An Extended Kalman Filter for RTF Estimation in Dual-Microphone Smartphones

Juan M. Martín-Doñas[1], Iván López-Espejo[2], Angel M. Gomez[1], Antonio M. Peinado[1]
[1]Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain
[2]VeriDas | das-Nano, Spain
E-mail: {mdjuamart,amgg,amp}@ugr.es, ilopez@das-nano.com

***Abstract*—The performance of speech beamformers relies on a good estimation of the relative transfer function (RTF) between the captured clean speech at each microphone. Most of the proposed RTF estimators make assumptions about the clean speech statistics or need a joint estimation of the RTF and the signal statistics. In this work we propose a minimum mean square error (MMSE) estimation of the RTF in an extended Kalman filter (eKF) framework. Our method exploits the knowledge about the RTF and noise statistics with no assumptions about the clean speech statistics. The proposed approach is evaluated when employed in combination with minimum variance distortionless response (MVDR) beamforming in a dual-microphone smartphone. To this end, a database of simulated dual-channel noisy speech recordings on a smartphone was used. Experimental results show that our approach achieves the most accurate RTF estimates among the evaluated methods, yielding less speech distortion and better intelligibility while competitive perceptual quality performance is obtained.**

***Index Terms*—Relative Transfer Function, Extended Kalman Filter, Beamforming, Dual-microphone, Smartphone**

I. Introduction

Beamforming algorithms are widely used in devices with multiple microphones to enhance speech signals [1]. These algorithms usually require the estimation of the relative transfer function (RTF) between the clean speech on a reference microphone and the other ones. The simplest model assumes a multiplicative RTF in the short-time Fourier transform (STFT) domain, the so-called *narrowband approximation* [2]. When time-invariant acoustic impulse responses (AIRs) are considered, this model assumes that the RTF only depends on the considered frequency bin. This approximation is no longer valid when a finite analysis window is employed, especially in the case of reverberant environments, so inter-band and inter-frame correlations should be considered [2].

To overcome this problem, and also be able to consider possible time variations of the RTF, one solution is to model the second order statistics of the RTFs [2]. In a single-speaker scenario, most of these models assumes that the clean speech at the different microphones is perfectly correlated. Based on this, two widely used RTF estimators are the covariance subtraction (CS) [3] and covariance whitening (CW) [4] methods. Although CW provides a more accurate estimation of the RTF than CS [5], the computational complexity of solving a generalized eigenvalue (GEV) problem makes that method inappropriate for real-time applications. More recently, the eigenvalue decomposition (EVD) method was proposed in [6]. The advantage of EVD is its similar performance to the CW method but with a lower complexity [7], which makes it suitable for real-time applications. An additional issue with the former methods is that they make assumptions about the clean speech statistics that can be inaccurate, especially in reverberant environments. Also, their performance relies on a good estimation of the noise statistics [7].

Alternative methods formulate the RTF estimation as a weighted least-square (WLS) problem, where both speech and RTF sparsities are exploited [8]. These methods do not use prior statistics of the speech and noise signals. They estimate both the RTF and the noise statistics, and different constraints are imposed to the solutions. Moreover, they have higher complexity and slower convergence because they need many more frames and/or sparsity constraints to be accurate enough. In [9], an unscented Kalman filter is proposed in the time domain to jointly estimate the AIRs and the clean speech, but it assumes an autoregressive model for the speech signal and fixed-length AIRs. Other works formulate the joint estimation of the RTF and the clean speech and noise statistics in an expectation-maximization (EM) framework [10]. The EM algorithm can be used along with a Kalman filter that follows the temporal variations of the speech signal [11]. The problem of these algorithms is that they require the estimation of both the RTF and the signal statistics in an EM framework, which is computationally unfeasible on small devices.

In this work, we propose a novel minimum mean square error (MMSE) estimation of the RTF in an extended Kalman filter (eKF) framework capable to track the RTF evolution. Our method uses a priori knowledge about the RTF and noise statistics along with the observed noisy signals. The main advantage with respect to the aforementioned methods is that no assumptions about the clean speech are needed. To avoid dealing with non-zero mean complex variables [12], [13], our method works with vectors of the real and imaginary parts of the involved variables. For evaluation, we consider the estimation of the RTF associated to a dual-microphone smartphone. This is a quite common device whose typical utilization positions allow us the estimation of well-defined RTF a priori statistics, required for eKF tracking. Two typical

This work has been supported by the Spanish MINECO/FEDER Project TEC2016-80141-P and the Spanish Ministry of Education through the National Program FPU (grant reference FPU15/04161). We also acknowledge the support of NVIDIA Corporation with the donation of a Titan X GPU.

ISBN 978-90-827970-1-5 © EURASIP 2018

positions and several acoustic environments are evaluated.

The reminder of this paper is organized as follows. In Section II, the proposed method is described. In Section III, we describe the development of the dual-microphone smartphone database employed in our experiments. Then, in Section IV, our proposal is evaluated with different quality and intelligibility measures when combined with minimum variance distortionless response (MVDR) beamforming. Finally, conclusions are summarized in Section V.

II. Extended Kalman Filter-based RTF Estimation

Let us consider an additive noise distortion model in the STFT domain, namely,

$$Y_m(f,t) = X_m(f,t) + N_m(f,t), \tag{1}$$

where $Y_m(f,t)$, $X_m(f,t)$ and $N_m(f,t)$ represent, respectively, noisy speech, clean speech and noise STFT coefficients at the m-th microphone ($m = 1, 2$), f is the frequency bin and t the frame index. Without loss of generality, we consider $m = 1$ as the reference microphone and write the *narrowband* model for the secondary microphone as

$$Y_2(f,t) = A_{21}(f,t)\left(Y_1(f,t) - N_1(f,t)\right) + N_2(f,t), \tag{2}$$

where $A_{21}(f,t) = \frac{X_2(f,t)}{X_1(f,t)}$ is the relative transfer function (RTF) between the two microphones. The aim of this paper is the estimation of this RTF.

All the previous complex variables can be written as two-dimensional vectors with their corresponding real and imaginary parts. For example, we can define $\mathbf{y}_m^{(t)}$ as

$$\mathbf{y}_m^{(t)} = \begin{bmatrix} Re(Y_m(t)) & Im(Y_m(t)) \end{bmatrix}^\top, \tag{3}$$

where the index f is omitted for clarity. Similarly, $\mathbf{a}_{21}^{(t)}$ and $\mathbf{n}_m^{(t)}$ can also be defined. In order to develop an estimator for $\mathbf{a}_{21}^{(t)}$, we first assume that the RTF changes across frames according to a perturbed constant model,

$$\mathbf{a}_{21}^{(t)} = \mathbf{a}_{21}^{(t-1)} + \mathbf{w}^{(t)}, \tag{4}$$

where the random variables involved are assumed to be Gaussian-distributed, i.e., $\mathbf{a}_{21}^{(t)} \sim \mathcal{N}\left(\boldsymbol{\mu}_{A_{21}}, \boldsymbol{\Sigma}_{A_{21}}\right)$ and $\mathbf{w}^{(t)} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}\right)$. The parameters $\boldsymbol{\mu}_{A_{21}}$ and $\boldsymbol{\Sigma}_{A_{21}}$ represent, respectively, the overall mean and covariance of the RTF, as will be explained in Subsection II-A. On the other hand, $\mathbf{w}^{(t)}$ models the intra-utterance variability. The existence of this perturbation has a twofold meaning. First, the possible temporal variations of the AIRs, but also the inaccuracy of the *narrowband approximation* [2].

Also, by using (2) and (3), we propose the following non-linear observation model for $\mathbf{y}_2^{(t)}$ given $\mathbf{y}_1^{(t)}$:

$$\begin{aligned}\mathbf{y}_2^{(t)} &= \mathbf{h}\left(\mathbf{a}_{21}^{(t)}, \mathbf{n}_1^{(t)}; \mathbf{y}_1^{(t)}\right) + \mathbf{n}_2^{(t)} \\ &= \left[\mathbf{C}\left(\mathbf{y}_1^{(t)} - \mathbf{n}_1^{(t)}\right) \quad \mathbf{D}\left(\mathbf{y}_1^{(t)} - \mathbf{n}_1^{(t)}\right)\right]\mathbf{a}_{21}^{(t)} + \mathbf{n}_2^{(t)},\end{aligned} \tag{5}$$

where $\mathbf{D} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $\mathbf{C} = \mathbf{I}_2$ (2×2 identity matrix). This is an *ad-hoc* model given the specific signal $\mathbf{y}_1^{(t)}$ captured by the reference microphone. The noises are also assumed to be Gaussian-distributed, i.e., $\mathbf{n}_m^{(t)} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{N_m,t}\right)$.

Under these assumptions, we propose an MMSE estimator of $\mathbf{a}_{21}^{(t)}$ using an extended Kalman filter (eKF) framework [14]. The RTF estimate, $\widehat{\mathbf{a}}_{21}^{(t)}$, and its error covariance matrix,

$$\mathbf{P}_t = E\left[\left(\mathbf{a}_{21}^{(t)} - \widehat{\mathbf{a}}_{21}^{(t)}\right)\left(\mathbf{a}_{21}^{(t)} - \widehat{\mathbf{a}}_{21}^{(t)}\right)^\top\right], \tag{6}$$

where $(\cdot)^\top$ indicates matrix transposition, are calculated on a frame-by-frame basis through a two-step procedure. The *prediction step* uses the model of (4) to make a first estimation,

$$\widehat{\mathbf{a}}_{21}^{(t|t-1)} = \widehat{\mathbf{a}}_{21}^{(t-1)}, \tag{7}$$

$$\mathbf{P}_{t|t-1} = \mathbf{P}_{t-1} + \mathbf{Q}, \tag{8}$$

with an initialization given by $\widehat{\mathbf{a}}_{21}^{(0)} = \boldsymbol{\mu}_{A_{21}}$ and $\mathbf{P}_0 = \boldsymbol{\Sigma}_{A_{21}}$. Next, we apply the *updating step* to obtain the estimation of the RTF by using the *ad-hoc* observation model of (5). The non-linear function $\mathbf{h}$ is approximated by a first-order vector Taylor series (VTS) linearization, yielding the following eKF:

$$\widehat{\mathbf{a}}_{21}^{(t)} = \widehat{\mathbf{a}}_{21}^{(t-1)} + \mathbf{K}_t\left(\mathbf{y}_2^{(t)} - \mathbf{h}\left(\widehat{\mathbf{a}}_{21}^{(t-1)}, \mathbf{0}; \mathbf{y}_1^{(t)}\right)\right), \tag{9}$$

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_{t|t-1}, \tag{10}$$

where $\mathbf{K}_t$ is the Kalman gain, defined as

$$\begin{aligned}\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}_t^\top\big(&\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^\top + \mathbf{J}_{N_1,t}\boldsymbol{\Sigma}_{N_1,t}\mathbf{J}_{N_1,t}^\top \\ &+\boldsymbol{\Sigma}_{N_2,t} + \mathbf{J}_{N_1,t}\boldsymbol{\Sigma}_{N_1N_2,t} + \boldsymbol{\Sigma}_{N_2N_1,t}\mathbf{J}_{N_1,t}^\top\big)^{-1},\end{aligned} \tag{11}$$

and $\boldsymbol{\Sigma}_{N_1N_2,t} = \boldsymbol{\Sigma}_{N_2N_1,t}^\top$ is a spatial cross-covariance matrix of the noise. Finally,

$$\mathbf{H}_t = \left.\frac{\partial \mathbf{h}}{\partial \mathbf{a}_{21}^{(t)}}\right|_{\mathbf{n}_1^{(t)}=\mathbf{0}} = \left[\mathbf{C}\mathbf{y}_1^{(t)} \quad \mathbf{D}\mathbf{y}_1^{(t)}\right], \tag{12}$$

$$\mathbf{J}_{N_1,t} = \left.\frac{\partial \mathbf{h}}{\partial \mathbf{n}_1^{(t)}}\right|_{\mathbf{a}_{21}^{(t)}=\widehat{\mathbf{a}}_{21}^{(t-1)}} = -\left[\mathbf{C}\widehat{\mathbf{a}}_{21}^{(t-1)} \quad \mathbf{D}\widehat{\mathbf{a}}_{21}^{(t-1)}\right], \tag{13}$$

are the Jacobian matrices required for the VTS approximation.

A. Estimation of the noise and RTF statistics

The proposed eKF algorithm requires knowledge of the noise and RTF statistics previously defined. The noise statistics are obtained at each frame using the *Multichannel Speech Presence Probability* (MC-SPP) noise tracking algorithm proposed in [15]. This algorithm estimates the noise spatial correlation matrix

$$\boldsymbol{\Phi}_{nn}(f,t) = \begin{bmatrix} \Phi_{11}(f,t) & \Phi_{12}(f,t) \\ \Phi_{21}(f,t) & \Phi_{22}(f,t) \end{bmatrix}. \tag{14}$$

Assuming a zero-mean, symmetric circular complex Gaussian distribution for $N_m(f,t)$, it can be shown [13] that $\boldsymbol{\Sigma}_{N_1} = \frac{1}{2}\Phi_{11}\mathbf{I}_2$, $\boldsymbol{\Sigma}_{N_2} = \frac{1}{2}\Phi_{22}\mathbf{I}_2$, and

$$\boldsymbol{\Sigma}_{N_1N_2} = \frac{1}{2}\begin{bmatrix} Re(\Phi_{12}) & -Im(\Phi_{12}) \\ Im(\Phi_{12}) & Re(\Phi_{12}) \end{bmatrix}, \tag{15}$$

where the indices f and t are omitted for clarity.

ISBN 978-90-827970-1-5 © EURASIP 2018

On the other hand, the RTF statistics (i.e., $\boldsymbol{\mu}_{A_{21}}$, $\boldsymbol{\Sigma}_{A_{21}}$ and $\mathbf{Q}$) are estimated in advance. For the estimation of these statistics, a development set of dual-channel clean speech recordings at different acoustic environments has been employed. The dynamic range of the utterance is first normalized to its mean square value. Then, for each utterance, we compute $A_{21}(f,t)$ at those time-frequency bins where the reference channel is greater than 20 dB in order to ensure speech presence and avoid outliers. For a given frequency f, and using the sequence of selected $A_{21}(f,t)$ values, we obtain a mean value for each utterance l, $\boldsymbol{\mu}_{A_{21}}^{(l)}$, and an overall average over all the utterances, $\boldsymbol{\mu}_{A_{21}}$. Finally, the RTF covariances are computed according to (4),

$$\begin{aligned}\boldsymbol{\Sigma}_{A_{21}} &= E\left[\left(\mathbf{a}_{21}^{(t)}-\boldsymbol{\mu}_{A_{21}}\right)\left(\mathbf{a}_{21}^{(t)}-\boldsymbol{\mu}_{A_{21}}\right)^{\top}\right] \\ &= \boldsymbol{\Sigma}_{A_{21},r}+\boldsymbol{\Sigma}_{A_{21},v},\end{aligned} \tag{16}$$

$$\mathbf{Q} = E\left[\left(\mathbf{a}_{21}^{(t)}-\mathbf{a}_{21}^{(t-1)}\right)\left(\mathbf{a}_{21}^{(t)}-\mathbf{a}_{21}^{(t-1)}\right)^{\top}\right] = 2\boldsymbol{\Sigma}_{A_{21},v}. \tag{17}$$

where

$$\boldsymbol{\Sigma}_{A_{21},r} = E\left[\left(\boldsymbol{\mu}_{A_{21}}^{(l)}-\boldsymbol{\mu}_{A_{21}}\right)\left(\boldsymbol{\mu}_{A_{21}}^{(l)}-\boldsymbol{\mu}_{A_{21}}\right)^{\top}\right] \tag{18}$$

is the covariance of the utterance-dependent means $\boldsymbol{\mu}_{A_{21}}^{(l)}$, whose mean value is $\boldsymbol{\mu}_{A_{21}}$. Therefore, it accounts for the inter-utterance variability due to acoustic changes. On the other hand,

$$\boldsymbol{\Sigma}_{A_{21},v} = E\left[\left(\mathbf{a}_{21}^{(t)}-\boldsymbol{\mu}_{A_{21}}^{(l)}\right)\left(\mathbf{a}_{21}^{(t)}-\boldsymbol{\mu}_{A_{21}}^{(l)}\right)^{\top}\right] \tag{19}$$

represents an intra-utterance variability mainly due to the inaccuracy of the *narrowband approximation*. This covariance matrix has been estimated by averaging the particular intra-utterance sample covariances $\boldsymbol{\Sigma}_{A_{21},v}^{(l)}$.

B. RTF updating

The observation model of (5) assumes that speech is present. Moreover, the updating information provided by (5), and applied in (9), will be more accurate as long as speech more clearly stands above noise. Thus, the RTF at each frequency will be updated only in those frames where the SNR is large enough. Otherwise the previous value will be preserved. In order to estimate the updating binary mask, which indicates the bins where the RTF is to be updated, we use the following parameter proportional to the SNR at the considered time-frequency bin [15],

$$\beta(f,t) = \mathbf{Y}^{H}(f,t)\boldsymbol{\Phi}_{nn}^{-1}(f,t)\boldsymbol{\Phi}_{xx}(f,t)\boldsymbol{\Phi}_{nn}^{-1}(f,t)\mathbf{Y}(f,t), \tag{20}$$

where $\mathbf{Y}(f,t) = \begin{bmatrix}Y_1(f,t) & Y_2(f,t)\end{bmatrix}^{\top}$ is the dual-channel noisy speech vector, and $\boldsymbol{\Phi}_{xx}(f,t)$ is the clean speech spatial correlation matrix. This matrix is obtained from the noise tracking algorithm [15]. In our implementation, the values of the binary mask are set to true when $10\log_{10}\beta(f,t) > 30$ dB.

III. Dual-microphone Smartphone Database

The proposed algorithm will be applied to the estimation of the RTF between the primary and secondary microphones of a smartphone. With this purpose, we simulated dual-channel noisy speech recordings on a smartphone. We considered two different modes of use: close-talk (CT, when the loudspeaker of the smartphone is placed at the ear of the user) and far-talk (FT, when the user holds the device at a distance from her/his face). The methodology followed is similar to the one considered in [16], [17]. The smartphone employed is a Motorola Moto G, which has a primary microphone at its bottom and a secondary one at its top.

First, we recorded dual-channel noise signals at eight different noisy environments, both for CT and FT modes. The noise signals recorded were divided into two sets. Set A includes the noises car (CAR), street (STR), babble (BAB) and mall (MLL). Set B includes the noises bus (BUS), cafe (CAF), pedestrian street (PST) and bus station (BST).

Next, we obtained several dual-microphone AIRs, both for CT and FT, at four different reverberant acoustic environments. Each type of noise was assigned to a specific type of acoustic environment, according to its expected reverberation level. The reverberant environments (and their corresponding noises) are the following: no reverberation (CAR, STR, PST), low (BUS, CAF), medium (BAB, BST) and high (MLL). We recorded both close-talk high quality cardioid microphone and smartphone recordings of clean speech, which were synchronized later. A sampling frequency of 48 kHz was selected. Then, the AIRs were estimated using these clean speech recordings. First, the high quality microphone recording is assumed to be the true-ground clean speech signal $s(n)$. Then, the smartphone recordings $x_m(n)$ are approximated as filtered versions of $s(n)$ using FIR filters $a_m(n)$, which model both the environment and the microphone responses. The estimation of $a_m(n)$ is formulated as a least-square (LS) problem with sparse coefficients enforced by using $\mathcal{L}_1$-norm. First, the LS-based cost function is defined as

$$J(\mathbf{a}_m) = \mathbf{a}_m^{\top}\mathbf{R}_s\mathbf{a}_m - \mathbf{a}_m^{\top}\mathbf{r}_{x_m s} - \mathbf{r}_{x_m s}^{\top}\mathbf{a}_m, \tag{21}$$

where $\mathbf{a}_m$ is an $L\times 1$ vector with the AIR coefficients, $\mathbf{R}_s$ is the $L\times L$ autocorrelation matrix of $s(n)$ and $\mathbf{r}_{x_m s}$ is the $L\times 1$ cross-correlation vector between $x_m(n)$ and $s(n)$. We define $\mathbf{a}_m^*$ as the value of the AIR that analytically minimizes $J(\mathbf{a}_m)$. Finally, $\mathbf{a}_m$ is obtained as

$$\mathbf{a}_m = \underset{\mathbf{a}_m}{\operatorname{argmin}}\left\{(1-\lambda)\tfrac{J(\mathbf{a}_m)-J(\mathbf{a}_m^*)}{|J(\mathbf{a}_m^*)|} + \lambda\tfrac{\|\mathbf{a}_m\|_1}{\|\mathbf{a}_m^*\|_1}\right\}, \tag{22}$$

where $\|\cdot\|_1$ means $\mathcal{L}_1$-norm and λ is a trade-off factor between LS minimization and filter sparseness. Finally, the estimated AIRs are downsampled to 16 kHz. During our experiments, we set $\lambda = 0.15$ and different values of L were considered for each reverberant environment (320, 960, 2560, 5120). The minimization problem in (22) has not a closed-form solution, but it is a convex equation, so it can be solved using either convex optimization or gradient-based methods.

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE I
PESQ AND STOI RESULTS FOR THE NOISY SPEECH FROM THE REFERENCE MICROPHONE (NOISY) AND THE DIFFERENT RTF-ESTIMATION METHODS IN CLOSE-TALK (CT) AND FAR-TALK (FT) CONDITIONS.

	CT								FT							
SNR	**Noisy**		**EVD**		**CW**		**eKF (Prop.)**		**Noisy**		**EVD**		**CW**		**eKF (Prop.)**	
(dB)	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
20	1.803	0.882	1.985	0.888	**1.994**	**0.889**	1.916	**0.889**	1.943	0.910	2.172	0.911	**2.177**	0.910	2.093	**0.915**
15	1.516	0.843	1.664	**0.853**	**1.671**	**0.853**	1.618	**0.853**	1.615	0.872	1.866	0.879	**1.870**	0.878	1.744	**0.881**
10	1.309	0.791	1.423	0.804	**1.427**	**0.805**	1.396	**0.805**	1.370	0.811	1.575	0.822	**1.578**	0.822	1.471	**0.824**
5	1.180	0.717	1.257	0.731	**1.260**	0.732	1.246	**0.736**	1.211	0.727	1.359	0.726	**1.361**	0.726	1.284	**0.737**
0	1.116	0.626	1.162	0.636	**1.163**	0.637	1.158	**0.648**	1.129	0.623	**1.221**	0.600	**1.221**	0.601	1.171	**0.630**
-5	1.099	0.532	1.127	0.549	**1.130**	0.549	1.112	**0.561**	1.122	0.517	1.179	0.496	**1.182**	0.497	1.124	**0.532**

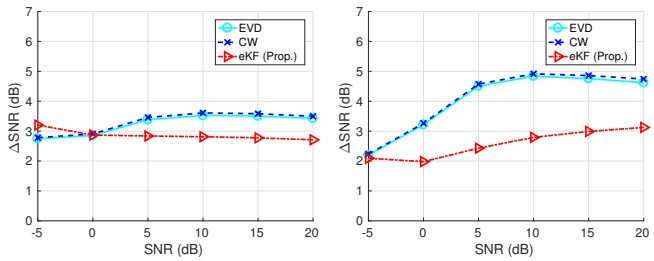

Fig. 1. Performance of the evaluated methods in terms of ΔSNR for close-talk (a) and far-talk (b) positions.

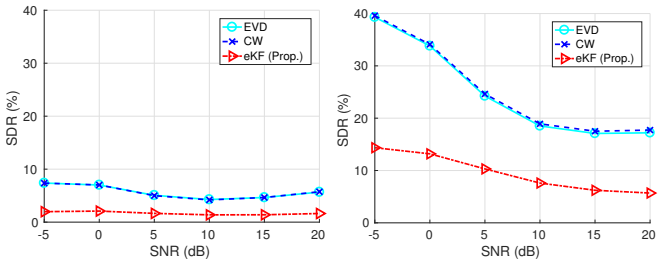

Fig. 2. Performance of the evaluated methods in terms of SDR (%) for close-talk (a) and far-talk (b) positions.

Finally, the CT and FT databases were set up. Clean speech signals were obtained from the VCTK database [18], downsampled to 16 kHz. The utterances from the 108 speakers of the VCTK were split into three sets: training (72 speakers), development (18 speakers) and test (18 speakers) sets. For each noisy environment, the clean speech is filtered using dual-channel AIRs randomly selected from the set of available AIRs for the corresponding reverberant environment. The noise signals are added at six different SNR levels from -5 dB to 20 dB. The training and development sets only include noises from set A, while the test set includes both set A and set B noises.

IV. EXPERIMENTAL RESULTS

The proposed eKF method is compared with the EVD and CW methods for the estimation of the RTF. The estimated RTF is employed, along with the noise spatial correlation matrix obtained using MC-SPP, to enhance the noisy speech from the reference microphone using an MVDR beamformer. Both EVD and CW make use of the clean speech spatial correlation matrix $\mathbf{\Phi}_{xx}$ obtained for MC-SPP. Also, EVD and CW use the RTF updating described in Subsection II-B for a fair comparison. For STFT computation, we choose a 25 ms square-root Hann window with 75% overlap.

The resulting enhanced signals are evaluated using different objective measures. We use PESQ (Perceptual Evaluation of the Speech Quality) [19] and STOI (Short-Time Objective Intelligibility) [20] metrics for evaluating, respectively, perceptual quality and intelligibility. Clean speech from the reference microphone is taken as a reference for these performance metrics. The results for close-talk (CT) and far-talk (FT) conditions are shown in Table I. The results obtained for the noisy speech from the reference microphone are included as our baseline. We also evaluate the SNR increment (array gain, ΔSNR) [1] and the speech distortion ratio (SDR) [1] for the different methods with respect to the noisy and clean speech from the reference microphone, respectively. For both measures, as in [7], we use 30 ms non-overlapping voice segments and the final measure is averaged over all segments. The ΔSNR and SDR results are shown in Figures 1 and 2, respectively.

These results show that the proposed method achieves the best results in terms of intelligibility and speech distortion, especially in far-talk conditions. The improvements are more significant when the SNR is lower, where our method is more robust against noise. On the other hand, in terms of perceptual quality and noise reduction, the CW and EVD methods generally achieve better results, especially in far-talk and medium/high SNRs. As expected, CW and EVD perform similarly, while EVD has a lower computational burden. While CW and EVD focus on noise reduction, our method achieves the best performance in terms of distortion over the speech signal, finally yielding better intelligibility scores. This can be explained by the fact that speech distortion plays a more important role in intelligibility metrics than in perceptual ones (as PESQ) [21]. Finally, it can be also observed that smaller improvements are achieved for the CT position. This is due to the lower performance of the MVDR beamformer when speech is meaningfully attenuated at the secondary microphone with respect to the primary one (as in CT mode).

ISBN 978-90-827970-1-5 © EURASIP 2018

In order to better understand these results, let us consider the enhancement algorithm employed. MVDR beamforming tries to minimize the noise power with a distortionless condition for the reference speech signal, which could be achieved through a perfect estimation of the RTF. In general, this is not feasible, so speech distortion is introduced. The fact that our method introduces less distortion indicates that our RTF estimation is more accurate. CW and EVD make assumptions about the clean speech spatial correlation that can turn out inaccurate, especially in reverberant environments, while our method does not need any assumption about the clean speech signal. On the other hand, CW and EVD exhibit a strong dependency on the estimation of the noise statistics. Thus, they likely tend to reduce noise at the expense of increasing distortion on the speech signal. Also, since the estimation of the noise might be inaccurate and may include speech components, this could introduce speech cancellation on the beamformer. Our method neither has such a strong dependence on the noise nor needs the estimation of the noisy speech statistics as the other methods. This is due to the fact that our method directly works with the RTF statistics, which are much easier to model than the clean speech ones.

In view of the results and the previous analysis, we can conclude that the proposed method achieves more accurate RTF estimates than EVD and CW, which yields better results on speech distortion introduced. Despite the lower noise reduction obtained when used in combination with MVDR beamforming, our method achieves an advantageous trade-off between noise reduction and speech distortion, which leads on better intelligibility results. That is, we obtain a more accurate RTF estimator that also achieves a competitive perceptual quality performance, much less speech distortion and better intelligibility scores when employed with MVDR beamforming.

V. Conclusions

In this paper we have proposed an eKF-based estimation of the RTF between two microphones. The proposed approach exploits the statistics of the RTF and its variability without any assumption about the clean speech signal. We have evaluated the proposed estimator when employed with MVDR beamforming on a dual-microphone smartphone to enhance the speech signal. With this purpose, we have used a database of simulated dual-channel noisy speech recordings on a smartphone to test our method. The experimental results indicate that our method achieves more accurate RTF estimates than other methods, leading to less speech distortion and better intelligibility of the enhanced speech signal. As future work, we will extend the proposed methodology to other specific dual-microphone enhancement techniques as well as to a more general case with more microphones, where the use of a more accurate RTF is expected to yield significant improvements.

References

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

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

[3] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[4] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[5] 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.

[6] 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 Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 785–799, 2014.

[7] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. HSCMA*, 2017, pp. 11–15.

[8] Z. Koldovský, J. Malek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 8, pp. 1335–1347, 2015.

[9] S. Gannot and M. Moonen, "On the application of the unscented Kalman filter to speech processing," in *Proc. IWAENC*, 2003.

[10] D. Schmid, S. Malik, and G. Enzner, "An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain," in *Proc. ICASSP*, 2012, pp. 17–20.

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

[12] D. H. Dini and D. P. Mandic, "Class of widely linear complex Kalman filters," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 5, pp. 775–786, 2012.

[13] G. R. Ducharme, P. Lafaye de Micheaux, and B. Marchina, "The complex multinormal distribution, quadratic forms in complex random vectors and an omnibus goodness-of-fit test for the complex normal distribution," *Annals of the Institute of Statistical Mathematics*, vol. 68, no. 1, pp. 77–104, 2016.

[14] F. L. Lewis, D. Popa, and L. Xie, *Optimal and robust estimation*, 2nd ed., ser. Automation and control engineering. CRC Pr, 2008, vol. 26.

[15] M. Souden, J. Benesty, S. Affes, and J. Chen, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2159–2169, 2011.

[16] I. López-Espejo, A. M. Gomez, J. A. González, and A. M. Peinado, "Feature enhancement for robust speech recognition on smartphones with dual-microphone," in *Proc. EUSIPCO*, 2014.

[17] I. López-Espejo, A. M. Peinado, A. M. Gomez, and J. A. González, "Dual-channel spectral weighting for robust speech recognition in mobile devices," *Digital Signal Processing*, vol. 75, pp. 13–24, 2018.

[18] J. Yamagishi. (2012) English multi-speaker corpus for CSTR voice cloning toolkit. [Online]. Available: http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

[19] "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.

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

[21] J. Ma and P. C. Loizou, "SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech," *Speech Communication*, vol. 53, no. 3, pp. 340–354, 2011.

ISBN 978-90-827970-1-5 © EURASIP 2018