Online Multichannel Speech Enhancement Based on Recursive EM and DNN-based Speech Presence Estimation

Juan M. Martín-Doñas, Jesper Jensen, Zheng-Hua Tan, *Senior Member, IEEE,* Angel M. Gomez, and Antonio M. Peinado, *Senior Member, IEEE*

Abstract

This paper presents a recursive expectation-maximization algorithm for online multichannel speech enhancement. A deep neural network mask estimator is used to compute the speech presence probability, which is then improved by means of statistical spatial models of the noisy speech and noise signals. The clean speech signal is estimated using beamforming, single-channel linear postfiltering and speech presence masking. The clean speech statistics and speech presence probabilities are finally used to compute the acoustic parameters for beamforming and postfiltering by means of maximum likelihood estimation. This iterative procedure is carried out on a frame-by-frame basis. The algorithm integrates the different estimates in a common statistical framework suitable for online scenarios. Moreover, our method can successfully exploit spectral, spatial and temporal speech properties. Our proposed algorithm is tested in different noisy environments using the multichannel recordings of the CHiME-4 database. The experimental results show that our method outperforms other related state-of-the-art approaches in noise reduction performance, while allowing low-latency processing for real-time applications.

Index Terms

Recursive expectation-maximization, multichannel speech enhancement, deep neural networks, speech presence probability, Kalman filter

This work was supported in part by the Spanish MICINN/FEDER Project PID2019-104206GB-I00 and in part by the Spanish Ministry of Universities through the research stay grants of the National Program FPU under reference FPU15/04161.

J. M. Martín-Doñas, A. M. Peinado and A. M. Gomez are with the Department of Signal Theory, Telematics and Communications, Universidad de Granada, 18071 Granada, Spain (e-mail: mdjuamart@ugr.es; amgg@ugr.es; amp@ugr.es).

J. Jensen is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark, and also with Oticon A/S, 2765 Smørum, Denmark (e-mail: jje@es.aau.dk).

Z.-H. Tan is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: zt@es.aau.dk).

Online Multichannel Speech Enhancement Based on Recursive EM and DNN-based Speech Presence Estimation

I. Introduction

Multichannel speech enhancement techniques have gained an increasing interest in the last decade mainly due to the availability of communication devices with microphone arrays and increasing computational capabilities [1]. These techniques improve the performance on different speech processing tasks, such as noise reduction, dereverberation or source separation, and can be applied in different scenarios: mobile communications, far-field speech recognition, hearing aid devices, etc.

The most common multichannel technique is beamforming [2], which applies a spatial filter to the multichannel speech signal for interference reduction. Different beamformers with different goals can be found in the literature. For example, the minimum variance distortionless response (MVDR) [3], [4] preserves the target signal, while minimizing the power of the noise. The multichannel Wiener filter (MWF) [5] estimates the clean speech signal in terms of linear minimum mean square error (MMSE) and can be decomposed into an MVDR beamformer and a single-channel Wiener postfilter. This beamformer-plus-postfilter architecture has been also explored for non-linear postfilters [6], [7]. Recently, a multichannel Kalman filter (MKF) was proposed in [8], showing improvements with respect to the classical MWF. The authors also showed that the algorithm is equivalent to an MVDR beamformer followed by a single-channel modulation-domain Kalman filter (KF) [9]. The time-domain modeling capability of KF has also been exploited in other speech processing tasks, as multichannel online dereverberation [10], [11] or noise power spectral density tracking [12].

The previous methods require knowledge of acoustic parameters: the second-order statistics of clean speech and noise, the relative transfer function (RTF) between acoustic channels [13] and, in the case of KF, a linear prediction model for the clean speech spectra amplitudes. Frequently, certain assumptions about the noisy speech signal are made to estimate these parameters. For example, the estimation of the noise spatial statistics can be addressed assuming a specific noise field [14]–[16] or the availability of knowledge about the RTF [17], [18]. A possible solution to overcome the need for assumptions is that of exploiting the sparsity of the clean speech signal in the time-frequency (TF) domain. The problem is then the estimation of the speech presence probability (SPP) [19], [20] in each TF bin, so that the noise spatial statistics can be estimated by an SPP-controlled recursive procedure [21]. Therefore, these statistics are computed only in bins where speech is absent. Then, the RTF can be estimated from the clean speech spatial statistics e.g. using sub-space methods [22]. Alternatively, some recent deep neural network (DNN) mask estimators [23]–[28] obtain speech and noise dominant soft masks for the estimation of the clean speech and noise spatial statistics needed for beamforming.

Another approach for parameter estimation is that derived from the adoption of a Bayesian framework. In this case, the acoustic parameters can be obtained using a maximum likelihood estimation (MLE), which requires knowledge about the clean speech statistics. The expectation-maximization (EM) iterative algorithm can be used to jointly estimate the clean speech signal and the acoustic parameters. This method has been applied to different offline multichannel speech enhancement, dereverberation and source separation problems [29]–[32]. In [33], an EM multichannel source separation technique using MWF, where the source spectra are estimated using DNNs models, was proposed. On the other hand, to deal with online scenarios, a recursive version of the EM algorithm (REM) [34] can be used instead. This approach was used for speech dereverberation [10] and speech enhancement [35].

Other EM approaches are based on the estimation of the active speech source in each TF bin during the E-step [36], [37], which corresponds to the SPP in the one-speaker case. This clustering approach allows for the estimation of the acoustic parameters during the M-step. Finally, an MVDR beamformer or MWF can be applied for the enhancement task. In [38] an online estimation of the SPP under the EM framework is proposed, establishing a relationship between the MLE approach and the recursive procedure in [21]. The integration of DNNs and statistical models has been also evaluated in [39], [40]. A new approach that estimates both the clean speech statistics and the predominant speaker was proposed in [41] for offline blind source separation.

In this work we propose a novel REM framework for multichannel speech enhancement that incorporates a DNN-based SPP estimator into the framework. Our proposal uses beamforming and postfiltering to enhance the noisy speech signal, employing two different postfilters, Wiener and Kalman, in its formulation. In this way, the SPP estimation performed by a DNN spectral model is further refined using the estimated spatial statistics. Moreover, the obtained clean speech and SPP estimates allow for a re-estimation of the acoustic parameters of the model. The Kalman-based posfiltering variation of our proposal is inspired by the Switching Kalman filter framework proposed in [42], but we simplify it into two models for speech presence and absence, respectively, and the transition probabilities between states are replaced by the estimated probabilities given by a DNN. A remarkable advantage of our approach is the joint estimation, in an online fashion, of the different statistics and parameters required for beamforming as well as postfiltering. This allows for better performance, while keeping the algorithm suitable for real-time applications. The evaluation of the algorithm and its comparison with other state-of-the-art approaches show the benefits of the proposal.

The remainder of this paper is organized as follows. First, the main contributions of our proposal and their relation with other existing techniques are discussed in Section II. The statistical model and the introduction of the recursive EM algorithm is formulated in Section III. The estimation of the clean speech statistics and SPP (E-step), and the acoustic parameters (M-step) is derived in Section IV and V, respectively. In Section VI, the integration of the DNN SPP estimator in the algorithm is described, and some practical considerations are shown in Section VII. Finally, the experimental framework and results are presented in Section VIII, and conclusions are drawn in Section IX.

II. Contributions and related work

In this paper we develop a complete REM framework for multichannel speech enhancement, which considers the SPP of each TF bin, and integrates DNNs to improve the robustness against noise. Conveniently gathering a number of techniques from the state-of-the-art, we propose a novel framework which outperforms other existing approaches. The main contributions of our proposal are:

1) The derivation of a novel REM algorithm for speech enhancement that fully integrates a joint estimation of the clean speech statistics, the SPPs and the different acoustic parameters of the noisy speech signal.
2) The use of Kalman filtering to model the temporal properties of the clean speech signal in the REM framework, allowing for MLE estimation of the KF parameters.
3) An estimator of the clean speech power spectral density that avoids the distortion introduced by the SPPs in the clean speech signal at the system output.
4) The use of DNN prior estimates in the REM framework in order to improve the robustness in non-stationary noisy scenarios.

The use of EM frameworks for offline speech processing tasks has been explored in different works [29]–[32]. The EM source separation approach in [41] extended most of these previous ideas with the use of speech presence posteriors to better discriminate between the active source at each TF bin. However, this approach has several drawbacks, including the distortion introduced by the SPP in the estimated speech signals or the potentially high number of iterations needed. Alternatively, a REM framework for multichannel speech enhancement was proposed in [35]. However, this approach only considers a simple distortion model with delayed responses between microphones, it does not include the SPP in its analysis, and a priori knowledge of some acoustic parameters (RTFs or noise statistics) is needed. As shown in the following sections, our approach overcomes these limitations.

The proposed REM approach extends and adapts the idea of MKF in the STFT amplitude domain [8] by changing the filtering model so that it can better represent the abrupt changes of natural clean speech. In addition, the required acoustic parameters are more accurately computed thanks to the MLE approach employed in our proposal, which considers the clean speech estimates and the SPP probabilities. Although KFs have previously been used in a REM framework [10] and a multichannel linear prediction framework [11], these approaches model the effect of a convolutive transfer function in scenarios with reverberations. As such, the proposed KF uses a completely different state-space model, which describes the temporal correlations in the clean speech amplitude coefficients and the multichannel distortion model.

Finally, we must point out that the use of DNNs in an EM framework for multichannel speech enhancement has only been very little explored in existing works. Unlike other works like [33], [39], [40], we propose an approach that exploits the DNN outputs as a priori SPP estimates to jointly obtain the clean speech statistics, the a posteriori SPPs and the acoustic parameters. Specifically, the approach in [33] uses DNNs to estimate the speech sources spectra and does not consider the speech presence. Furthermore, [39] and [40] combine DNNs with spatial statistical models to compute speech and noise dominant masks, which are then used to obtain the beamformer coefficients. Thus, a clustering approach is used where the TF bins are classified as speech or noise bins, and complex angular

models are applied to model the distribution of the normalized noisy vectors. The use of the clean speech estimates in our statistical models improves the estimation of the a posteriori SPP and the acoustic parameters. As a result, and in contrast to previous methods, our model allows noise estimation even in speech presence frames, which increases the robustness against non-stationary noises.

III. Formulation of the statistical model

Let us first express the multichannel noisy speech signal, captured by a microphone array, in the short-time Fourier transform (STFT) domain under a narrowband assumption [2] as,

$$\mathbf{y}_{t,f} = \mathbf{h}_{t,f} X_{1,t,f} + \mathbf{n}_{t,f}, \tag{1}$$

where $X_{1,t,f}$ is the clean speech signal at the reference microphone (in the following, we use the microphone with index 1 as the reference microphone, for convenience) and

$$\mathbf{y}_{t,f} = \begin{bmatrix} Y_{1,t,f} & Y_{2,t,f} & \cdots & Y_{M,t,f} \end{bmatrix}^\top, \tag{2}$$

$$\mathbf{n}_{t,f} = \begin{bmatrix} N_{1,t,f} & N_{2,t,f} & \cdots & N_{M,t,f} \end{bmatrix}^\top, \tag{3}$$

$$\mathbf{h}_{t,f} = \begin{bmatrix} 1 & H_{21,t,f} & \cdots & H_{M1,t,f} \end{bmatrix}^\top, \tag{4}$$

represent the noisy speech, the noise and the set of relative transfer functions, respectively. The model in (1) is defined under a speech presence hypothesis ($\mathcal{H}_x$). When speech is absent ($\mathcal{H}_n$), this model can be simplified to

$$\mathbf{y}_{t,f} = \mathbf{n}_{t,f}. \tag{5}$$

From now on, with no loss of generality, we will omit the frequency index f for the sake of simplicity.

We assume that the noise signal follows a circularly symmetric complex normal distribution, $\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Phi}_{N,t})$, and that it is uncorrelated with the speech signal. The clean speech $X_{1,t}$ is a zero mean circularly symmetric complex random variable, whose variance, under speech presence assumption, can be defined as

$$\phi_{x,t} = E\left\{ |X|_{1,t}^2 \middle| \mathcal{H}_x \right\}, \tag{6}$$

where we define $|X|_{1,t} \triangleq |X_{1,t}|$. In addition, let

$$q_{x,t} = P\left(\mathcal{D}_t = \mathcal{H}_x\right) \tag{7}$$

denote the a priori SPP, where $\mathcal{D}_t = \{\mathcal{H}_x, \mathcal{H}_n\}$ is a discrete random variable, which indicates speech presence/absence in a time-frequency bin.

Additionally, we employ a single-channel temporal linear prediction model over the clean speech amplitudes [8],

$$|X|_{1,t} = \mathbf{a}_t^\top \mathbf{x}_{t-1} + V_t, \tag{8}$$

where

$$\mathbf{x}_{t-1} \triangleq \begin{bmatrix} |X|_{1,t-1} & |X|_{1,t-2} & \cdots & |X|_{1,t-p} \end{bmatrix}^\top \tag{9}$$

is a vector of clean speech amplitudes from previous time frames,

$$\mathbf{a}_t = \begin{bmatrix} A_{t,1} & A_{t,2} & \cdots & A_{t,p} \end{bmatrix}^\top \tag{10}$$

is the vector of linear prediction coefficients (LPC), $V_t \sim \mathcal{N}(0, \phi_{v,t})$ is the prediction error and p is the prediction order.

Using the aforementioned models, we can define the likelihood of the data sequence until time t as $f(\mathbf{y}_{1:t}, X_{1,1:t}, \mathcal{D}_{1:t}; \mathbf{\Theta}_{1:t})$, where $\mathbf{y}_{1:t}$ is the observable data (until time t), $X_{1,1:t}$ and $\mathcal{D}_{1:t}$ are the latent variables, and $\mathbf{\Theta}_t = \{\mathbf{a}_t, \phi_{v,t}, \mathbf{h}_t, \mathbf{\Phi}_{N,t}, q_{x,t}\}$ are the required model parameters. Assuming a Markov process, this likelihood can be developed as

$$\begin{aligned} &f(\mathbf{y}_{1:t}, X_{1,1:t}, \mathcal{D}_{1:t}; \mathbf{\Theta}_{1:t}) = \\ &f(\mathbf{x}_0) \prod_{\tau=1}^{t} P(\mathcal{D}_\tau) \cdot f\left(|X|_{1,\tau} \middle| \mathbf{x}_{\tau-1}, \mathcal{D}_\tau; \mathbf{a}_t, \phi_{v,t} \right) \cdot \\ &f\left(X_{1,\tau} \middle| |X|_{1,\tau}, \mathcal{D}_\tau \right) \cdot f\Big(\mathbf{y}_\tau | X_{1,\tau}, \mathcal{D}_\tau; \mathbf{h}_t, \mathbf{\Phi}_{N,t} \Big). \end{aligned} \tag{11}$$

This separates the linear prediction model for the clean speech amplitudes, $f\left(|X|_{1,\tau} \middle| \mathbf{x}_{\tau-1}, - \right)$, and the multichannel noisy observation model given the clean speech signal, $f(\mathbf{y}_\tau | X_{1,\tau}, -)$. The phase term $f\left(X_{1,\tau} \middle| |X|_{1,\tau}, \mathcal{D}_\tau \right)$ will be ignored as the phase estimation is not considered in our REM framework (we keep the phase provided by the beamformer, see Section IV). We are interested in an online estimation of the clean speech signal at the reference microphone. To this end, we define the exponentially-weighted log-likelihood of the data sequence at time t,

$$\mathcal{L}_{\lambda,t} = \sum_{\tau=1}^{t} \lambda^{t-\tau} \log f(\mathbf{y}_\tau, X_{1,\tau}, \mathcal{D}_\tau; \mathbf{\Theta}_\tau), \tag{12}$$

where $\lambda \in (0, 1]$ is a forgetting factor. Our objective is to obtain an MLE estimate of the model parameters at each time step. There is no closed form solution to this problem, as we have to estimate at the same time the latent variables and the model parameters. Instead, we can use the REM algorithm [34] to achieve a good approximation. This is a frame-wise procedure which is repeated for a given frame until a number of iterations is reached. Given the computed model parameters $\mathbf{\Theta}_t^l$ at iteration l, the parameters are re-computed by means of the following two-step procedure:

- **E-step**: An auxiliary function is calculated taking the conditioned expectation of the log-likelihood $\mathcal{L}_{\lambda,t}$ given the observations and the current parameters,

$$Q\left(\mathbf{\Theta}_t | \, \mathbf{\Theta}_t^l \right) = E\left\{ \mathcal{L}_{\lambda,t} | \, \mathbf{y}_t; \mathbf{\Theta}_t^l \right\}, \tag{13}$$

  This results in a function that depends on the conditional expectations over the latent variables $X_{1,t}$ and $\mathcal{D}_t$. Therefore, we will need to estimate the first- and second-order moments of $X_{1,t}$, and the a posteriori SPP,

$$p_{x,t}^l = P\left(\mathcal{D}_t = \mathcal{H}_x | \, \mathbf{y}_t; \mathbf{\Theta}_t^l \right), \tag{14}$$

  in order to be able to derive the M-step by using the Q function.

- **M-step**: Once the expectations over the latent variables are computed, a new set of parameters is obtained by maximizing the auxiliary function,

$$\mathbf{\Theta}_t^{l+1} = \underset{\mathbf{\Theta}_t}{\operatorname{argmax}} \; Q\left(\mathbf{\Theta}_t | \, \mathbf{\Theta}_t^l \right). \tag{15}$$

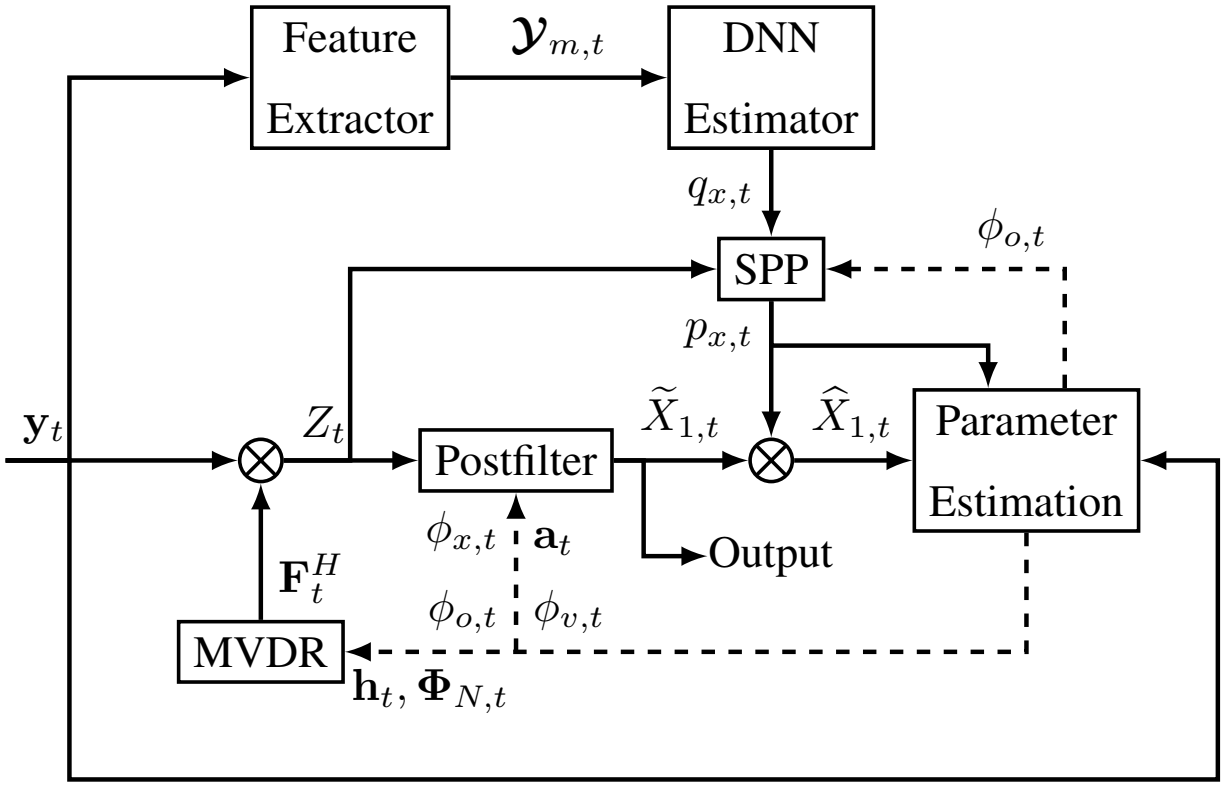


Fig. 1. Block diagram of the proposed REM algorithm for multichannel speech enhancement. Only the most relevant parts of the algorithm are indicated in the flowchart for clarity purposes. The dashed lines mean the feedback due to the M-step of the algorithm.

Fig. 1 depicts a diagram of our proposed REM algorithm for speech enhancement. In the next sections we will detail each block, the parameters involved in each one and the latent variables they depend on. For simplicity, we will omit the iteration index l in the following two sections, where the E-step and M-step of the algorithm are particularized.

IV. Estimation of the latent variables

The E-step of the algorithm considers the computation of the expectation in Eq. (13), which gives the auxiliary Q function. This Q function, and its maximization for the computation of the acoustic parameters (M-step), depends on the a posteriori SPP $p_{x,t}$ and the statistics of the clean speech signal $X_{1,t}$ (see Appendix A for this derivation). In this section we will address the estimation of these expectations and the a posteriori SPP.

The first- and second-order expectations of the clean speech signal $X_{1,t}$ conditioned on the observations are obtained using the E-step in the REM framework as indicated in [41],

$$\widehat{X}_{1,t} = E\left\{ X_{1,t} | \, \mathbf{y}_t; \boldsymbol{\Theta}_t \right\} = p_{x,t} \widetilde{X}_{1,t}, \tag{16}$$

$$S_{x,t} = E\left\{ |X|_{1,t}^2 \Big| \, \mathbf{y}_t; \boldsymbol{\Theta}_t \right\} = \widehat{|X|}_{1,t}^2 + P_t, \tag{17}$$

where

$$\widetilde{X}_{1,t} = E\left\{ X_{1,t} | \, \mathbf{y}_t, \mathcal{H}_x; \boldsymbol{\Theta}_t \right\} \tag{18}$$

is the filtered clean speech signal (i.e. MMSE estimate and system output) when speech presence is assumed [38], and

$$P_t = E\left\{ \left| X_{1,t} - \widehat{X}_{1,t} \right|^2 \middle| \mathcal{H}_x \right\} \tag{19}$$

is the error variance for the estimated clean speech signal when speech presence is assumed (i.e. $\widehat{X}_{1,t} = \widetilde{X}_{1,t}$).

The expectations in (18) and (19) are obtained using a multichannel MMSE estimator, which can be implemented by concatenating a spatial filtering stage with a single-channel linear postfilter. We first apply an MVDR beamformer to the noisy speech signal,

$$Z_t = \mathbf{F}_t^H \mathbf{y}_t, \tag{20}$$

with the MVDR beamformer coefficients given by [4],

$$\mathbf{F}_t = \frac{\mathbf{\Phi}_{N,t}^{-1} \mathbf{h}_t}{\mathbf{h}_t^H \mathbf{\Phi}_{N,t}^{-1} \mathbf{h}_t}. \tag{21}$$

Under the distortionless constraint of the MVDR beamformer, the signal at the beamformer output is given by

$$Z_t = X_{1,t} + O_t, \tag{22}$$

where $O_t \sim \mathcal{N}\left(0, \phi_{o,t}\right)$ is the residual noise, with variance

$$\phi_{o,t} = \left(\mathbf{h}_t^H \mathbf{\Phi}_{N,t}^{-1} \mathbf{h}_t \right)^{-1}. \tag{23}$$

Then, the single-channel postfilter is applied to the beamformer output Z_t to obtain $\widetilde{X}_{1,t}$ and its error P_t. This postfilter only modifies the amplitude of the signal, while the phase remains the same as that of the MVDR output.

We use $\widetilde{X}_{1,t}$ as the output signal in our REM framework instead of $\widehat{X}_{1,t}$. This is because, in practice, the SPP masking in (16) introduces speech distortions that deteriorate the speech quality and intelligibility. Nevertheless, the estimation $\widehat{X}_{1,t}$ is still required to obtain the acoustic parameters in the M-step.

In the next subsections we will describe the two different linear postfilters that we use in our REM framework. Also, we will define an estimator for the a posteriori SPP $p_{x,t}$.

A. Wiener filter

The Wiener filter (WF) is a linear MMSE estimator that only considers the variance of the clean speech and the noise at the current time-frequency bin. The filtered clean speech signal is obtained as [43]

$$\widetilde{X}_{1,t}^{(\mathrm{WF})} = W_t Z_t \tag{24}$$

where

$$W_t = \frac{\phi_{x,t}}{\phi_{x,t} + \phi_{o,t}} = \frac{\xi_t}{\xi_t + 1}. \tag{25}$$

is the Wiener gain, and $\xi_t = \phi_{x,t}/\phi_{o,t}$ is the a priori signal-to-noise ratio (SNR). The error variance in (19) is computed as

$$P_t^{(\mathrm{WF})} = \left(1 - W_t\right) \phi_{x,t}. \tag{26}$$

B. Kalman filter

The Kalman filter (KF) takes into account the time-domain linear prediction model for the clean speech amplitudes described by Eq. (8). We slightly modify the modulation-domain Kalman filter proposed in [8] by estimating only the filtered speech signal at the current time step given the estimated clean speech from previous frames, which differs from the standard Kalman filtering for vector states [43]. First, we consider an estimate $\widehat{\mathbf{x}}_{t-1}$ of the vector of clean speech amplitudes in the previous time steps, $\mathbf{x}_{t-1}$ (9). Similarly, we consider a vector $\widetilde{\mathbf{x}}_{t-1}$ with the filtered amplitudes of previous time steps, without the SPP masking. In addition, we define

$$\mathbf{P}_{t-1} = E\left\{\left(\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_{t-1}\right)\left(\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_{t-1}\right)^{\top}\right\}, \tag{27}$$

as the error covariance matrix of this filtered version of the previous clean speech amplitudes.

The temporal prediction model in (8) is used to obtain a prediction of the clean speech amplitude,

$$\widetilde{|X|}_{1,t|t-1} = \mathbf{a}_t^{\top}\widehat{\mathbf{x}}_{t-1}, \tag{28}$$

and its corresponding error variance,

$$P_{t|t-1} = \mathbf{a}_t^{\top}\mathbf{P}_{t-1}\mathbf{a}_t + \phi_{v,t}. \tag{29}$$

The Kalman filter combines the previous prediction and the MVDR output (modeled according to Eqs. (20)-(23)), which gives the following linear MMSE estimator [43],

$$\widetilde{|X|}_{1,t}^{(\mathrm{KF})} = \widetilde{|X|}_{1,t|t-1} + K_t\left(|Z_t| - \widetilde{|X|}_{1,t|t-1}\right) \tag{30}$$

where

$$K_t = \frac{P_{t|t-1}}{P_{t|t-1} + \phi_{o,t}} \tag{31}$$

is the Kalman gain. The error variance in (19) is computed as

$$P_t^{(\mathrm{KF})} = \left(1 - K_t\right)P_{t|t-1}. \tag{32}$$

Moreover, we can compute the cross-covariance error vector between the current frame and the previous frames as

$$\begin{aligned}\mathbf{p}_{t,t-1} &= E\left\{\left(\,|X|_{1,t} - \widetilde{|X|}_{1,t}\right)\left(\,\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_{t-1}\right)^{\top}\right\} = \\ &= \left(1 - K_t\right)\mathbf{a}_t^{\top}\mathbf{P}_{t-1}.\end{aligned} \tag{33}$$

Finally, the filtered estimation of the clean speech signal $\widetilde{X}_{1,t}^{(\mathrm{KF})}$ is obtained by using $\widetilde{|X|}_{1,t}^{(\mathrm{KF})}$ and the phase of Z_t.

The values needed for the next frame, $\widehat{\mathbf{x}}_t$ and $\mathbf{P}_t$, are obtained by using the values of the previous frames and the new estimations,

$$\widehat{\mathbf{x}}_t = \mathbf{U}\widehat{\mathbf{x}}_{t-1} + \mathbf{u}\widehat{|X|}_{1,t}^{(\mathrm{KF})}, \tag{34}$$

$$\mathbf{P}_t = \mathbf{U}\mathbf{P}_{t-1}\mathbf{U}^{\top} + \mathbf{U}\mathbf{p}_{t,t-1}\mathbf{u}^{\top} + \mathbf{u}\mathbf{p}_{t,t-1}^{\top}\mathbf{U}^{\top} + \mathbf{u}P_t^{(\mathrm{KF})}\mathbf{u}^{\top}, \tag{35}$$

where

$$\mathbf{u} = \begin{bmatrix}1 & \mathbf{0}_{1\times p-1}\end{bmatrix}^{\top}, \tag{36}$$

$$\mathbf{U} = \begin{bmatrix} \mathbf{0}_{1\times p} \\ \mathbf{I}_{p-1\times p-1} \quad \mathbf{0}_{p-1\times 1} \end{bmatrix}, \tag{37}$$

are a structure vector and matrix, respectively, $\mathbf{0}$ is a zero vector and $\mathbf{I}$ is the identity matrix.

In the case that there is no prediction of the clean speech amplitudes (the coefficients $\mathbf{a}_t$ are zero), so that $P_{t|t-1} = \phi_{x,t}$, both filters, WF and KF, are equivalent. Therefore, the Kalman filter is a generalization that includes the Wiener filter as a specific case.

C. A posteriori speech presence probability

The a posteriori SPP $p_{x,t}$ defined in (14) can, using the Bayes' rule, be re-written as

$$p_{x,t} = \frac{q_{x,t} f\left(\mathbf{y}_t | \mathcal{H}_x; \mathbf{\Theta}_t\right)}{q_{x,t} f\left(\mathbf{y}_t | \mathcal{H}_x; \mathbf{\Theta}_t\right) + \left(1 - q_{x,t}\right) f\left(\mathbf{y}_t | \mathcal{H}_n; \mathbf{\Theta}_t\right)}. \tag{38}$$

This expression takes into account the a priori SPP and the likelihood of speech presence and absence given the observed data. In our proposed model, the likelihoods in (38) are multivariate Gaussian distributions. They can also be expressed directly from the MVDR output, which simplifies into the following Gaussian likelihoods,

$$f\left(\mathbf{y}_t | \mathcal{H}_x; \mathbf{\Theta}_t\right) = \mathcal{N}\left(Z_t; 0, \phi_{z,t}\right), \tag{39}$$

$$f\left(\mathbf{y}_t | \mathcal{H}_n; \mathbf{\Theta}_t\right) = \mathcal{N}\left(Z_t; 0, \phi_{o,t}\right), \tag{40}$$

where

$$\phi_{z,t} = p_{x,t} S_{x,t} + \phi_{o,t}. \tag{41}$$

is the variance of Z_t given the residual noise variance $\phi_{o,t}$, the second-order statistics of the clean speech signal from (17) and the a posteriori SPP [41].

In our iterative procedure, the a posteriori SPP is first initialized as $p_{x,t}^{l=0} = q_{x,t}$ and, after applying the postfiltering step, it is updated at each E-step iteration.

V. Estimation of the model parameters

In this section we will describe the estimation of the different acoustic parameters by means of the estimated latent variables during the previous E-step. The acoustic parameters for the beamforming and Kalman filtering are obtained using the M-step in our REM framework. These parameters need an estimate of the speech variance under speech presence. Therefore, we will first describe how to obtain this variance.

A. Speech variance

Although the speech variance can be estimated under the REM framework (M-step), this procedure has two problems. First, the REM framework assumes slowly time-variant parameters [34], which is not necessarily true for the speech variance. Secondly, the resulting variance estimate takes into consideration the SPP [41]. Initial experiments reveal that this yields a filtered signal with a high degree of sparsity, which can degrade the perceptual quality and intelligibility of the enhanced speech signal. Therefore, we propose a specific estimation of the speech variance under the speech presence assumption (Eq. (6)).

To this end, we adapted the estimation proposed in [10]. Specifically, we estimate the speech variance directly from the signal at the beamformer output as

$$\phi_{x,t} = G_{x,t} \left| Z \right|_t^2 , \tag{42}$$

where

$$G_{x,t} = \frac{\xi_t}{1+\xi_t} \left(\frac{1}{\gamma_t} + \frac{\xi_t}{1+\xi_t} \right) \tag{43}$$

is a gain function derived as in [44], and $\gamma_t = \left| Z \right|_t^2 / \phi_{o,t}$ is the a posteriori SNR. The advantage of this gain estimator is that it approximates the Wiener suppression rule at high instantaneous SNR, while lessens the severity of the attenuation otherwise [44].

However, the a priori SNR is not available (we would need knowledge about the speech variance). Therefore, we propose the following estimate,

$$\widehat{\xi}_t = \frac{R_{z,t}}{\phi_{o,t}}, \tag{44}$$

where

$$R_{z,t} = \frac{1-\lambda}{1-\lambda^t} \sum_{\tau=1}^{t} \lambda^{t-\tau} p_{x,\tau} \left| Z \right|_\tau^2 \tag{45}$$

is a smoothed estimate of the clean speech squared magnitude spectrum given the MVDR output and the a posteriori SPP.

B. Kalman filter parameters

The parameters of the prediction model can also change quickly, so their estimation should be done in a frame-wise fashion. The LPC coefficients and the prediction error variance can be obtained in the M-step as

$$\mathbf{a}_t = \mathbf{R}_{x,t-1}^{-1} \mathbf{r}_{x,t,t-1}, \tag{46}$$

$$\phi_{v,t} = \phi_{x,t} - \mathbf{a}_t^\top \mathbf{R}_{x,t-1} \mathbf{a}_t, \tag{47}$$

where

$$\mathbf{R}_{x,t-1} = E\left\{ \mathbf{x}_{t-1} \mathbf{x}_{t-1}^\top \middle| \mathbf{y}_t; \boldsymbol{\Theta}_t \right\} = \widehat{\mathbf{x}}_{t-1} \widehat{\mathbf{x}}_{t-1}^\top + \mathbf{P}_{t-1}, \tag{48}$$

$$\mathbf{r}_{x,t,t-1} = E\left\{ \left| X \right|_{1,t} \mathbf{x}_{t-1} \middle| \mathbf{y}_t; \boldsymbol{\Theta}_t \right\} = \widehat{\left| X \right|}_{1,t} \widehat{\mathbf{x}}_{t-1} + \mathbf{p}_{t,t-1} \tag{49}$$

are MMSE estimates of the speech signal correlations obtained during the E-step. The complete derivation of the previous expressions can be found in Appendix A. The subtraction in (47) could produce negative values. In such cases, the LPC coefficients are set to zero and $\phi_{v,t} = \phi_{x,t}$, so the Kalman filter reduces to the Wiener filter.

C. Beamformer parameters

We assume that the MVDR beamforming parameters, that is, the RTF and the spatial covariance matrix of the noise, are slowly variant. These parameters can be obtained in the M-step as follows,

$$\mathbf{h}_t = \mathbf{r}_{yx,t} R_{x,t}^{-1}, \tag{50}$$

$$\mathbf{\Phi}_{N,t} = \mathbf{\Phi}_{Y,t} - \mathbf{h}_t R_{x,t} \mathbf{h}_t^H, \tag{51}$$

where

$$R_{x,t} = \frac{1-\lambda}{1-\lambda^t} \sum_{\tau=1}^{t} \lambda^{t-\tau} p_{x,\tau} S_{x,\tau} \tag{52}$$

is a smoothed estimate of the clean speech power spectrum obtained from Eqs. (16)-(17) in the E-step,

$$\mathbf{r}_{yx,t} = \frac{1-\lambda}{1-\lambda^t} \sum_{\tau=1}^{t} \lambda^{t-\tau} p_{x,\tau} \mathbf{y}_\tau \widehat{X}_{1,\tau}^* \tag{53}$$

is a smoothed cross-correlation estimate between noisy and clean speech, and

$$\mathbf{\Phi}_{Y,t} = \frac{1-\lambda}{1-\lambda^t} \sum_{\tau=1}^{t} \lambda^{t-\tau} \mathbf{y}_\tau \mathbf{y}_\tau^H \tag{54}$$

is a smoothed estimate of the spatial covariance matrix of the noisy speech. The complete derivation of the previous expressions can be found in Appendix A.

The advantage of the noise estimator in (51) is that it can be updated even in speech presence bins, which allows for a quicker adaptation, especially in non-stationary noisy scenarios.

VI. A priori SPP estimation based on Deep Neural Network

We have described the estimation of the different model parameters under the EM framework. Nevertheless, this procedure is not convenient for some parameters that can change quickly over time, as in the case of the speech variance. The same problem arises with the a priori SPP. Taseska *et al.* [38] analyzed the estimation of this parameter under a REM framework for noise estimation. They concluded that, although elegant, the estimation is not robust enough when the noise is not stationary. Several algorithms have been proposed to compute this parameter: the SNR-based single-channel SPP estimator [19] and its multichannel version [21] or the coherence to diffuse ratio estimator [38] are some examples. Also, we have recently proposed an a priori SPP estimator for dual-channel smartphones [45] combining spatial properties and the power level difference between microphones.

In this work we propose to estimate the a priori SPP using a deep neural network (DNN)-based mask estimator [23]. These DNN estimators have been successfully applied in speech and noise covariance estimation for multi-channel speech enhancement [24], [25] in both offline and online scenarios [26]–[28]. They often work in each microphone channel individually using only spectral information, so they are called spectral models. Recently, these models have been successfully combined with statistical spatial models [39], [40] to improve the performance of the estimator. These combinations have also shown promising results in blind source separation [46]. Therefore, our approach integrates the use of statistical signal processing with deep learning for the difficult situation where classical assumptions are no longer valid.

Our mask estimator is based on the one proposed in [24], [47]. The model consists of a recurrent neural network followed by two fully connected layers with ReLU activations and an output layer with sigmoid activation. We use a unidirectional long-short term memory (LSTM) recurrent neural network, so the mask estimator can be used in an online scenario. The input feature vector is the noisy log magnitude spectrum,

$$\boldsymbol{\mathcal{Y}}_{m,t} \triangleq \left[\log|Y|_{m,t,0} \quad \cdots \quad \log|Y|_{m,t,F-1}\right]^{\top}, \tag{55}$$

where m refers to the microphone channel index and F is the number of frequency bins. A time-recursive mean normalization is applied on the input features before feeding them into the network [47]. A single speech presence mask is obtained for each channel. The target features used during the training phase are ideal binary masks (IBMs) for the speech signal,

$$\text{IBMx}_{m,t,f} = \begin{cases} 1 & \text{if } \frac{|X|^2_{m,t,f}}{|N|^2_{m,t,f}} > 10^{\mu_f}, \\ 0 & \text{otherwise,} \end{cases} \tag{56}$$

where μ_f are frequency-dependent thresholds [24]. During evaluation, the output masks of each channel are combined in a single mask by means of a median operation, thus providing the final a priori SPP estimate $q_{x,t}$.

VII. Implementation issues

The proposed REM algorithm for multichannel speech enhancement with DNN-based SPP estimation is summarized in Algorithm 1. In the following, we discuss some practical aspects that must be considered for the implementation of the algorithm.

A. Recursive estimation

In the computation of the acoustic parameters we have to deal with several expressions that include a sum over the time frames, with the following structure,

$$R_{\mathcal{B},t} = \frac{1-\lambda}{1-\lambda^t} \sum_{\tau=1}^{t} \lambda^{t-\tau} \mathcal{B}_\tau, \tag{57}$$

where $\mathcal{B}_t$ is any expression computed at time instant t. For an efficient computation, we can translate the previous expression to a recursive estimation,

$$R_{\mathcal{B},t} = (1-\alpha_t)\, R_{\mathcal{B},t-1} + \alpha_t \mathcal{B}_t \tag{58}$$

where

$$\alpha_t = \frac{1-\lambda}{1-\lambda^t} \tag{59}$$

is a time-dependent recursive parameter. Therefore, we only need to save the values from the previous frame to update the recursions.

Algorithm 1 REM algorithm with DNN-based SPP estimation

Initialize variables and parameters
for each t in T (total frames) **do**
  Update $\mathbf{\Phi}_{Y,t}$ using $\mathbf{y}_t$ (54)
  Update $\mathbf{h}_t$ (60) and $\mathbf{\Phi}_{N,t}$ (62) if needed
  Compute $q_{x,t}$ using DNN and initialize $p_{x,t}^0 = q_{x,t}$
  for $l = 1$ to l_{max} **do**
    Beamformer: Compute Z_t (20) and $\phi_{o,t}$ (23) **(E-step)**
    Compute speech variance $\phi_{x,t}$ (42)
    if using Kalman filter **then**
      Compute $\mathbf{a}_t$ (46) and $\phi_{v,t}$ (47) **(M-step)**
    end if
    Postfilter: Estimate $\widetilde{X}_{1,t}$ (18) and P_t (19) **(E-step)**
    Estimate $\widehat{X}_{1,t}$ (16) and $S_{x,t}$ (17) **(E-step)**
    Estimate $p_{x,t}$ (38) **(E-step)**
    Update Λ_t (61)
    Compute $\mathbf{h}_t$ (50) and $\mathbf{\Phi}_{N,t}$ (51) **(M-step)**
  end for
  Update variables for next frame
end for

B. Initialization of the relative transfer function

The estimation of the RTF in (50) is not possible until the first speech frames are processed. This causes the problem that the MVDR beamformer may not be correctly steered towards the target speaker during the first speech frames, which leads to a poor performance of the algorithm. The same problem arises after long speech inactivity periods. To prevent this, we propose an initialization of the RTF in iteration $l = 0$ (before the MVDR beamforming step) in those bins where there has been no recent speech activity. This initialization can be done by eigenvalue decomposition (EVD) [48] of an estimate of the speech covariance matrix,

$$\mathbf{h}_t^{l=0} = \mathcal{P}\left(\mathbf{\Phi}_{Y,t} - \mathbf{\Phi}_{N,t}\right), \tag{60}$$

where $\mathcal{P}(\cdot)$ gives the eigenvector corresponding to the maximum eigenvalue of the matrix and $\mathbf{\Phi}_{Y,t}$ and $\mathbf{\Phi}_{N,t}$ are computed according to Eqs. (51)-(54). To quantify the speech activity in the previous frames, we propose a weighted recursive sum of the SPP in the previous frames,

$$\Lambda_{t,f} = \lambda \Lambda_{t-1,f} + p_{x,t,f}, \tag{61}$$

with $\Lambda_{0,f} = 0$. At time t, we use the EVD-based initialization in those bins where $\Lambda_{t-1,f}$ is below a threshold value Λ_{thr}.

TABLE I
HYPERPARAMETER VALUES USED IN OUR ALGORITHM

Param.	λ	$l_{\max}$	p	Λ_{thr}	T_{init}
Value	0.9	2	2	1.0	10

C. Initialization of the noise covariance matrix

A good initialization of the spatial covariance matrix of the noise in the first frames can improve the convergence of the REM algorithm. Moreover, these are usually noise-only frames or frames with low speech activity. Therefore, we can use the noisy observations of these frames to update the spatial noise statistics. During the first T_{init} frames of the signal, we initialize the spatial noise matrix at iteration $l = 0$ using the following recursion,

$$\mathbf{\Phi}_{N,t}^{l=0} = \beta_t \mathbf{\Phi}_{N,t-1} + (1 - \beta_t)\, \mathbf{y}_\tau \mathbf{y}_\tau^H \tag{62}$$

where

$$\beta_t = 1 + (q_{x,t} - 1)\, \alpha_t. \tag{63}$$

is a recursive factor that uses the a priori SPP to prevent updating if speech presence bins are found. This procedure can be seen as an adaptation of the minima controlled recursive averaging (MCRA) method in [21]. For the next iterations or the next time frames, the noise spatial covariance matrix is computed using (51), as indicated in Algorithm 1.

D. Updating the Kalman filter parameters

The LPC coefficients and the error prediction variance should be updated before using the Kalman postfilter to track the speech variability. The problem is that this computation requires the computation of $\mathbf{r}_{x,t,t-1}$ using (49), which depends on the Kalman filter output. Therefore, in the first EM iteration we propose to compute a Wiener filter with an SPP masking to approximate it as $\mathbf{r}_{x,t,t-1} \simeq \widehat{|X|}_{1,t} \widehat{\mathbf{x}}_{t-1}$. Once the parameters are obtained, the Kalman filter is applied. In the next iterations, we directly compute these parameters as indicated in Section V-B by using the estimates obtained in the previous iteration.

VIII. EXPERIMENTAL RESULTS

A. Experimental framework

To test the performance of the proposed algorithm, we evaluate it in the simulated set for the CHiME-4 database [49]. This database comprises six-channel tablet recordings in different noisy environments: cafe (CAF), street (STR), pedestrian (PED) and bus (BUS). The SNR of the noisy speech signals is in the range between 0 dB and 15 dB. The training subset has 7138 utterances from 83 speakers, while the development and evaluation subset consists of 1640 and 1320 utterances, respectively, from four different speakers in each subset. The audio signals are sampled at 16 kHz. The fifth microphone of the tablet is used as the reference channel for the different algorithms and the evaluations.

For STFT computation, a 512-point DFT is applied using a 32 ms square-root Hann window with 50% overlap. This results in a total of 257 frequency bins for each time frame. The values of the different parameters used in our algorithm are summarized in Table I, where the same value of λ is used in the different equations of our algorithm.

The DNN model has an LSTM layer with 512 units, two fully connected layers with 512 units each and an output layer with 257 units. We train the model using the training subset. The loss function used for training and validation is the binary cross-entropy between the estimated and target masks, as in [23]. During the training phase, a batch size of five utterances and the ADAM optimizer [50] are used. To prevent overfitting, dropout is applied in the hidden layers with a de-activation probability factor of 0.5. After each epoch, the DNN is validated using the development subset. The training is stopped after 20 epochs without improvement on the development subset and the best model obtained is saved. We use Pytorch as deep learning framework.

We evaluate the performance of the proposed REM framework either using a Wiener postfilter (WF) or using a Kalman postfilter (KF). For comparison purposes, the enhanced signal before (REMWF-BF and REMKF-BF) and after (REMWF and REMKF) the postfilter is considered and evaluated for both algorithms (i.e. Z_t and $\widetilde{X}_{1,t}$ outputs).

B. Evaluation results

In this subsection we use three objective performance measures to assess the quality of the enhanced signal: the wideband Perceptual Evaluation of the Speech Quality (PESQ) [51], the Extended Short-Time Objective Intelligibility (ESTOI) [52], [53], and the scale-invariant Signal to Distortion Ratio (SDR) [54]. We compare the four variants of our REM framework with three state-of-the-art methods. For a fair comparison, all of these methods use the SPP masks provided by the DNN to obtain the beamformer parameters. The recursive procedure used in the MCRA method [21] is applied over these masks to obtain the noise spatial covariance matrices, while the RTF is computed by means of EVD decomposition [48]. The reference methods are:

- MVDR beamforming (MVDR) as described in [21].
- Multichannel Wiener filter (MWF) using the rank-1 approximation for the speech spatial covariance matrix, as described in [55].
- Multichannel Kalman filter (MKF) as proposed in [8]. This method uses the baseline MVDR and a Kalman postfilter. The LPC parameters are computed from the baseline MWF output via LPC analysis. For a first comparison, we update these parameters each frame, and a total of five previous frames are used to compute them.

Tables II, III and IV show, respectively, the results obtained by the tested methods for each metric. We include the average results for each noise type, and the average result of each technique with the 95% confidence intervals indicated. The results of the unprocessed noisy speech signals are also included to compare the gains of the different methods. The evaluation subset of the simulated set from the CHiME-4 database is used in our evaluations. As can be observed, the proposed REMWF and REMKF outperform the reference methods in terms of quality, intelligibility and signal distortion. Moreover, the results obtained for REMWF-BF and REMKF-BF improve the baseline MVDR beamformer. The postfiltering implies an increase on PESQ and SDR metrics, while ESTOI keeps similar when

TABLE II

PESQ RESULTS FOR THE DIFFERENT EVALUATED ALGORITHMS. RESULTS ARE BROKEN DOWN BY NOISE ENVIRONMENT.

Method	**Noise**				**Avg.**
	BUS	**CAF**	**PED**	**STR**	
Noisy	1.32	1.24	1.26	1.28	1.27 $\pm$ 0.01
MVDR	1.71	1.51	1.58	1.57	1.59 $\pm$ 0.01
REMWF-BF	1.87	1.67	1.74	1.69	1.74 $\pm$ 0.02
MWF	2.07	1.83	1.94	1.90	1.94 $\pm$ 0.01
REMWF	2.19	1.96	2.07	1.98	2.05 $\pm$ 0.02
REMKF-BF	1.89	1.68	1.75	1.70	1.76 $\pm$ 0.02
MKF	1.97	1.68	1.79	1.78	1.81 $\pm$ 0.01
REMKF	**2.22**	**1.98**	**2.10**	**2.01**	**2.08 $\pm$ 0.02**

comparing with the beamformer output. The results using KF in our REM framework are slightly better than its WF counterpart, both for the beamforming and postfiltering. Regarding the reference methods, both MWF and MKF perform better than a basic MVDR beamformer for PESQ and SDR. Nevertheless, MKF does not outperform MWF. Finally, the gains obtained are consistent for the different noises analyzed, with REMKF as the best approach.

These results show that the use of the REM framework with DNN-based SPP estimation improves the performance of the multichannel speech enhancement approaches. This is observed in the gains in both beamforming and postfiltering when comparing with the reference methods. The estimation of the beamforming parameters benefits both on the use of the estimated clean speech statistics and an improved SPP estimation using spectral (DNN) and spatial (statistical) models. This shows the advantage of using a DNN mask estimator, which does not need explicit assumptions about the a priori SPP, in combination with a statistical spatial model for the noisy speech. The postfilter further enhances the speech signal at the beamformer output, increasing the noise reduction at the cost of a slight degradation of speech intelligibility. This improves speech quality metrics as PESQ and SDR when comparing with the beamforming approaches (REMWF-BF and REMKF-BF), while speech intelligibility metrics such as ESTOI are not severely affected. The separation of postfiltering and SPP masking has the advantage that additional speech distortion is not introduced in the filtered signal, which could degrade PESQ and ESTOI metrics, while the algorithm still benefits from this masking to obtain the acoustic parameters.

By comparing the Wiener and Kalman postfiltering approaches, we conclude that the REM framework benefits from using the Kalman filter to take into account temporal correlations, improving the Wiener filter estimation in speech presence bins. This also allows a better estimation of the beamforming parameters during the M-step. The same behavior is not observed when comparing MWF and MKF, where the use of the temporal model degrades the metrics. Although these results do not match with those from [8], this could be explained in part by the fact that we use longer analysis windows and smaller overlapping between windows than the original implementation, which could affect the MKF performance. We choose the same window length and overlapping in the different

TABLE III

ESTOI (x100) RESULTS FOR THE DIFFERENT EVALUATED ALGORITHMS. RESULTS ARE BROKEN DOWN BY NOISE ENVIRONMENT.

Method	**Noise**				**Avg.**
	BUS	**CAF**	**PED**	**STR**	
Noisy	70.9	65.9	68.7	67.1	68.2 $\pm$ 0.5
MVDR	82.9	77.2	79.6	78.5	79.5 $\pm$ 0.4
REMWF-BF	86.3	82.2	83.8	82.4	83.7 $\pm$ 0.4
MWF	83.3	78.0	80.1	79.1	80.1 $\pm$ 0.4
REMWF	86.1	81.7	83.2	82.1	83.3 $\pm$ 0.4
REMKF-BF	86.8	**82.6**	**84.3**	**82.9**	**84.1 $\pm$ 0.4**
MKF	80.6	74.7	76.8	76.3	77.1 $\pm$ 0.4
REMKF	**86.9**	82.5	84.0	**82.9**	**84.1 $\pm$ 0.4**

TABLE IV

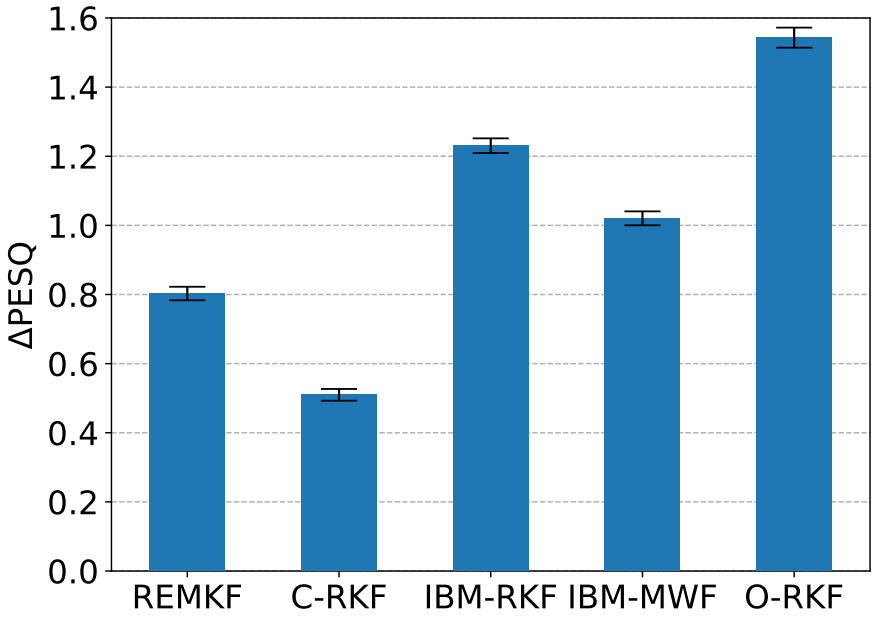
SDR RESULTS (IN DB) FOR THE DIFFERENT EVALUATED ALGORITHMS. RESULTS ARE BROKEN DOWN BY NOISE ENVIRONMENT.

Method	**Noise**				**Avg.**
	BUS	**CAF**	**PED**	**STR**	
Noisy	6.79	7.77	8.60	6.86	7.51 $\pm$ 0.11
MVDR	11.42	11.12	11.94	10.88	11.34 $\pm$ 0.14
REMWF-BF	12.49	12.26	12.83	11.74	12.33 $\pm$ 0.14
MWF	13.72	12.44	13.05	12.83	13.01 $\pm$ 0.15
REMWF	15.22	14.06	14.47	14.08	14.46 $\pm$ 0.16
REMKF-BF	12.63	12.36	12.96	11.88	12.46 $\pm$ 0.15
MKF	13.58	11.71	12.39	12.45	12.53 $\pm$ 0.16
REMKF	**15.75**	**14.38**	**14.90**	**14.54**	**14.89 $\pm$ 0.16**

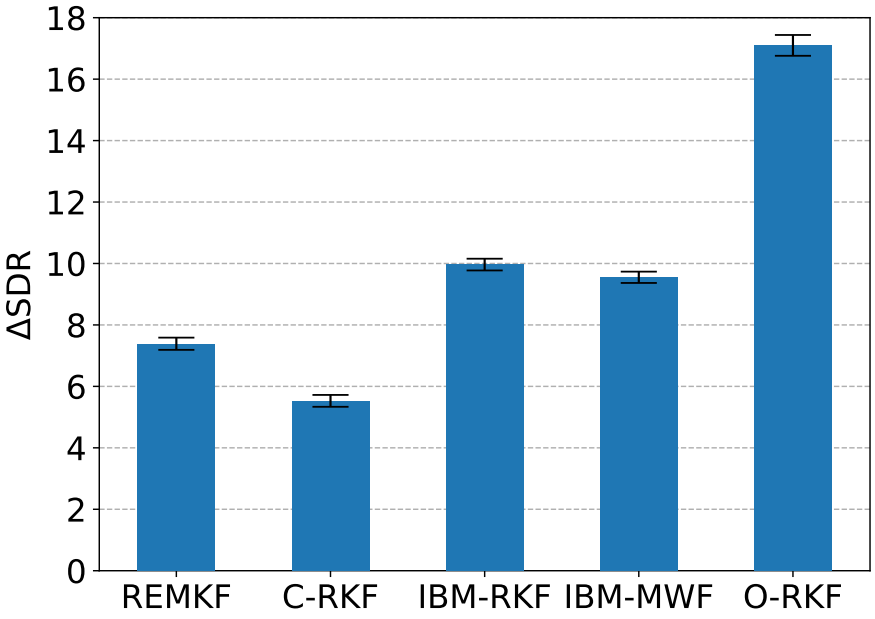
methods we compare to use the same DNN architecture in them. On the other hand, MKF uses LPC analysis on the enhanced signal obtained by MWF, which contains residual noise, and several frames are needed to compute the matrices to solve the linear equations system. Our proposal directly uses the estimated clean speech statistics of the current frame in a maximum-likelihood estimation of these parameters. Moreover, the SPP is considered in this procedure, which allows to better discriminate between clean speech and noise bins. Thus, our REM framework allows for a better estimation of the LPC parameters needed for the Kalman filter.

C. Analysis using oracle and baseline estimates for the SPP and the acoustic parameters

The upper-bound performance of our proposed algorithm can also be analyzed using oracle estimates for the SPP and the acoustic parameters. Fig. 2 compares the results in terms of PESQ and SDR scores obtained using REMKF



(a)



(b)

Fig. 2. Results for the evaluation using baseline and oracle estimates for the SPP and the acoustic parameters. The improvement with respect to the noisy speech results is showed along with the 95% confidence intervals: (a) PESQ results, (b) SDR results.

and three experiments with oracle estimates: the REMKF and MWF approaches using oracle IBM masks instead of the DNN estimation (IBM-RKF and IBM-MWF), and REMKF using oracle estimates of the RTF and the noise spatial covariance matrix (O-RKF), but with DNN estimates for the a priori SPP. The goal of this last experiment is to clearly distinguish the contribution of the acoustic parameters from SPPs. Oracle noise estimates can be obtained using Eq. (54) with the noise signal instead of the noisy speech signal. On the other hand, the oracle RTF can be

derived from Eq. (50), where the correlations are now obtained directly from the clean speech signal, avoiding the use of the clean speech estimates and the a posteriori SPP. We also include an additional experiment, named C-RKF, which uses the REMKF framework along with the a priori SPP estimates obtained from the coherent-to-diffuse ratio (CDR)-based a priori SPP estimator proposed in [38]. The motivation for this is to compare the performance using a priori SPP estimates from the DNN (which rely on spectro-temporal signal characteristics) with a scheme that finds a priori SPP estimates based on assumptions about the spatial signal characteristics.

As can be observed, the REMKF approach outperforms C-RKF, which indicates that the DNN estimates yield a better SPP initialization than the CDR-based approach. That is, the DNN estimator provides more discriminative SPP estimates than classic signal processing methods. This increases the performance of the REM framework which takes advantage of a good initialization for the a priori SPP. Regarding the oracle estimators, IBM-RKF performs better than IBM-MWF, especially in the case of the PESQ metric. This suggests that the robustness of the REM framework is not only due to the availability of accurate SPP estimates, but also to the use of the clean speech expectations and SPPs to compute the acoustic parameters in non-stationary environments. On the other hand, O-RKF outperforms the rest of the oracle estimators, showing that a good estimation of the acoustic parameters has a larger contribution to the performance than the use of oracle SPP estimates. This highlights the importance of integrating DNN estimators with statistical spatial models, which, in addition, improves the estimation of both the a posteriori SPP and the clean speech signal.

D. Performance of the SPP estimators

In order to compare the a priori SPP estimates given by the DNN and the a posteriori SPP estimates obtained using our framework, we consider a binary detector as [38]

$$\widehat{\mathcal{M}}_{x,t,f} = \begin{cases} 1 & \text{if } p_{x,t,f} > p_{\text{thr}}, \\ 0 & \text{otherwise}, \end{cases} \tag{64}$$

where p_{thr} is a selected threshold. In the case of the DNN output, we use $q_{x,t,f}$ instead of $p_{x,t,f}$. We can also define a ground truth detector $\mathcal{M}_{x,t,f} = 1$ when speech is present ($\mathcal{H}_x$) and zero otherwise. The values of this ideal detector are chosen from the IBM masks used to train the DNN. Then, we define the true positive rate (TPR) and the false positive rate (FPR) of the detector for a given utterance as

$$\text{TPR} = \frac{\sum_{t,f}\left[\left(\widehat{\mathcal{M}}_{x,t,f} = 1\right)\&\left(\mathcal{M}_{x,t,f} = 1\right)\right]}{\sum_{t,f}\left[\mathcal{M}_{x,t,f} = 1\right]} \tag{65}$$

$$\text{FPR} = \frac{\sum_{t,f}\left[\left(\widehat{\mathcal{M}}_{x,t,f} = 1\right)\&\left(\mathcal{M}_{x,t,f} = 0\right)\right]}{\sum_{t,f}\left[\mathcal{M}_{x,t,f} = 0\right]} \tag{66}$$

We can evaluate the performance of the binary detector by means of the Receiver Operating Characteristics (ROC) curve [56], which is a representation of TPR vs. FPR for different threshold values. The higher the area under the curve, the best the performance of the binary detector.

Fig. 3 shows the ROC curves obtained using the DNN output and the SPP estimates from the REMWF and REMKF algorithms in different noisy environments. We use the values $p_{\text{thr}} \in [0.2, 0.8]$ with 0.05 step to focus on

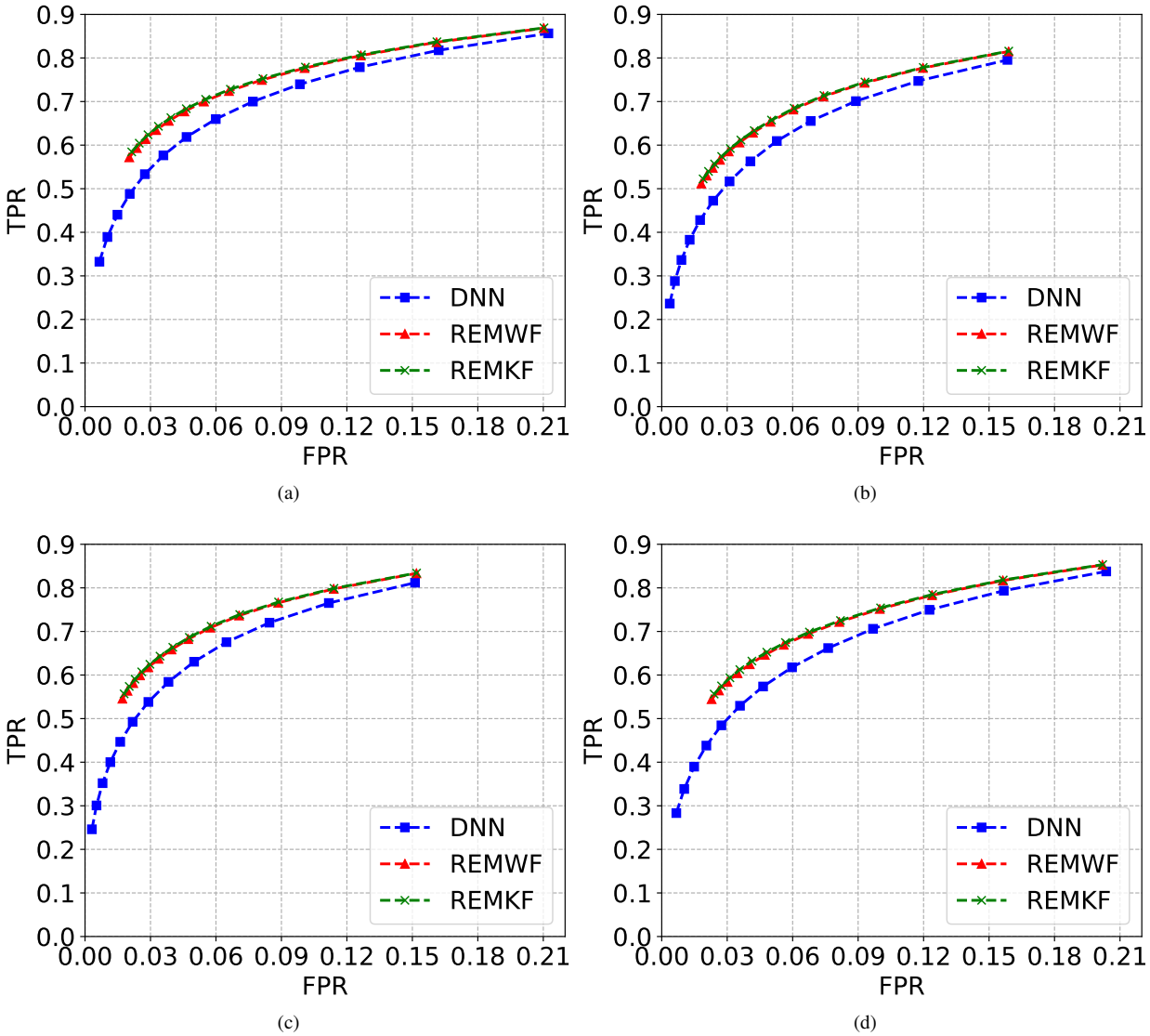
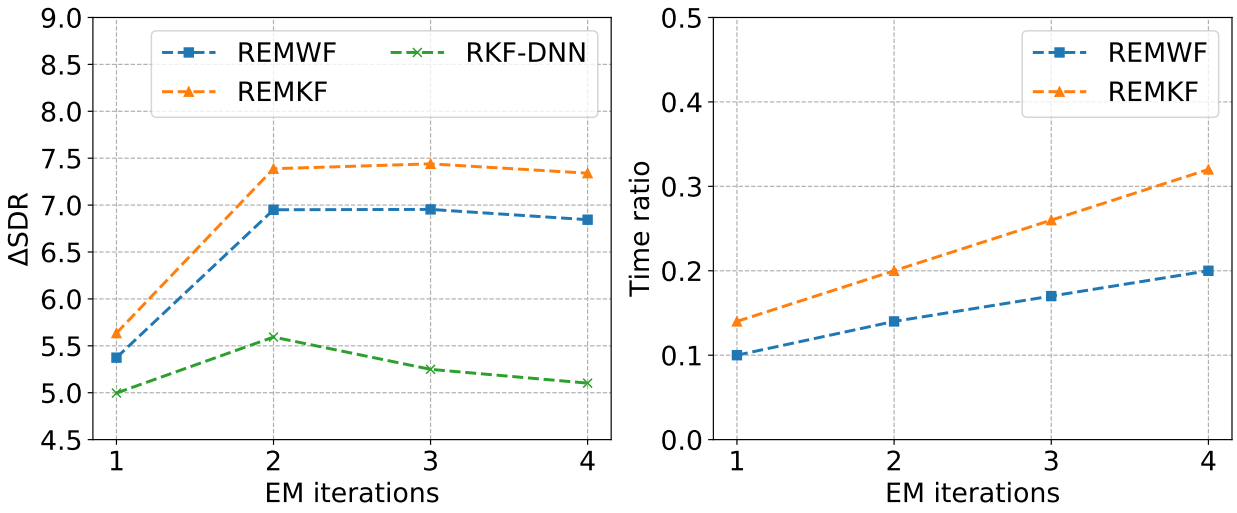


Fig. 3. ROC curves of detectors obtained using the DNN output (a priori SPP) and the a posteriori SPP estimates of the REMWF and REMKF algorithms. The values of the threshold p_{thr} are chosen between 0.2 and 0.8 with a step of 0.05. The noisy environments analyzed are: (a) bus, (b) cafeteria, (c) pedestrian street and (d) street.

the regions of the curve where the differences are more noticeable. These ratios are computed using the evaluation subset. A single value per noisy environment is obtained. The results show that the proposed algorithms achieve better TPR than the DNN for same values of FPR, which indicates that the use of probabilistic spatial models helps to better discriminate between speech presence and speech absence bins. On the other hand, the performance of both REMWF and REMKF is comparable in terms of the SPP estimation. In addition, the performance is similar across the different noisy environments.



(a) (b)

Fig. 4. Performance evaluation of our REMWF and REMKF approaches in terms of the number of EM iterations: (a) Improvements on SDR metric, (b) Time needed to process each second of the noisy speech signal. The analysis of the REMKF using a posteriori SPP estimates directly obtained from the DNN (RKF-DNN) is also included.

E. EM iterations and computational latency

In this subsection, we evaluate the performance in terms of the number of EM iterations for the REMWF and REMKF approaches. First, we evaluate the improvements in terms of SDR with respect to the noisy speech signal. Fig. 4a shows that the performance stabilizes after two or more iterations. This fast convergence has been previously observed in other works [21], [38]. It can be explained by the fact that the estimation of the a posteriori SPP in the first iteration uses the acoustic parameters obtained in the previous frames, but from the second iteration the acoustic parameters are updated using information for the current time frame, which allows for a better estimation of the a posteriori SPP. Thus, we have chosen two EM iterations for our evaluations. Moreover, we also analyze the performance of the REMKF algorithm across the EM iterations, when the DNN estimates are bypassed and directly used as a posteriori SPP (RKF-DNN), that is, Eq. (38) is no longer used to this end. Our goal here is to test whether the REM framework is able to improve the SPP estimates from the DNN yielding an increase on the final performance. It is observed in Fig. 4a that the estimation of the a posteriori SPP in the REM framework outperforms the RKF-DNN approach for the different EM iterations. Moreover, the RKF-DNN approach does not show significant improvements from the first iteration as in the case of the REMKF approach. This highlights how both frameworks, REM and DNN, are successfully integrated and help each other to improve the performance.

In addition, we evaluate the computational latency of our implementations in terms of the number of iterations. It must be noted that the algorithms are implemented using Python, while we run our implementations on an Intel Core i7-4790 CPU at 3.6 GHz with four cores, 16 GB of RAM, and an Nvidia GeForce GTX 1060 GPU with 6 GB of memory. The GPU is only used for DNN inference of the a priori SPP. The algorithms are evaluated over 220 files from the evaluation subset, and the ratio between the time needed to process the file and its total

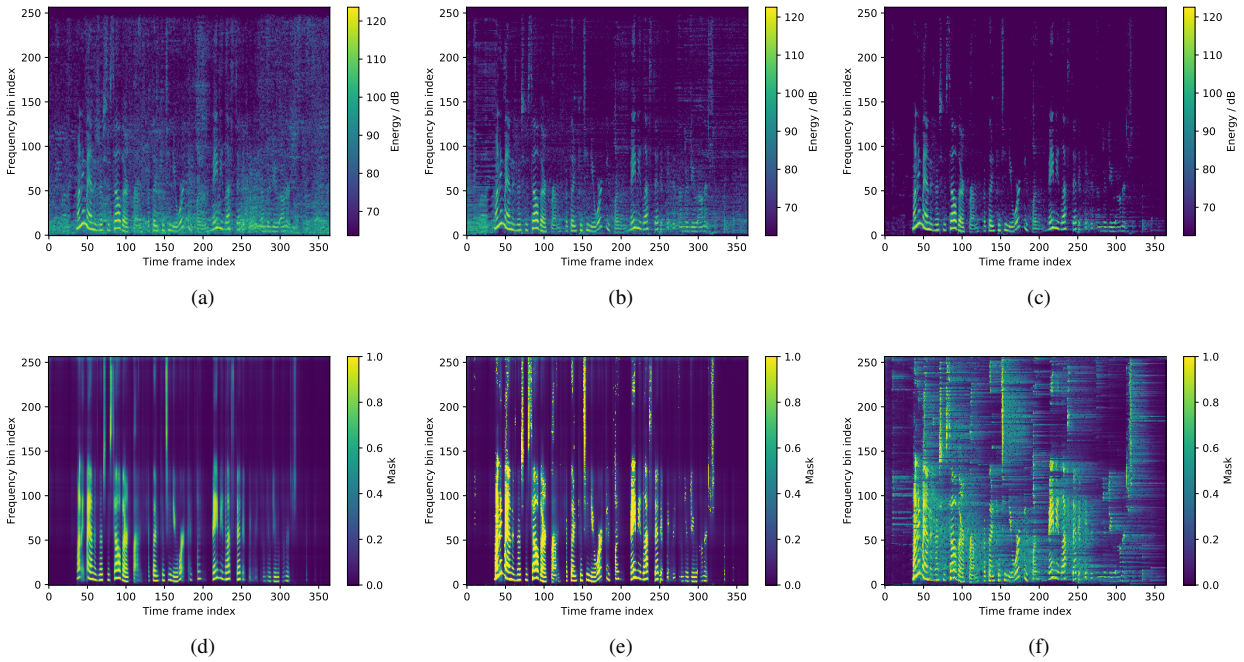


Fig. 5. Example of different noisy and enhanced spectrograms and estimated masks when the REMKF approach is applied to the audio file F05_444C0214_CAF (cafeteria noise, SDR = 7.23 dB) from the CHiME-4 database. (a) Noisy speech spectrogram at the reference channel. (b) Beamformer output spectrogram. (c) Filtered speech signal spectrogram. (d) A priori SPP obtained by the DNN. (e) A posteriori SPP obtained using the REM framework. (f) Kalman gain used in the postfilter.

duration is computed. The obtained average ratios are showed in Fig. 4b. The results show that the computational time increases almost linearly with the number of iterations and that both algorithms can be executed faster than real-time on a computer with similar settings.

F. Example results

For a qualitative evaluation of the proposed framework, Fig. 5 shows example spectrograms of the noisy speech signal (Fig. 5a), the enhanced signal at the beamformer (Fig. 5b) and the enhanced signal at the postfilter output (Fig. 5c) for the REMKF approach. In addition, the corresponding a priori (Fig. 5d) and a posteriori SPP (Fig. 5e), and the Kalman gain (Fig. 5f) are also shown. It can be observed that efficient noise reduction is achieved, especially at the postfilter output (Fig. 5c), where much of the noise is removed at medium and high frequencies. Furthermore, the speech formants appear preserved, which is in line with the objective results obtained. The a priori SPP (Fig. 5d) obtained by the DNN-based mask estimator shows the suitability of these deep learning models for accurately estimating the speech presence without the need of any assumptions about the spectral properties of the speech signal. The a posteriori SPP (Fig. 5e) obtained using the REM framework can improve this estimation by using statistical models on the multichannel noisy observations, which helps to differentiate more clearly between speech presence and absence bins. This turns out in more discriminative SPP masks as shown in the example. Finally, it is observed that the Kalman gain (Fig. 5f) does not show the sparse pattern of the a posteriori SPP, but it presents

higher values in high SNR bins, as expected, and a smooth decay when speech is absent. The same behavior was observed for the Wiener gain in the REMWF variant. These gains depend on the speech variance, whose estimation is addressed without using the M-step to avoid the problem with the sparsity of the SPP. Thus, these results suggest the decoupling achieved between the postfiltering and the SPP masking used for the estimation of the acoustic parameters.

IX. Conclusion

In this paper we have proposed a recursive expectation-maximization algorithm with a deep neural network-based speech presence probability estimation for multichannel speech enhancement. Our proposal combines a statistical framework for the joint estimation of the clean speech signal and the acoustic parameters with the powerful modeling capabilities of deep learning models for the estimation of speech presence probabilities. The combined use of beamforming and postfiltering, based on a Wiener or a Kalman filter, is proposed to take advantage of the spatial, spectral and temporal properties of the speech signal for noise reduction. Moreover, the sparsity of the speech signal is exploited for the estimation of the different acoustic parameters needed for beamforming and postfiltering. The main advantages of the proposal are the use of statistical spatial models to improve the DNN estimation and the separation between the postfiltering step and the SPP masking, which prevents severe speech distortion. The experimental results show that the proposed framework helps to outperform other state-of-the-art approaches in terms of speech quality and intelligibility, and noise reduction, with the KF-based approach achieving the best results. Moreover, our proposal allows for online processing with low-latency.

As future work, we will address the estimation of the clean speech phase and the use of this framework in other challenging scenarios including speech dereverberation and blind source separation with multiple speakers.

Appendix A
Derivation of the recursive EM algorithm

In order to derive the acoustic parameters, we have to reformulate the Q function previously defined in Eqs. (11)-(13). Using them, we can rewrite the Q function in (13) as follows,

$$\begin{aligned} &Q\left(\boldsymbol{\Theta}_t|\,\boldsymbol{\Theta}_t^l\right) = C+ \\ &\sum_{\tau=1}^{t}\lambda^{t-\tau}\sum_{\mathcal{D}_\tau} p_{\mathcal{D}_\tau} E\left\{\log f\left(|X|_{1,\tau}\Big|\,\mathbf{x}_{\tau-1},\mathcal{D}_\tau;\mathbf{a}_t,\phi_{v,t}\right)\right\}+ \\ &\sum_{\tau=1}^{t}\lambda^{t-\tau}\sum_{\mathcal{D}_\tau} p_{\mathcal{D}_\tau} E\left\{\log f\left(\mathbf{y}_\tau\,|X_{1,\tau},\mathcal{D}_\tau;\mathbf{h}_t,\boldsymbol{\Phi}_{N,t}\right)\right\}, \end{aligned} \tag{67}$$

where C refers to the sum of terms that are independent of the parameters of interest and therefore can be neglected, and $p_{\mathcal{D}_t} = P\left(\mathcal{D}_t|\,\mathbf{y}_t\right)$. The expectation depends on the noisy speech signal and the current acoustic parameters (omitted for clarity purposes). Let us start with the second non-trivial term, which depends on the RTF and the noise

spatial statistics. Given that the noise signal follows a multivariate complex Gaussian distribution, the expectation can be developed as

$$E\left\{\log f\left(\mathbf{y}_\tau\left|X_{1,\tau},\mathcal{H}_x;\mathbf{h}_t,\mathbf{\Phi}_{N,t}\right.\right)\right\}=C-\frac{1}{2}\log\left|\mathbf{\Phi}_{N,t}\right|-\\ \frac{1}{2}E\left\{\left[\mathbf{y}_\tau-\mathbf{h}_tX_{1,\tau}\right]^H\mathbf{\Phi}_{N,t}^{-1}\left[\mathbf{y}_\tau-\mathbf{h}_tX_{1,\tau}\right]\right\}, \tag{68}$$

when speech is present, and

$$E\left\{\log f\left(\mathbf{y}_\tau\left|\mathcal{H}_n;\mathbf{\Phi}_{N,t}\right.\right)\right\}=C-\frac{1}{2}\log\left|\mathbf{\Phi}_{N,t}\right|-\frac{1}{2}\mathbf{y}_\tau^H\mathbf{\Phi}_{N,t}^{-1}\mathbf{y}_\tau, \tag{69}$$

when speech is absent (C is used for the independent terms). The RTF is derived directly from the speech presence assumption. Computing the derivative of Q with respect to the RTF yields

$$\frac{\partial Q}{\partial\mathbf{h}_t}=\sum_{\tau=1}^{t}\lambda^{t-\tau}p_{x,\tau}E\left\{\mathbf{\Phi}_{N,t}^{-1}\left[\mathbf{y}_\tau-\mathbf{h}_tX_{1,\tau}\right]X_{1,\tau}^*\right\}, \tag{70}$$

and by making this expression equals to zero, we obtain the following MLE estimate,

$$\mathbf{h}_t=\frac{\sum_{\tau=1}^{t}\lambda^{t-\tau}p_{x,\tau}\mathbf{y}_\tau\widehat{X}_{1,\tau}^*}{\sum_{\tau=1}^{t}\lambda^{t-\tau}p_{x,\tau}S_{x,\tau}}, \tag{71}$$

which is equivalent to that in Eq. (50). The same procedure can be used to derive $\mathbf{\Phi}_{N,t}$ in Eq. (51), now taking both hypotheses $\mathcal{H}_n$ and $\mathcal{H}_x$ into consideration, which yields

$$\frac{\partial Q}{\partial\mathbf{\Phi}_{N,t}^{-1}}=-\frac{1}{2}\sum_{\tau=1}^{t}\lambda^{t-\tau}\Big[\mathbf{y}_\tau\mathbf{y}_\tau^H-p_{x,\tau}\Big(\mathbf{h}_t\widehat{X}_{1,\tau}\mathbf{y}_\tau^H+\\ \mathbf{y}_\tau\mathbf{h}_t^H\widehat{X}_{1,\tau}^*-\mathbf{h}_tS_{x,\tau}\mathbf{h}_t^H\Big)\Big]+\frac{1}{2}\mathbf{\Phi}_{N,t}\sum_{\tau=1}^{t}\lambda^{t-\tau}. \tag{72}$$

Given the previous definition of $\mathbf{h}_t$ and by using $\sum_{\tau=1}^{t}\lambda^{t-\tau}=\frac{1-\lambda^t}{1-\lambda}$, the noise covariance matrix can be finally obtained using the following MLE estimate,

$$\mathbf{\Phi}_{N,t}=\frac{1-\lambda}{1-\lambda^t}\sum_{\tau=1}^{t}\lambda^{t-\tau}\left(\mathbf{y}_\tau\mathbf{y}_\tau^H-p_{x,\tau}\mathbf{h}_tS_{x,\tau}\mathbf{h}_t^H\right). \tag{73}$$

To obtain the Kalman filter parameters, we use the first non-trivial term of the Q function, when speech is present. Given the stochastic process in (8), which follows a Gaussian distribution, the expectation can be expanded as,

$$E\left\{\log f\left(\left|X\right|_{1,\tau}\middle|\mathbf{x}_{\tau-1},\mathcal{H}_x;\mathbf{a}_t,\phi_{v,t}\right)\right\}=C-\frac{1}{2}\log\phi_{v,t}-\\ \frac{1}{2}E\left\{\left[\left|X\right|_{1,\tau}-\mathbf{a}_t^\top\mathbf{x}_{\tau-1}\right]^\top\phi_{v,t}^{-1}\left[\left|X\right|_{1,\tau}-\mathbf{a}_t^\top\mathbf{x}_{\tau-1}\right]\right\}. \tag{74}$$

In order to consider the fact that these parameters can change quickly, we compute the derivatives in the case $\lambda=0$,

$$\frac{\partial Q_{\lambda=0}}{\partial\mathbf{a}_t^\top}=p_{x,t}E\left\{\phi_{v,t}^{-1}\left[\left|X\right|_{1,t}-\mathbf{a}_t^\top\mathbf{x}_{t-1}\right]\mathbf{x}_{t-1}^\top\right\}, \tag{75}$$

$$\frac{\partial Q_{\lambda=0}}{\partial\phi_{v,t}^{-1}}=-\frac{1}{2}p_{x,t}\Big(E\left\{\left|X\right|_{1,t}^2-\mathbf{a}_t^\top\mathbf{x}_{t-1}\left|X\right|_{1,t}-\\ \mathbf{x}_{t-1}^\top\mathbf{a}_t\left|X\right|_{1,t}+\mathbf{a}_t^\top\mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top\mathbf{a}_t\right\}-\phi_{v,t}\Big), \tag{76}$$

so only the instantaneous statistics are used. Thus, the following MLE estimates can be obtained for the LPC coefficients,

$$\mathbf{a}_t=E\left\{\mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top\right\}^{-1}E\left\{\left|X\right|_{1,t}\mathbf{x}_{t-1}\right\}, \tag{77}$$

and the error prediction variance,

$$\phi_{v,t} = E\left\{|X|_{1,t}^2\right\} - \mathbf{a}_t^\top E\left\{\mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top\right\}\mathbf{a}_t. \tag{78}$$

We use the speech variance under speech presence, $\phi_{x,t}$, instead of $E\left\{|X|_{1,t}^2\right\}$ in the above expression to avoid the problem of the sparsity in the filtered signal $\widetilde{X}_{1,t}$. This yields to the expressions of Eqs. (46) and (47).

REFERENCES

[1] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Process. Magaz.*, vol. 29, no. 6, pp. 127–140, 2012.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[3] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[4] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Process.* Springer, 2008, vol. 1.

[5] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, 2002.

[6] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech and Audio Process.*, vol. 12, no. 6, pp. 561–571, 2004.

[7] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Comm.*, vol. 49, no. 7-8, pp. 657–666, 2007.

[8] W. Xue, A. Moore, M. Brookes, and P. Naylor, "Modulation-domain multichannel Kalman filtering for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1833–1847, 2018.

[9] S. So and K. Paliwal, "Modulation-domain kalman filtering for single-channel speech enhancement," *Speech Comm.*, vol. 53, no. 6, pp. 818–829, 2011.

[10] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 2, pp. 394–406, 2015.

[11] S. Braun and E. Habets, "Linear prediction-based online dereverberation and noise reduction using alternating Kalman filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1115–1125, 2018.

[12] I. Batina, J. Jensen, and R. Heusdens, "Noise power spectrum estimation for speech enhancement using an autoregressive model for speech power spectrum dynamics," in *Proc. ICASSP*, 2006, pp. 1064–1067.

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

[14] M. Rahmani, A. Akbari, B. Ayad, and B. Lithgow, "Noise cross PSD estimation using phase information in diffuse noise field," *Signal Process.*, vol. 89, no. 5, pp. 703–709, 2009.

[15] N. Ito, E. Vincent, T. Nakatani, N. Ono, S. Araki, and S. Sagayama, "Blind suppression of nonstationary diffuse acoustic noise based on spatial covariance matrix decomposition," *Journal Signal Process. Systems*, vol. 79, no. 2, pp. 145–157, 2015.

[16] W. Jin, M. J. Taghizadeh, K. Chen, and W. Xiao, "Multi-channel noise reduction for hands-free voice communication on mobile phones," in *Proc. ICASSP*, 2017, pp. 506–510.

[17] R. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 211–221, 2012.

[18] A. Kuklasinski, S. Doclo, S. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1595–1608, 2016.

[19] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[20] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 910–919, 2008.

[21] M. Souden, J. Benesty, S. Affes, and J. Chen, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, 2011.

[22] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. ICASSP*, 2015, pp. 544–548.

[23] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[24] ——, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Comput. Speech and Language*, vol. 46, pp. 374–385, 2017.

[25] L. Pfeifenberger, M. Zhrer, and F. Pernkopf, "Eigenvector-based speech mask estimation for multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2162–2172, 2019.

[26] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming," in *Proc. ICASSP*, 2018, pp. 531–535.

[27] S. Chakrabarty and E. Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal Selec. Topics Signal Process.*, vol. 13, no. 4, pp. 787–799, 2019.

[28] J. M. Martín-Doñas, J. Heitkaemper, R. Haeb-Umbach, A. M. Gomez, and A. M. Peinado, "Multi-channel block-online source extraction based on utterance adaptation," in *Proc. InterSpeech*, 2019, pp. 96–100.

[29] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[30] M. Togami and Y. Kawaguchi, "Simultaneous optimization of acoustic echo reduction, speech dereverberation, and noise reduction against mutual interference," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 11, pp. 1612–1623, 2014.

[31] O. Schwartz, S. Gannot, and E. Habets, "An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1491–1506, 2016.

[32] F. Gu, H. Zhang, W. Wang, and S. Wang, "An expectation-maximization algorithm for blind separation of noisy mixtures using gaussian mixture model," *Circ., Systems, Signal Process.*, vol. 36, no. 7, pp. 2697–2726, 2017.

[33] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[34] O. Cappé and E. Moulines, "Online EM algorithm for latent data models," *Journal Royal Statist. Soc.: B*, vol. 71, no. 3, pp. 593–613, 2009.

[35] O. Schwartz and S. Gannot, "A recursive expectation-maximization algorithm for online multi-microphone noise reduction," in *Proc. EUSIPCO*, 2018, pp. 1542–1546.

[36] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. EUSIPCO*, 2016, pp. 1153–1157.

[37] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 780–793, 2017.

[38] M. Taseska and E. Habets, "Nonstationary noise PSD matrix estimation for multichannel blind speech extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 11, pp. 2223–2236, 2017.

[39] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating dnn-based and spatial clustering-based mask estimation for robust mvdr beamforming," in *Proc. ICASSP*, 2017, pp. 286–290.

[40] Y. Matsui, T. Nakatani, M. Delcroix, K. Kinoshita, N. Ito, S. Araki, and S. Makino, "Online integration of DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *Proc. IWAENC*, 2018, pp. 71–75.

[41] B. Schwartz, S. Gannot, and E. Habets, "Two model-based EM algorithms for blind source separation in noisy environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 11, pp. 2209–2222, 2017.

[42] K. P. Murphy, "Switching Kalman filters," U. C. Berkeley, Tech. Rep., 1998.

[43] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. USA: Prentice-Hall, Inc., 1993.

[44] P. Wolfe and S. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *Eurasip Journal Applied Signal Process.*, no. 10, pp. 1043–1051, 2003.

[45] J. M. Martín-Doñas, A. M. Peinado, I. López-Espejo, and A. Gomez, "Dual-channel speech enhancement based on extended Kalman filter relative transfer function estimation," *Applied Sciences*, vol. 9, no. 12, 2019.

[46] L. Drude and R. Haeb-Umbach, "Integration of neural networks and probabilistic spatial models for acoustic blind source separation," *IEEE Journal Selec. Topics Signal Process.*, vol. 13, no. 4, pp. 815–826, 2019.

[47] J. Heitkaemper, J. Heymann, and R. Haeb-Umbach, "Smoothing along frequency in online neural network supported acoustic beamforming," in *ITG, Oldenburg, Germany*, 2018.

[48] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 785–799, 2014.

[49] E. Vincent, S. Watanabe, A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech and Lang.*, vol. 46, pp. 535–557, 2017.

[50] D. P. Kingma and J. L. Ba, "ADAM: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–13.

[51] "P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codec," ITU-T Std. P.862.2, 2007.

[52] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[53] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[54] J. Roux, S. Wisdom, H. Erdogan, and J. Hershey, "SDR - Half-baked or Well Done?" in *Proc. ICASSP*, 2019, pp. 626–630.

[55] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *Computer Speech and Lang.*, vol. 49, pp. 37–51, 2018.

[56] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.