





Article

# Dual-Channel Speech Enhancement Based on Extended Kalman Filter Relative Transfer Function Estimation <sup>†</sup>

Juan M. Martín-Doñas <sup>1,\*</sup>, Antonio M. Peinado <sup>1</sup>, Iván López-Espejo <sup>2</sup> and Angel Gomez <sup>1</sup>

<sup>1</sup> Department of Signal Theory, Telematics and Communications, University of Granada, 18071 Granada, Spain; amp@ugr.es (A.M.P.); amgg@ugr.es (A.G.)

<sup>2</sup> Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark; ivl@es.aau.dk

\* Correspondence: mdjuamart@ugr.es

<sup>†</sup> This paper is an extended version of our paper published in the conference IberSPEECH2018.

Received: 4 May 2019; Accepted: 17 June 2019; Published: 20 June 2019



**Abstract:** This paper deals with speech enhancement in dual-microphone smartphones using beamforming along with postfiltering techniques. The performance of these algorithms relies on a good estimation of the acoustic channel and speech and noise statistics. In this work we present a speech enhancement system that combines the estimation of the relative transfer function (RTF) between microphones using an extended Kalman filter framework with a novel speech presence probability estimator intended to track the noise statistics' variability. The available dual-channel information is exploited to obtain more reliable estimates of clean speech statistics. Noise reduction is further improved by means of postfiltering techniques that take advantage of the speech presence estimation. Our proposal is evaluated in different reverberant and noisy environments when the smartphone is used in both close-talk and far-talk positions. The experimental results show that our system achieves improvements in terms of noise reduction, low speech distortion and better speech intelligibility compared to other state-of-the-art approaches.

**Keywords:** dual-microphone smartphone; beamforming; relative transfer function; speech presence probability; postfiltering

## 1. Introduction

Speech-related services are pervasively available on mobile devices such as smartphones or tablets. However, reverberant and noisy environments, where these devices are frequently used, often degrade speech signal quality and/or intelligibility [1]. Many current devices include several microphones, so that multi-channel speech processing techniques can be applied to reduce the distortions, which improves the noise reduction performance compared to single-channel approaches. This is our research focus.

The most common multi-channel speech processing technique is beamforming [2], which applies spatial filtering to the noisy speech signals captured by several microphones. One of these beamformers is the well-known Minimum Variance Distortionless Response (MVDR) beamformer [3], which has the advantage of being able to reduce the noise power without introducing speech distortion. The performance of MVDR depends on an accurate estimation of the noise spatial characteristics and the acoustic transfer function (ATF) between the target speaker and the microphones. For the estimation of the noise spatial statistics, methods based on multi-channel speech presence probability (SPP) [4,5] are commonly used when there is no prior knowledge of the signal propagation or the spatial structure of the noise. These techniques update the noise information when speech absence is detected, which allows for the tracking of time-varying noise signals.

However, the estimation of ATFs is a more challenging task, as these depend on the speaker's location, room acoustics and microphone responses. One possible solution is based on the estimation of the beamformer weights for a certain reference microphone. As a result, the problem becomes that of estimating a relative transfer function (RTF) between the reference microphone and the rest of microphones [6]. The most common RTF estimators are based on sub-space search, using estimates of the noisy speech and the noise spatial correlations. This category includes techniques such as covariance subtraction [7], covariance whitening [8,9] and eigenvalue decomposition [10,11]. A weighted least-squares RTF estimator was proposed in Reference [12] but it is unfeasible for real-time applications. Other approaches based on online expectation-maximization [13,14] have been analyzed to jointly estimate the RTF and the clean speech and noise statistics under the assumption of a spatially-white noise field. Recently, we proposed an extended Kalman filter (eKF) framework for the estimation of the RTF between two microphones [15] and evaluated its performance on a dual-microphone smartphone under different noisy and reverberant environments. We showed that this technique obtains better estimation accuracy than the sub-space-based estimators while allowing for the tracking of the RTF variability, especially in highly reverberant scenarios.

Despite beamforming algorithms are often used in multi-microphone devices, their performance is quite limited on dual-microphone smartphones mainly due to the reduced number of microphones, their particular placement on the device and the short separation between them [16]. Therefore, alternatives to beamforming are necessary to obtain a good performance in these situations. One possible approach is the use of single-channel filters for the reference microphone using statistical information obtained from the dual-channel signal. For example, the power level difference (PLD) algorithm proposed in Reference [17] exploits the clean speech power difference between microphones when the smartphone is used in close-talk (CT) conditions (i.e., when the loudspeaker of the smartphone is placed at the ear of the user). An estimate of the single-channel noise statistics is obtained for both microphones and a Wiener filter is calculated from them. In the case of smartphones used in far-talk (FT) conditions (i.e., when the user holds the device at a distance from her/his face), the algorithm proposed in Reference [18] exploits the spatial properties of the noisy speech and noise signals. This spatial information is used along with a single-channel SPP detector to estimate the noise at the reference channel and the signal-to-noise ratio (SNR), which are needed for the filter design. This proposal was later extended to general multi-channel devices in Reference [19].

While the techniques mentioned above apply a single-channel filtering at the reference microphone, better performance can be achieved when the filter is designed to operate at the beamformer output, what is known as postfiltering [1]. Moreover, the multi-channel Wiener filter can be expressed as an MVDR beamformer followed by a single-channel Wiener filter [6]. Several postfilters based on the Wiener filter have been proposed in the literature [20–22], mainly differing in the assumptions made about the noise field. The authors of Reference [23] also evaluate the use of non-linear postfilters, showing improvements with respect to the linear approaches. The work in Reference [24] evaluates the performance of a generalized sidelobe canceler (GSC) beamformer along with an SPP estimator and a non-linear postfilter, showing that the SPP information is useful in the postfiltering design. The postfiltering approach has also been studied on dual-microphone scenarios [25,26]. For example, in Reference [27] we extended our eKF-RTF framework with the use of postfilters along with MVDR beamforming and the SPP estimator of Reference [4] in order to improve noise reduction. These postfilters exploit the SPP and the RTFs previously obtained to estimate the required single-channel statistics. Both linear and non-linear postfilters were evaluated, showing better performance than a standalone MVDR and other state-of-the-art enhancement algorithms intended to be used in smartphones.

In this work we analyze and further extend our dual-microphone speech enhancement eKF-RTF framework presented in References [15,27]. This extension is developed in a threefold way. First, we present a more detailed derivation of the eKF-RTF estimator where we improve the estimation of the *a priori* RTF statistics without any simplification in the estimation of the covariance matrices.

Second, we analyze the *a posteriori* SPP estimation and the importance of the *a priori* speech absence probability (SAP). Then, we propose a novel SAP estimator suitable for dual-microphone smartphones, which exploits the spatial structure of the noise and the power differences between microphones, showing improvements with respect to the estimator used in our previous work. Third, we take advantage of the better estimates of SPP, RTFs and single-channel statistics to redefine the postfilters described in Reference [27] and propose new ones. As a result, we end up with a comprehensive dual-microphone speech enhancement system for smartphones that shows a great performance in terms of both speech quality and intelligibility when compared to other state-of-the-art approaches.

The rest of the paper is organized as follows: in Section 2, an overview of our dual-microphone speech enhancement system is given, presenting the constituent elements of the algorithm. Next, in Section 3, the eKF-RTF framework is detailed. The estimation of the noise statistics and the SPP along with the new proposals for SAP estimation on smartphones are developed in Section 4. In Section 5, the different postfilters are described and the estimation of the required single-channel statistics is addressed. In Section 6, the experimental framework and results are presented and discussed, while Section 7 finally summarizes the conclusions.

## 2. System Overview

The proposed enhancement system for dual-microphone smartphones is depicted in Figure 1. The microphones capture the noisy speech signals  $y_m(n)$ , where  $m$  indicates the microphone index ( $m = 1, 2$ ). We assume an additive noise distortion model which in the short-time Fourier transform (STFT) domain can be expressed as

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

where  $Y_m(t, f)$ ,  $X_m(t, f)$  and  $N_m(t, f)$  represent, respectively, noisy speech, clean speech and noise signal STFTs,  $t$  is the frame index and  $f$  the frequency bin. The two channel components can be stacked in a vector,

$$\mathbf{Y}(t, f) = [Y_1(t, f) \quad , \quad Y_2(t, f)]^T, \tag{2}$$

where  $[\cdot]^T$  indicates matrix transposition. Similarly, we define vectors  $\mathbf{X}(t, f)$  and  $\mathbf{N}(t, f)$ . In the following, we will consider that each frequency component can be processed independently from the others, which is commonly referred to as narrowband approximation [6].

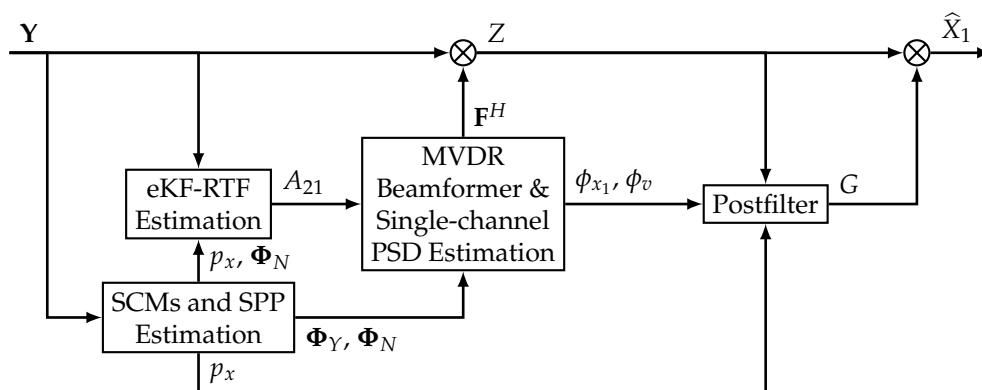


Figure 1. Overview of the dual-microphone enhancement system for smartphones.

As shown in Figure 1, the beamforming block requires two previous procedures. First, a speech presence probability (SPP)-based algorithm is employed for the estimation of the noisy speech and noise spatial correlation matrices (SCM),  $\Phi_Y(t, f)$  and  $\Phi_N(t, f)$ , respectively, and the SPP,  $p_x(t, f)$ .

Then, an extended Kalman filter (eKF)-based estimator obtains the relative transfer function (RTF) between the secondary microphone ( $m = 2$ ) and the reference microphone ( $m = 1$ ), defined as

$$A_{21}(f, t) = \frac{X_2(f, t)}{X_1(f, t)}. \tag{3}$$

Once the SCM matrices and RTF have been computed, the dual-channel noisy speech vector  $\mathbf{Y}(t, f)$  is processed using a Minimum Variance Distortionless Response (MVDR) beamformer, which applies spatial filtering yielding an output signal  $Z(t, f)$  defined as

$$Z(t, f) = \mathbf{F}^H(t, f)\mathbf{Y}(t, f), \tag{4}$$

where  $(\cdot)^H$  indicates a Hermitian transpose. The coefficients of the beamformer,  $\mathbf{F}(t, f)$ , are estimated as [2],

$$\mathbf{F}(t, f) = \frac{\Phi_N^{-1}(t, f)\mathbf{d}(t, f)}{\mathbf{d}^H(t, f)\Phi_N^{-1}(t, f)\mathbf{d}(t, f)}, \tag{5}$$

where

$$\mathbf{d}(t, f) = \left[ 1 \quad , \quad A_{21}(t, f) \right]^T \tag{6}$$

is the steering vector normalized to the reference channel.

Finally, the speech signal at the beamformer output is enhanced by a single-channel postfilter for additional noise reduction. A spectral gain  $G(t, f)$  is obtained using the power spectral density (PSD) of the speech at the reference microphone  $\phi_{x_1}(t, f)$ , the PSD of the residual noise  $\phi_v(t, f)$  and the SPP  $p_x(t, f)$ . The above single-channel statistics ( $\phi_{x_1}(t, f)$  and  $\phi_v(t, f)$ ) are estimated from the noisy speech and noise SCMs and the RTF. The gain function is further processed by a musical noise reduction algorithm [28] and finally applied to the beamformer output, thus obtaining

$$\hat{X}_1(t, f) = G(t, f)Z(t, f), \tag{7}$$

where  $\hat{X}_1(t, f)$  is the estimate of the clean speech signal at the reference microphone. In the following sections, the description of the different parts of the system is addressed.

### 3. Extended Kalman Filter-Based Relative Transfer Function Estimation

The proposed system requires knowledge of the RTF between the two microphones, namely  $A_{21}(f, t)$ . Among other approaches, the computation of this RTF can be addressed as the estimation of a variable that changes across frames in terms of a stochastic model. Given this dynamic model, the noise statistics and the noisy observations, we proposed in [15] the tracking of the RTF using an extended Kalman filter (eKF), showing a better estimation performance in comparison with other state-of-the-art approaches. Next, we describe in detail the derivation of this eKF-based RTF estimation.

First, we formulate the narrowband approximation for a given frequency of the noisy speech signal at the secondary microphone in terms of the reference microphone as

$$Y_2(t) = A_{21}(t) (Y_1(t) - N_1(t)) + N_2(t), \tag{8}$$

where the frequency index  $f$  has been omitted for the sake of simplicity. In addition, complex variables are expressed as stacked vectors of their real and imaginary parts. For example, given  $Y_m(t) = Y_{m,t}^r + jY_{m,t}^i$  with  $m = 1, 2$ , we can define

$$\mathbf{y}_{m,t} = \left[ Y_{m,t}^r \quad , \quad Y_{m,t}^i \right]^T, \tag{9}$$

and, in a similar way,

$$\mathbf{a}_{21,t} = \left[ A_{21,t}^r \quad , \quad A_{21,t}^i \right]^\top, \tag{10}$$

$$\mathbf{n}_{m,t} = \left[ N_{m,t}^r \quad , \quad N_{m,t}^i \right]^\top. \tag{11}$$

Using these definitions for the model variables in terms of vectors, we propose the following dynamic and observation models needed by the Kalman filter:

- Dynamic model for the RTF  $\mathbf{a}_{21,t}$ : We assume that the state vector  $\mathbf{a}_{21,t}$  is a random walk stochastic process which can be expressed as

$$\mathbf{a}_{21,t} = \mathbf{a}_{21,t-1} + \mathbf{w}_t, \tag{12}$$

where

$$\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}) \tag{13}$$

is a zero-mean multivariate white Gaussian noise that models the variability of the RTF. A detailed discussion on this model is provided in Section 3.2.

- Observation model for the noisy speech at the secondary microphone,  $\mathbf{y}_{2,t}$ : It is defined using the distortion model in (8) as

$$\begin{aligned} \mathbf{y}_{2,t} &= \mathbf{h}(\mathbf{a}_{21,t}, \mathbf{n}_{1,t}; \mathbf{y}_{1,t}) + \mathbf{n}_{2,t} \\ &= \left( \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} (Y_{1,t}^r - N_{1,t}^r) + \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} (Y_{1,t}^i - N_{1,t}^i) \right) \mathbf{a}_{21,t} + \mathbf{n}_{2,t}, \end{aligned} \tag{14}$$

where the noises are assumed to be zero-mean multivariate Gaussian variables,

$$\begin{bmatrix} \mathbf{n}_{1,t} \\ \mathbf{n}_{2,t} \end{bmatrix} \sim \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} \Sigma_{n_{11},t} & \Sigma_{n_{12},t} \\ \Sigma_{n_{21},t} & \Sigma_{n_{22},t} \end{bmatrix} \right), \tag{15}$$

where  $\Sigma_{n_{ij},t} = E \left[ \mathbf{n}_{i,t} \mathbf{n}_{j,t}^\top \right]$  is a noise covariance matrix whose estimation is addressed in Section 4. Additionally, we assume that there is no correlation between  $\mathbf{a}_{21,t}$  and  $\mathbf{n}_{m,t}$  while  $\mathbf{n}_{1,t}$  and  $\mathbf{n}_{2,t}$  are correlated. The model  $\mathbf{h}$  is non linear because of the product between the variables  $\mathbf{a}_{21,t}$  and  $\mathbf{n}_{1,t}$ . This model depends on the observation  $\mathbf{y}_{1,t}$ , which acts as a model parameter.

Finally, the Kalman filter framework is applied to obtain a (recursive) minimum mean square error (MMSE) estimate of  $\mathbf{a}_{21,t}$ . This is a two-step procedure that is applied frame-by-frame for all frequencies:

1. The *prediction step*, using the model (12), is applied for every frame  $t > 0$ ,

$$\hat{\mathbf{a}}_{21,t|t-1} = \hat{\mathbf{a}}_{21,t-1}, \tag{16}$$

$$\mathbf{P}_{t|t-1} = \mathbf{P}_{t-1} + \mathbf{Q}, \tag{17}$$

where

$$\mathbf{P}_t = E \left[ (\mathbf{a}_{21,t} - \hat{\mathbf{a}}_{21,t}) (\mathbf{a}_{21,t} - \hat{\mathbf{a}}_{21,t})^\top \right], \tag{18}$$

$$\mathbf{P}_{t|t-1} = E \left[ \left( \mathbf{a}_{21,t} - \hat{\mathbf{a}}_{21,t|t-1} \right) \left( \mathbf{a}_{21,t} - \hat{\mathbf{a}}_{21,t|t-1} \right)^\top \right] \tag{19}$$

are error covariance matrices. The Kalman filter is initialized using the overall mean and covariance of the RTF, that is,  $\hat{\mathbf{a}}_{21,0} = \boldsymbol{\mu}_{a_{21}}$  and  $\mathbf{P}_0 = \Sigma_{a_{21}}$ . Further details about these parameters are provided in Section 3.2.

- The *updating step* is applied to correct the previous estimation with the observations  $\mathbf{y}_{1,t}$  and  $\mathbf{y}_{2,t}$  (whose relationship is given by Equation (14)),

$$\hat{\mathbf{a}}_{21,t} = \hat{\mathbf{a}}_{21,t|t-1} + \mathbf{K}_t \left( \mathbf{y}_{2,t} - \boldsymbol{\mu}_{y,t} \right), \tag{20}$$

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{S}_{y,t} \mathbf{K}_t^\top, \tag{21}$$

where

$$\mathbf{K}_t = \mathbf{C}_{ay,t} \mathbf{S}_{y,t}^{-1} \tag{22}$$

is the Kalman gain, and

$$\boldsymbol{\mu}_{y,t} = E \left[ \mathbf{y}_{2,t} \right], \tag{23}$$

$$\mathbf{S}_{y,t} = E \left[ \left( \mathbf{y}_{2,t} - \boldsymbol{\mu}_{y,t} \right) \left( \mathbf{y}_{2,t} - \boldsymbol{\mu}_{y,t} \right)^\top \right], \tag{24}$$

$$\mathbf{C}_{ay,t} = E \left[ \left( \mathbf{a}_{21,t} - \hat{\mathbf{a}}_{21,t|t-1} \right) \left( \mathbf{y}_{2,t} - \boldsymbol{\mu}_{y,t} \right)^\top \right] \tag{25}$$

are necessary statistics, the estimation of which is developed in the next subsection. These equations correspond to the most general Kalman filter model as described in Reference [29]. We define the equations in terms of  $\boldsymbol{\mu}_{y,t}$ ,  $\mathbf{S}_{y,t}$  and  $\mathbf{C}_{ay,t}$  because the non-linearity of  $\mathbf{h}$  in Equation (14) makes their estimation non-trivial.

To deal with non-linear functions, two variants of the Kalman filter are widely used: the eKF and the unscented Kalman filter (uKF) [29]. In the case of the model of (14), it could be demonstrated that both eKF and uKF yield the same closed-form expressions for  $\boldsymbol{\mu}_{y,t}$ ,  $\mathbf{S}_{y,t}$  and  $\mathbf{C}_{ay,t}$ . We choose the eKF approach because it gives a more stable computational solution and it is easier to implement than the uKF one. In the next subsection the eKF approach is presented.

### 3.1. Vector Taylor Series Approximation

The eKF is based on applying a linearization over the prediction and observation models using first-order vector Taylor series (VTS). In our case, the prediction model is linear, so the linearization is only applied to function  $\mathbf{h}$  in the observation model. The first-order VTS approximates the model of Equation (14) as

$$\mathbf{y}_{2,t} \simeq \mathbf{h} \left( \hat{\mathbf{a}}_{21,t|t-1}, \mathbf{0}; \mathbf{y}_{1,t} \right) + \mathbf{H}_t \left( \mathbf{a}_{21,t} - \hat{\mathbf{a}}_{21,t|t-1} \right) + \mathbf{J}_{n_1,t} \mathbf{n}_{1,t} + \mathbf{n}_{2,t}, \tag{26}$$

where

$$\mathbf{H}_t = \left. \frac{\partial \mathbf{h}}{\partial \mathbf{a}_{21,t}} \right|_{\mathbf{n}_{1,t}=\mathbf{0}} = \begin{bmatrix} Y_{1,t}^r & -Y_{1,t}^i \\ Y_{1,t}^i & Y_{1,t}^r \end{bmatrix}, \tag{27}$$

$$\mathbf{J}_{n_1,t} = \left. \frac{\partial \mathbf{h}}{\partial \mathbf{n}_{1,t}} \right|_{\mathbf{a}_{21,t}=\hat{\mathbf{a}}_{21,t|t-1}} = - \begin{bmatrix} \hat{A}_{21,t|t-1}^r & -\hat{A}_{21,t|t-1}^i \\ \hat{A}_{21,t|t-1}^i & \hat{A}_{21,t|t-1}^r \end{bmatrix} \tag{28}$$

are the Jacobian matrices required for the VTS approach.

Finally, using (26), the noisy speech statistics can be estimated as [29]

$$\boldsymbol{\mu}_{y,t} \simeq \mathbf{h} \left( \hat{\mathbf{a}}_{21,t|t-1}, \mathbf{0}; \mathbf{y}_{1,t} \right), \tag{29}$$

$$\mathbf{S}_{y,t} \simeq \mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^\top + \mathbf{J}_{n_1,t} \boldsymbol{\Sigma}_{n_{11},t} \mathbf{J}_{n_1,t}^\top + \boldsymbol{\Sigma}_{n_{22},t} + \mathbf{J}_{n_1,t} \boldsymbol{\Sigma}_{n_{12},t} + \boldsymbol{\Sigma}_{n_{21},t} \mathbf{J}_{n_1,t}^\top, \tag{30}$$

$$\mathbf{C}_{ay,t} \simeq \mathbf{P}_{t|t-1} \mathbf{H}_t^\top. \quad (31)$$

### 3.2. A Priori RTF Statistics

The dynamic model for the RTF vector  $\mathbf{a}_{21,t}$  presented in Equation (12) accounts for the process variability across time in terms of a white noise  $\mathbf{w}_t$ . This variability has a twofold meaning: first, the possible temporal variations of the acoustic channel due to environment changes, head or smartphone movements, etc., and, second, the inaccuracy of the narrowband approximation of Equation (8) [6]. This is due to the fact that, especially in reverberant environments where the analysis window of the STFT is shorter than the impulse response of the acoustic channel, the transfer function is not multiplicative but convolutive. This convolutive transfer function expands in both time and frequency dimensions. Thus, nearby frames and frequencies are correlated, which violates the narrowband approximation [6]. In order to account for the variability associated to this effect, we will assume the statistical model proposed in (12) so that, although the channel could be time-invariant, the RTF, as defined in Equation (3), is time-variant.

The eKF-based RTF estimator requires the *a priori* statistics of the RTF. These statistics are the overall mean vector and covariance matrix of the RTF, respectively,

$$\boldsymbol{\mu}_{a_{21}} = E[\mathbf{a}_{21,t}], \quad (32)$$

$$\boldsymbol{\Sigma}_{a_{21}} = E\left[\left(\mathbf{a}_{21,t} - \boldsymbol{\mu}_{a_{21}}\right)\left(\mathbf{a}_{21,t} - \boldsymbol{\mu}_{a_{21}}\right)^\top\right], \quad (33)$$

and also the covariance of the RTF variability across frames,

$$\mathbf{Q} = E\left[\mathbf{w}_t \mathbf{w}_t^\top\right] = E\left[\left(\mathbf{a}_{21,t} - \mathbf{a}_{21,t-1}\right)\left(\mathbf{a}_{21,t} - \mathbf{a}_{21,t-1}\right)^\top\right]. \quad (34)$$

Additionally, we can define the overall correlation matrix of the RTF as

$$\mathbf{R}_{a_{21}} = E\left[\mathbf{a}_{21,t} \mathbf{a}_{21,t}^\top\right] = \boldsymbol{\Sigma}_{a_{21}} + \boldsymbol{\mu}_{a_{21}} \boldsymbol{\mu}_{a_{21}}^\top. \quad (35)$$

The previous statistics can be estimated in advance using a training set of dual-channel clean speech utterances in different acoustic environments and device positions. In order to avoid outliers which might yield useless estimates, we select, on an utterance basis, time-frequency bins where the speech power at the reference channel is large enough (higher than the maximum power at that frequency in the utterance minus 3 dB). For those bins, we estimate the RTF using (3), which yields a set of RTFs  $A_{21}^{(l)}(t, f)$  for the  $l$ -th utterance, that are later converted to RTF vectors  $\mathbf{a}_{21,t}^{(l)}$ . For each utterance, a sample mean vector  $\hat{\boldsymbol{\mu}}_{a_{21}}^{(l)}$  and a sample correlation matrix  $\hat{\mathbf{R}}_{a_{21}}^{(l)}$  are estimated using those  $\mathbf{a}_{21,t}^{(l)}$  vectors, while a sample covariance matrix  $\hat{\mathbf{Q}}^{(l)}$  is computed from consecutive RTF vectors. Finally, the global sample statistics  $\hat{\boldsymbol{\mu}}_{a_{21}}$ ,  $\hat{\mathbf{R}}_{a_{21}}$  and  $\hat{\mathbf{Q}}$  are obtained by averaging the particular utterance-dependent statistics. The sample covariance  $\hat{\boldsymbol{\Sigma}}_{a_{21}}$  can then be estimated using  $\hat{\boldsymbol{\mu}}_{a_{21}}$  and  $\hat{\mathbf{R}}_{a_{21}}$  in (35).

### 4. Speech Presence Probability-Based Noise Statistics Estimation

The eKF-RTF estimator and the beamforming algorithm also require knowledge of the noise spatial correlation matrix (SCM)  $\boldsymbol{\Phi}_N(t, f)$ . In this work, we follow the multi-channel speech presence probability (SPP) approach described in Reference [4]. This estimation method is based on the recursive updating of the noise SCM in those time-frequency bins where speech is absent. First, two hypotheses,  $H_x$  and  $H_n$ , are considered for speech presence and speech absence, respectively,

$$H_x : \mathbf{Y}(t, f) = \mathbf{X}(t, f) + \mathbf{N}(t, f), \quad (36)$$

$$H_n : \mathbf{Y}(t, f) = \mathbf{N}(t, f). \quad (37)$$



Assuming zero-mean random variables, the SCM of the noise will be updated by means of the following recursion,

$$\widehat{\Phi}_N(t, f) = \alpha(t, f)\widehat{\Phi}_N(t - 1, f) + (1 - \alpha(t, f))\mathbf{Y}(t, f)\mathbf{Y}^H(t, f), \tag{38}$$

where the forgetting factor is computed as

$$\alpha(t, f) = \tilde{\alpha} + p_x(t, f)(1 - \tilde{\alpha}), \tag{39}$$

where  $\tilde{\alpha} \in [0, 1)$  is an updating constant and

$$p_x(t, f) = P(H_x|\mathbf{Y}(t, f)) \tag{40}$$

is the *a posteriori* SPP. The estimation of this probability is detailed in Subsection 4.1.

From now on, the time and frequency indices are omitted whenever possible. The noise statistics for the RTF estimator (required by (30)) are directly derived from  $\Phi_N$ . Assuming a zero-mean, symmetric circular complex Gaussian distribution for  $\mathbf{N}$  [30] and

$$\Phi_N = \begin{bmatrix} \phi_{N11} & \phi_{N12} \\ \phi_{N21} & \phi_{N22} \end{bmatrix}, \tag{41}$$

with  $\phi_{Nij} = E[N_i N_j^*]$ , the following relations can be demonstrated,

$$\Sigma_{n11} = \frac{1}{2}\phi_{N11}\mathbf{I}_2, \tag{42}$$

$$\Sigma_{n22} = \frac{1}{2}\phi_{N22}\mathbf{I}_2, \tag{43}$$

$$\Sigma_{n12} = \frac{1}{2} \begin{bmatrix} \phi_{N12}^r & -\phi_{N12}^i \\ \phi_{N12}^i & \phi_{N12}^r \end{bmatrix}, \tag{44}$$

$$\Sigma_{n21} = \Sigma_{n12}^\top, \tag{45}$$

where  $\mathbf{I}_2$  is the 2-dimensional identity matrix.

#### 4.1. A Posteriori SPP Estimation

The *a posteriori* SPP  $p_x(t, f)$  allows us to control the updating procedure of Equation (38) for the computation of the noise statistics. Nevertheless,  $p_x(t, f)$  is also exploited in two additional parts of our system:

- The estimation of the RTF presented in the previous section is only accurate in time-frequency bins where speech is present. The *a posteriori* SPP indicates those bins where speech presence is more likely. Therefore, in our implementation we only update the eKF in those bins where  $p_x(t, f) > p_{\text{thr}}$ , with  $p_{\text{thr}}$  being a predefined probability threshold. Otherwise, the previous values are preserved.
- The postfiltering performance can be improved if additional information about SPP is provided, as shown later in Section 5.

The estimation of the *a posteriori* SPP can be addressed assuming complex multivariate Gaussian distributions for the noisy speech  $\mathbf{Y}$  [4,5], according to the two hypotheses previously formulated (see (36) and (37)). Using the Bayes' rule, the *a posteriori* SPP at each time-frequency bin can be calculated as

$$p_x = \frac{(1 - q_n) f(\mathbf{Y}|H_x)}{(1 - q_n) f(\mathbf{Y}|H_x) + q_n f(\mathbf{Y}|H_n)}, \tag{46}$$



where  $q_n = P(H_n)$  is the *a priori* speech absence probability (SAP) and

$$f(\mathbf{Y}|H_x) = \frac{e^{-\mathbf{Y}\Phi_Y^{-1}\mathbf{Y}^H}}{\pi^M \det[\Phi_Y]}, \tag{47}$$

$$f(\mathbf{Y}|H_n) = \frac{e^{-\mathbf{Y}\Phi_N^{-1}\mathbf{Y}^H}}{\pi^M \det[\Phi_N]}, \tag{48}$$

are the likelihoods of observing the noisy speech signal under the different hypotheses, with  $\det[\cdot]$  being the matrix determinant operator and  $M$  the number of microphones. Equation (46) can be redefined using these likelihoods, yielding the following expression for the SPP,

$$p_x = \left( 1 + \frac{q_n \det[\Phi_Y] e^{-\mathbf{Y}\Phi_N^{-1}\mathbf{Y}^H}}{1 - q_n \det[\Phi_N] e^{-\mathbf{Y}\Phi_Y^{-1}\mathbf{Y}^H}} \right)^{-1}. \tag{49}$$

Then, the *a posteriori* SPP can be estimated at each frame for all frequencies according to the following two-iteration algorithm:

- *Initialization:* Estimate the noisy SCM with a recursive updating,

$$\hat{\Phi}_Y(t, f) = \tilde{\alpha} \hat{\Phi}_Y(t - 1, f) + (1 - \tilde{\alpha}) \mathbf{Y}(t, f) \mathbf{Y}^H(t, f), \tag{50}$$

where  $\tilde{\alpha}$  is an updating constant as in (39). Also, estimate the *a priori* SAP  $q_n(t, f)$  (see Section 4.2).

- *1st iteration:* Estimate  $p_x(t, f)$  using  $\hat{\Phi}_N(t - 1, f)$  in (49). Then, estimate  $\hat{\Phi}_N(t, f)$  using  $p_x(t, f)$  in (38).
- *2nd iteration:* Re-estimate  $p_x(t, f)$  using now  $\hat{\Phi}_N(t, f)$  in (49). Finally, re-estimate  $\hat{\Phi}_N(t, f)$  using  $p_x(t, f)$ .

#### 4.2. A Priori SAP Estimation

To obtain an accurate *a posteriori* SPP that allows for robust tracking of the noise statistics, the *a priori* SAP is a key parameter. Methods on single-channel noise tracking, as the fixed *a priori* SNR algorithm of Reference [31] or the minima controlled recursive averaging (MCRA) algorithm [32], estimate this SAP in terms of the *a priori* SNR. The MCRA framework was extended to multi-channel speech signals in Reference [4]. This is the SPP estimator that we used in our previous works [15,27]. A major drawback of the MCRA scheme is the lack of robustness in case of a time-varying SNR, which makes noise changes to be detected as speech presence. More recently, the authors of [5] proposed to use spatial information in the SAP estimation, specifically the coherent-to-diffuse ratio (CDR). Alternatively, the power level difference (PLD) between microphones is used in Reference [17] to update the noise statistics, as it is a good indicator of speech presence in CT conditions. In this work, we propose a novel SAP estimator for dual-microphone smartphones that combines the CDR spatial information with the PLD between the reference and the secondary microphones.

First, an SCM  $\tilde{\Phi}_Y(t, f)$  is estimated using a rectangular window of eight past frames as in Reference [5]. Then, we calculate (1) the power spectral density (PSD) ratio between microphones as

$$\tilde{\sigma}_{Y_{21}}(t, f) = \frac{\tilde{\phi}_{Y_{22}}(t, f)}{\tilde{\phi}_{Y_{11}}(t, f)}, \tag{51}$$

where  $\tilde{\phi}_{Y_{ij}}$  is an estimate of  $E[Y_i Y_j^*]$ , and (2) the short-term complex coherence between microphones,

$$\hat{\psi}_{Y_{12}}(t, f) = \frac{\tilde{\phi}_{Y_{12}}(t, f)}{\sqrt{\tilde{\phi}_{Y_{11}}(t, f) \tilde{\phi}_{Y_{22}}(t, f)}}. \tag{52}$$

Both terms are needed by the SAP estimators.

The PLD-based *a priori* SAP is based on the PLD noise estimator for CT conditions described in Reference [17]. In that work, a normalized difference of the noisy speech PSDs was defined as

$$\Delta\hat{\phi}_{\text{nPLD}}(t, f) = \frac{\tilde{\phi}_{Y_{11}}(t, f) - \tilde{\phi}_{Y_{22}}(t, f)}{\tilde{\phi}_{Y_{11}}(t, f) + \tilde{\phi}_{Y_{22}}(t, f)} = \frac{1 - \hat{\sigma}_{Y_{21}}(t, f)}{1 + \hat{\sigma}_{Y_{21}}(t, f)}. \quad (53)$$

The noise statistics are updated by this parameter, as it gives information about speech presence. That is, assuming similar noise power at both microphones and that speech is more attenuated at the secondary microphone with respect to the primary one,  $\Delta\hat{\phi}_{\text{nPLD}}(t, f)$  is close to one when speech is present and tends to zero otherwise. Thus, an *a priori* SAP based on this indicator can be estimated as

$$q_{\text{PLD}}(t, f) = 1 - \Delta\hat{\phi}_{\text{nPLD}}(t, f) = \frac{2\hat{\sigma}_{Y_{21}}(t, f)}{1 + \hat{\sigma}_{Y_{21}}(t, f)}, \quad (54)$$

where  $q_{\text{PLD}}(t, f)$  is upper-bounded by 1. This estimator is valid for CT conditions, but it can also be useful in FT conditions for frequencies where speech at the secondary microphone is attenuated.

On the other hand, the CDR is another good indicator of speech presence [33], which is defined as

$$\Gamma_{Y_{12}}(t, f) = \frac{\psi_{\text{diff}}(f) - \psi_{Y_{12}}(t, f)}{\psi_{Y_{12}}(t, f) - \psi_{X_{12}}(t, f)}, \quad (55)$$

where  $\psi_{X_{12}}(t, f)$  is the clean speech short-term complex coherence (defined as in (52)) and

$$\psi_{\text{diff}}(f) = \text{sinc}\left(\frac{2\pi f F_s d_{\text{mic}}}{K c_v}\right) \quad (56)$$

is the diffuse noise field complex coherence, with  $F_s$  the sampling frequency,  $K$  the number of frequency bins,  $d_{\text{mic}}$  the distance between microphones and  $c_v$  the speed of sound. While higher values of the CDR indicate the presence of a strong coherent component, often a speech signal, lower values indicate that a diffuse component is dominant, which is more common for noise signals. In practice, the CDR is obtained using the estimator proposed in Reference [33],

$$\hat{\Gamma}_{Y_{12}}(t, f) = \Re\left(\frac{\psi_{\text{diff}}(f) - \hat{\psi}_{Y_{12}}(t, f)}{\hat{\psi}_{Y_{12}}(t, f) - e^{j\angle\hat{\phi}_{Y_{12}}(t, f)}}\right), \quad (57)$$

where  $\angle\hat{\phi}_{Y_{12}}(t, f)$  is the phase of  $\hat{\phi}_{Y_{12}}(t, f)$ . Although the CDR is positive and real-valued, estimation errors yield complex-valued results, so the real-part  $\Re(\cdot)$  is taken in (57). Additionally, a frequency-averaged CDR with a normalized Hamming window  $h_W$  is computed as

$$\bar{\Gamma}_{Y_{12}}(t, f) = \sum_{i=-w}^w h_W(i) \hat{\Gamma}_{Y_{12}}(t, f - i), \quad (58)$$

with  $w = 10$  (the window width is  $2w + 1$ ). Then, the local *a priori* SAP estimate is obtained as in Reference [5] using  $\hat{\Gamma}_{Y_{12}}(t, f)$ ,

$$q_{\text{local}}(t, f) = q_{\text{min}} + (q_{\text{max}} - q_{\text{min}}) \frac{10^{\frac{c\rho}{10}}}{10^{\frac{c\rho}{10}} + \hat{\Gamma}_{Y_{12}}^\rho(t, f)}, \quad (59)$$

where  $q_{\text{min}}$  and  $q_{\text{max}}$  define the minimum and maximum values of the function, respectively,  $c$  is an offset parameter and  $\rho$  controls the steepness of the transition region. Similarly, the global *a priori* SAP

estimate  $q_{\text{global}}(t, f)$  is computed using  $\bar{\Gamma}_{Y_{12}}(t, f)$  in (59). The CDR-based *a priori* SAP is then obtained as in Reference [5],

$$q_{\text{CDR}}(t, f) = 1 - (1 - q_{\text{local}}(t, f))(1 - q_{\text{global}}(t, f)). \tag{60}$$

This estimator is the same as that proposed in Reference [5], except for the term representing the frame *a priori* SAP. This term was neglected as it did not increase the performance in our preliminary experiments.

Finally, the *a priori* SAP estimates obtained by PLD and CDR are combined to achieve a more robust joint decision. Assuming statistical independence between both estimators, the final *a priori* SAP estimate for dual-microphone smartphones is calculated as the joint probability of the speech absence decision by each estimator,

$$q_n(t, f) = q_{\text{CDR}}(t, f)q_{\text{PLD}}(t, f). \tag{61}$$

The above estimator can be used in both CT and FT conditions to obtain prior information about speech presence. Moreover, as it only allows for noise statistics updating when both estimates indicate speech absence, this estimator is expected to be more robust.

### 5. Postfiltering Approaches for Dual-Microphone Smartphones

As introduced above, the performance achieved by beamforming algorithms is limited in our scenario mainly due to the reduced number of microphones and the particular placement on the smartphone. In this section we propose different alternatives for single-channel postfiltering at the beamformer output. These postfilters make use of the previously estimated parameters to design a gain function which is applied to the beamformer output signal for further noise reduction. In the next subsections we describe several postfiltering techniques as well as our proposals for the estimation of the clean speech and noise single-channel statistics required by the former ones.

#### 5.1. Parametric Wiener Filtering

Under the distortionless constraint of MVDR, if we assume an accurate estimation of the RTF, we can approximate the single-channel speech signal  $Z(t, f)$  at the beamformer output as

$$Z(t, f) \simeq X_1(t, f) + V(t, f), \tag{62}$$

where  $V(t, f)$  is the residual noise at the beamformer output,

$$V(t, f) = \mathbf{F}^H(t, f)\mathbf{N}(t, f). \tag{63}$$

The PSDs of the clean speech signal  $X_1(t, f)$  and the residual noise  $V(t, f)$  are given by  $\phi_{x_1}(t, f)$  and  $\phi_v(t, f)$ , respectively. The corresponding Wiener filter (WF) for  $X_1(t, f)$  is then defined as

$$G(t, f) = \frac{\xi(t, f)}{1 + \xi(t, f)}, \tag{64}$$

where

$$\xi(t, f) = \frac{\phi_{x_1}(t, f)}{\phi_v(t, f)} \tag{65}$$

is the *a priori* SNR.

The noise reduction performance of WF can be improved if we consider the *a posteriori* SPP  $p_x(t, f)$  in the postfiltering design. Our goal is to decrease the gain factor for time-frequency bins where speech is absent. This yields the parametric WF (pWF) [34]

$$G(t, f) = \frac{\xi(t, f)}{\mu(t, f) + \xi(t, f)}, \tag{66}$$

where  $\mu(t, f)$  is an SPP-driven trade-off parameter. To obtain this parameter, we use the same mapping function as in Reference [5],

$$\mu(t, f) = \mu_{min} + (\mu_{max} - \mu_{min}) \frac{10^{\frac{c_p}{10}}}{10^{\frac{c_p}{10}} + \left(\frac{p_x(t, f)}{1 - p_x(t, f)}\right)^\rho}, \tag{67}$$

which is similar to the mapping function of (59). We can see that  $\mu(t, f) \in [\mu_{min}, \mu_{max}]$  and therefore, the lower the *a posteriori* SPP, the higher  $\mu(t, f)$ , thus improving the noise reduction performance.

### 5.2. Optimally Modified Log-Spectral Amplitude Estimator

Instead of estimating a WF filter over the signal statistics, other approaches try to directly make an MMSE estimation of the clean speech signal amplitude  $|X_1(t, f)|$ . One of these is the log-spectral amplitude (LSA) estimator proposed in Reference [35] and defined as

$$\left| \widehat{X}_1(t, f) \right| = \exp(E[\log(|X_1(t, f)|) | Z(t, f)]) = G_{H_x}(t, f) |Z(t, f)|, \tag{68}$$

where

$$G_{H_x}(t, f) = \frac{\zeta(t, f)}{1 + \zeta(t, f)} \exp\left(\frac{1}{2} \int_{\frac{\zeta(t, f)}{1 + \zeta(t, f)}}^\infty \frac{e^{-u}}{u} \gamma(t, f) du\right) \tag{69}$$

is the LSA gain function, and

$$\gamma(t, f) = \frac{|Z(t, f)|^2}{\phi_v(t, f)} \tag{70}$$

is the *a posteriori* SNR. As can be seen, this estimator is equivalent to applying a gain function over the noisy speech signal  $Z(t, f)$  that depends not only on the *a priori* SNR but also on the *a posteriori* SNR. The obtained gain function is similar to WF for high *a priori* SNRs and has lower values than WF when the *a priori* SNR is lower and the *a posteriori* SNR is higher (i.e., when noise tends to dominate).

The estimator of (68) is only valid under the assumption of speech presence. To improve the performance, the authors of Reference [36] proposed a different estimator that also takes into account the *a posteriori* SPP and different gain functions for speech presence and absence. That is the optimally modified LSA (OMLSA) estimator, which is defined as

$$G(t, f) = G_{H_x}(t, f)^{p_x(t, f)} G_{H_v}(t, f)^{1 - p_x(t, f)}, \tag{71}$$

where  $G_{H_v}$  is a constant attenuation applied when speech is absent, whose value is usually  $-25$  dB [36].

### 5.3. Single-Channel Speech and Noise PSD Estimators

The computation of the gain functions previously presented relies on the estimation of the single-channel PSDs of the clean speech at the reference channel  $\phi_{x_1}(t, f)$  and the residual noise at the beamformer output  $\phi_v(t, f)$ . The estimation of the residual noise PSD is straightforward if estimates of the SCM  $\Phi_N(t, f)$  and the steering vector  $\mathbf{d}(t, f)$  are available. In this case, the residual noise PSD can be obtained at each time-frequency bin as

$$\widehat{\phi}_v = \left( \widehat{\mathbf{d}}^H \widehat{\Phi}_N^{-1} \widehat{\mathbf{d}} \right)^{-1}. \tag{72}$$

On the other hand, a clean speech PSD estimate is more difficult to obtain because of its higher variability. In Reference [27], we proposed two different estimators that make use of the noisy speech and noise statistics and the RTF between microphones. In the following, we describe both clean speech PSD estimators, omitting time and frequency indices for simplicity.

### 5.3.1. Power Level Difference-Based Estimation

The PLD-based estimator is derived from the method in Reference [17], which exploits the PLD between the microphones of a smartphone used in CT conditions; that is, a more attenuated clean speech component is expected at the secondary microphone with respect to the primary one. This PLD is defined in terms of the noisy speech PSDs at the microphone inputs,

$$\Delta\hat{\phi}_{\text{PLD}} = \max(\hat{\phi}_{Y_{11}} - \hat{\phi}_{Y_{22}}, 0), \tag{73}$$

where it is assumed that the power at the reference microphone is always higher than the one at the secondary microphone. Assuming that the noise PSD difference ( $\Delta\phi_N = \phi_{N_{11}} - \phi_{N_{22}}$ ) can be neglected when compared to  $\Delta\hat{\phi}_{\text{PLD}}$ , the clean speech PSD can be estimated as in [17],

$$\hat{\phi}_{x_1} = \frac{\Delta\hat{\phi}_{\text{PLD}}}{1 - |\hat{A}_{21}|^2}. \tag{74}$$

Although this estimator offers good performance in CT conditions, the previous assumptions are not longer valid in FT conditions.

### 5.3.2. Minimum Variance Distortionless Response-Based Estimation

This estimator calculates the PSD directly at the beamformer output by spectral subtraction, taking into account the distortionless property of MVDR. The estimator is defined as

$$\hat{\phi}_{x_1} = \mathbf{F}^H (\hat{\Phi}_Y - \hat{\Phi}_N) \mathbf{F}, \tag{75}$$

so that it fully exploits the spatial information by using the SCM matrices of the noisy speech and noise signals. The combination of both channels by means of the beamformer weights ( $\mathbf{F}$ ) allows for a more robust estimation than directly taking the first element of  $\hat{\Phi}_Y - \hat{\Phi}_N$ . Negative PSDs may be obtained due to the estimator variance, so (75) is bounded by 0.

## 6. Experimental Evaluation

The performance of the different estimators and speech enhancement algorithms discussed along this paper is evaluated by means of objective speech quality and intelligibility metrics. Two different well-known objective metrics are used:

- The Perceptual Evaluation Speech Quality (PESQ) [37] metric is utilized to evaluate the speech quality of the enhanced speech signal. This metric gives a mean opinion score between one and five. The higher the PESQ values, the better the speech quality.
- The Short-Time Objective Intelligibility (STOI) [38] metric is intended to evaluate the speech intelligibility of the enhanced speech signal. The resulting score is a value between zero and one. The higher the STOI value, the better the speech intelligibility.

PESQ and STOI are both intrusive metrics, which means that they need a clean reference. As a reference, we use the clean speech signal at the reference microphone,  $x_1(n)$ .

Additionally, in order to evaluate the RTF estimation accuracy, we use the speech distortion (SD) index [2]. This index measures the distortion level on the clean speech signal at the beamformer output, namely  $\tilde{x}_1(n)$  (inverse STFT of  $\mathbf{F}^H(t, f)\mathbf{X}(t, f)$ ). The idea behind using this metric is that, because of the distortionless property of MVDR, more accurate RTF estimates should yield lower speech distortion. The SD index is measured segmentally across the speech signal, in such a way that the SD value at the  $i$ -th segment is obtained as

$$SD(i) = \frac{\sum_{n=(i-1)T}^{iT-1} |x_1(n) - \tilde{x}_1(n)|^2}{\sum_{n=(i-1)T}^{iT-1} |x_1(n)|^2}, \quad (76)$$

where  $T$  is the number of samples per segment. The segmental SD values are averaged to achieve the final SD index, which is a value between zero and one. The lower the SD value, the lower the speech distortion. As in Reference [5], silence frames are excluded from this evaluation by means of calculating the median of the segment-wise signal power and removing those segments with a power 15 dB lower than that median.

We evaluate the proposed algorithms by using simulated dual-channel noisy speech recordings from a dual-microphone smartphone. Two different databases were developed for each device use mode: close-talk (CT) and far-talk (FT). To simulate the recordings, clean speech signals are filtered using dual-channel acoustic impulse responses and real dual-channel environmental noise is added at different signal-to-noise ratios (SNRs). We evaluate four different reverberation environments and eight different noises, which are matched as indicated in Table 1, yielding eight different acoustic environments (including both reverberation and noise). Details about the methodology used to obtain the acoustic impulse responses and the noises can be found in References [15,39].

**Table 1.** Predefined acoustic environments: each environment combines a reverberation environment with a given noise.

Reverberation	Noise (Test Only)
(A) No reverb.	Car, Street, Pedestrian street
(B) Low	Bus, Cafe
(C) Medium	Babble, Bus station
(D) High	Mall

Clean speech signals are obtained from the TIMIT database [40,41] and downsampled to 16 kHz. In particular, a total of 850 clean speech utterances from different speakers are employed. All the utterances have a length of around seven seconds, which is achieved by same-speaker utterance concatenation. Two different sets are then defined, namely training and test. Speakers do not overlap across sets. Moreover, the number of utterances from female and male speakers is balanced. The distribution of utterances and the number of speakers in the training and test sets are indicated in Table 2.

**Table 2.** Distribution of clean speech utterances and speakers across training and test sets.

	N° Utterances	N° Speakers
<b>Training set</b>	700	440
<b>Test set</b>	150	93

The training set consists of reverberated clean speech utterances and is only used to estimate the *a priori* statistics for the eKF-RTF estimator. Such utterances were obtained by filtering each clean speech utterance with a set of dual-channel acoustic impulse responses which model four reverberant environments, thereby yielding a total of 2800 training utterances. Sixteen different acoustic impulse responses were considered for each acoustic environment. For each utterance, the impulse response was randomly selected.

On the other hand, the test set consists of utterances contaminated according to the eight noisy environments defined in Table 1. This set is intended to evaluate the different algorithms proposed in this work. Noises were added to the reverberated speech at six different SNRs from  $-5$  dB to 20 dB, so a total of 7200 test utterances was obtained. In order to simulate reverberation, ten different

acoustic impulse responses, in turn also different from those of the training set, were considered for each acoustic environment and, again, randomly applied.

For STFT computation, a 512-point DFT was 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 noisy speech and noise SCMs were estimated using an updating constant  $\tilde{\alpha} = 0.9$ . The parameters for CDR-based *a priori* SAP calculation were set as [5]:  $d_{mic} = 13$  cm,  $q_{min} = 0.1$ ,  $q_{max} = 0.998$ ,  $c = 3$  and  $\rho = 2.5$ . The SPP threshold for RTF updating was set as  $p_{thr} = 0.9$ . Finally, for the SPP-driven trade-off parameter of the parametric Wiener filter, the following parameter values were used [5]:  $\mu_{min} = 1$ ,  $\mu_{max} = 4$ ,  $c = -3$  and  $\rho = 4$ .

The algorithm was implemented in Python and it is available at [42]. The computational burden of the implementation was evaluated on a PC with an Intel Core i7-4790 CPU. The algorithm works on a frame-by-frame basis, so that the algorithmic delay is in this case the duration of a frame, that is, 32 ms. The average performance of the whole system (i.e., including SPP and RTF estimation, MVDR beamforming and postfiltering) on this machine achieved approximately 8x faster than real-time.

In the following, the performance of the different techniques is tested. In order to simplify the reading of the results tables, an acronym is provided (in parentheses) for each considered technique after its description.

### 6.1. Experimental Results: Performance of SAP Estimators

First, we compare the different *a priori* SAP estimators when used along with our eKF-RTF estimator for MVDR beamforming (i.e., no postfilter is applied yet). This comparison is shown in Table 3 in terms of PESQ and STOI. The techniques evaluated are the multi-channel version of MCRA (MCRA) [4], CDR-based SAP estimation (CDR) [5] and our proposed PLD-based SAP estimator (PLDn) and its combination with CDR-based estimation (P&C), both presented in Section 4.2. Results for the noisy speech signal at the reference microphone (Noisy) are given as a baseline. In addition, we show the results achieved by the eKF-RTF estimator with an oracle estimation of the noise SCM (eKF-OracleN) as a performance upper-bound. This estimation was obtained using the true noise signals in a recursive procedure similar to Equation (50). The speech presence probability obtained from clean speech was used in the eKF-RTF estimator to obtain these oracle results.

For CT conditions, the best results are obtained for the eKF-PLDn system. The speech power difference between microphones for CT conditions makes that the PLD-based SAP estimator can easily detect those bins where speech is absent. This power difference reduces the CDR ratio, defined in Equation (57), of the multi-channel signal in the presence of speech, leading the CDR-based SAP method to underestimate the speech presence and decrease the performance of the noise tracking algorithm. Therefore, the combination of both approaches does not improve the single decision based on PLD between channels.

On the other hand, for FT conditions, speech power at both channels is more similar and CDR increases under the presence of speech. This is especially true at higher SNRs, where the CDR-based SAP detector outperforms the PLD-based one. However, the performance of the CDR-based detector degrades more severely at lower SNRs, while the PLD detector is more robust in these conditions. Finally, the combination of both detectors increases the performance in terms of both noise reduction and speech intelligibility, keeping a performance similar to the PLD one at lower SNRs and improving at higher SNRs.

To sum up, our proposals improve the tracking of the noise statistics in dual-microphone smartphones. The eKF-PLDn proposal is the best solution in CT conditions, with eKF-P&C having a similar performance. The joint decision proposed in eKF-P&C achieves the best results in FT conditions at higher SNRs, while eKF-PLDn performs slightly better at lower SNRs.



**Table 3.** Perceptual Evaluation Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) results for different speech absence probability (SAP) estimators when combined with speech presence probability (SPP)-based extended Kalman filter - relative transfer function (eKF-RTF) estimation for Minimum Variance Distortionless Response (MDVR) beamforming. Results are broken down by both signal-to-noise ratio (SNR) and device placement.

Place.	Method	PESQ						STOI (%)					
		SNR (dB)						SNR (dB)					
		20	15	10	5	0	−5	20	15	10	5	0	−5
CT	Noisy	2.26	1.80	1.46	1.23	1.11	1.07	95.36	91.01	83.82	74.05	62.48	51.46
	eKF-MCRA	2.28	1.84	1.49	1.26	1.13	1.08	95.58	91.91	85.37	75.57	63.65	52.20
	eKF-CDR	2.42	2.00	1.63	1.35	1.18	<b>1.12</b>	93.37	89.66	83.13	73.80	62.11	50.90
	eKF-PLDn	<b>2.60</b>	<b>2.09</b>	<b>1.67</b>	<b>1.38</b>	<b>1.20</b>	1.11	<b>96.99</b>	<b>93.77</b>	<b>87.96</b>	<b>79.12</b>	<b>67.56</b>	<b>55.61</b>
	eKF-P&C	2.59	2.07	1.66	1.37	1.19	1.11	96.90	93.59	87.60	78.56	66.85	54.84
	eKF-OracleN	2.76	2.21	1.76	1.44	1.23	1.12	97.76	95.18	90.26	82.35	71.43	59.49
FT	Noisy	2.38	1.89	1.51	1.26	1.11	1.07	94.65	89.91	82.52	72.69	61.09	50.09
	eKF-MCRA	2.35	1.90	1.52	1.27	1.13	1.07	94.47	90.48	83.71	73.65	61.11	49.69
	eKF-CDR	2.57	2.08	1.66	1.36	1.16	1.08	94.80	90.79	83.77	73.75	61.24	49.41
	eKF-PLDn	2.43	2.03	1.65	<b>1.37</b>	<b>1.19</b>	<b>1.10</b>	92.62	89.34	83.41	74.64	63.46	<b>52.20</b>
	eKF-P&C	<b>2.65</b>	<b>2.11</b>	<b>1.67</b>	<b>1.37</b>	1.18	1.09	<b>95.78</b>	<b>91.96</b>	<b>85.45</b>	<b>76.01</b>	<b>64.01</b>	52.03
	eKF-OracleN	2.99	2.41	1.88	1.51	1.26	1.13	97.25	94.68	89.88	82.29	71.64	59.85

6.2. Experimental Results: Performance of RTF Estimators

In this subsection, we compare our eKF-RTF estimator with the well-known eigenvalue decomposition (EVD) [11] and covariance whitening (CW) [9] sub-space methods for RTF estimation. The results are shown in Table 4. In addition, we show the results obtained with an oracle estimation of the RTF (OracleC) as a performance upper-bound. This oracle RTF was obtained from the clean speech signals using Equation (3) for time-frequency bins where speech presence was detected (using the speech presence probability obtained from clean speech), while the RTF of the previous frame was reused for the remaining ones. The evaluation is carried out in terms of speech distortion (SD) and the speech intelligibility metric STOI. We evaluate SD instead of PESQ because here we are only interested in the distortion introduced over the reference clean speech due to RTF estimation errors when using a distortionless beamformer.

**Table 4.** Speech distortion (SD) and STOI results for different RTF estimators when used for MVDR beamforming. Results are broken down by both SNR and device placement.

Place.	Method	SD (%)						STOI (%)					
		SNR (dB)						SNR (dB)					
		20	15	10	5	0	−5	20	15	10	5	0	−5
CT	EVD-PLDn	<b>0.52</b>	0.66	0.98	1.53	2.46	3.64	96.96	93.72	87.76	78.70	66.89	54.88
	CW-PLDn	<b>0.52</b>	0.63	0.92	1.41	2.27	3.43	<b>97.00</b>	<b>93.77</b>	87.87	78.89	67.10	55.08
	eKF-PLDn	<b>0.52</b>	<b>0.58</b>	<b>0.72</b>	<b>0.90</b>	<b>1.16</b>	<b>1.44</b>	96.99	<b>93.77</b>	<b>87.96</b>	<b>79.12</b>	<b>67.56</b>	<b>55.61</b>
	OracleC-PLDn	0.07	0.11	0.16	0.22	0.28	0.34	97.33	94.32	88.72	80.05	68.57	56.60
FT	EVD-P&C	3.64	3.56	4.03	5.12	7.11	10.04	95.56	91.84	85.38	75.80	63.69	51.69
	CW-P&C	3.96	3.79	4.19	5.22	7.15	10.01	95.54	91.88	<b>85.46</b>	75.94	63.85	51.78
	eKF-P&C	<b>2.09</b>	<b>2.63</b>	<b>3.32</b>	<b>4.23</b>	<b>5.57</b>	<b>7.49</b>	<b>95.78</b>	<b>91.96</b>	<b>85.45</b>	<b>76.01</b>	<b>64.01</b>	<b>52.03</b>
	OracleC-P&C	0.24	0.45	0.81	1.26	1.82	2.43	97.05	93.89	88.23	79.66	68.18	56.17

The comparison is performed using the best SAP estimator obtained for each device placement according to Table 3 (eKF-PLDn for CT and eKF-P&C for FT). A similar improvement due to our proposed SAP estimators, as the one observed in Table 3 for the eKF-RTF estimator, is expected for the other RTF estimators (due to the fact that these estimators would also take advantage of more accurate estimates of the noise SCM). We compare the RTF estimators using only one SAP estimator in order

to narrow the number of possible system combinations. For a fair comparison, we use the same RTF initialization and the same updating scheme based on SPP for the different systems.

From Table 4, we can see that our eKF proposal obtains slightly better results in terms of STOI and much lower speech distortion than the other approaches, particularly in FT conditions. As we previously analyzed in a former study [15], the distortionless property of MVDR involves none or very low speech distortion if an accurate estimate of the RTF is available. Therefore, we can conclude that our estimator tracks better the RTF variability than the other approaches. This is more noticeable in FT conditions, where the reverberation level increases and the secondary microphone captures similar power to the reference one, thereby making the RTF more variable and its tracking more challenging.

Furthermore, Table 5 shows the SD results for the different approaches grouped by reverberant environment (averaging by noise environment and SNR level). The oracle results for each reverberant environment are also shown, as they give an upper-bound reference of the performance of the different approaches. This can be useful when different reverberant environments are compared, as some of them include more challenging noise environments (e.g., cafe in low reverberation). It can be observed that the increase of the reverberation level makes the RTF estimation more difficult. The performance of the different algorithms degrades with the reverberation level, as the variations of the RTF are harder to track, although our proposal is more robust against reverberant environments than sub-space approaches.

**Table 5.** SD results for different RTF estimators when used for MVDR beamforming. Results are broken down by both reverberation environment and device placement. The noise environments are grouped in terms of the reverberant environment as in Table 1: A (Car, Street, Pedestrian street), B (Bus, Cafe), C (Bus station, Babble) and D (Mall).

Place.	Method	SD (%)			
		Environment			
		A	B	C	D
CT	EVD-PLDn	1.16	1.97	1.73	2.17
	CW-PLDn	1.08	1.87	1.59	2.08
	eKF-PLDn	<b>0.59</b>	<b>1.24</b>	<b>0.87</b>	<b>1.09</b>
	OracleC-PLDn	0.03	0.39	0.17	0.36
FT	EVD-P&C	3.62	7.29	5.54	8.13
	CW-P&C	3.62	7.78	5.55	8.26
	eKF-P&C	<b>2.92</b>	<b>5.12</b>	<b>4.32</b>	<b>6.13</b>
	OracleC-P&C	0.45	1.80	1.06	2.30

### 6.3. Experimental Results: Performance of Single-Channel Clean Speech PSD Estimators

Next, we evaluate different clean speech PSD estimators when combined with Wiener postfiltering (Equation (64)) applied to the MVDR beamformer output. Table 6 shows a comparison between the estimator based on the power difference (WF-Ps) of Equation (74), the one based on the distortionless constraint of MVDR (WF-Ms) of Equation (75) and a standalone MVDR beamformer (i.e., with no postfiltering) as a reference, all of them with the best SAP configurations determined above. This comparison is done in terms of speech quality and intelligibility.

Results show that WF-Ms performs better than WF-Ps, obtaining slightly better results in CT conditions and clearly outperforming it in FT conditions. The WF-Ms estimation does not make any assumptions about the noise power similarity between microphones as the other approach does. This makes the WF-Ms estimation procedure, which also exploits the cross-correlation elements of the noisy speech and noise SCMs, more robust. Moreover, the power difference assumption is no longer valid in FT conditions, leading the WF-Ps approach to degrade the performance in this scenario, particularly at higher SNRs.

**Table 6.** PESQ and STOI results for different clean speech power spectral density (PSD) estimators when combined with Wiener postfiltering applied to the MVDR beamformer output. Results are broken down by SNR and device placement.

Place.	Method	PESQ						STOI (%)					
		SNR (dB)						SNR (dB)					
		20	15	10	5	0	−5	20	15	10	5	0	−5
CT	eKF-PLDn	2.60	2.09	1.67	1.38	1.20	1.11	96.99	93.77	87.96	79.12	67.56	<b>55.61</b>
	WF-Ps-eKF-PLDn	2.79	2.32	1.91	1.56	1.31	1.16	97.17	94.18	88.66	80.02	67.95	54.88
	WF-Ms-eKF-PLDn	<b>2.81</b>	<b>2.34</b>	<b>1.95</b>	<b>1.62</b>	<b>1.36</b>	<b>1.20</b>	<b>97.18</b>	<b>94.20</b>	<b>88.72</b>	<b>80.11</b>	<b>67.97</b>	54.51
FT	eKF-P&C	2.65	2.11	1.67	1.37	1.18	1.09	95.78	91.96	85.45	76.01	64.01	<b>52.03</b>
	WF-Ps-eKF-P&C	2.59	2.22	1.85	1.54	1.31	1.17	92.59	89.56	83.75	74.60	62.70	50.25
	WF-Ms-eKF-P&C	<b>2.95</b>	<b>2.45</b>	<b>1.99</b>	<b>1.64</b>	<b>1.36</b>	<b>1.21</b>	<b>96.10</b>	<b>92.64</b>	<b>86.51</b>	<b>76.91</b>	<b>64.20</b>	50.80

6.4. Experimental Results: Performance of Postfiltering Approaches

Finally, in Table 7 we compare the pWF and OMLSA postfilters with a basic WF postfilter and two other related state-of-the-art dual-channel speech enhancement algorithms intended for smartphones: the PLD-based single-channel WF filter of Reference [17] (PLDwf) and the SPP- and coherence-based single-channel WF filter of Reference [18] (SPPCwf) for CT and FT conditions, respectively. The MVDR-based clean speech PSD estimator (Ms) is used for the different proposed postfilters (WF, pWF and OMLSA). The comparison is done in terms of PESQ and STOI metrics.

**Table 7.** PESQ and STOI results for different postfilters applied to the MVDR beamformer output and for other related state-of-the-art approaches. Results are broken down by SNR and device placement.

Place.	Method	PESQ						STOI (%)					
		SNR (dB)						SNR (dB)					
		20	15	10	5	0	−5	20	15	10	5	0	−5
CT	PLDwf	2.81	2.38	1.98	1.64	1.36	1.20	95.94	92.21	85.70	76.11	63.53	50.42
	WF-Ms-eKF-PLDn	2.81	2.34	1.95	1.62	1.36	1.20	97.18	94.20	88.72	80.11	67.97	54.51
	pWF-Ms-eKF-PLDn	2.86	2.40	2.00	1.64	1.37	1.20	97.24	94.35	89.00	80.53	68.41	54.54
	OMLSA-Ms-eKF-PLDn	<b>2.96</b>	<b>2.49</b>	<b>2.06</b>	<b>1.68</b>	<b>1.40</b>	<b>1.23</b>	<b>97.25</b>	<b>94.45</b>	<b>89.24</b>	<b>80.86</b>	<b>68.72</b>	<b>55.11</b>
FT	SPPCwf	2.74	2.26	1.81	1.48	1.25	1.12	94.43	90.26	83.27	73.28	61.16	49.34
	WF-Ms-eKF-P&C	2.95	2.45	1.99	<b>1.64</b>	<b>1.36</b>	<b>1.21</b>	96.10	92.64	86.51	76.91	<b>64.20</b>	<b>50.80</b>
	pWF-Ms-eKF-P&C	<b>2.99</b>	<b>2.49</b>	<b>2.01</b>	1.63	<b>1.36</b>	<b>1.21</b>	<b>96.12</b>	<b>92.73</b>	<b>86.68</b>	<b>77.08</b>	64.18	50.36
	OMLSA-Ms-eKF-P&C	2.85	2.38	1.94	1.60	1.35	1.20	95.98	92.64	86.63	77.03	64.13	50.54

For CT conditions, both pWF and OMLSA outperform the WF and PLDwf approaches, with OMLSA achieving the best results in noise reduction performance and speech intelligibility. PLDwf achieves more noise reduction (better PESQ) than our WF due to the fact that the former introduces an overestimation of the noise. Such an overestimation also means more speech distortion, so the speech intelligibility is lower compared to our WF approach. The use of SPP information in our postfilters allows for larger noise reduction in frequency bins where speech is absent without additional speech distortion. The availability of accurate SPPs in CT conditions makes OMLSA the best approach in this case, as the LSA estimator performs better than the WF when bins where speech is present are clearly differentiated from those where speech is absent.

On the other hand, the SPPs obtained in FT conditions are less accurate for the postfiltering task, thereby degrading the performance of the OMLSA estimator. Nevertheless, this SPP information is still useful for the pWF, which, in general, outperforms the basic WF, especially at higher SNRs. Moreover, both WF and pWF outperform SPPCwf in terms of noise reduction and speech intelligibility. In summary, the availability of accurate RTF and SPP estimates, as those provided by our proposal, clearly helps to improve the performance of postfiltering in FT mode.

## 7. Conclusions

In this paper we have proposed a dual-channel speech enhancement algorithm intended for dual-microphone smartphones. Our proposal brings the use of a novel speech presence probability calculation method for the estimation of noise statistics needed by our extended Kalman filter RTF estimator, allowing for online processing of the noisy speech signal. The noise and acoustic channel information is used for speech enhancement through MVDR beamforming. To improve the noise reduction performance with low speech distortion, we also proposed several postfiltering techniques which make use of both single-channel statistics estimated at the beamformer output and information about speech presence. Moreover, our system can take advantage of information about the user position in relation to the smartphone (available through the smartphone's sensors). The experimental results indicate that our proposal achieves more accurate estimates of RTFs and SPPs than other related state-of-the-art algorithms, which yields low speech distortion and better speech quality and intelligibility. Furthermore, the proposed postfilters improve the noise reduction performance compared to other algorithms specifically intended for dual-microphone smartphones without degrading the speech intelligibility.

Regarding future work, we think that the large margin of improvement revealed by the oracle results, with respect to the estimation of the noise statistics and the acoustic channel, is a boost to this research. In particular, we will generalize our extended Kalman filter framework to general multi-microphone devices and also investigate the estimation of the speech presence probability and the development of postfiltering techniques for multi-channel speech enhancement.

**Author Contributions:** Conceptualization, J.M.M.-D., A.M.P. and I.L.-E.; Methodology, J.M.M.-D. and A.M.P.; Software, J.M.M.-D.; Validation, J.M.M.-D., A.M.P. and A.G.; Formal analysis, J.M.M.-D., A.M.P. and I.L.-E.; Investigation, J.M.M.-D. and A.M.P.; Resources, A.G.; Data curation, J.M.M.-D. and A.G.; Writing—original draft preparation, J.M.M.-D.; Writing—review and editing, A.M.P., I.L.-E. and A.G.; Visualization, J.M.M.-D.; Supervision, A.M.P. and A.G.; Project administration, A.M.P. and A.G.; Funding acquisition, A.G.

**Funding:** This research was funded by the Spanish MINECO/FEDER Project TEC2016-80141-P and the Spanish Ministry of Education through the National Program FPU under Grant FPU15/04161.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Parchami, M.; Zhu, W.P.; Champagne, B.; Plourde, E. Recent developments in speech enhancement in the short-time Fourier transform domain. *IEEE Circuits Syst. Mag.* **2016**, *16*, 45–77. [[CrossRef](#)]
2. Benesty, J.; Chen, J.; Huang, Y. *Microphone Array Signal Processing*; Springer: Berlin, Germany, 2008; Volume 1.
3. Kumatani, K.; McDonough, J.; Raj, B. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *IEEE Signal Process. Mag.* **2012**, *29*, 127–140. [[CrossRef](#)]
4. Souden, M.; Benesty, J.; Affes, S.; Chen, J. An Integrated Solution for Online Multichannel Noise Tracking and Reduction. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2159–2169. [[CrossRef](#)]
5. Taseska, M.; Habets, E. Nonstationary noise PSD matrix estimation for multichannel blind speech extraction. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2223–2236.
6. Gannot, S.; Vincent, E.; Markovich-Golan, S.; Ozerov, A. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 692–730. [[CrossRef](#)]
7. Cohen, I. Relative transfer function identification using speech signals. *IEEE Trans. Speech Audio Process.* **2004**, *12*, 451–459. [[CrossRef](#)]
8. Markovich, S.; Gannot, S.; Cohen, I. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 1071–1086. [[CrossRef](#)]
9. Markovich-Golan, S.; Gannot, S. Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, Australia, 19–24 April 2015; pp. 544–548.

10. Serizel, R.; Moonen, M.; Van Dijk, B.; Wouters, J. Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants. *IEEE Trans. Audio Speech Lang. Process.* **2014**, *22*, 785–799. [[CrossRef](#)]
11. Varzandeh, R.; Taseska, M.; Habets, E.A.P. An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation. In Proceedings of the 2017 Hands-Free Speech Communications and Microphone Arrays (HSCMA), San Francisco, CA, USA, 1–3 March 2017; pp. 11–15.
12. Koldovský, Z.; Malek, J.; Gannot, S. Spatial source subtraction based on incomplete measurements of relative transfer function. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 1335–1347. [[CrossRef](#)]
13. Schmid, D.; Malik, S.; Enzner, G. An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 17–20.
14. Schwartz, B.; Gannot, S.; Habets, E.A.P. Online Speech Dereverberation Using Kalman Filter and EM Algorithm. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 394–406. [[CrossRef](#)]
15. Martín-Doñas, J.M.; López-Espejo, I.; Gomez, A.M.; Peinado, A.M. An Extended Kalman Filter for RTF Estimation in Dual-Microphone Smartphones. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 2488–2492.
16. Tashev, I.; Mihov, S.; Gleghorn, T.; Acero, A. Sound capture system and spatial filter for small devices. In Proceedings of the Interspeech, Brisbane, Australia, 22–26 September 2008; pp. 435–438.
17. Jeub, M.; Herglotz, C.; Nelke, C.; Beaugeant, C.; Vary, P. Noise reduction for dual-microphone mobile phones exploiting power level differences. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 1693–1696.
18. Nelke, C.M.; Beaugeant, C.; Vary, P. Dual microphone noise PSD estimation for mobile phones in hands-free position exploiting the coherence and speech presence probability. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7279–7283.
19. Jin, W.; Taghizadeh, M.J.; Chen, K.; Xiao, W. Multi-channel noise reduction for hands-free voice communication on mobile phones. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 506–510.
20. Zelinski, R. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In Proceedings of the ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing, New York, NY, USA, 11–14 April 1988; pp. 2578–2581.
21. Marro, C.; Mahieux, Y.; Simmer, K.U. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Trans. Speech Audio Process.* **1998**, *6*, 240–259. [[CrossRef](#)]
22. McCowan, I.A.; Bourslard, H. Microphone Array Post-Filter Based on Noise Field Coherence. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 709–716. [[CrossRef](#)]
23. Lefkimmatis, S.; Maragos, P. A generalized estimation approach for linear and nonlinear microphone array post-filters. *Speech Commun.* **2007**, *49*, 657–666. [[CrossRef](#)]
24. Gannot, S.; Cohen, I. Speech enhancement based on the general transfer function GSC and postfiltering. *IEEE Trans. Speech Audio Process.* **2004**, *12*, 561–571. [[CrossRef](#)]
25. Habets, E.; Gannot, S.; Cohen, I. Dual-microphone speech dereverberation in a noisy environment. In Proceedings of the 2006 IEEE International Symposium on Signal Processing and Information Technology, Vancouver, BC, Canada, 27–30 August 2006; pp. 651–655.
26. Zheng, C.; Liu, H.; Peng, R.; Li, X. A statistical analysis of two-channel post-filter estimators in isotropic noise fields. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 336–342. [[CrossRef](#)]
27. Martín-Doñas, J.M.; López-Espejo, I.; Gomez, A.M.; Peinado, A.M. A postfiltering approach for dual-microphone smartphones. In Proceedings of the IberSpeech, Barcelona, Spain, 21–23 November 2018; pp. 142–146.
28. Esch, T.; Vary, P. Efficient musical noise suppression for speech enhancement systems. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 4409–4412.
29. Julier, S.; Uhlmann, J. A new extension of the Kalman filter to nonlinear systems. In Proceedings of the SPIE, Orlando, FL, USA, 28 July 1997; pp. 182–193.



30. Ducharme, G.R.; Lafaye de Micheaux, P.; Marchina, B. The complex multinormal distribution, quadratic forms in complex random vectors and an omnibus goodness-of-fit test for the complex normal distribution. *Ann. Inst. Stat. Math.* **2016**, *68*, 77–104. [CrossRef]
31. Gerkmann, T.; Breithaupt, C.; Martin, R. Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 910–919. [CrossRef]
32. Cohen, I. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 466–475. [CrossRef]
33. Schwarz, A.; Kellermann, W. Coherent-to-Diffuse Power Ratio Estimation for Dereverberation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 1006–1018. [CrossRef]
34. Doclo, S.; Spriet, A.; Wouters, J.; Moonen, M. Speech Distortion Weighted Multichannel Wiener Filtering Techniques for Noise Reduction. In *Speech Enhancement*; Springer: Berlin, Germany, 2005; pp. 199–228.
35. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error-log-spectral amplitude estimator. *IEEE Trans. Acoustics Speech Signal Process.* **1985**, *33*, 443–445. [CrossRef]
36. Cohen, I.; Berdugo, B. Speech enhancement for non-stationary noise environments. *Signal Process.* **2001**, *81*, 2403–2418. [CrossRef]
37. 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; International Telecommunication Union: Geneva, Switzerland, 2007.
38. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]
39. López-Espejo, I.; Peinado, A.M.; Gomez, A.M.; González, J.A. Dual-channel spectral weighting for robust speech recognition in mobile devices. *Digit. Signal Process.* **2018**, *75*, 13–24. [CrossRef]
40. Garofolo, J. *Getting Started With the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*; NIST Tech. Rep.; National Institute of Standards and Technology (NIST): Gaithersburgh, MD, USA, 1988.
41. Lamel, L.; Kassel, R.; Seneff, S. Speech database development: Design and analysis of the acoustic-phonetic corpus. In Proceedings of the DARPA Speech Recognition Workshop, Noordwijkerhout, The Netherlands, 20–23 September 1989; pp. 2161–2170.
42. Martín-Doñas, J.M.; Peinado, A.M.; López-Espejo, I.; Gomez, A.M. Dual-Channel Postfiltering and eKF-RTF Estimation: Source Code and Audio Examples. 2019. Available online: <http://sigmat.ugr.es/dc-ekf-rtf> (accessed on 1 May 2019).



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).