PARAMETRIC MULTICHANNEL NOISE REDUCTION ALGORITHM UTILIZING TEMPORAL CORRELATIONS IN REVERBERANT ENVIRONMENT

Yu Gwang Jin, Jong Won Shin, Chul Min Lee, Soo Hyun Bae and Nam Soo Kim*

Department of Electrical and Computer Engineering and INMC
Seoul National University, Seoul, Korea
*School of Information and Communications
Gwangju Institute of Science and Technology, Gwangju, Korea
ygjin@hi.snu.ac.kr, jwshin@gist.ac.kr, {cmlee, shbae}@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

In this paper, we propose a parametric multichannel noise reduction algorithm utilizing temporal correlations in a noisy and reverberant environment. Under the reverberant condition, the received acoustic signal becomes highly correlated in the time domain and it makes successful noise reduction quite difficult. The proposed parametric noise reduction method takes account of interdependencies between components observed from different frames. Extended speech and noise power spectral density (PSD) matrices are estimated containing additional temporal information, and the parametric multichannel noise reduction filter based on these PSD matrices is applied to the input microphone array signal. According to the experimental results, the proposed algorithm has been found to show better performances compared with the conventional multiplicative filtering technique which considers the current input signals only.

Index Terms— Multichannel noise reduction, microphone array, parameterized non-causal multichannel Wiener filter, reverberant environment.

1. INTRODUCTION

It is well-known that the quality of a speech signal significantly deteriorated when additive noise is present in the background, and obviously it becomes much worse in many real situations with reverberation. Over the last few decades, several multichannel noise reduction approaches have been proposed [1]-[5]. In many approaches to multichannel noise reduction, the clean speech is estimated by multiplying a proper gain to the input signal, and the gain is obtained from the channel transfer function (TF) estimates between the speech source and the microphone array. Theoretically the utilization of multiple microphones could make effective noise reduction possible without much speech distortion when the exact channel TFs are known. In the real environments, however, it is quite difficult to estimate the unknown channel TFs, and inaccurate estimates lead to a serious degradation in noise reduction performance.

Recently, an optimal filtering technique which depends on the statistics of the speech and noise signals was proposed [4] with new simplified expressions of the parameterized multichannel non-causal Wiener filter (PMWF), the minimum variance distortionless response (MVDR) filter and the generalized sidelobe canceller (GSC). This approach estimates the clean speech component by multiplying a parametric filter to the input signal observed in the current frame without considering the temporal correlations at all. In case of a real reverberant environment, however, if the effective length of reverberation is longer than the frame length, there exist strong temporal correlations that help to estimate the clean speech from the received data. Hence it is possible to further improve the performance of the conventional noise reduction techniques by utilizing not only the spatial correlations but also the temporal correlations.

In this paper, we propose a parametric multichannel noise reduction algorithm utilizing the temporal correlations. In order to take advantage of the temporal correlations of the input signals, the proposed method performs a grouping of the noisy observations not only in the same frame but also in adjacent frames, and estimates the extended power spectral density (PSD) matrices of the speech and noise data including cross-correlations between grouped components. The parametric filter based on these PSD matrices is applied to the augmented input observations for multichannel noise reduction and interference rejection. Through a number of experiments, we have observed that the proposed parametric noise reduction approach shows performance that is superior to that of the noise reduction algorithm which considers the current input signals only.

This research was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2012R1A2A2A01045874) and the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2013-H0301-13-4005) supervised by the NIPA(National IT Industry Promotion Agency).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. PROBLEM STATEMENT

Let $y_i(k,t)$, $x_i(k,t)$ and $v_i(k,t)$ denote the short-time Fourier transform (STFT) coefficients of the noisy speech, clean speech and noise signal, respectively, for the k-th frequency bin at frame t observed from the i-th microphone. When an array of N microphones with an arbitrary geometry is applied, the STFT component of the noisy speech $y_i(k,t)$ is expressed as

$$y_i(k,t) = x_i(k,t) + v_i(k,t), \quad i = 1, 2, \cdots, N. \quad (1)$$

In many conventional approaches to multichannel noise reduction, the clean speech component in the i-th channel is estimated by multiplying a filter gain to the input microphone array signal in the following way:

$$\hat{x}_i(k,t) = \sum_{j=1}^{N} g_{ij}^*(k,t) y_j(k,t) = \mathbf{g}_i^H(k,t) \cdot \mathbf{y}(k,t) \quad (2)$$

where $\mathbf{g}_i(k,t) = [g_{i1}(k,t), \cdots, g_{iN}(k,t)]^T$ and $\mathbf{y}(k,t) = [y_1(k,t), \cdots, y_N(k,t)]^T$. The superscripts ^{H}, ^{T} and * denote the transpose-conjugate, transpose and conjugate operators, respectively. Additionally we use the notation $\hat{(\cdot)}$ to denote "the estimate of".

3. PARAMETRIC FILTERING FOR MULTICHANNEL NOISE REDUCTION UTILIZING TEMPORAL CORRELATIONS

3.1. Relation to prior work

To find a proper filter gains $\{\mathbf{g}_i(k,t)\}$ in (2) for noise reduction, many different approaches have been employed. Compared with the single channel based approaches, those filtering techniques using multiple microphones have achieved competitive advantage by utilizing additional information about the spatial properties of the speech and noise components. However, the multiplicative filtering method considers only the spatial correlation while ignoring the temporal correlations even though it is generally known that the actual speech signals have high level of temporal correlations. Moreover, temporal correlations of the received signals become stronger in a reverberant environment. Therefore it could be an important information to reduce noise components and provide accurate estimates of the clean speech.

In this paper, in order to utilize the temporal correlations of the input signals for multichannel noise reduction, we propose an extended parametric filtering technique as follows:

$$\begin{aligned}\hat{x}_i(k,t) &= \sum_{j=1}^{N} \sum_{\tau=0}^{L-1} g_{ij}^*(k,t,\tau) y_j(k,t-\tau) \\ &= \sum_{j=1}^{N} \bar{\mathbf{g}}_{ij}^H(k,t) \bar{\mathbf{y}}_j(k,t) = \mathbf{G}_i^H(k,t) \cdot \mathbf{Y}(k,t) \quad (3)\end{aligned}$$

where $\bar{\mathbf{g}}_{ij}(k,t) = [g_{ij}(k,t,L-1), \cdots, g_{ij}(k,t,0)]^T$, $\bar{\mathbf{y}}_j(k,t) = [y_j(k,t-L+1), \cdots, y_j(k,t)]^T$, $\mathbf{G}_i(k,t) = [\bar{\mathbf{g}}_{i1}(k,t), \cdots, \bar{\mathbf{g}}_{iN}(k,t)]^T$ and $\mathbf{Y}(k,t) = [\bar{\mathbf{y}}_1(k,t), \cdots, \bar{\mathbf{y}}_N(k,t)]^T$. Compared with a conventional multiplicative filter $\mathbf{g}_i(k,t)$ in (2), the dimension of the noise reduction filter $\mathbf{G}_i(k,t)$ is extended from N to NL when we take L neighbor frames from t to $t-L+1$ into account, and it is easily found that (3) is a generalized form of (2) which is a special case with $L = 1$.

3.2. Parametric noise reduction filtering

When $\mathbf{X}(k,t)$ and $\mathbf{V}(k,t)$ are obtained by stacking the speech components $\{x_i(k,t)\}$ and the noise components $\{v_i(k,t)\}$, respectively in the same order as $\mathbf{Y}(k,t)$ with $\{y_i(k,t)\}$, we can generalize the definition of the PSD matrices [4] in the following way:

$$\begin{aligned}\mathbf{\Phi}_{yy}(k,t) &\triangleq E\{\mathbf{Y}(k,t)\mathbf{Y}^H(k,t)\}, \\ \mathbf{\Phi}_{xx}(k,t) &\triangleq E\{\mathbf{X}(k,t)\mathbf{X}^H(k,t)\}, \\ \mathbf{\Phi}_{vv}(k,t) &\triangleq E\{\mathbf{V}(k,t)\mathbf{V}^H(k,t)\},\end{aligned} \quad (4)$$

where the dimension of each PSD matrix becomes $NL \times NL$. These PSD matrices contain additional temporal information while preserving spatial information of the conventional PSD matrices.

In this paper, we have extended and applied the PMWF, which includes GSC and MVDR beamformer as special cases [4], for multichannel noise reduction. The optimal filter gain $\mathbf{G}_i(k,t)$ is given by

$$\mathbf{G}_i(k,t) = [\widehat{\mathbf{\Phi}}_{xx}(k,t) + \beta \widehat{\mathbf{\Phi}}_{vv}(k,t)]^{-1} \widehat{\mathbf{\Phi}}_{xx}(k,t) \mathbf{u}_i \quad (5)$$

where $\mathbf{u}_i = [0 \cdots 0 \underbrace{1}_{(iL)-th} 0 \cdots 0]^T$ is an NL-dimensional vector and $\beta \geq 0$ is a factor that allows for tuning the noise reduction and speech distortion. It is noted that $\mathbf{G}_i(k,t)$ depends on $\widehat{\mathbf{\Phi}}_{xx}(k,t)$ and $\widehat{\mathbf{\Phi}}_{vv}(k,t)$ only, and an explicit estimation of the channel TFs in real environments is not required any longer. Consequently the clean speech component of the i-th channel $\hat{x}_i(k,t)$ is obtained by applying (5) to (3).

3.3. Estimation of speech and noise PSD matrices

Since the parametric noise reduction filter depends on the clean speech and noise PSD matrices, $\mathbf{\Phi}_{xx}(k,t)$ and $\mathbf{\Phi}_{vv}(k,t)$, accurate estimation of these PSD matrices is quite important for successful noise reduction. For robust estimation of the statistics of signals, we update the estimates for $\mathbf{\Phi}_{yy}(k,t)$ and $\mathbf{\Phi}_{vv}(k,t)$ recursively as follows:

$$\begin{aligned}\widehat{\mathbf{\Phi}}_{yy}(k,t) &= \alpha_y(k,t) \widehat{\mathbf{\Phi}}_{yy}(k,t-1) \\ &\quad + [1 - \alpha_y(k,t)] \mathbf{Y}(k,t) \mathbf{Y}^H(k,t), \quad (6)\end{aligned}$$

$$\begin{aligned}\widehat{\mathbf{\Phi}}_{vv}(k,t) &= \widetilde{\alpha}_v(k,t) \widehat{\mathbf{\Phi}}_{vv}(k,t-1) \\ &\quad + [1 - \widetilde{\alpha}_v(k,t)] \mathbf{Y}(k,t) \mathbf{Y}^H(k,t) \quad (7)\end{aligned}$$

where $0 \leq \alpha_y(k,t) \leq 1$ and $0 \leq \widetilde{\alpha}_v(k,t) \leq 1$ are two forgetting factors.

Contrary to $\alpha_y(k,t)$ which is usually fixed to a constant value, $\widetilde{\alpha}_v(k,t)$ should be regard as a time-varying smoothing factor depending on the speech presence probability (SPP). In order to control the noise tracking speed according to whether the speech is present or absent, $\widetilde{\alpha}_v(k,t)$ is updated as

$$\widetilde{\alpha}_v(k,t) = \alpha_v(k,t) + (1 - \alpha_v(k,t))p(k,t) \tag{8}$$

where $0 \leq \alpha_v(k,t) \leq 1$. The multichannel SPP $p(k,t)$ in (8) is obtained by [6] under the assumption that the speech and noise components are multivariate Gaussian and their real and imaginary parts are uncorrelated and identically distributed.

In a single channel scenario, the minima controlled recursive averaging (MCRA) algorithm [7] is one of the most popular approaches to estimate noise statistics in adverse conditions, and recently it is generalized to the multichannel case [5]. In this work, therefore, the multichannel MCRA approach is applied to track the time-varying noise PSD matrix in (7).

Finally, the PSD matrix of clean speech is obtained as the following form [4], [5], [6]

$$\widehat{\mathbf{\Phi}}_{xx}(k,t) = \widehat{\mathbf{\Phi}}_{yy}(k,t) - \widehat{\mathbf{\Phi}}_{vv}(k,t) \tag{9}$$

under the assumption that the speech and noise components are uncorrelated.

4. EXPERIMENTAL RESULTS

In order to verify the performance of the proposed parametric multichannel noise reduction algorithm, a number of objective quality measurements were performed under various noisy conditions with different values of L, the number of previous frames we took account of. We considered a simulation setup in which a target speech and the interference sources were located in a reverberant room with dimensions of 6.7 m $\times$ 6.1 m $\times$ 2.9 m. The image method [8], [9] was used to generate the impulse responses for the room with the reverberation time T_{60} = 300 ms. The speech and the interference sources were located at (1.737 m, 4.6 m, 1.4 m) and (3.337 m, 4.6 m, 1.4 m), respectively, and we adopted a scenario with two microphones which are placed at (2.437 m, 5.6 m, 1.4 m) and (2.637 m, 5.6 m, 1.4 m). The test material consisted of ten utterances from the TIMIT database which were sampled at 16 kHz, and corrupted by three different types of interference signals. For the experiments, babble, factory and F-16 noises from the NOISEX-92 database were applied with 0, 5, 10, 15 and 20 dB signal-to-interference ratio (SIR). The multichannel noise reduction filter in (3) was implemented with various values of L, and obviously the case of $L = 1$ is equivalent to the conventional multiplicative filtering method. The update factor for the noise PSD matrix was set to $\alpha_v(k,t) = 0.99$ to implement the proposed algorithm, and the values of $\alpha_y(k,t)$ and β were experimentally determined depending on L and an estimate of the input SIR.

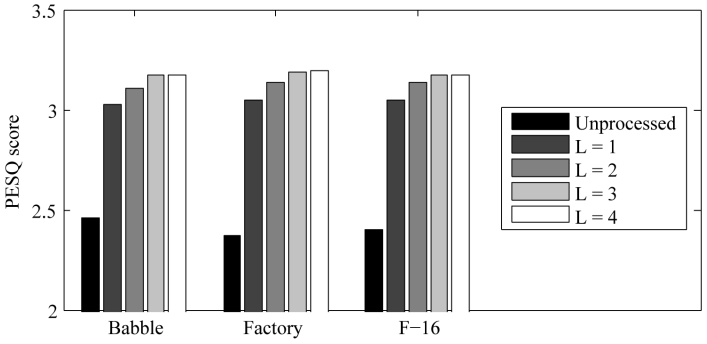

Fig. 1. Results of the perceptual evaluation of speech quality under different noisy conditions with various values of L.

Table 1. Results of the cepstrum distance and the output SIR under different noisy conditions with various values of L.

[dB]	Noise type	Unprocessed	L			
			1	2	3	4
CD	Babble	6.57	6.11	5.96	5.90	5.88
	Factory	7.08	6.62	6.39	6.29	6.26
	F-16	7.29	6.70	6.49	6.39	6.35
SNR_o	Babble	10.37	20.24	21.53	22.22	22.59
	Factory	11.75	22.14	23.61	24.12	24.75
	F-16	11.45	22.93	24.37	25.44	25.70

In order to evaluate the performance of the proposed multichannel parametric filtering approach, we calculated the perceptual evaluation of speech quality (PESQ) [10], the cepstrum distance (CD) [11] and the output SIR (SIR_o) [4], [12] which is defined by

$$\begin{aligned} SIR_o &\triangleq \frac{E\{||\mathbf{X}_{filtered}(k,t)||^2\}}{E\{||\mathbf{V}_{residual}(k,t)||^2\}} \\ &= \frac{E\{\mathbf{G}_i{}^H(k,t)\mathbf{X}(k,t)\mathbf{X}^H(k,t)\mathbf{G}_i(k,t)\}}{E\{\mathbf{G}_i{}^H(k,t)\mathbf{V}(k,t)\mathbf{V}^H(k,t)\mathbf{G}_i(k,t)\}}. \end{aligned} \tag{10}$$

The results of PESQ obtained under different noisy conditions with various values of L are shown in Fig. 1, and the results of CD and SIR_o are summarized in Table 1. All experimental results are averaged over all of the input SIRs. From the results in Fig. 1, we can see that the multichannel noise reduction performance which is measured in terms of PESQ improved as the value of L increased, and the proposed approach outperformed the case with $L = 1$, which is equivalent to the conventional multiplicative algorithm. Furthermore as seen in Table 1, the noise reduction performance which is measured in terms of CD and SIR_o also improved as

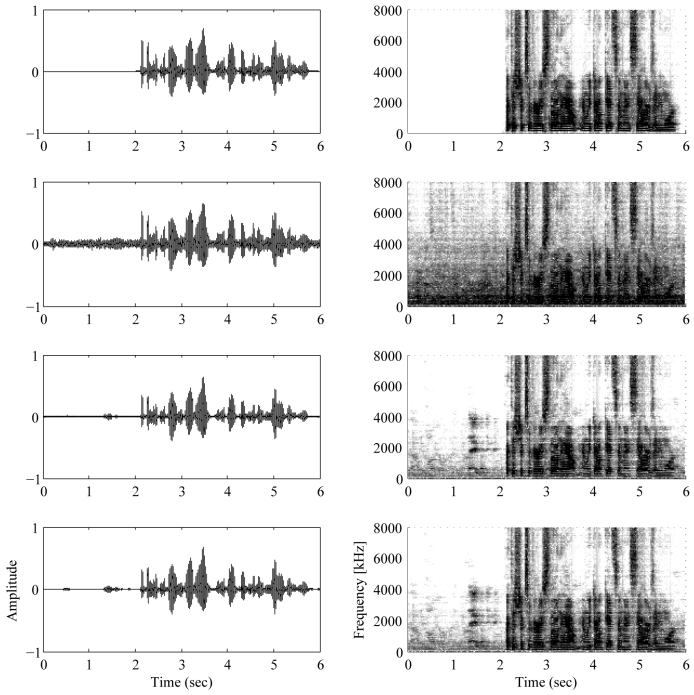

Fig. 2. Waveform and spectrogram of the (a) first microphone noise-free speech, (b) speech corrupted with babble noise (SIR = 10 dB) (c) output of the conventional multiplicative filter (L = 1), and (d) output of the proposed filter utilizing temporal correlation (L = 4).

the value of L increased, and the proposed filtering technique showed better quality than the conventional one.

For the convenience of comparing the experimental results, we show the waveforms and spectrograms of an example of the noise-free, noisy, and filtered signals in Fig. 2. From the result we can see that the proposed approach slightly improved the performance of multichannel noise reduction in the reverberant environments.

5. CONCLUSIONS

In this paper, we have proposed a parametric filtering algorithm for multichannel noise reduction utilizing temporal correlations. In contrast to the conventional multiplicative filtering method, the proposed filtering technique considers correlations between signal components in adjacent frames. An extended form of speech and noise PSD matrices is introduced, and their estimates are recursively updated for successful noise reduction. To reject interferences from the received data, the parametric filter based on these PSD matrices is applied to the augmented input observations. Performances of the proposed approach have been evaluated by a number of objective quality measurements under various conditions, and experimental results have shown that the proposed algorithm outperforms the conventional multiplicative filtering technique not utilizing temporal correlations at all.

6. REFERENCES

[1] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. AP-30, no. 1, pp. 27-34, Jan. 1982.

[2] S. Gannot, D. Burstein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614-1626, Aug. 2001.

[3] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, Y. Huang, and M. M. Sondhi, Eds. New York: Springer-Verlag, 2007, ch. 47, pp. 945-978.

[4] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260-276, Feb. 2010.

[5] M. Souden, J. Chen, J. Benesty, and S. Affes, "An Integrated Solution for Online Multichannel Noise Tracking and Reduction" *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159-2169, Sep. 2011.

[6] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1072-1077, Jul. 2010.

[7] I. Cohen, "Noise Spectrum Estimation in Adverse Environments : Improved Minima Controlled Recursive Averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466-475, Sep. 2003.

[8] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943-950, Apr. 1979.

[9] E. A. Lehmann, "Image-source method for room acoustics," http://www.eric-lehmann.com/ism_code.html.

[10] ITU, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Rec. P.862, 2000.

[11] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Boca Raton, CRC press, 2007.

[12] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Berliln, Germany: Springer-Verlag, 2008.