BLIND SOURCE SEPARATION IN A DISTRIBUTED MICROPHONE MEETING ENVIRONMENT FOR IMPROVED TELECONFERENCING

Jacek P. Dmochowski[*]

Université du Québec, INRS-EMT
800 rue de la Gauchetière Ouest
Montréal, Québec, Canada, H2X 3R4

Zicheng Liu, Philip A. Chou

Microsoft Research
One Microsoft Way
Redmond, WA, 98052, United States

ABSTRACT

From an audio perspective, the present state of teleconferencing technology leaves something to be desired; speaker overlap is one of the causes of this inadequate performance. To that end, this paper presents a frequency-domain implementation of convolutive BSS specifically designed for the nature of the teleconferencing environment. In addition to presenting a novel depermutation scheme, this paper presents a least-squares post-processing scheme, which exploits segments during which only a subset of all speakers are active. Experiments with simulated and real data demonstrate the ability of the proposed methods to provide SIRs at or near that of the adaptive noise cancellation (ANC) solution which is obtained under idealistic assumptions that the ANC filters are adapted with one source being on at a time.

Index Terms— Microphone arrays, blind source separation, independent components analysis.

1. INTRODUCTION

Due to the environmental impact of everyday commuting and the busy lifestyles of today's professionals, audio- and video-conferencing (collectively termed "teleconferencing") is expected to become the primary method of multi-party interaction in the future. In terms of audio, the present quality of the teleconferencing experience is inadequate for many reasons; one of these is the presence of overlapped speech from multiple participants, resulting in poor intelligibility for the remote listener. With the decreasing cost of personal communication devices, it is common for participants to be accompanied by a device with an embedded microphone. It is possible to connect the devices of the participants over a network, allowing for multi-channel processing of the microphone recordings. Ad hoc microphone arrays differ from centralized arrays in several aspects: the inter-microphone spacing is generally large, leading to spatial aliasing. Since the various microphones are not connected to the same clock, some form of network synchronization is necessary. Notice also that each speaker is usually much nearer to his/her microphone than that of the other participants; thus the input signal-to-interference ratio (SIR) is high.

The problem of blindly separating instantaneous mixtures of independent signals is covered in [1], [2], and prominent algorithms include [3], [4]. However, the speech mixtures received at an array of microphones are not instantaneous but convolutive; the convolutive blind source separation (BSS) [5] task may be tackled in either the time- or frequency-domain. Notable time-domain solutions are detailed in [6] and [7]. The frequency-domain approach [8] decomposes the signals at the array into narrowband frequency bins and processes each bin using well-established instantaneous BSS methods. The separation of speech sources with distributed microphones is a recent idea described e.g. in [9].

It is interesting to note that the vast majority of BSS algorithm evaluations are performed on recordings during which two or more speakers are unnaturally speaking simultaneously and without any regard for the other speaker(s). This scenario does not represent reality. Even in a "cocktail-party" environment, it is unreasonable to assume that all signals are active across all time. In a multi-party conversation, only a single speaker is active during many segments. Non-blind techniques such as adaptive noise cancellation (ANC) [10] exploit the single-source case to estimate the transfer functions from the interference to the primary sensor. While blind techniques do not have knowledge of the on-times of the various sources, such information may be estimated from the separated signals.

This paper presents a frequency-domain approach to blind separation of speech that is tailored to the nature of the teleconferencing environment. In addition to presenting a novel permutation-solving scheme, we propose a least-squares post-processing of the frequency-domain independent components analysis (ICA) outputs. The presence of single-speaker segments (and in general, any segments during which the set of active speakers is a subset of the set of all speakers) is exploited to compute more accurate estimates of the frequency-domain mixing matrices.

2. SIGNAL AND SEPARATION MODELS

Consider an array of M microphones where the output of the mth microphone is denoted by $x_m(k)$, where k is the discrete-time sample index. Assuming N sources whose signals are given by $s_n(k)$, the output of the mth microphone is the convolutive mixture

$$x_m(k) = \sum_{n=1}^{N} \sum_{l=0}^{L_h-1} h_{mn}(l) s_n(k-l) + v_m(k),\ m = 1, \ldots, M, \tag{1}$$

where h_{mn} is the finite impulse response (FIR) channel from source n to microphone m, L_h is the length of the longest impulse response, and $v_m(k)$ is the additive sensor noise at microphone m. It is generally assumed that the source signals are mutually independent. The task of blind source separation (BSS) in such convolutive mixtures is to recover the source signals $s_n(k)$ given only the microphone recordings $x_m(k)$. Generally speaking, this requires $N \leq M$.

The separation of the signals is achieved by applying a FIR filter to each sensor's output and then summing across the sensors:

$$y_n(k) = \sum_{m=1}^{M} \sum_{l=0}^{L_w-1} w_{nm}(l) x_m(k-l),\ n = 1, \ldots, N, \tag{2}$$

[*]The first author performed the work while at Microsoft Research.

where $y_n(k)$ is the estimate of $s_n(k)$, $w_{nm}(k)$ is the filter applied to microphone m in order to separate source n, and L_w is the length of the longest separation filter.

Taking the Fourier transform of (1) and rewriting in matrix notation, we obtain the instantaneous mixture model

$$\mathbf{x}(\omega) = \sum_{n=1}^{N} \mathbf{h}_{:n}(\omega) S_n(\omega) + \mathbf{v}(\omega) = \mathbf{H}(\omega)\mathbf{s}(\omega) + \mathbf{v}(\omega), \qquad (3)$$

where

$$\mathbf{x}(\omega) = \begin{bmatrix} X_1(\omega) & X_2(\omega) & \dots & X_M(\omega) \end{bmatrix}^T,$$
$$\mathbf{h}_{:n}(\omega) = \begin{bmatrix} H_{1n}(\omega) & H_{2n}(\omega) & \dots & H_{Mn}(\omega) \end{bmatrix}^T,$$
$$\mathbf{v}(\omega) = \begin{bmatrix} V_1(\omega) & V_2(\omega) & \dots & V_M(\omega) \end{bmatrix}^T,$$
$$\mathbf{H}(\omega) = \begin{bmatrix} H_{11}(\omega) & H_{12}(\omega) & \dots & H_{1N}(\omega) \\ H_{21}(\omega) & H_{22}(\omega) & \dots & H_{2N}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ H_{M1}(\omega) & H_{M2}(\omega) & \dots & H_{MN}(\omega) \end{bmatrix}$$
$$\mathbf{s}(\omega) = \begin{bmatrix} S_1(\omega) & S_2(\omega) & \dots & S_N(\omega) \end{bmatrix}^T,$$

and $X_m(\omega)$, $H_{mn}(\omega)$, $S_n(\omega)$, and $V_m(\omega)$ are the discrete-time Fourier transforms of $x_m(k)$, $h_{mn}(k)$, $s_n(k)$, and $v_m(k)$, respectively. In the frequency-domain, the separation model becomes

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega), \qquad (4)$$

where $\mathbf{y}(\omega) = \begin{bmatrix} Y_1(\omega) & Y_2(\omega) & \cdots & Y_N(\omega) \end{bmatrix}^T$ is a vector of the Fourier transformed separated signals $y_n(k)$ and $\mathbf{W}(\omega)$ is the separation matrix with $[\mathbf{W}(\omega)]_{nm} = W_{nm}(\omega)$.

3. CONVOLUTIVE BSS IN FREQUENCY DOMAIN

To enable frequency-domain processing, the time-domain microphone signals $x_m(k)$ are transformed to the frequency-domain via the short-time Fourier transform:

$$X_m(\omega, \tau) = \sum_{l=-\infty}^{\infty} x_m(l)\mu(l-\tau)e^{-j\omega l}, \qquad (5)$$

where $\mu(l)$ is a windowing function with $\mu(l) = 0,\ |l| > W$, and τ is the time frame index. Similar definitions hold for $V_m(\omega, \tau)$, $S_n(\omega, \tau)$, $\mathbf{x}(\omega, \tau)$, $\mathbf{v}(\omega, \tau)$, and $\mathbf{s}(\omega, \tau)$. Hence (3) and (4) become

$$\mathbf{x}(\omega, \tau) = \mathbf{H}(\omega)\mathbf{s}(\omega, \tau) + \mathbf{v}(\omega, \tau), \qquad (6)$$

$$\mathbf{y}(\omega, \tau) = \mathbf{W}(\omega)\mathbf{x}(\omega, \tau). \qquad (7)$$

For each frequency ω, the complex-valued ICA procedure computes a matrix $\mathbf{W}(\omega)$ such that the components of the output $\mathbf{y}(\omega, \tau)$ are as mutually independent as possible. This is achieved through either a complex version of the FastICA algorithm [11] or a complex version of InfoMax [3] along with the natural gradient procedure of [4], [8].

Assuming that the components of $\mathbf{s}(\omega, \tau)$ are mutually independent and that the microphone noise $\mathbf{v}(\omega, \tau)$ is zero, the separation matrix $\mathbf{W}(\omega)$ derived by ICA will be equal to the pseudo-inverse of the underlying mixing matrix $\mathbf{H}(\omega)$ up to a permuation and scaling, namely, $\mathbf{W}(\omega) = \mathbf{\Lambda}(\omega)\mathbf{P}(\omega)\mathbf{H}^{+}(\omega)$, where $\mathbf{\Lambda}(\omega) = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$ is a diagonal matrix and $\mathbf{P}(\omega)$ is a permutation matrix. Thus $\mathbf{y}(\omega, \tau) = [\lambda_1 s_{\Pi_\omega^{-1}(1)}(\omega, \tau), \dots, \lambda_N s_{\Pi_\omega^{-1}(N)}(\omega, \tau)]^T$, where $\Pi_\omega(i) = j$ is the permutation mapping between the ith source and the jth separated signal at frequency ω. Moreover, denoting $\mathbf{W}^{+}(\omega) = \mathbf{H}(\omega)\mathbf{P}^{-1}(\omega)\mathbf{\Lambda}^{-1}(\omega) = \begin{bmatrix} \mathbf{a}_{:1} & \mathbf{a}_{:2} & \cdots & \mathbf{a}_{:N} \end{bmatrix}$, it is easy to show that $\mathbf{a}_{:n}(\omega) = \mathbf{h}_{:\Pi_\omega^{-1}(n)}(\omega)/\lambda_n$. The challenge in convolutive BSS is to determine $\mathbf{P}(\omega)$ and $\mathbf{\Lambda}(\omega)$ at each frequency.

4. SOLVING THE PERMUTATION PROBLEM

Recently, a permutation solving scheme that is applicable to distributed microphones has been proposed in [9]. Unlike the methods based on source localization that utilize the phases of the columns $\mathbf{a}_{:n}(\omega)$ (which cannot be used with distributed mics due to aliasing), only the magnitudes are taken into account. For ease of presentation, if $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_{N_u} \end{bmatrix}^T$ is a complex vector, then $\mathbf{u}' = \begin{bmatrix} |u_1| & |u_2| & \cdots & |u_{N_u}| \end{bmatrix}^T$ is the vector $\mathbf{u}$ but with the phases of each element discarded. In order to remove the scaling ambiguity that appears in the columns $\mathbf{a}'_{:n}(\omega)$, [9] proposes to divide successive elements of each column in order to remove the scaling factor. In this paper, a different approach is taken at solving the permutation problem. At each frequency, the magnitudes of the vectors $\mathbf{a}_{:n}(\omega)$ are normalized to unit norm:

$$\hat{\mathbf{a}}'_{:n}(\omega) = \frac{\mathbf{a}'_{:n}(\omega)}{\|\mathbf{a}'_{:n}(\omega)\|} = \frac{\mathbf{h}'_{:\Pi_\omega^{-1}(n)}(\omega)}{\|\mathbf{h}'_{:\Pi_\omega^{-1}(n)}(\omega)\|}, \qquad (8)$$

thus removing the scaling factor, which is constant over the entries of a fixed column $\mathbf{a}_{:n}(\omega)$. The resulting normalized column vectors reflect the *relative* energy attenuation experienced between source $\Pi_\omega^{-1}(n)$ and the array of microphones. Each source is identified by its own vector of relative attenuation values, which are independent of frequency and may be used to solve the permutation ambiguity.

Notice that in the teleconferencing environment, the attenuation experienced by a speaker at his/her microphone will be significantly less than that experienced by the same speaker at the other participants' microphones. Thus, we propose a de-permutation scheme that assigns the vector $\hat{\mathbf{a}}'_{:n}(\omega)$ to the speaker identified by the largest element of $\hat{\mathbf{a}}'_{:n}(\omega)$. Specifically, $\mathbf{h}'_{:j}(\omega) = \sum_{i=1}^{N} p_{ij}(\omega)\mathbf{a}'_i(\omega)$, where $p_{ij}(\omega) = 1$ if $j = \arg\max_n \hat{a}'_{ni}(\omega)$ and $p_{ij}(\omega) = 0$ otherwise. Notice that with this scheme (termed "maximum-magnitude" or MM), if two columns exhibit a maximum at the same row, the synthesized signals will contain components from multiple source signals at a particular frequency. However, a more detrimental swapping of the coefficients from different sources will not occur.

It is also possible to cluster the relative attenuation vectors $\hat{\mathbf{a}}'_{:i}(k)$ of all frequencies into N clusters, and perform a least-squares optimization across all possible one-to-one permutations of $1, 2, \cdots, N$. In that case, the depermutation is equivalent to that of [9]. The clustering operation is symbolically written as:

$$\{\mathbf{c}_1, \sigma_1, \cdots, \mathbf{c}_N, \sigma_N\} = \mathrm{cluster}\left[\forall\omega, \hat{\mathbf{a}}'_{:1}(\omega), \cdots, \hat{\mathbf{a}}'_{:N}(\omega)\right] \qquad (9)$$

where $\mathbf{c}_n$ as the centroid of the nth cluster C_n: $\mathbf{c}_n = \sum_{\mathbf{a}\in C_n} \frac{\mathbf{a}}{|C_n|}$, where $|C_n|$ is the number of elements in C_n, and $\sigma_n^2 = \sum_{\mathbf{a}\in C_n} \frac{\|\mathbf{c}_n - \mathbf{a}\|^2}{|C_n|}$ is the variance of cluster n. A common clustering algorithm is the k-means algorithm [12]. Once the clusters are determined, the permutation mapping is computed using the least-squares optimization:

$$\Pi_\omega = \arg\min_{\Pi} \sum_{n=1}^{N} \|\mathbf{c}_n - \hat{\mathbf{a}}'_{:\Pi(n)}(\omega)\|^2, \qquad (10)$$

where Π denotes all permutations of $1, 2, \cdots, N$.

5. LEAST-SQUARES POST-PROCESSING

Conventional ICA-based convolutive BSS does not explicitly take the microphone noise into account in its solution. We rewrite (6) to include F frames:

$$\mathbf{X}(\omega) = \mathbf{H}(\omega)\mathbf{S}(\omega) + \mathbf{V}(\omega), \tag{11}$$

where

$$\begin{aligned} \mathbf{X}(\omega) &= \left[\begin{array}{ccc} \mathbf{x}(\omega,1) & \cdots & \mathbf{x}(\omega,F) \end{array} \right], \\ \mathbf{S}(\omega) &= \left[\begin{array}{ccc} \mathbf{s}(\omega,1) & \cdots & \mathbf{s}(\omega,F) \end{array} \right], \\ \mathbf{V}(\omega) &= \left[\begin{array}{ccc} \mathbf{v}(\omega,1) & \cdots & \mathbf{v}(\omega,F) \end{array} \right]. \end{aligned}$$

We seek an approximate factorization of microphone measurements $\mathbf{X}(\omega)$ into matrices $\mathbf{H}(\omega)$ and $\mathbf{S}(\omega)$ such that the squared error of the microphone noise $||\mathbf{V}(\omega)||^2$ is minimized. This is clearly trivial to achieve if there are no constraints on $\mathbf{S}(\omega)$. For example, if there are $N = M$ simultaneously active sources, then we may set $\mathbf{H}(\omega) = I$ and $\mathbf{S}(\omega) = \mathbf{X}(\omega)$ to obtain zero error. However, if it is known that for some frames of $\mathbf{S}(\omega)$ a subset of the sources are inactive, then the mixing matrix $\mathbf{H}(\omega)$ becomes constrained. For example, if only sources n_1 and n_2 are active in frames $\tau \in A_{12}$, then the set of vectors $\{\mathbf{X}(\omega,\tau) : \tau \in A_{12}\}$ determines the subspace spanned by the columns $\mathbf{h}_{:n_1}(\omega)$ and $\mathbf{h}_{:n_2}(\omega)$, while if only sources n_1 and n_3 are active in frames $\tau \in A_{13}$, then $\{\mathbf{X}(\omega,\tau) : \tau \in A_{13}\}$ determines the subspace spanned by the columns $\mathbf{h}_{:n_1}(\omega)$ and $\mathbf{h}_{:n_3}(\omega)$. Intersecting these subspaces determines the column $\mathbf{h}_{:n_1}(\omega)$ (up to scale). Thus this least squares approach can refine $\mathbf{H}(\omega)$ using knowledge of the frames during which a subset of the sources are inactive.

We start with an estimate of which speakers are inactive by applying speaker activity detection (SAD) to the ICA outputs (7). In this paper, a simple energy-based threshold detection is used. For each source $n \in \{1,...,N\}$, let B_n denote the set of frames for which source n is inactive according to SAD. By adding speech inactivity constraints to (11), we obtain the following optimization problem:

$$\begin{aligned} \mathrm{M}inimize \quad & ||\mathbf{V}(\omega)||^2 \\ s.t. \quad & \\ & \mathbf{X}(\omega) = \mathbf{H}(\omega)\mathbf{S}(\omega) + \mathbf{V}(\omega) \\ & S_n(\omega,\tau) = 0, \forall \tau \in B_n, n \in \{1,...,N\}. \end{aligned} \tag{12}$$

To solve this problem, we start with an estimate of $\mathbf{H}(\omega)$ as the pseudo-inverse of the ICA result, i.e., $\mathbf{H}(\omega) = \mathbf{W}(\omega)^+$. Then it is straightforward to solve for $\mathbf{S}(\omega)$. Specifically, considering each column of $\mathbf{S}(\omega)$ separately, let $\tilde{\mathbf{s}}(\omega,\tau)$ be the subvector of $\mathbf{s}(\omega,\tau)$ consisting of only the active sources $S_n(\omega,\tau)$ where $\tau \in \{1,...,F\} \setminus B_n$, and let $\tilde{\mathbf{H}}(\omega)$ be the submatrix of $\mathbf{H}(\omega)$ consisting of only the corresponding columns. Then

$$\tilde{\mathbf{s}}(\omega,\tau) = \tilde{\mathbf{H}}^+(\omega)\mathbf{x}(\omega,\tau)$$

minimizes the norm of $\mathbf{v}(\omega,\tau)$ under the speaker inactivity constraints. Performing this for all frames τ minimizes the squared error $||\mathbf{V}(\omega)||^2$ under the inactivity constraints.

Suppose now we fix the $\mathbf{S}(\omega)$ just determined, and re-solve for $\mathbf{H}(\omega)$ in (12) to minimize $||\mathbf{V}(\omega)||^2$ still further. Since $\mathbf{S}(\omega)$ is fixed, (12) becomes an unconstrained least square problem:

$$\mathrm{M}inimize_{\mathbf{H}(\omega)} ||\mathbf{X}(\omega) - \mathbf{H}(\omega)\mathbf{S}(\omega)||^2. \tag{13}$$

Its solution is

$$\mathbf{H}(\omega) = \mathbf{X}(\omega)\mathbf{S}^H(\omega)(\mathbf{S}(\omega)\mathbf{S}^H(\omega))^{-1} \tag{14}$$

where $\mathbf{S}^H$ is the conjugate transpose of $\mathbf{S}$.

Iterating this procedure (solving $\mathbf{S}(\omega)$ for fixed $\mathbf{H}(\omega)$ and then solving $\mathbf{H}(\omega)$ for fixed $\mathbf{S}(\omega)$) is clearly a descent algorithm that minimizes the same metric $||\mathbf{V}(\omega)||^2$ in each step and hence it converges. This potentially improves the mixing matrix $\mathbf{H}(\omega) = \mathbf{W}^+(\omega)$ obtained by ICA, under the constraint that some of the sources are inactive in some of the frames. Note that if all sources are active in all frames, then the initial mixing matrix $\mathbf{H}(\omega)$ determined from ICA remains unchanged by these iterations.

Given speech activity detection outputs, one can solve (12) to obtain an improved mixing matrix $\mathbf{H}(\omega)$, an improved separation matrix $\mathbf{W}(\omega) = \mathbf{H}^+(\omega)$, and an improved source separation (7). The newly separated sources can be used to re-estimate the inactive sources in each frame, thus resulting in a new optimization problem (12) where the inactivity constraints are modified. One could once again solve this problem by using the same procedure as described above. If the newly obtained speech activity detection outputs are more accurate, the results obtained by solving (12) will also be improved. Unfortunately there is no effective way to determine whether the speech activity detection is improving. To be conservative, we usually perform $2-3$ iterations of speech activity detection in our experiments.

This section has described a post-processing procedure to minimize the norm of the error in the mixing model (12). A corresponding algorithm may also be developed to minimize the norm of an error in the separation model,

$$\mathbf{Y}(\omega) = \mathbf{W}(\omega)\mathbf{X}(\omega) + \mathbf{U}(\omega),$$

where $\mathbf{U}(\omega)$ is the error under constraints that some components of $\mathbf{Y}(\omega)$ are zero. The principles are similar, but the resulting separation filters will be different. Due to space constraints, the development is not shown in this paper; however, the performance of the proposed scheme under both models is given in the next section.

6. EXPERIMENTAL EVALUATION

The proposed algorithms are evaluated using both synthetic data – generated using the image method [13] – as well as real recordings taken in an actual conference room with moderate reverberation.

For the synthetic data, the signal-to-noise ratio (SNR) is 20 dB and the additive noise is spatially uncorrelated and temporally white. The simulated room's reverberation time is $T_{60} = 300$ms. The simulations employ an $M = 2$ element array with the two microphones located at $(203.2, 228.6, 101.6)$ m and $(101.6, 228.6, 101.6)$ m, respectively. The two speakers are located at $(254, 228.6, 101.6)$ m and $(50.8, 228.6, 101.6)$ m, respectively. The sampling rate is 8kHz, with a framelength and frame-shift of 4096 and 1024 samples, respectively. Each speaker speaks for 4 seconds, with the level of overlap between the two speakers being varied. Four levels of overlap are simulated: $0, 25, 50$, and 100%. The least-squares algorithm parameter is $\delta = -3$dB, determined heuristically.

For the real data, the room dimensions are 4-by-7-by-3 m. The two microphones (which are synchronized in hardware) are placed on a table of height 0.8m. The $x-y$ coordinates of the two microphones are $(0.15, -0.42)$m and $(0.23, 0.42)$m (the table center is the origin of the coordinate system). The speakers sit approximately 35cm behind each microphone. In order to allow for SIR computation, each speaker is recorded individually and the recordings are "merged" (i.e., added) accordingly with the four desired levels of

overlap. The sampling rate is 16kHz, with a frame-length and frame-shift of 8192 and 2048 samples, respectively. Each speaker talks for 10 seconds. The least-squares algorithm parameters are $p = 0.005$ and $\delta = 3$dB.

The frequency-domain data is separated using the FastICA algorithm of [11]. The algorithms are evaluated in terms of the SIRs of the separated signals. In total, 6 algorithms are evaluated: the two depermutation schemes without least-squares post-processing, and the least-squares scheme (under both $\mathbf{Y} = \mathbf{WX}+\mathbf{U}$ and $\mathbf{X} = \mathbf{HS}+\mathbf{V}$ models) with two initial conditions each (the initial conditions correspond to the two depermutation schemes). The SIRs of the 6 algorithms are compared to those that would be yielded by the ANC under the idealistic assumption that the separation filters are adapted with one source being on at a time. The separation filters obtained in the 0% overlap case with perfect knowledge of the on-times of the sources are used to measure the ANC SIR. It is important to understand that the ANC solution is not realizable in practice since it requires each source to be on one-at-a-time and exact knowledge of the various on-times.

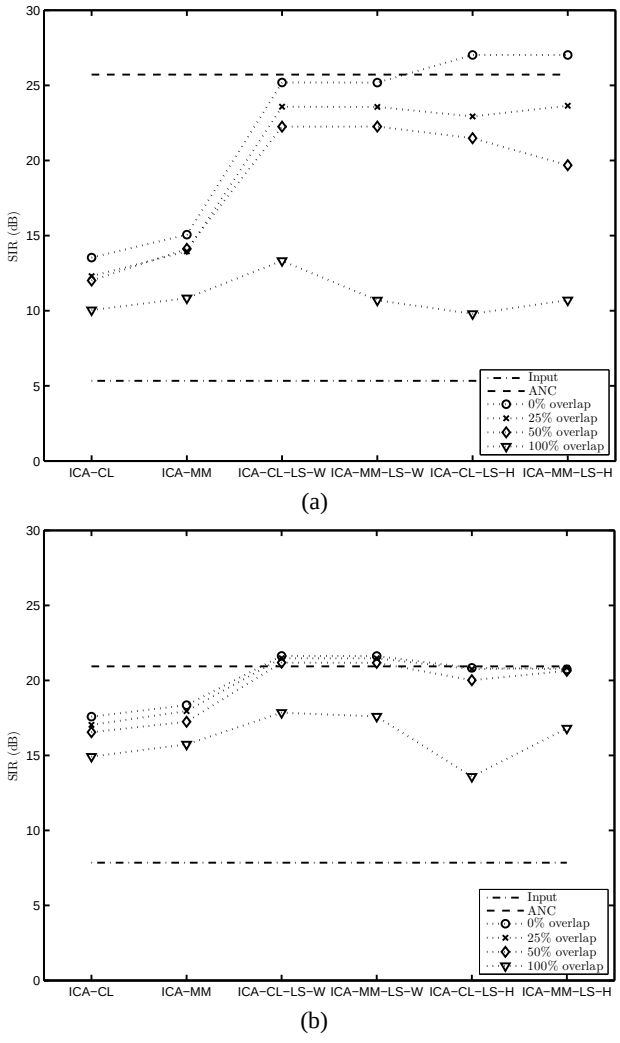

Fig. 1. Output SIRs: simulated data (a), real data (b).

The resulting SIRs are found in Fig. 1, with the following notation: "ICA-CL" (the scheme of [9]), "ICA-MM" (proposed depermutation scheme), "ICA-CL-LS-W" (ICA-CL with least-squares post-processing under the $\mathbf{Y} = \mathbf{WX} + \mathbf{U}$ model), "ICA-MM-LS-W" (ICA-MM with least-squares post-processing under the $\mathbf{Y} = \mathbf{WX}+\mathbf{U}$ model), "ICA-CL-LS-H" (ICA-CL with least-squares post-processing under the $\mathbf{X} = \mathbf{HS}+\mathbf{V}$ model), "ICA-MM-LS-H" (ICA-MM with least-squares post-processing under the $\mathbf{X} = \mathbf{HS} + \mathbf{V}$ model).

From the figure, it is evident that the proposed depermutation scheme provides greater signal separation than the clustering technique in all cases. This is attributed to the fact that the MM scheme is less likely to erroneously swap the two Fourier coefficients. Moreover, the least-squares post-processing provides a tremendous boost in the output SIRs, with the improvement increasing as the level of overlap decreases. As the level of overlap decreases, the least-squares scheme has more time to learn the constrained values of the separation filters. Since in the real recordings the speakers talk for 10 sec, the ANC solution is attained even for the 50% overlap case.

7. CONCLUSION

This paper has presented a formulation of convolutive BSS tailored to the nature of the teleconferencing environment. A novel depermutation scheme and a least-squares post-processing method were developed. Experiments with simulated and real data demonstrated the potential of the proposed schemes to provide SIRs at or near that of the ANC solution obtained under idealistic assumptions.

8. ACKNOWLEDGMENTS

The authors would like to thank Robert Aichner and Zhengyou Zhang for helpful discussions, and Li-Wei He for help with the data collection.

9. REFERENCES

[1] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*. John Wiley & Sons, 2000.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Components Analysis*, John Wiley & Sons, 2001.

[3] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[4] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251–276, 1998.

[5] M. S. Pedersen, J. Larsen, U. Kjems, L. C. Parra, "A survey of convolutive blind source separation methods", in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi and Y. Huang, eds., Springer-Verlag, Berlin, 2007.

[6] H. Buchner, R. Aichner, W. Kellermann, "Blind source separation for convolutive mixtures: a unified treatment, " in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty (eds.), Kluwer Academic Publishers, Boston, Feb. 2004.

[7] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics," *IEEE Trans. on Speech and Audio Processing*, Vol. 13, pp. 120-134 , Jan. 2005.

[8] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-Domain Blind Source Separation," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, eds., pp. 299–327, Springer-Verlag, 2005.

[9] E. Robledo-Arnuncio and Biing-Hwang (Fred) Juang, "Blind source separation of acoustic mixtures with distributed microphones," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, 2007, pp. III 949 – III 952.

[10] B. Widrow et al., "Adaptive noise cancelling: principles and applications," in *Proc. IEEE*, vol. 63, pp. 1692–1717, 1975.

[11] E. Bingham and A. Hyvarinen, "A fast fixed-point algorithm for independent components analysis of complex valued signals," *International Journal of Neural Systems*, vol. 10, pp. 1–8, Feb. 2000

[12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., Wiley Interscience, 2000.

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