A GMM post-filter for residual crosstalk suppression in blind source separation

Khong, Andy Wai Hoong; Liu, Benxu; Reju, Vaninirappuputhenpurayil Gopalan; Reddy, Vinod Veera

2014

Liu, B., Reju, V. G., Khong, A. W. H., & Reddy, V. V. (2014). A GMM Post-Filter for Residual Crosstalk Suppression in Blind Source Separation. IEEE Signal Processing Letters, 21(8), 942-946.

https://hdl.handle.net/10356/79667

https://doi.org/10.1109/LSP.2014.2317761

A GMM Post-Filter for Residual Crosstalk Suppression in Blind Source Separation

Benxu Liu, Vaninirappuputhenpurayil Gopalan Reju, *Member, IEEE,* Andy W. H. Khong, *Member, IEEE,* and Vinod Veera Reddy *Member, IEEE,*

***Abstract*—Existing algorithms employ the Wiener filter to suppress residual crosstalk in the outputs of blind source separation algorithms. We show that, in the context of BSS, the Wiener filter is optimal in the maximum likelihood (ML) sense only for normally-distributed signals. We then propose to model the distribution of speech signals using the Gaussian mixture model (GMM) and then derive a post-filter in the ML sense using the expectation-maximization algorithm. We show that the GMM introduces a probabilistic sample weight that is able to emphasize speech segments that are free of crosstalk components in the BSS output and this results in a better estimate of the post-filter. Simulation results show that the proposed post-filter achieves better crosstalk suppression than the Wiener filter for BSS.**

***Index Terms*—Blind source separation, residual crosstalk suppression, Gaussian mixture model, maximum likelihood, expectation-maximization**

I. Introduction

Blind source separation (BSS) refers to the recovery of source signals from their mixtures without any knowledge of the mixing process or the source signals [1]–[5]. Although existing convolutive BSS techniques can achieve effective separation under low reverberant conditions, their performance may deteriorate due to late reflection components of the room impulse responses [1]. For a scenario with two speech sources and two microphones, the *partially separated* BSS outputs can be expressed as

$$\begin{aligned} y_1(n) &= g_{1,1} * s_1(n) + g_{1,2} * s_2(n), \\ y_2(n) &= g_{2,1} * s_1(n) + g_{2,2} * s_2(n), \end{aligned} \tag{1}$$

where $g_{p,q} * s_q(n)$ is the filtered version of $s_q(n)$ contained in the pth BSS output $y_p(n)$, $1 \leq p \leq 2$ and $1 \leq q \leq 2$. Note that partial separation implies that one of the sources is dominant in each of the partially separated outputs. We therefore assume $\widetilde{s}_1(n) \triangleq g_{1,1} * s_1(n)$ is dominant in $y_1(n)$ while $\widetilde{s}_2(n) \triangleq g_{2,2} * s_2(n)$ is dominant in $y_2(n)$. Defining the crosstalk filter as h_q such that $g_{1,2} = h_2 * g_{2,2}$ and $g_{2,1} = h_1 * g_{1,1}$, (1) can be reformulated as

$$\begin{aligned} y_1(n) &= \widetilde{s}_1(n) + h_2 * \widetilde{s}_2(n), \\ y_2(n) &= h_1 * \widetilde{s}_1(n) + \widetilde{s}_2(n). \end{aligned} \tag{2}$$

The problem is therefore to suppress the crosstalk components $h_q * \widetilde{s}_q(n)$ using only $y_1(n)$ and $y_2(n)$ so as to improve the separation performance.

The authors are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore - 639798, e-mail: LIUB0008@e.ntu.edu.sg, {reju, AndyKhong}@ntu.edu.sg, vinodreddy@pmail.ntu.edu.sg.

Post-filters have been proposed for the effective suppression of these crosstalk components [6]–[10]. These techniques decompose (2) into two adaptive noise cancelation (ANC) problems as

$$\begin{aligned} y_1(n) &= \widetilde{s}_1(n) + h_2 * \widetilde{s}_2(n), \\ y_2(n) &\approx \widetilde{s}_2(n), \end{aligned} \tag{3}$$

and

$$\begin{aligned} y_2(n) &= h_1 * \widetilde{s}_1(n) + \widetilde{s}_2(n), \\ y_1(n) &\approx \widetilde{s}_1(n). \end{aligned} \tag{4}$$

The unknown filter h_q can then be estimated using the well-known Wiener filter [6]–[8], [11] or its adaptive least-mean-square (LMS) version [9]. As will be explained in Section II, it is important to note that the Wiener filter is optimal in the maximum likelihood (ML) sense only for normally distributed $\widetilde{s}_q$. This implies that for non-Gaussian speech signals, the use of Wiener filter at the outputs of BSS algorithms may not be well-suited. In this work, we propose to describe the non-Gaussian speech signals using the Gaussian mixture model (GMM) and proceed to develop an expectation-maximization (EM) algorithm for estimating the corresponding filter h_q in the ML-sense. Note that existing works related to GMM based filters assume the ANC reference input to be known [12], [13]. However, in the case of BSS post-filtering application, the reference input is not fully known and only an approximation is available due to partial separation effect. As will be shown in Section III, this work provides a theoretical analysis to show that the probabilistic weight introduced in the GMM filter can enhance its robustness against any interfering residual crosstalk present in the reference input to ANC.

II. Maximum Likelihood Estimation of h_q for Normally Distributed Signals

We discuss the estimation of h_2 for the first sub-problem (3), while the approach holds true for the estimation of h_1 as well. Denoting $\widehat{h}_2$ as the estimate of h_2, the refined estimate of the source s_1 can be obtained as

$$\begin{aligned} \widetilde{s}_1(n) &\approx y_1(n) - \widehat{h}_2 * y_2(n) \\ &\approx \widetilde{s}_1(n) - (\widehat{h}_2 - h_2) * \widetilde{s}_2(n). \end{aligned} \tag{5}$$

The above implies that a good estimate of h_2 is essential for the performance of crosstalk suppression.

We next focus on obtaining the ML estimate of h_2 for (3) [14]. In order to establish the relationship between the ML estimate for normally distributed signals and the conventional Wiener solution, we assume that $\widetilde{s}_1(n)$ is drawn from a zero

mean Gaussian distribution with variance σ_1^2 such that its probability density function (pdf) is defined as

$$p(\widetilde{s}_1(n)) = \mathcal{N}(\widetilde{s}_1(n); 0, \sigma_1^2), \tag{6}$$

where $\mathcal{N}(s;\mu,\sigma^2) = (1/\sqrt{2\pi\sigma^2})\exp\left(-(s-\mu)^2/(2\sigma^2)\right)$. Denoting the L-length filter $\mathbf{h}_2 = [h_2(0), h_2(1), \ldots, h_2(L-1)]^T$ and $\mathbf{y}_2(n) = [y_2(n), y_2(n-1), \ldots, y_2(n-L+1)]^T$, the probability of $y_1(n)$ conditioned on $\mathbf{h}_2$ and $\mathbf{y}_2(n)$ can then be formulated as

$$p(y_1(n)|\mathbf{h}_2, \mathbf{y}_2(n)) = \mathcal{N}\left(y_1(n); \mathbf{h}_2^T\mathbf{y}_2(n), \sigma_1^2\right). \tag{7}$$

The corresponding log-likelihood of $\mathbf{h}_2$ for observing $y_1(n)$ over N i.i.d. snapshots can then be expressed as

$$\log\mathcal{L}(\mathbf{h}_2) = \sum_{n=1}^{N}\left(\mathbf{h}_2^T\mathbf{y}_2(n)y_1(n) - \frac{1}{2}\mathbf{h}_2^T\mathbf{y}_2(n)\mathbf{y}_2^T(n)\mathbf{h}_2\right) + c, \tag{8}$$

where c contains terms that are independent of $\mathbf{h}_2$. The ML estimate of h_2 can be obtained by solving $\partial\log\mathcal{L}(\mathbf{h}_2)/\partial\mathbf{h}_2 = \mathbf{0}$. The corresponding solution is

$$\widehat{\mathbf{h}}_2^{\mathrm{GD}} = \mathbf{R}_{\mathbf{y}_2\mathbf{y}_2}^{-1}\mathbf{r}_{\mathbf{y}_2 y_1}, \tag{9}$$

where the superscript $^{\mathrm{GD}}$ denotes for the normally distributed signal and

$$\mathbf{R}_{\mathbf{y}_2\mathbf{y}_2} = \sum_{n=1}^{N}\mathbf{y}_2(n)\mathbf{y}_2^T(n), \tag{10}$$

$$\mathbf{r}_{\mathbf{y}_2 y_1} = \sum_{n=1}^{N}\mathbf{y}_2(n)y_1(n). \tag{11}$$

It is interesting to note that the solution obtained by (9) is identical to the Wiener solution [14]. However, since speech signals are generally non-Gaussian distributed, the corresponding log-likelihood expression will not be of the form shown in (8). Therefore in the above case, the Wiener solution will not correspond to the ML estimate.

III. The Proposed GMM Post-Filter

It is well-known that any non-Gaussian distribution can be approximated using the Gaussian mixture model (GMM) [15]. We define the distribution of $\widetilde{s}_1(n)$ using a K-order GMM as

$$p(\widetilde{s}_1(n)) = \sum_{k=1}^{K}\alpha_{1k}\mathcal{N}(\widetilde{s}_1(n); 0, \sigma_{1k}^2), \tag{12}$$

where α_{1k} is the weight and σ_{1k}^2 is the variance of the kth Gaussian component. All the components are assumed to be zero mean. The probability of $y_1(n)$ conditioned on $\mathbf{h}_2$ and $\mathbf{y}_2(n)$ is then given by

$$p\left(y_1(n)|\mathbf{h}_2, \mathbf{y}_2(n)\right) = \sum_{k=1}^{K}\alpha_{1k}\mathcal{N}\left(y_1(n); \mathbf{h}_2^T\mathbf{y}_2(n), \sigma_{1k}^2\right). \tag{13}$$

Denoting the set of unknown parameters as $\theta = \{\mathbf{h}_2, \alpha_{1k}, \sigma_{1k}^2, 1 \le k \le K\}$, the likelihood of θ over N i.i.d. snapshots can be written as

$$\mathcal{L}^{\mathrm{GMM}}(\theta) = \prod_{n=1}^{N}\left(\sum_{k=1}^{K}\alpha_{1k}\mathcal{N}\left(y_1(n); \mathbf{h}_2^T\mathbf{y}_2(n), \sigma_{1k}^2\right)\right). \tag{14}$$

Note that since $\partial\log\mathcal{L}^{\mathrm{GMM}}(\theta)/\partial\mathbf{h}_2$ cannot be expressed by a closed-form solution, the EM algorithm is therefore exploited to solve for the ML estimate of $\mathbf{h}_2$.

The EM algorithm iteratively achieves the ML estimate of θ by assuming the existence of the missing data [15]. Similar to [15], this data is denoted by $z_k(n)$ and its value corresponds to whether $\widetilde{s}_1(n)$ is generated by the kth GMM component $\mathcal{N}_k$, i.e.,

$$z_k(n) = \begin{cases} 1, & \text{if } \widetilde{s}_1(n) \text{ is generated by } \mathcal{N}_k, \\ 0, & \text{others.} \end{cases} \tag{15}$$

Denoting $\mathcal{Z} = \{z_k(n), 1 \le n \le N, 1 \le k \le K\}$ and $\mathcal{Y}_1 = \{y_1(n), 1 \le n \le N\}$, the probability of the complete data $\{\mathcal{Y}_1, \mathcal{Z}\}$ can then be expressed as

$$p(\mathcal{Y}_1, \mathcal{Z}|\theta) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left(\alpha_{1k}\mathcal{N}(y_1(n); \mathbf{h}_2^T\mathbf{y}_2(n), \sigma_{1k}^2)\right)^{z_k(n)}, \tag{16}$$

while the log-likelihood of θ for the complete data is

$$\log\mathcal{L}_{\mathrm{compl}}^{\mathrm{GMM}}(\theta) = \sum_{n=1}^{N}\sum_{k=1}^{K}z_k(n)\Big(\log\alpha_{1k} + \log\mathcal{N}(y_1(n); \mathbf{h}_2^T\mathbf{y}_2(n), \sigma_{1k}^2)\Big). \tag{17}$$

The E-step in the iterative EM algorithm finds the expectation of (17) conditioned on $\mathcal{Y}_1$ and the previous estimate of θ given by $\theta^{(l-1)} = \{\widehat{\mathbf{h}}_2^{\mathrm{GMM}^{(l-1)}}, {\alpha_{1k}}^{(l-1)}, {\sigma_{1k}^2}^{(l-1)}\}$. This results in

$$\begin{aligned}\mathcal{Q}(\theta, \theta^{(l-1)}) &= E\left[\log\mathcal{L}_{\mathrm{compl}}^{\mathrm{GMM}}(\theta)|\mathcal{Y}_1, \theta^{(l-1)}\right] \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_k^{(l)}(n)\Big(\log\alpha_{1k} - \frac{1}{2}\log\sigma_{1k}^2 \\ &\quad - \frac{1}{2}\frac{(y_1(n) - \mathbf{h}_2^T\mathbf{y}_2(n))^2}{\sigma_{1k}^2}\Big) + c,\end{aligned} \tag{18}$$

where $\gamma_k^{(l)}(n) = E\left[z_k(n)|\mathcal{Y}_1, \theta^{(l-1)}\right]$. The variable $\gamma_k^{(l)}(n)$ can be computed using Bayes' theorem as

$$\gamma_k^{(l)}(n) = \frac{{\alpha_{1k}}^{(l-1)}\mathcal{N}\left(y_1(n); \mathbf{y}_2^T(n)\widehat{\mathbf{h}}_2^{\mathrm{GMM}^{(l-1)}}, {\sigma_{1k}^2}^{(l-1)}\right)}{\sum_{k'=1}^{K}{\alpha_{1k'}}^{(l-1)}\mathcal{N}\Big(y_1(n); \mathbf{y}_2^T(n)\widehat{\mathbf{h}}_2^{\mathrm{GMM}^{(l-1)}}, {\sigma_{1k'}^2}^{(l-1)}\Big)}. \tag{19}$$

The M-step in the lth iteration is to find a new θ which maximizes the conditional expectation $\mathcal{Q}(\theta, \theta^{(l-1)})$, i.e.,

$$\theta^{(l)} = \max_{\theta}\mathcal{Q}(\theta, \theta^{(l-1)}). \tag{20}$$

Using the Lagrange method, the refined values of α_{1k} and σ_{1k}^2 maximizing $\mathcal{Q}(\theta, \theta^{(l-1)})$ can be estimated using

$$\alpha_{1k}^{(l)} = \frac{1}{N}\sum_{n=1}^{N}\gamma_k^{(l)}(n) \tag{21}$$

and

$${\sigma_{1k}^2}^{(l)} = \frac{\sum_{n=1}^{N}\gamma_k^{(l)}(n)\left(y_1(n) - \mathbf{y}_2^T(n)\widehat{\mathbf{h}}_2^{\mathrm{GMM}^{(l-1)}}\right)^2}{\sum_{n=1}^{N}\gamma_k^{(l)}(n)}, \tag{22}$$

respectively. The refined value of $\mathbf{h}_2$ can be obtained by solving $\partial\mathcal{Q}/\partial\mathbf{h}_2 = \mathbf{0}$ which results in

$$\widehat{\mathbf{h}}_2^{\mathrm{GMM}^{(l)}} = \left(\mathbf{R}_{\mathbf{y}_2\mathbf{y}_2}^{\mathrm{GMM}}\right)^{-1}\mathbf{r}_{\mathbf{y}_2 y_1}^{\mathrm{GMM}}, \tag{23}$$

where

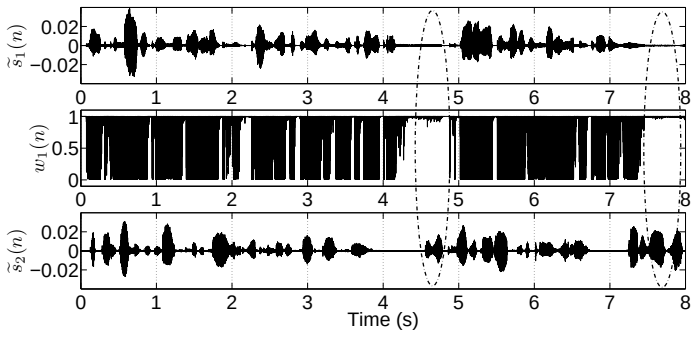

Fig. 1. Plots of $\widetilde{s}_1(n)$, estimated $w_1(n)$ and $\widetilde{s}_2(n)$.

$$\mathbf{R}_{\mathbf{y}_2\mathbf{y}_2}^{\mathrm{GMM}} = \sum_{n=1}^{N} w_1(n)\left[\mathbf{y}_2(n)\mathbf{y}_2^T(n)\right], \tag{24}$$

$$\mathbf{r}_{\mathbf{y}_2 y_1}^{\mathrm{GMM}} = \sum_{n=1}^{N} w_1(n)\left[\mathbf{y}_2(n) y_1(n)\right], \tag{25}$$

$$w_1(n) = \sum_{k=1}^{K} \frac{\gamma_k^{(l)}(n)}{\sigma_{1k}^{2\ (l)}}. \tag{26}$$

Note that, compared with (9), the proposed GMM filter in (23) weighs the ANC inputs with $w_1(n)$, and hence it can be considered as a generalized Wiener filter. Since $w_1(n)$ varies with n, the inverse operation in (23) will not cancel the effect of $w_1(n)$. The proposed method is expected to be robust against the approximation in (3) if $w_1(n)$ varies inversely with the magnitude of $\widetilde{s}_1(n)$. This ensures that the segments over which the reference signal is free of crosstalk will be emphasized for post-filter estimation. Even though (26) defines an inverse relationship between $w_1(n)$ and σ_{1k}^2, the instantaneous values of $\widetilde{s}_1(n)$ varies according to (12) for a given σ_{1k}^2. We proceed to prove that $w_1(n)$ is indeed inversely related to the magnitude of the instantaneous values of $\widetilde{s}_1(n)$.

Proof: For clarity of presentation, we remove iteration index $^{(l)}$ and assume $K = 2$ while the same logic can be extended to higher value of K. For $K = 2$, (26) can be reformulated as

$$\begin{aligned} w_1(n) &= \frac{\gamma_1(n)}{\sigma_{11}^2} + \frac{\gamma_2(n)}{\sigma_{12}^2} \\ &= \frac{1}{\sigma_{12}^2} - \frac{\sigma_{11}^2 - \sigma_{12}^2}{\sigma_{11}^2\sigma_{12}^2}\gamma_1(n), \quad \text{since} \sum_{k=1}^{2}\gamma_k(n) = 1. \end{aligned} \tag{27}$$

Without loss of generality, we assume $\sigma_{11}^2 > \sigma_{12}^2$. Using (19) and since $\widetilde{s}_1(n) = y_1(n) - \mathbf{h}_2^T\mathbf{y}_2(n)$, we have

$$\begin{aligned} \gamma_1(n) &= \frac{\alpha_{11}\mathcal{N}\left(\widetilde{s}_1(n); 0, \sigma_{11}^2\right)}{\alpha_{11}\mathcal{N}\left(\widetilde{s}_1(n); 0, \sigma_{11}^2\right) + \alpha_{12}\mathcal{N}\left(\widetilde{s}_1(n); 0, \sigma_{12}^2\right)} \\ &= \frac{\alpha_{11}}{\alpha_{11} + \alpha_{12}\frac{\sigma_{11}}{\sigma_{12}}\exp\left(-\frac{\sigma_{11}^2-\sigma_{12}^2}{\sigma_{11}^2\sigma_{12}^2}\frac{\widetilde{s}_1^2(n)}{2}\right)}. \end{aligned} \tag{28}$$

This implies that if $|\widetilde{s}_1(n_1)| > |\widetilde{s}_1(n_2)|$, we have $\gamma_1(n_1) > \gamma_1(n_2)$, which in turn leads to $w_1(n_1) < w_1(n_2)$ according to (27). Hence a larger $|\widetilde{s}_1(n)|$ implies a lower $w_1(n)$. □

We next illustrate this inverse relationship by way of simulation. Figure 1 shows $\widetilde{s}_1(n)$, the estimated $w_1(n)$ and $\widetilde{s}_2(n)$ in a simulation. Details of this simulation setup will be provided in the first two paragraphs of Section IV. Since (23) is invariant to the scale of $w_1(n)$, it has been scaled to a maximum value of one for clarity of presentation. It can be seen that

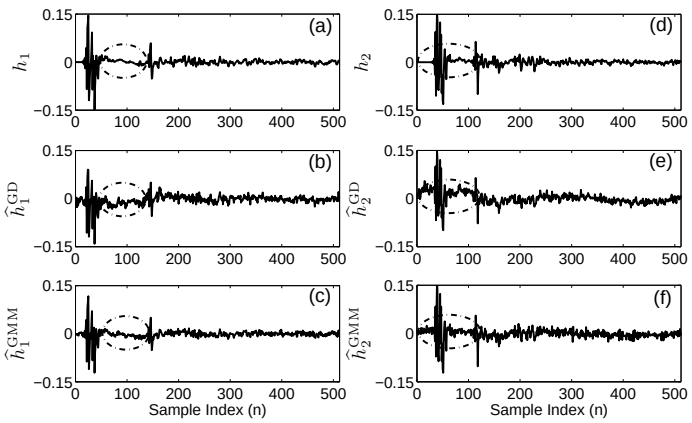

Fig. 2. The true h_q and estimated post-filter coefficients using Wiener method ($\widehat{h}_q^{\mathrm{GD}}$) and the proposed GMM method ($\widehat{h}_q^{\mathrm{GMM}}$).

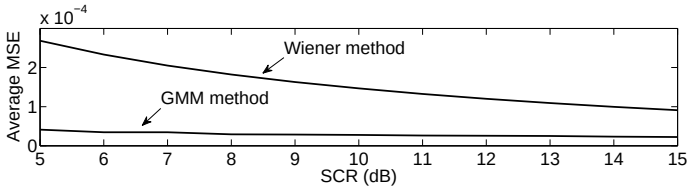

Fig. 3. Average MSE in $\widehat{h}_q$ estimated using the Wiener and GMM methods for SCR = 5 to 15 dB.

$w_1(n)$ varies inversely with the magnitude of $\widetilde{s}_1(n)$, especially for the highlighted regions where $\widetilde{s}_1(n) \approx 0$ corresponds to $w_1(n) \approx 1$. Note that the EM algorithm requires initialization of $\alpha_{qk}{}^{(0)}$, $\sigma_{qk}^{2\ (0)}$ and $\widehat{\mathbf{h}}_2^{\mathrm{GMM}^{(0)}}$. Since $\widetilde{s}_q(n)$ is dominant in $y_q(n)$, we initialize $\alpha_{qk}{}^{(0)}$ and $\sigma_{qk}^{2\ (0)}$ as the resulting weights and variances obtained by fitting y_q into a K-order GMM model while $\widehat{\mathbf{h}}_2^{\mathrm{GMM}^{(0)}}$ is initialized as $\mathbf{0}$.

While the proposed algorithm has been presented for the two-source case, the derivation can be extended for multiple sources. For $Q > 2$ sources, (5) can be generalized as

$$\widetilde{s}_q(n) \approx y_q(n) - \sum_{i=1, i\neq q}^{Q} h_{q,i} * y_i(n), \tag{29}$$

where $h_{q,i}$ denotes the crosstalk filter between $\widetilde{s}_i(n)$ and $y_q(n)$. Equation (29) shows that $\widetilde{s}_q(n)$ is approximated using the filtered sum of all partially separated signals $y_q(n)$, $q = 1, \cdots, Q$. This implies, according to central limit theorem, that as the number of sources and/or room reverberation time increases, the non-Gaussianity of $\widetilde{s}_q(n)$ will reduce and the performance of the proposed algorithm will converge to that of the conventional Wiener filter approach.

IV. Simulation Results

Three simulations are conducted to compare the performance of the proposed GMM filter and conventional Wiener filter described by (23) and (9), respectively. Since the proposed algorithm is in time-domain, the time-domain Wiener post-filtering method will be compared. For these simulations, we have used two microphones and two sources, where the maximum number of EM iterations is set to 10, $K = 2$ and $L = 512$. The average performance was computed over 50

simulation trials where, for each trial, the source signals are randomly selected from a set of sixteen speech utterances of 8 s each. The sampling frequency used is 16 kHz. The simulations are performed in offline mode unless specified.

We first investigate the robustness of the algorithm to the approximation of (3) when estimating h_q. This approximation is quantified using the signal-to-crosstalk ratio (SCR), where

$$\mathrm{SCR}(y_q(n)) = 10\log_{10}\frac{\sum_n \widetilde{s}_q^2(n)}{\sum_{q'\neq q}\sum_n \left[h_{q'} * \widetilde{s}_{q'}(n)\right]^2}. \qquad (30)$$

Therefore, a low $\mathrm{SCR}(y_q(n))$ implies an increased residual crosstalk in $y_q(n)$ which reduces the degree of validity of the approximation in (3). In this simulation, we first use $\widetilde{s}_q(n)$ and h_q to generate the partially separated signals $y_q(n)$ according to (2). We have used 8 s speech utterances for $\widetilde{s}_q(n)$ and the publicly available recorded room impulse responses (RIRs) [16] (which have been truncated to 512 samples) to form h_q. With reference to (2), h_1 and h_2 are then scaled such that the resulting $y_2(n)$ and $y_1(n)$ will have the predefined SCR. The generated $y_q(n)$ are used as inputs to the ANC for the estimation of h_q using both the proposed GMM and conventional Wiener methods. The true h_q and its estimates in a simulation are illustrated in Fig. 2, where the corresponding $\widetilde{s}_1(n)$, estimated $w_1(n)$ and $\widetilde{s}_2(n)$ are presented in Fig. 1. As shown in Fig. 1, the inverse relationship between $w_1(n)$ and $\widetilde{s}_1(n)$ will cause the algorithm to emphasize segments that are free of crosstalk. Hence, the estimated $\widehat{h}_q^{\mathrm{GMM}}$ is more accurate compared to $\widehat{h}_q^{\mathrm{GD}}$, as illustrated in Fig. 2. It can be seen from (5) that the difference $\widehat{h}_q - h_q$ determines the amount of residual crosstalk after post-filtering. Hence the corresponding mean-squared error (MSE) of the estimate $\widehat{h}_q$,

$$\mathrm{MSE} = \frac{1}{Q}\sum_{q=1}^{Q}\left(\frac{1}{L}\sum_{l=0}^{L-1}\left|\widehat{h}_q(l) - h_q(l)\right|^2\right) \qquad (31)$$

is utilized to quantify the post-filter estimation error. Figure 3 shows the average MSE for the Wiener and GMM methods when SCR is varied from 5 to 15 dB. We note that the performance of the Wiener method improves with SCR as expected. More importantly, the average MSE of the proposed method is robust against low SCR and is significantly lower than that of the Wiener method.

The second simulation studies the convergence performance of the Wiener and GMM methods in an online setup. The $y_q(n)$ is generated using the same procedure as that of the first simulation with $\mathrm{SCR} = 8$ dB. For both methods, the first 0.3 s of $y_q(n)$ is used to obtain an initial estimate of h_q. The number of samples used in each of the following incremental update corresponds to 0.01 s without overlap and the adaptation constant is set to 0.1. The GMM method performs one EM iteration for each new sample frame. Figure 4 shows the average MSE for the Wiener and GMM methods plotted against time. It can be seen that the performance of both methods improve with time as expected while the GMM method exhibits better performance than the Wiener method.

In the third simulation, the Wiener and the proposed GMM filters are used as post-filters for the time-domain BSS (TDBSS) algorithm [2]. Here, we have used the same simulation setup as described in [2]. This setup consists of two sources positioned at $\pm 45^\circ$ and 1.2 m away from the centroid of the microphone pair. The RIRs are generated

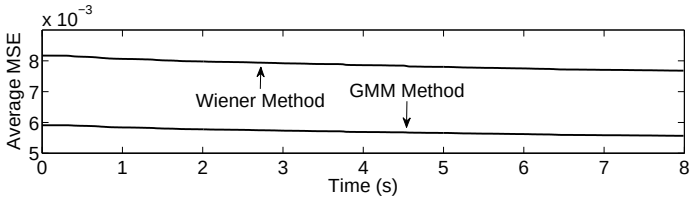

Fig. 4. Online convergence performance in h_q estimation using the Wiener and GMM methods at SCR = 8 dB.

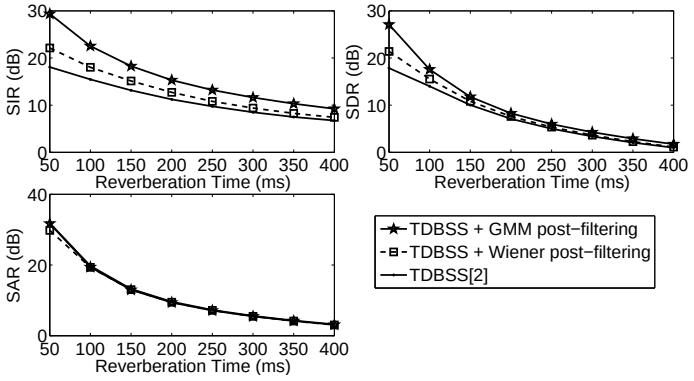

Fig. 5. Performance of the TDBSS algorithm [2], TDBSS followed by Wiener post-filtering and TDBSS followed by the proposed GMM post-filtering.

using the method of images [17] with T_{60} varying from 50 to 400 ms. We have employed the commonly used signal-to-interference ratio (SIR), signal-to-distortion ratio (SDR) and signal-to-artifacts ratio (SAR) measures [18] for performance evaluation. Figure 5 shows the average separation performance of the TDBSS with and without the post-filters. It is evident that the use of the proposed GMM filter can achieve better performance than the Wiener filter. For $T_{60} = 50$ ms, compared with the Wiener filter, the GMM filter can achieve an improvement of 7.5 and 5.5 dB in terms of SIR and SDR, respectively. When T_{60} is increased to 250 ms, this improvement corresponds to 2.5 and 0.6 dB respectively in terms of SIR and SDR. Since post-filtering techniques are linear, only modest amount of artifacts will be introduced and this is validated by the SAR plot in Fig. 5.

V. Conclusion

It is shown, in the context of BSS, that the Wiener filter used by existing ANC-based BSS crosstalk suppression algorithms are in the ML sense only for Gaussian distributed source signals. Since speech signals are non-Gaussian, the GMM is used to model the signal distribution. We then derived an EM algorithm to estimate the corresponding ML-sense post-filter. The resulting GMM filter is a weighted version of the Wiener filter and has proven to be more robust against interfering crosstalk components in the reference input to ANC. Simulation results show that the proposed GMM filter achieves better performance than the Weiner filter method.

References

[1] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 2, pp. 109–116, 2003.

[2] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 1, pp. 120–134, 2005.
[3] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, 1997.
[4] K. Kokkinakis and A. K. Nandi, "Multichannel blind deconvolution for source separation in convolutive mixtures of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 200 – 212, 2006.
[5] B. Liu, V. G. Reju, and A. W. H. Khong, "Underdetermined instantaneous blind source separation of sparse signals with temporal structure using the state-space model," in *Proc. ICASSP*, May 2013, pp. 81–85.
[6] R. Aichner, M. Zourub, H. Buchner, and W. Kellermann, "Post-processing for convolutive blind source separation," in *Proc. ICASSP*, May 2006, vol. 5, pp. 37–41.
[7] K. S. Park, J. S. Park, K. S. Son, and H. T. Kim, "Postprocessing with Wiener filtering technique for reducing residual crosstalk in blind source separation," *IEEE Signal Process. Letters*, vol. 13, no. 12, pp. 749–751, 2006.
[8] T. Noohi and M. H. Kahaei, "Residual cross-talk suppression for convolutive blind source separation," in *Proc. ICCET*, Apr. 2010, vol. 1, pp. 543–547.
[9] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual crosstalk components in blind source separation using LMS filters," in *Proc. NNSP*, 2002, pp. 435–444.
[10] S. Y. Low, S. Nordholm, and R. Togneri, "Convolutive blind signal separation with post-processing," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 5, pp. 539–548, 2004.
[11] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual cross-talk components in blind source separation using time-delayed spectral subtraction," in *Proc. ICASSP*, May 2002, vol. 2, pp. 1789–1792.
[12] V. Bhatia and B. Mulgrew, "Non-parametric likelihood based channel estimator for Gaussian mixture noise," *Signal Process.*, vol. 87, no. 11, pp. 2569–2586, 2007.
[13] R. J. Kozick and B. M. Sadler, "Maximum-likelihood array processing in non-Gaussian noise with Gaussian mixtures," *IEEE Trans. Signal Process.*, vol. 48, no. 12, pp. 3520–3535, 2000.
[14] J. K. Tugnait, L. Tong, and Z. Ding, "Single-user channel estimation and equalization," *IEEE Signal Process. Mag.*, vol. 17, no. 3, pp. 16–28, 2000.
[15] G. McLachlan and D. Peel, *Finite mixture models*, Wiley-Interscience, 2004.
[16] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. DSP*, 2009, pp. 1–5.
[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am*, vol. 65, no. 4, pp. 943–950, 1979.
[18] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1462 –1469, 2006.