number of multi delays is given by I, and the reduction ratio ρ is given by $\rho = (1+\alpha^2)/2$. We also assume that the complexity of $2F$-point FFT of the real signal is $(2F/4)\log_2(2F)$, and that the complexity of $P-$point FFT of the complex signal is $(P/2)\log_2(P)$. Based on the symmetry that the FFT output of a real signal satisfies, the results for the higher F bands in the frequency domain are given as the complex conjugate of the lower F bands.

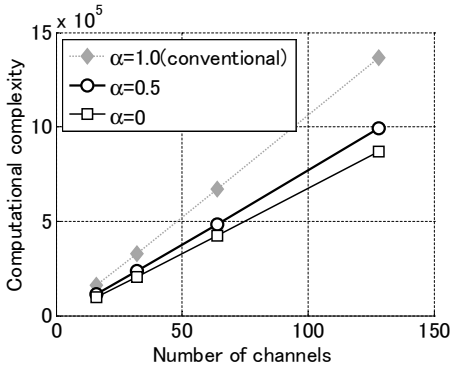

Fig. 5. Overall complexities of conventional and proposed wave-domain adaptive algorithm for various reduction ratios α.

Figure 5 plots the overall complexities of the conventional ($\alpha = 1$) and proposed wave-domain adaptive filtering ($\alpha = 0.5$ and 0), where $F = 128$, $\sigma = 2$, and $I = 3$. For $\alpha = 0$, the proposed method reduces the total complexity to 65%, and for $\alpha = 0.5$, the proposed method reduces the total complexity to 71%.

5. SIMULATION

We confirmed the validity of our proposed method by conducting a computer simulation of a 32-channel wave-domain echo canceller (P=32) with measured impulse responses. We investigated how close the behaviours of the conventional and proposed adaptive filtering were in single-talk conditions.

Figure 6 shows the configuration of the linear arrays of loudspeakers and microphones used in the simulation. We used white noise and voices as the source signals in the transmitting room. These signals were band-limited to 2.8 kHz to avoid aliasing due to spatial sampling by the linear microphone array. The distance from each source to the linear microphone array was 1 [m]. The 32-channel transmitting signals were generated by convolving each source signal with 32 impulse responses between the sound source and microphones in the transmitting room. These signals were transformed into the driving signals by the wave field reconstruction filter (WFRF) [11]. In the receiving room, the p-th microphone signal was obtained by computing the sound propagation of the driving signals from 32 loudspeakers to the p-th microphone with corresponding impulse responses. All impulse responses were measured at the sampling frequency of 8 kHz in an actual room with the reverberation time T_{60} = 300 ms and were truncated to 1536 taps.

We used the frame size $F = 128$ and an $F/2$ frame shift. We used $\sigma = 2$, that is, a 5-input 1-output system as the wave-

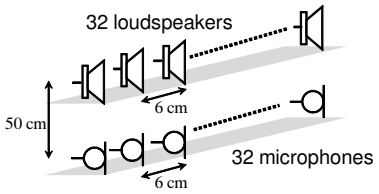

Fig. 6. Configuration of linear loudspeaker array and linear microphone array.

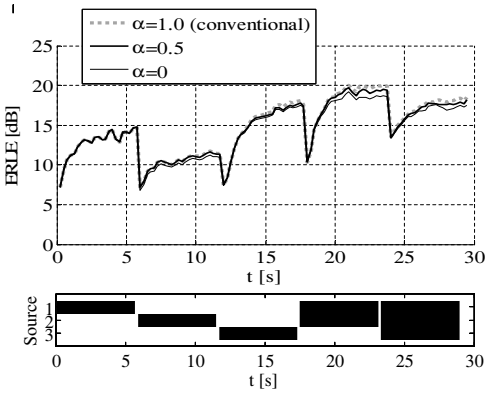

Fig. 7. Behaviour of channel 6 for white noise sources in single-talk conditions. Upper plot: ERLE of conventional (gray dotted line) and proposed methods with $\alpha = 0.5$ (thick line) and $\alpha = 0$ (thin line). Lower plot: activity of source signals; darker area indicates more signal power.

domain decoupled system. We used 256-point temporal FFT and 32-point spatial FFT. The step size was set to $\mu = 0.1$. The forgetting factor for normalization β was set to 0.85.

White noise sources

Figure 7 plots the echo return loss enhancement (ERLE) of the conventional wave-domain adaptive filtering (gray dotted line), and of the proposed method with $\alpha = 0.5$ (thick line) and $\alpha = 0$ (thin line). The activity of the sources is shown below the ERLE curves. Figure 8 plots the average ERLE of $t = 26$–29 s of each channel. In the proposed method

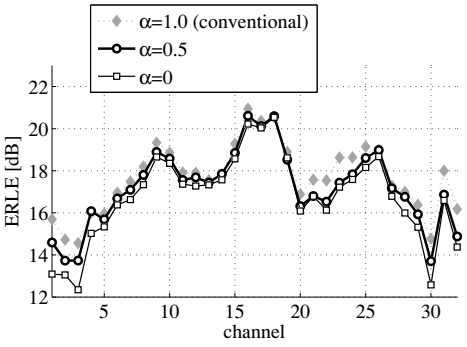

Fig. 8. Average ERLE between $t = 26$–29 s of each channel for white noise sources: conventional (gray dotted line) and proposed method with $\alpha = 0.5$ (thick line) and $\alpha = 0$ (thin line).

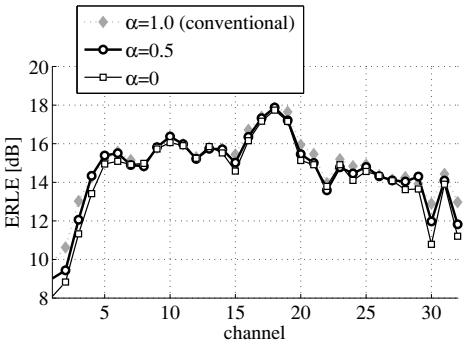

Fig. 9. Average ERLE between $t = 26$–29 s of each channel for voice sources: conventional (gray dotted line) and proposed method with $\alpha = 0.5$ (thick line) and $\alpha = 0$ (thin line).

with $\alpha = 0$, ERLE was degraded about 1 to 2 dB from the conventional method. By using $\alpha = 0.5$, this degradation was reduced (channels 1–20) or almost disappeared (channels 26–29).

Voice sources

We used voice sources whose activity pattern was similar to that of the white noise sources depicted in Fig. 7. Figure 9 plots the average ERLE of $t = 26$–29 s of each channel. In the proposed method with $\alpha = 0$, ERLE was degraded about 1 dB from the conventional method. By using $\alpha = 0.5$, this degradation was well reduced (channels 15–32) or almost disappeared (channels 1–7).

6. RELATION TO PRIOR WORK

The work presented here shows that the complexities of wave-domain adaptive filtering algorithms proposed by Bucher et al. [5] and Schneider et al. [6] can be further reduced by exploiting the property of the 2D Fourier transform of sound field analyzed by Ajdler et al. [7][8]. This work focused on the evanescent mode of the waves which was not considered in the wave-domain residual echo processing in [9].

7. CONCLUSION

We proposed a method to further reduce the computational complexity of wave-domain adaptive filtering by exploiting the property of the sound field in the space-time-frequency representation. A simulation using the measured impulse responses and voices revealed that, when the evanescent mode of waves was ignored, the proposed method was able to reduce the overall complexity to 65% of the conventional multi-delay wave-domain adaptive filtering with ERLE degradation of 1 dB. By taking the evanescent mode into account, it was shown that this degradation was well reduced on some channel and almost disappeared on other channels. The proposed method enables the adaptive filter to handle the echo of a massive-channel acoustic system more efficiently.

8. REFERENCES

[1] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. SAP*, vol. 6, pp. 156 –165, 1998.

[2] H. Buchner, J. Benesty, T. Gaensler, and W. Kellermann, "Robust extended multidelay filter and double-talk detector for acoustic echo cancellation," *IEEE Trans. ASLP*, vol. 14, pp. 1633 –1644, 2006.

[3] S. Emura, Y. Haneda, and S. Makino, "Enhanced frequency-domain adaptive algorithm for stereo echo cancellation," *Proc. ICASSP2002*, 2002.

[4] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, pp. 2764 –2778, 1993.

[5] H. Buchner, S. Spors, and W. Kellermann, "Wave-domain adaptive filtering: Acoustic echo cancellation for full-duplex systems based on wave-field synthesis," *Proc. ICASSP2004*, 2004.

[6] M. Schneider and W. Kellermann, "A wave-domain model for acoustic mimo systems with reduced complexity," *Joint Workshop on Hands-free Speech Communication and Microphone arrays*, 2011.

[7] T. Ajdler, L. Sbaiz, and M. Vetterli, "Dynamic measurement of room impulse responses using a moving microphone," *J. Acoust. Soc. Am.*, pp. 1636 –1645, 2007.

[8] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenoacoustic function and its sampling," *IEEE Trans SP*, pp. 3790 –3804, 2006.

[9] S. Emura, S. Koyama, K. Furuya, and Y. Haneda, "Posterior residual echo canceling and its complexity reduction in the wave domain," *Proc. IWAENC2012*, 2012.

[10] E. Moulines, O. Ait Amrane, and Y. Grenier, "The generalized multidelay adaptive filter: Structure and convergence analysis," *IEEE Trans. SP*, pp. 14 –28, 1995.

[11] S. Koyama, Y. Hiwasaki, K. Furuya, and Y. Haneda, "Design of transform filter for sound field reproduction using microphone array and loudspeaker array," *IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA2011)*, 2011.