A MULTICHANNEL MMSE-BASED FRAMEWORK FOR JOINT BLIND SOURCE SEPARATION AND NOISE REDUCTION

Mehrez Souden, Shoko Araki, Keisuke Kinoshita, Tomohiro Nakatani, Hiroshi Sawada

NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

ABSTRACT

In this paper, we propose a new framework to separate multiple speech signals and reduce the additive acoustic noise using multiple microphones. In this framework, we start by formulating the minimum-mean-square error (MMSE) criterion to retrieve each of the desired speech signals from the observed mixtures of sounds and outline the importance of multi-speaker activity detection. The latter is modeled by introducing a latent variable whose posterior probability is computed via expectation maximization (EM) combining both the spatial and spectral cues of the multichannel speech observations. We experimentally demonstrate that the resulting joint blind source separation (BSS) and noise reduction solution performs remarkably well in reverberant and noisy environments.

Index Terms— Microphone arrays, blind source separation, multichannel Wiener filter, noise reduction.

1. INTRODUCTION

In real world acoustic environments, background noise and multiple competing speakers can coexist in the same reverberant enclosure (e.g., teleconferencing rooms with multiple participants and noise sources). Retrieving speech signals of interest from the observed sound mixtures turns out to be quite challenging in this context due to the detrimental effects of reverberation and noise, yet highly desirable due to the diversity of its applications.

Traditionally, blind source separation (BSS) is achieved by exploiting the mutual independence between source signals via the celebrated independent component analysis (ICA). Information maximization (InfoMax) and FastICA are state of the art algorithms that have been shown to be very efficient in separating speech signals [1, 2, 3]. Besides, the speech representation in the time-frequency (t-f) domain reveals the important property of sparseness following which the major speech components of simultaneously active speakers rarely overlap [4, 5]. This has led to the development of clustering-based BSS approaches where t-f masking is applied once the speech mixture is well clustered. In [6], for instance, Sawada et al. proposed a powerful method that uses the spatial signatures of simultaneously active speakers in the absence of noise to cluster and separate them via binary masking. In contrast to BSS, noise reduction approaches have been essentially developed to recover a single speech signal which is corrupted by some acoustic noise. It is generally assumed that the noise is sufficiently stationary so that it can be tracked during the absence of speech and reduced using the Wiener filter or minimum variance distortionless response (MVDR), for instance (see [7] and references therein).

It is known that the performance of BSS deteriorates in the presence of acoustic noise. On the other hand, noise reduction algorithms are commonly designed to recover a single speech signal. To overcome both limitations, we propose a new framework that achieves simultaneous multiple speech sources separation and noise reduction. This framework is based on the minimum-mean-square error (MMSE) criterion whose formulation in the current context requires the detection and tracking of the activity of every speaker. Hence, we introduce a latent variable to model the multi-speaker activity and use the spatial and spectral cues of the observed mixtures of sounds via expectation maximization (EM) to estimate its posterior probability. The present work extends our proposal in [8] where only the spatial information is used to track the speech sources.

2. DATA MODEL AND ACTIVITY PATTERN OF MULTIPLE SPEAKERS

Let us consider the case of $N \geq 1$ speakers and an array of M microphones located in the same acoustic enclosure. The recorded signals are chopped into frames and transformed into the frequency domain via short time Fourier transform (STFT). At time frame l and frequency $k = 1, ..., K$, where K is the number of frequency components, we have

$$\mathbf{y}(k,l) \approx \sum_{n=1}^{N} \mathbf{x}_n(k,l) + \mathbf{v}(k,l), \tag{1}$$

where $\mathbf{y}(k,l) = [Y_1(k,l) \cdots Y_M(k,l)]^T$ and $\mathbf{x}_n(k,l) = \mathbf{h}_n(k)S_n(k,l)$. These vectors contain the M noisy sound mixtures and reverberant microphone observations of the nth speech signal, respectively. $\mathbf{h}_n(k) = [H_{1n}(k) \cdots H_{Mn}(k)]^T$ contains the channel transfer functions between the nth source, $S_n(k,l)$, and all microphone elements, and $\mathbf{v}(k,l) = [V_1(k,l) \cdots V_M(k,l)]^T$ contains all additive acoustic noise components. It is assumed that the analysis window is longer than the channel impulse responses. For the sake of simplicity, we omit mentioning the explicit dependence on the frequency, k, in our following notations since all our processing is done frequency-bin-wise.

It is known that speech signals are sparse and "only a small percentage of the time-frequency coefficients in the Gabor (STFT, in particular) expansion of speech capture a large percentage of the overall energy" [4]. Furthermore, it was established that it is very unlikely that the major energy components of different simultaneously active speech signals overlap in the t-f representation [4, 5]. It is, then, reasonable to assume the disjointness (or approximate disjointness in a less restrictive sense) of the STFT components of speech signals [4].

Now, to track the dominance of each of the N speech sources within the observed mixtures, we define the hidden variable, $\mathcal{H}$. Following the discussion above, $\mathcal{H}$ can take $N+1$ discrete states denoted as $\mathcal{H}_1, ..., \mathcal{H}_N, \mathcal{H}_{N+1}$: in state $\mathcal{H}_n$, $n = 1, ..., N$, the nth speaker dominates the mixture while in state $\mathcal{H}_{N+1}$, the noise is dominant. Next, we use the shorthand notation $p\left[\mathcal{H}_n|\mathbf{y}(l)\right]$

instead of $p\left[\mathcal{H}=\mathcal{H}_n|\mathbf{y}(l)\right]$ for the posterior probability that the nth signal dominates the mixture, which plays a fundamental role in the proposed framework.

3. MMSE-BASED MULTI-SOURCE/MULTICHANNEL FILTER

Our objective is to design an MMSE-based filter that extracts the nth speech source up to some frequency-dependent scalar coefficient. Since we are only interested in BSS and noise reduction, we define our objective as extracting $\tilde{S}_n(l) = X_{1n}(l) = H_{1n}S_n(l),\ n = 1,...,N$. In other words, we consider the MMSE solution $\hat{\tilde{S}}_n(l) = E\left\{X_{1n}(l)|\mathbf{y}(l)\right\}$ which is written as

$$\hat{\tilde{S}}_n(l) = p\left[\mathcal{H}_n|\mathbf{y}(l)\right] E\left\{X_{1n}(l)|\mathbf{y}(l),\mathcal{H}_n\right\} + \mathcal{E}_n(l) \tag{2}$$

where $\mathcal{E}_n(l) = \sum_{n'=1,n'\neq n}^{N+1} p\left[\mathcal{H}_{n'}|\mathbf{y}(l)\right] E\left\{X_{1n}(l)|\mathbf{y}(l),\mathcal{H}_{n'}\right\}$. Empirically, we found that setting $\mathcal{E}_n(l) \approx 0$ does not affect much the estimation accuracy of the sources[1]. We assume that all signals' complex spectra are Gaussian. Hence, solving for the expectation term on the right-hand side of (2) amounts to looking for the linear filter that minimizes the quadratic error $E\left\{\left|\mathbf{w}^H\mathbf{y}(l) - X_{1n}(l)\right|^2\right\}$, defined for a variable $\mathbf{w}$, which is known to be the Wiener filter. By defining the undesired signals covariance matrix as $\mathbf{R}_{\mathrm{u_n}} = \mathbf{R}_{\mathbf{yy}} - \mathbf{R}_{\mathbf{x}_n\mathbf{x}_n}$ where $\mathbf{R}_{\mathbf{yy}} = E\left\{\mathbf{y}(l)\mathbf{y}^H(l)\right\}$ and $\mathbf{R}_{\mathbf{x}_n\mathbf{x}_n} = E\left\{\mathbf{x}_n(l)\mathbf{x}_n^H(l)\right\}$, the Wiener filter can be modified by emphasizing or de-emphasizing the suppression of the undesired signals [7]

$$\mathbf{w}_n^{(\lambda)} = \frac{\mathbf{R}_{\mathrm{u_n}}^{-1}\mathbf{R}_{\mathbf{x}_n\mathbf{x}_n}\mathbf{u}_1}{\lambda + \mathrm{trace}\left(\mathbf{R}_{\mathrm{u_n}}^{-1}\mathbf{R}_{\mathbf{x}_n\mathbf{x}_n}\right)}, \tag{3}$$

where $\mathbf{u}_1 = [1\ 0\ ...\ 0]^T$ and $\lambda \geq 0$. $\lambda = 1$ and 0 correspond to the traditional Wiener filter and MVDR, respectively. Finally, the nth source estimate depends on λ and is given by

$$\hat{\tilde{S}}_n^{(\lambda)}(l) \quad = \quad p\left[\mathcal{H}_n|\mathbf{y}(l)\right]\mathbf{w}_n^{(\lambda)H}\mathbf{y}(l). \tag{4}$$

To implement (4), we need to estimate $p\left[\mathcal{H}_n|\mathbf{y}(l)\right]$ as we will show in Section 4. Besides, $\mathbf{R}_{\mathbf{yy}}$ can be directly obtained from the microphone observations and the estimation of $\mathbf{R}_{\mathbf{x}_n\mathbf{x}_n}$ will be detailed next.

Statistics Estimation: the covariance matrix of the recorded mixtures of sounds, $\mathbf{R}_{\mathbf{yy}} = \int_{\mathbf{y}} \mathbf{y}\mathbf{y}^H p\left(\mathbf{y}\right) d\mathbf{y}$, can be estimated as $\hat{\mathbf{R}}_{\mathbf{yy}} = \frac{1}{T}\sum_{l=1}^{T}\mathbf{y}(l)\mathbf{y}^H(l)$ using a block of T data samples. Now, to estimate the desired and undesired signals' statistics, we decompose the covariance matrix of the observations as $\mathbf{R}_{\mathbf{yy}} = \sum_{n=1}^{N+1}\mathbf{R}_n$, where

$$\mathbf{R}_n = \int_{\mathbf{y}} \mathbf{y}\mathbf{y}^H p\left(\mathcal{H}_n|\mathbf{y}\right) p(\mathbf{y}) d\mathbf{y}. \tag{5}$$

$\mathbf{R}_{N+1}$ corresponds to the noise covariance matrix (i.e., $\mathbf{R}_{N+1} = \mathbf{R}_{\mathbf{vv}}$) if we assume that the noise is stationary enough –which is commonly the case (see [7] and references therein, for instance)– and neglect the presence of speech when the noise dominates the observed mixture. For $n = 1,...N$, the nth marginal term is given by

$$\mathbf{R}_n = \mathbf{R}_{\mathbf{vv}} + \mathbf{R}_{\mathbf{x}_n\mathbf{x}_n}, \tag{6}$$

meaning that $\mathbf{R}_n$ is to the covariance matrix of the noise plus the nth speech source. Now, it is clear that the multi-speaker activity detection and tracking (i.e., the estimation of the posterior probabilities of $\mathcal{H}_1,...,\mathcal{H}_{N+1}$) is critical to the utilization of the MMSE to perform joint BSS and noise reduction. Having these posterior probabilities at one's disposal, it becomes possible to calculate the following in practice: (i) the noise covariance matrix $\mathbf{R}_{\mathbf{vv}}$, which is empirically well approximated as

$$\hat{\mathbf{R}}_{\mathbf{vv}} = \frac{1}{T}\sum_{l=1}^{T}\mathbf{y}(l)\mathbf{y}^H(l)p\left[\mathcal{H}_{N+1}|\mathbf{y}(l)\right] \tag{7}$$

and (ii) the nth source covariance matrix $\mathbf{R}_{\mathbf{x}_n\mathbf{x}_n}$ for $n = 1,...N$, which is empirically well approximated as

$$\hat{\mathbf{R}}_{\mathbf{x}_n\mathbf{x}_n} = \frac{1}{T}\sum_{l=1}^{T}\mathbf{y}(l)\mathbf{y}^H(l)p\left[\mathcal{H}_n|\mathbf{y}(l)\right] - \hat{\mathbf{R}}_{\mathbf{vv}}. \tag{8}$$

4. POSTERIOR PROBABILITY ESTIMATION

To estimate $p\left[\mathcal{H}_n|\mathbf{y}(l)\right]$, $n = 1,...,N+1$, we first recall that the vector of observations bears two types of information: the desired speech spectra and the spatial information (propagation environment, source location, and array geometry). In our work, we assume that both types of information can be captured by a scalar and a vector variables denoted as $\mathcal{Y}(l)$ and $\boldsymbol{\psi}(l)$, respectively, and we have $p\left[\mathcal{H}_n|\mathbf{y}(l)\right] = p\left[\mathcal{H}_n|\boldsymbol{\psi}(l),\mathcal{Y}(l)\right],\ n = 1,...,N+1$. By defining $Q_n(l) = p\left[\boldsymbol{\psi}(l)|\mathcal{H}_n\right]$ and $P_n(l) = p\left[\mathcal{Y}(l),\mathcal{H}_n\right]$, we can demonstrate that [9]

$$p\left[\mathcal{H}_n|\boldsymbol{\psi}(l),\mathcal{Y}(l)\right] = \frac{Q_n(l)P_n(l)}{\sum_{n'=1}^{N+1}Q_{n'}(l)P_{n'}(l)}. \tag{9}$$

Here, it is important to point out that in contrast to [6, 9], we further include the noise contribution to the observed mixtures of sounds in the computation of the posterior probabilities.

4.1. Using the Spatial Cue

In [6], it was demonstrated that the spatial information of the source can be captured using the normalized vector

$$\boldsymbol{\psi}(l) = \frac{\mathbf{y}(l)}{\|\mathbf{y}(l)\|}. \tag{10}$$

Indeed, when the nth source is dominant, we have $\boldsymbol{\psi}(l) \approx \frac{\mathbf{h}_n}{\|\mathbf{h}_n\|}\frac{S_n(l)}{|S_n(l)|}$, thereby meaning that $\boldsymbol{\psi}(l)$ is located within the vicinity of the steering vector of the source up to a certain complex scaling term (the effect of additive noise is investigated experimentally). The distribution of $\boldsymbol{\psi}(l)$ can be well approximated by a complex Gaussian-like density function [6]

$$p\left[\boldsymbol{\psi}(l)|\mathcal{H}_n\right] = \frac{1}{\left(\pi\sigma_n^2\right)^{M-1}}\exp\left[-\frac{\left\|\boldsymbol{\psi}(l) - [\mathbf{a}_n^H\boldsymbol{\psi}(l)]\mathbf{a}_n\right\|^2}{\sigma_n^2}\right] \tag{11}$$

$\mathbf{a}_n$ is the centroid with unit norm of the nth cluster and σ_n^2 is the variance. A similar model can be forced for the normalized noise model. Hence, the density function of $\boldsymbol{\psi}(l)$ is

$$p\left[\boldsymbol{\psi}(l)|\theta\right] = \sum_{n=1}^{N+1}\alpha_n p\left[\boldsymbol{\psi}(l)|\mathcal{H}_n\right] \tag{12}$$

where $\theta = \{\mathbf{a}_1,\sigma_1,\alpha_1...,\mathbf{a}_{N+1},\sigma_{N+1},\alpha_{N+1}\}$, $\sum_n\alpha_n = 1$, and $0 \leq \alpha_n \leq 1$. We can demonstrate that [6], in an iterative EM scheme, for a given old estimate θ' and $n = 1,...,N+1$, $\mathbf{a}_n$ corresponds to the maximum eigenvector of the matrix $\mathbf{R} = \sum_{l=1}^{T}p\left[\mathcal{H}_n|\boldsymbol{\psi}(l),\theta'\right]\boldsymbol{\psi}(l)\boldsymbol{\psi}^H(l)$, $\sigma_n^2 = \frac{\sum_{l=1}^{T}p\left[\mathcal{H}_n|\boldsymbol{\psi}(l),\theta'\right]\left\|\boldsymbol{\psi}(l)-(\mathbf{a}_n^H\boldsymbol{\psi}(l))\mathbf{a}_n\right\|^2}{(M-1)\sum_{l=1}^{T}p\left[\mathcal{H}_n|\boldsymbol{\psi}(l),\theta'\right]}$, and $\alpha_n = \frac{1}{T}\sum_{l=1}^{T}p\left[\mathcal{H}_n|\boldsymbol{\psi}(l),\theta'\right]$. $Q_n(l)$ is computed as in (11).

[1] This can also be theoretically justified since the energy of the nth speech signal dominates when $\mathcal{H}_n$ is verified.

4.2. Using the Spectral Cue

In this section, we further take advantage of the spectral information by defining

$$\mathcal{Y}(l) = \log\left[\|\mathbf{y}(l)\|^2 / M\right]. \quad (13)$$

Note that the averaging operation over the M observations flattens the reverberant channel and reduces the additive noise. It is common to model the distribution of the log-spectra of speech, $\mathcal{S}(l)$, using a GMM, i.e.,

$$p\left[\mathcal{S}(l)\right] = \sum_{i=1}^{G} \gamma_i \beta_i\left[\mathcal{S}(l)\right] \quad (14)$$

where G is the number of Gaussian components, $\beta_i\left[\mathcal{S}(l)\right] = \mathcal{N}\left(\mathcal{S}(l), \mu_i, \sigma_i^2\right)$, and $(\gamma_i, \mu_i, \sigma_i^2)$ are trained off-line. Here we assume that we have a single model for all speech log-spectra even though different models can be used if the signals are taken from different databases. The cumulative distribution function (CDF) of the ith Gaussian component is denoted as $\Psi_i(x) = \int_{-\infty}^{x} \beta_i(s)\, ds$. Furthermore, we model the noise log-spectrum using a single Gaussian, $\beta^{(N+1)}(\cdot)$, with mean μ_v and covariance σ_v^2. The noise CDF is denoted $\Psi^{(N+1)}(\cdot)$. In contrast to (μ_i, σ_i^2) which are estimated using a training data set, (μ_v, σ_v^2) are obtained from the observed data by assuming that we have a primary estimate of $p\left[\mathcal{H}_{N+1}|\mathbf{y}(l)\right]$ and computing $\mu_v = \frac{1}{T}\sum_{l=1}^{T} p\left[\mathcal{H}_{N+1}|\mathbf{y}(l)\right]\mathcal{Y}(l)$ and $\sigma_v^2 = \frac{1}{T}\sum_{l=1}^{T} p\left[\mathcal{H}_{N+1}|\mathbf{y}(l)\right]\mathcal{Y}(l)^2 - \mu_v^2$.

To have a tractable formulation, it is convenient to consider the most significant Gaussian component of the speech log-spectra. For the nth source, this index is denoted $i^{(n)}$ and its selection from the G possible values will be detailed next. Using the log-max model as in [9], it is possible to demonstrate that a good approximation of the sound mixture distribution when the nth speech signal, $n = 1, ..., N$, is dominating is given by[2]

$$p\left[\mathcal{Y}(l), \mathcal{H}_n | i^{(n)}\right] = \beta_{i^{(n)}}\left[\mathcal{Y}(l)\right] \Psi^{(N+1)}\left[\mathcal{Y}(l)\right] \prod_{\substack{n=1 \\ n' \neq n}}^{N} \Psi_{i^{(n')}}\left[\mathcal{Y}(l)\right] \quad (15)$$

and $p\left[\mathcal{Y}(l), \mathcal{H}_{N+1}\right] = \beta^{(N+1)}\left[\mathcal{Y}(l)\right] \prod_{n=1}^{N} \Psi_{i^{(n)}}\left[\mathcal{Y}(l)\right]$. Now, to select the most significant Gaussian index of the nth source, $i^{(n)}$, we have to maximize the following likelihood function [9]

$$\begin{aligned}\mathcal{L}^{(n)}(i) &= p\left[\mathcal{H}_n|\mathbf{y}(l)\right] \log\left(\beta_i\left[\mathcal{Y}(l)\right]\right) \quad (16)\\ &\quad + \left(1 - p\left[\mathcal{H}_n|\mathbf{y}(l)\right]\right) \log\left(\Psi_i\left[\mathcal{Y}(l)\right]\right) + \log\left[p(i)\right].\end{aligned}$$

Finally, we implement our algorithm that combines all steps described above as: (a) use the approximation $p\left[\mathcal{H}_n|\mathbf{y}(l)\right] \approx p\left[\mathcal{H}_n|\boldsymbol{\psi}(l)\right] = \alpha_n Q_n(l)/p\left[\boldsymbol{\psi}(l)|\theta\right]$ as in [6, 8] to determine some initial estimates of the $N+1$ posterior probabilities, (b) for $n = 1, ..., N$, find the Gaussian component that maximizes the likelihood function in (16), (c) update the noise statistics using $p\left[\mathcal{H}_{N+1}|\mathbf{y}(l)\right]$, (d) for $n = 1, ..., N$, calculate $p\left[\mathcal{Y}(l), \mathcal{H}_n|i^{(n)}\right]$ and $p\left[\mathcal{Y}(l), \mathcal{H}_{N+1}\right]$, set $P_n(l) \approx p\left[\mathcal{Y}(l), \mathcal{H}_n|i^{(n)}\right]$ then calculate (9), (e) iterate few times steps (b) to (d).

5. EXPERIMENTAL RESULTS

We implement the proposed method to separate two speech signals in a reverberant and noisy environment. We investigate two methods to estimate the posterior probability: using only the space information, i.e., assuming $p\left[\mathcal{H}_n|\mathbf{y}(l)\right] = p\left[\mathcal{H}_n|\boldsymbol{\psi}(l)\right]$ as we proposed in [8] and using both the space and spectrum information, i.e., assuming $p\left[\mathcal{H}_n|\mathbf{y}(l)\right] = p\left[\mathcal{H}_n|\boldsymbol{\psi}(l), \mathcal{Y}(l)\right]$. Both posteriors are then combined with the MVDR and Wiener filters leading to the L-MVDR, L-Wiener, LS-MVDR, and LS-Wiener (L stands for location-based and LS stands for location-and-spectrum-based), respectively. The resulting four filters are compared to a very robust implementation of an ICA-based algorithm combining the FastICA and InfoMax algorithms (using higher-order statistics) [1, 2, 3]. We also implement the masking-based method in [6]. The results are given in terms of output signal-to-noise ratio (SNR), signal-to-interference ratio (SIR), signal to artificial distortion ratio (SAR), signal to distortion ratio (SDR) [10] and the perceptual evaluation of speech quality (PESQ).

In our experiments, we have 3 data sets each consisting of 10 pairs of speakers (30 combinations in total) from the test set of the TIMIT database: two female speakers, two male speakers, and one male and one female speakers. The speech signals are convolved with actual measurements of acoustic impulse responses which are measured using a uniform circular array of 16 microphones with radius $r = 0.15$ m in a reverberant room with $T_{60} = 0.31$ s. Both speakers are located on the same plane at a distance 2 m away from the microphone array center and with an angular separation of 160 degrees. Segments of babble noise taken from the noisex database [11] are added to each of the microphone signals. The long-term input SIR at every microphone is approximately 0 dB while the noise segments are added at different SNR values as specified below. To exploit the spectral cue of the speech sources, we train a GMM of $G = 256$ components using the training set of the TIMIT database. The 16 microphone recordings are chopped in 64 ms-long frames with 50% overlap and processed by all methods.

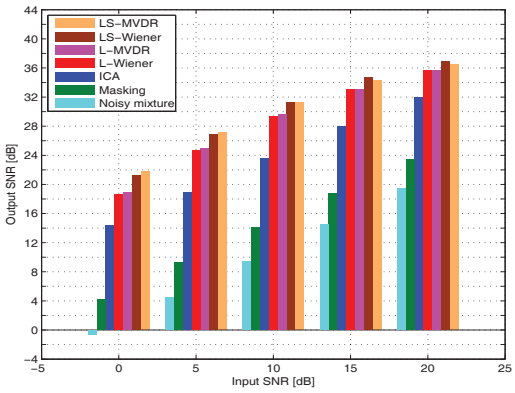

Fig. 1. Output SNR comparisons at different input SNR levels.

In Fig. 1, we see that the proposed MMSE-based filtering approaches (L-MVDR, L-Wiener, LS-MVDR, LS-Wiener) remarkably outperform the ICA and Masking. The reason is that the noise contribution is accounted for when extracting the speech sources and the space information is optimally exploited by the MMSE criterion. We also notice that by including the spectral information, the noise suppression becomes more significant as compared with the case where only the space information is used for posterior estimation. However, as we can see from Fig. 2, LS-MVDR and LS-Wiener achieve a slightly lower output SIR. In Fig. 3, we see that the new MMSE-based approaches outperform the ICA. We also observe that by using the spectral information in the computation of the posterior probabilities, we can

[2] In contrast to [9], we further include the noise in this model.

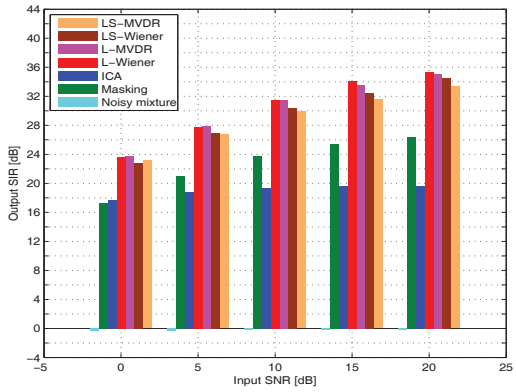

Fig. 2. SIR comparisons at different input SNR levels.

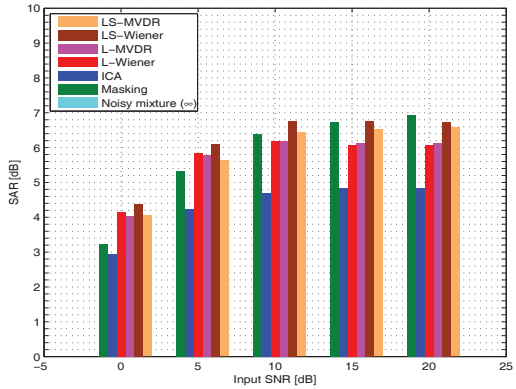

Fig. 3. SAR comparisons at different input SNR levels.

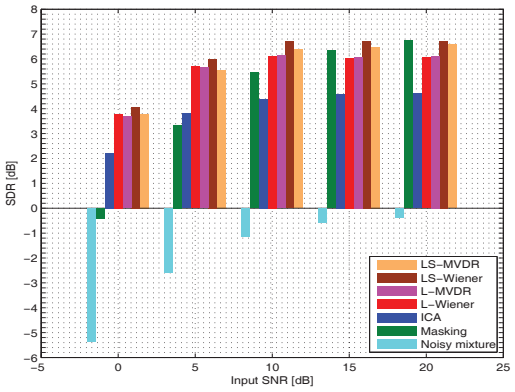

Fig. 4. SDR comparisons at different input SNR levels.

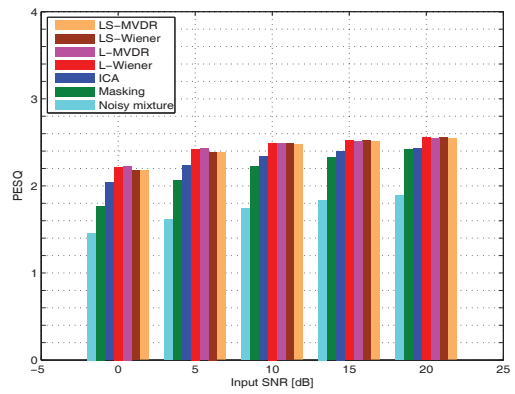

Fig. 5. PESQ comparisons at different input SNR levels.

reduce the level of speech distortion, especially with the MVDR filter. The same remarks hold for the output SDR as it is shown in Fig. 4. Finally, we can conclude from Fig. 5 that by using the proposed processing, even with the space information only, it is possible to achieve higher quality of the filtered speech signals. Our informal subjective evaluations corroborate this fact.

6. CONCLUSION

In this paper, we proposed a new multichannel MMSE-based framework for joint BSS and noise reduction. We demonstrated that it is possible to track the activities of multiple speakers using both spatial and spectral information contained in the recorded sound mixtures. Then, we estimated the posterior probabilities describing the activities of the speakers in an EM framework and used these probabilities to compute all statistics required to implement the MMSE-based estimator of every speech source. Our experiments demonstrated that our method performs remarkably well in reverberant and noisy environments.

7. REFERENCES

[1] A. J. Bell and T. J. Sejnowsky, "An imformation maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[2] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, pp. 626–634, 1999.

[3] 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. Springer, 2005.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, pp. 1830–1847, 2004.

[5] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda,"Sound source segregation based on estimating the incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical science& technology*, vol. 22, pp. 149–157, 2001.

[6] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignement," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, pp. 516–527, 2011.

[7] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 18, pp. 260–276, 2010.

[8] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "Simultaneous speech source separation and noise reduction via clustering and MMSE-based filtering," in *Proc. IEEE ICSPCC*, 2011, pp. 389–394.

[9] T. Nakatani, S. Araki, M. Delcroix, T. Yoshioka, and M. Fujimoto, "Reduction of highly nonstationary ambient noise by integrating spectral and locational characteristics of speech and noise," in *Proc. of INTERSPEECH*, 2011, pp. 1785–1788.

[10] E. Vincent, R. Gribonval, C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Speech, Audio Process.*, vol. 14, pp. 1462–1469, 2006.

[11] A. P. Varga, H. J. M. Steenekan, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," tech. rep., DRA Speech Research Unit, 1992.