Modulation-Domain Kalman Filtering for Monaural Blind Speech Denoising and Dereverberation

N. Dionelis, https://www.commsp.ee.ic.ac.uk/~sap/people-nikolaos-dionelis/ M. Brookes, Member, IEEE

Abstract—We describe a monaural speech enhancement algorithm based on modulation-domain Kalman filtering to blindly track the time-frequency log-magnitude spectra of speech and reverberation. We propose an adaptive algorithm that performs blind joint denoising and dereverberation, while accounting for the inter-frame speech dynamics, by estimating the posterior distribution of the speech log-magnitude spectrum given the log-magnitude spectrum of the noisy reverberant speech. The Kalman filter update step models the non-linear relations between the speech, noise and reverberation log-spectra. The Kalman filtering algorithm uses a signal model that takes into account the reverberation parameters of the reverberation time, $T_{60}$, and the direct-to-reverberant energy ratio (DRR) and also estimates and tracks the $T_{60}$ and the DRR in every frequency bin in order to improve the estimation of the speech log-magnitude spectrum. The Kalman filtering algorithm is tested and graphs that depict the estimated reverberation features over time are examined. The proposed algorithm is evaluated in terms of speech quality, speech intelligibility and dereverberation performance for a range of reverberation parameters and SNRs, in different noise types, and is also compared to competing denoising and dereverberation techniques. Experimental results using noisy reverberant speech demonstrate the effectiveness of the enhancement algorithm.

Index Terms—Speech enhancement, dereverberation, Kalman filter, minimum mean-square error (MMSE) estimator.

I. INTRODUCTION

NOWADAYS, technology is ever evolving with tremendous haste and the demand for speech enhancement systems is evident. Speech enhancement in noisy reverberant environments, for human listeners, is challenging. Speech is degraded by noise and reverberation when captured using a near-field or far-field distant microphone [1] [2]. A room impulse response (RIR) can include components at long delays, hence resulting in reverberation and echoes [3] [4]. Reverberation is a convolutive distortion that can be quite long with a reverberation time, $T_{60}$, of more than 0.8 s. Due to convolution, reverberation induces long-term correlation between consecutive observations. Reverberation and noise, which can be stationary or non-stationary, have a detrimental impact on speech quality and intelligibility. Reverberation, especially in the presence of non-stationary noise, damages the intelligibility of speech.

The direct to reverberant energy ratio (DRR) and the reverberation time, $T_{60}$, are the two main parameters of a reverberation model [5] [1]. The DRR describes reverberation in the space domain, depending on the positions of the sound source and the receiver. The $T_{60}$ is the time interval required for a sound level to decay 60 dB after ceasing its original stimulus. The reverberation time, when measured in the diffuse sound field, is independent of the source to microphone configuration and mainly depends on the room. The impact of reverberation on auditory perception depends on the $T_{60}$. If the $T_{60}$ is short, the environment reinforces the sound which may enhance the sound perception [6] [7]. On the contrary, if the $T_{60}$ is long, spoken syllables interfere with future spoken syllables. Reverberation spreads energy over time and this smearing across time has two effects: (a) the energy of individual phonemes spreads out in time and, hence, plosives have a delayed decay and fricatives are smoothed, and (b) preceding phonemes blur into the current phonemes.

The aim of speech enhancement is to reduce and ideally eliminate the effects of both noise and reverberation without distorting the speech signal [8]. Enhancement algorithms typically aim to suppress noise and late reverberation because early reverberation is not perceived as separate sound sources and usually improves the quality and intelligibility of speech. Noise is assumed to be uncorrelated with speech, early reverberation is correlated with speech and late reverberation is commonly assumed to be uncorrelated with speech [9] [6].

Speech enhancement can be performed in different domains. The ideal domain should be chosen such that (a) good statistical models of speech and noise exist in this domain, and (b) speech and noise are separable in this domain. Speech and noise are additive in the time domain and the Short Time Fourier Transform (STFT) domain [10] [11]. The relation between speech and noise becomes progressively more complicated in the amplitude, power and log-power spectral domains. Noise suppression algorithms usually operate in a time-frequency STFT domain and these techniques have been extended to address dereverberation. In [9], spectral enhancement methods based on a time-frequency gain, originally developed for noise suppression, have been modified and employed for dereverberation. Such algorithms suppress late reverberation assuming that the early and late reverberation components are uncorrelated. The spectral enhancement methods in [12] [9] estimate the late reverberant spectral variance (LRSV) and use it in the place of the noise spectral variance, reducing the problem of late reverberation suppression to that of estimating the LRSV [9]. In [7], blind spectral weighting is employed to reduce the overlap-masking effect of reverberation using an uncorrelated and additive assumption for late reverberation.

Dereverberation algorithms that leave the phase unaltered and operate in the amplitude, power or log-power spectral domains are relatively insensitive to minor variations in the spatial placement of sources [13]. Two criticisms of
spectral enhancement algorithms based on LRSV reverberation noise estimation are that they introduce musical noise and suppress speech onsets when they over-estimate reverberation [14]. The LRSV estimator in [15], which is a continuation of [6], models the RIR in the STFT domain and not in the time domain [6] [16], using the same model of the RIR that is attributed to J. Polack or J. Moorer [16]. Reverberation is estimated in [15] considering the STFT energy contribution of the direct path of speech and an external $T_{60}$ estimate.

Modelling the speech temporal dynamics is beneficial when the $T_{60}$ is long and the DRR is low [17] [2]. Joint denoising and dereverberation using speech and noise tracking is performed in [17]. The SPENDRED algorithm [18] [19], which is a model-based method with a convolution model for reverberation based on the $T_{60}$ and the DRR, considers the speech temporal dynamics. SPENDRED employs a parametric model of the RIR [20] and performs frequency-dependent and time-varying $T_{60}$ and DRR estimation. However, unless the source or the microphone are moving, the $T_{60}$ and the DRR will be constant throughout the recording. The SPENDRED algorithm assumes that DRR $\geq 1 - (10^{-6})\frac{L}{T_{60}}$ where $L$ is the acoustic frame increment. For example, when $T_{60} = 0.4$ s and $L = 5$ ms, then DRR $\geq -8$ dB is assumed. In addition, SPENDRED performs intra-frame correlation modelling, which can be beneficial in adverse conditions, while typical algorithms decouple different frequency dimensions [10].

Statistical-based models, such as the SPENDRED algorithm, describe reverberation by a convolution in the power spectral domain while LRSV models describe reverberation as an additive distortion in the power spectral domain [20] [21]. A model with an infinite impulse response is used either with the two parameters of the $T_{60}$ and the DRR, as in [18] [19], or with a finite number of parameters. The infinite-order convolution model of reverberation with the $T_{60}$ and the DRR is sparse and contrasts with the higher-order autoregressive processes in the complex STFT domain, used in [22] [23].

The algorithms described in [24], [20] and [21] create non-linear observation models of noisy reverberant speech in the log Mel-power spectral domain, using the reverberation-to-noise ratio (RNR). As discussed in [25], phase differences in Mel-frequency bands have different properties from phase differences in STFT bins. The phase factor between reverberant speech and noise is different from that between speech and noise [24]. In [20], the phase factor between reverberant speech and noise in Mel-frequency bands is examined.

In noisy reverberant conditions, finding the onset of speech phonemes and determining which frames are unvoiced/silence is difficult, due to the smearing across time, often leading to noise over-estimation. The concatenation of different techniques for denoising and dereverberation has lower performance than unified methods due to over-estimating noise when estimating noise and reverberation separately [17] [26].

Despite the claim that it is inefficient to perform a two step procedure that is comprised of denoising followed by dereverberation [17] [25], long-term linear prediction with pre-denosing can be used to suppress noise and reverberation. With the weighted prediction error (WPE) algorithm [27] [2], reverberation is represented as a one-dimensional convolution in each frequency bin. In [28], the WPE algorithm is discussed along with inter-frame correlation. In [29], the WPE algorithm is used in the complex STFT domain performing batch processing and iteratively estimating, first, the reverberation prediction coefficients and, then, the speech spectral variance. The WPE linear filtering approach, which can be employed in the power spectral domain [30] [31], takes into account past frames, from the 3-rd to the 40-th past frame [29] [30].

This paper presents an adaptive denoising and dereverberation Kalman filtering framework that tracks the speech and reverberation spectral log-magnitudes. In this paper, we extend the enhancer in [32] to include dereverberation. Enhancement is performed using a Kalman filter (KF) to model inter-frame correlations. We use an integrated structure of two parallel signal models to track speech, reverberation and the $T_{60}$ and DRR reverberation parameters. The $T_{60}$ and the DRR are updated in every frame to improve the estimation of the speech log-magnitude spectrum. We create an observation model and a series of non-linear KF update steps performing joint noise and reverberation suppression by estimating the first two moments of the posterior distribution of the speech log-spectrum given the noisy reverberant log-spectrum. The log-spectral domain is chosen, as in [33] [32], because good speech models exist in this domain. Modelling spectral log-amplitudes as Gaussian distributions leads to good speech modelling in noisy reverberant environments since super-Gaussian distributions that resemble the log-normal, such as the Gamma [11] [34], are used to model the speech amplitude spectrum. Mean squared errors (MSEs) in the log-spectral domain are a good measure to use for perceptual quality and speech log-spectra are well modelled by Gaussian distributions, as in [35] [36].

The structure of this paper is as follows. Section II describes the signal model and Sec. III presents the enhancement algorithm and its non-linear KF. The implementation and the validation of the algorithm are in Sec. IV. The algorithm’s evaluation is in Sec. V. Conclusions are drawn in Sec. VI.

II. SIGNAL MODEL AND NOTATION

In the complex STFT domain, the noisy speech, $Y_t(k)$, is given by $Y_t(k) = S_t(k) + R_t(k) + N_t(k)$ where $S_t(k)$ is the direct speech component, $R_t(k)$ is the reverberant speech component and $N_t(k)$ is the noise, as for example in [22] [23]. The time-frame index is $t$ and the frequency bin index is $k$. For clarity, we also define $Z_t(k) = R_t(k) + N_t(k)$. We drop the time and frequency indexes and we obtain $Y = S + Z = S + R + N$. We define the log-magnitude spectrum of $S$ as $s = \log(|S|)$ and we also define $r$, $n$, $y$ and $z$ similarly.

In the signal model, signal quantities with capital letters, such as $S_t$, are complex numbers with magnitude and phase values, $|S_t|$ and $\angle S_t$. In the complex STFT domain, using $a_t, b_t \in \mathbb{R}$, the reverberation signal model is given by

$$R_t = \sqrt{a_t} R_{t-1} \exp(j\theta_t) + \sqrt{b_t} S_{t-1} \exp(j\psi_t)$$

$$= \sum_{\tau=1}^{\infty} \prod_{i=1}^{\tau-1} (\sqrt{a_{t-i+1}} \exp(j\theta_{t-i+1})) \times \sqrt{b_{t-\tau+1}} S_{t-\tau} \exp(j\psi_{t-\tau+1})$$

This model is chosen, as in [33] [32], because good speech models exist in this domain. The log-spectral domain is chosen, as in [33] [32], because good speech models exist in this domain. Modelling spectral log-amplitudes as Gaussian distributions leads to good speech modelling in noisy reverberant environments since super-Gaussian distributions that resemble the log-normal, such as the Gamma [11] [34], are used to model the speech amplitude spectrum. Mean squared errors (MSEs) in the log-spectral domain are a good measure to use for perceptual quality and speech log-spectra are well modelled by Gaussian distributions, as in [35] [36].

The structure of this paper is as follows. Section II describes the signal model and Sec. III presents the enhancement algorithm and its non-linear KF. The implementation and the validation of the algorithm are in Sec. IV. The algorithm’s evaluation is in Sec. V. Conclusions are drawn in Sec. VI.
The expression in (1) is the convolution model for reverberation; the most common reverberation model is this single-pole filter that is described by the pole and zero positions that depend on the \( T_{60} \) and the DRR [18] [19]. Models that describe reverberation by a convolution are also discussed in and the DRR, to describe reverberation [18] [19]. Models that describe reverberation by a convolution are also discussed in [20] [21]. The signal model is defined by (1) and by

\[
Y_t = S_t + Z_t, \quad Z_t = R_t + N_t, \tag{3}
\]

\[
\gamma_t = 0.5 \log (a_t), \quad \delta_t = \gamma_t + r_{t-1} = \log (\sqrt{a_t} |R_{t-1}|), \tag{4}
\]

\[
\beta_t = 0.5 \log (b_t), \quad \epsilon_t = \beta_t + s_{t-1} = \log (\sqrt{b_t} |S_{t-1}|) \tag{5}
\]

where \( b_t > 0, 0 < a_t < 1 \) and \( \gamma_t < 0 \).

Figure 1 shows graphs of \( \beta \) against \( T_{60} \) for a fixed DRR and of \( \beta \) against DRR for a fixed \( T_{60} \). If DRR = 0 dB, then \( b_t = 1 - a_t \). If \( \beta_t = 0 \), then \( b_t = 1 \) and DRR = 1 - \( a_t \).

Figure 2 illustrates the flowchart of the signal model. The reverberation signal model in (1) uses \( \sqrt{a_t} \) and \( \sqrt{b_t} \) because the \( a_t \) and \( b_t \) reverberation parameters, in (2), and the DRR are defined in the power spectral domain, as in [18] [19]. The \( a_t \) and \( b_t \) parameters are mapped to \( \gamma_t \) and \( \beta_t \) using (4) and (5). The signals \( z_t, \delta_t \) and \( \epsilon_t \) are the total disturbance, the old (decaying) reverberation and the new reverberation, respectively. We note that \( z_t \) is defined in the first paragraph of this section and that \( \delta_t \) and \( \epsilon_t \) are defined in (4) and (5).

The signal model of how the reverberation parameters of \( \gamma_t \) and \( \beta_t \) change over time is a random walk model. This is used in the algorithm’s KF prediction step for \( \gamma_t \) and \( \beta_t \).

The signal model in Fig. 2 is directly linked to the alternating and interacting KFs of the enhancement algorithm. The algorithm is a collection of two KFs, the speech KF and the reverberation KF, that estimate the speech and reverberation log-amplitude spectra and the \( \gamma_t \) and \( \beta_t \) reverberation parameters. This KF algorithm is described in detail in Sec. III.

III. THE SPEECH ENHANCEMENT ALGORITHM

The KF algorithm operates in the log-magnitude spectral domain, tracking speech and reverberation. Figure 3 depicts the denoising and dereverberation algorithm that formulates a model of reverberation as a first-order autoregressive process and propagates the means and variances of the time variables. Almost all the signals follow a Gaussian distribution and the distribution of \( s_t \) conditioned on observations up to time \( \tau \) is given by \( \mathcal{N}(s_{\tau}, \Sigma_{\tau}) \). In Fig. 3 a Gaussian distribution is denoted by its mean, \( s_{\tau} \).

The core of the algorithm in Fig. 3 is the KF that is defined by the gray blocks in the flowchart diagram. The non-linear KF estimates and tracks the posterior distributions of the speech log-magnitude spectrum, \( s_t \), the reverberation log-magnitude spectrum, \( r_t \), and the reverberation parameters, \( \gamma_t \) and \( \beta_t \).

The input to the algorithm in Fig. 3 is the noisy reverberant speech in the time domain. The algorithm’s first step is to perform a STFT and obtain the signal in the complex STFT domain. The algorithm does not alter the noisy reverberant speech, \( Y_t \), and uses the noisy reverberant amplitude spectrum, \( |Y_t| \), in three ways: in the speech KF prediction step, in the KF update step and in the noise power modelling. The main part of the algorithm is the KF and the speech KF state, \( s_t \in \mathbb{R}^p \), is the speech log-spectrum from the previous \( p \) frames,

\[
s_t = (s_t \ s_{t-1} \ldots \ s_{t-p+1})^T. \tag{6}
\]

The speech KF prediction step is based on autoregressive (AR) modelling on the log-spectrum of pre-cleaned speech [33]. The reverberation KF state is \( r_t \) and the KF states of the reverberation parameters are \( \gamma_t \) and \( \beta_t \). The KF observation is the noisy reverberant speech log-spectrum, \( y_t \), which is used in the KF update step to compute the first two moments of the posterior of the speech log-spectrum. The mean of the speech log-spectrum posterior is used together with \( \angle Y \) to create the enhanced speech signal using the inverse STFT (ISTFT).

Apart from the speech log-spectrum, the non-linear KF also tracks the reverberation log-spectrum, \( r_t \), and the \( \gamma_t \) and \( \beta_t \) reverberation parameters. The KF, as defined by the gray blocks in Fig. 3, has a speech KF prediction step, a reverberation KF prediction step and a series of KF update steps. The reverberation KF is comprised of the blocks “Reverberation KF prediction”, “KF Update” and “\( \gamma_t, \beta_t \) KF Update”. These three blocks perform joint denoising and dereverberation and estimate \( s_{\tau t} \) and \( \Sigma_{\tau t} \) to enhance noisy reverberant speech.

The structure of the rest of this algorithm description section is as follows. Sections III.A and III.B present the speech and reverberation KF prediction steps, respectively. Section III.C describes the KF update step and Sec. III.D the priors for the \( \gamma_t \) and \( \beta_t \) parameters that are needed so that the KF (a) distinguishes between speech and reverberation, and (b) does
not diverge to non-realistic $T_{60}$ and DRR estimates. Section III.E describes the unshaded peripheral blocks in Fig. 3.

A. The Speech KF Prediction Step

The speech KF prediction step is linear and is related to the “Speech KF prediction”, “Decorrelate” and “Recorrelate” blocks in Fig. 3. The speech KF prediction step is described in [33] and in [32] [35] and is based on conditional distributions to model short-term dependencies. The decorrelation and recorrelation of the speech KF state in [35] are performed after and before the speech KF prediction step, respectively. The decorrelation and recorrelation operations in Fig. 3 which are performed so that the non-linear KF update step can be applied, perform vector-matrix and matrix-matrix multiplications for the speech KF state mean and its covariance matrix, respectively, using $B_t \in \mathbb{R}^{P \times P}$ [34] [33]. The outputs of the “Decorrelate” block are: (a) the first element of the speech KF state, and (b) the rest elements of the speech KF state.

The KF prediction step propagates the first and second moments of the speech KF state [33] [35]. Inter-frame linear relationships are used for the speech KF prediction step that uses AR modelling in the log-magnitude spectral domain. In the speech KF prediction step, $s_{t-1}$ is predicted as a linear combination of $s_{t-1}$ using the speech AR coefficients that are obtained from the “Speech AR(p)” block in Fig. 3 which uses pre-cleaned speech as an input. After the speech KF prediction step, $s_{t-1}$ is correlated; we decorrelate the speech KF state with a linear transformation (using $B_t$) to simplify the KF update step and impose the observation constraint [10]. The KF update step changes only the first element of the speech KF state and after the KF update step, recorrelation is applied with a linear transformation (using $B_{t-1}$) to continue the KF recursion.

B. The Reverberation KF Prediction Step

The presented algorithm uses a KF prediction step for $\gamma_t$ and $\beta_t$ that assumes that the variance of $\gamma_t$ and $\beta_t$ increases over time, preserving their mean. The KF algorithm implements a random-walk prediction step, performing the operations of

$$\gamma_{t|t-1} = \gamma_{t-1|t-1} + \Sigma_{\gamma|t-1}$$  \hspace{1cm} (7)  $$\beta_{t|t-1} = \beta_{t-1|t-1} + \Sigma_{\beta|t-1}$$  \hspace{1cm} (8)

where $Q_\gamma$ is a fixed error variance for $\gamma$ and $Q_\beta$ is a fixed error variance for $\beta$. The values used for the prediction error variances, $Q_\gamma$ and $Q_\beta$, depend on the rate at which the $T_{60}$ and the DRR are likely to change in a real situation.

After the reverberation KF prediction step, the algorithm computes and imposes priors on $\gamma_t$ and $\beta_t$ using Gaussian-Gaussian multiplication. The internally computed priors for $\gamma_t$ and $\beta_t$ in the “$\gamma_t$, $\beta_t$ priors” block in Fig. 3 are explained in Sec. III.D. After imposing the priors, the outputs are $\gamma_{t|t-1}$ and $\beta_{t|t-1}$. We note that a prime diacritic, $t$, is used in (7) and (8) to denote quantities before the priors.

The “Reverberation KF prediction” block in Fig. 3 estimates the first two moments of the prior distribution of the reverberation spectral log-amplitude, i.e. $r_{t|t-1}$ and its variance. The algorithm performs a reverberation KF prediction step based on the previous posterior of both speech and reverberation using the signal model in (1), where $\alpha$ is less than unity and this makes the reverberation KF prediction step stable.

From (1) and Fig. 2, the STFT-domain reverberation is the sum of two components arising, respectively, from the reverberation and speech components of the previous frame. The old reverberation, $\delta_t$, and the new reverberation, $\epsilon_t$, are defined in (4) and (5), respectively. The KF algorithm calculates the prior distributions of these two components in the log-amplitude spectral domain using $\delta_{t|t-1} = \gamma_{t|t-1} + \epsilon_{t-1|t-1}$ and $\epsilon_{t|t-1} = \beta_{t|t-1} + \gamma_{t-1|t-1}$. These equations are based on (4) and (5) with a common condition added to all terms. Assuming that $r_{t-1}$ and $s_{t-1}$ are uncorrelated with $\gamma_t$ and $\beta_t$, respectively, the means and variances of the two Gaussian distributions are added. The variances therefore add,
that the phase difference, \( \eta \), between the two disturbance sources, \( \delta \) and \( \varepsilon \), is uniformly distributed, i.e. \( \eta \sim U(-\pi, \pi) \), and independent of their magnitudes. We write \( e^{2r} = e^{2\delta} + e^{2\varepsilon} + 2\cos(\eta)e^{\delta+\varepsilon} \) and \( r = 0.5\log(e^{2\delta} + e^{2\varepsilon} + 2\cos(\eta)e^{\delta+\varepsilon}) \), which takes account of \( \eta \).

Next, we calculate

\[
\mathbb{E}\{ r \} = \int_{\eta=-\pi}^{\pi} p(\eta) \int_{\delta,\varepsilon} r p(\delta, \varepsilon) d\delta d\varepsilon d\eta
\]

\[
= \int_{\eta=0}^{\pi} \frac{1}{\pi} \int_{\delta,\varepsilon} 0.5\log(e^{2\delta} + e^{2\varepsilon} + 2\cos(\eta)e^{\delta+\varepsilon})
\]

\[
\times p(\delta) p(\varepsilon) d\delta d\varepsilon d\eta,
\]

\[
\mathbb{E}\{ r^2 \} = \int_{\eta=0}^{\pi} \frac{1}{\pi} \int_{\delta,\varepsilon} (0.5\log(e^{2\delta} + e^{2\varepsilon} + 2\cos(\eta)e^{\delta+\varepsilon}))^2 \times p(\delta) p(\varepsilon) d\delta d\varepsilon d\eta.
\]

where \( K_{(\delta,\varepsilon)} \) sigma points are used to evaluate the inner integral over \((\delta, \varepsilon)\) and \( K_\eta \) the outer integral over \( \eta \).

Equations (11) and (12) estimate the first two moments of the prior distribution of reverberation, \( r_{t|-1} \). In this algorithm description section, we provide expressions for either the second moment or the variance. We convert implicitly between them using, for example, \( \Sigma_{\delta,\varepsilon}^{(1)} = \mathbb{E}\{ r_t^2 \} - (\mathbb{E}\{ r_t \})^2 \) for \( r_t \).

### C. The Non-Linear KF Update Step

The KF algorithm decomposes the noisy reverberant observation, \( y_t \), into its component parts using distributions in the log-magnitude spectral domain. The decompositions are based on Fig. 2 and the signal model in (4) and (1). The KF algorithm performs a series of low-dimensional operations instead of a high-dimensional one in the KF update step. The adaptive KF algorithm propagates backwards through the non-linear KF update step, instead of a high-dimensional one in the KF update step. The non-linear dereverberation KF update step computes the first two moments of the posterior distributions for most signal quantities and, moreover, includes the prediction step as well for some quantities. Both means and variances are computed for the tracked Gaussian signals; for clarity, the variances, such as \( \Sigma_{\delta,\varepsilon}^{(1)} \) for \( r_t \), are not included in Table I.

#### Table I

| Inputs: | (a) \( s_{t|-1} \) from the speech KF prediction step, (b) \( n_{t|-1} \) from external noise estimation, (c) \( y_t \) from observation, and (d) \( s_{t|-1} \) from \( s_{t|t} \) in step 7 and \( B_t^{-1} \) from Sec. III.A. |
|---|---|
| 1: | \( \gamma_{t|-1} \) ← \( \gamma_{t|-1} \) from step 13 |
| 2: | \( \beta_{t|-1} \) ← \( \beta_{t|-1} \) from step 13 |
| 3: | \( \delta_{t|-1} \) ← \( \gamma_{t|-1} \), \( r_{t|-1} \), \( s_{t|-1} \) from steps 1, 13 |
| 4: | \( \epsilon_{t|-1} \) ← \( \beta_{t|-1} \), \( s_{t|-1} \), \( r_{t|-1} \) from steps 2, 13 |
| 5: | \( r_{t|-1} \) ← \( \delta_{t|-1} \), \( \epsilon_{t|-1} \) from steps 3, 4 |
| 6: | \( z_{t|-1} \) ← \( r_{t|-1} \), \( n_{t|-1} \) from step 5 and input (b) |
| 7: | \( s_{t|t} \) ← \( s_{t|-1} \), \( z_{t|-1} \), \( y_t \) from 6 and inputs (a), (c) |
| 8: | \( y_{t|t} \) ← \( s_{t|-1} \), \( n_{t|-1} \) from step 7 and input (b) |
| 9: | \( e_{t|t} \) ← \( \beta_{t|-1} \), \( s_{t|-1} \), \( r_{t|-1} \) from step 2 and input (d) |
| 10: | \( \delta_{t|t}, \epsilon_{t|t} \) ← \( \delta_{t|-1}, \epsilon_{t|-1} \) from steps 3, 8, 9 |
| 11: | \( \gamma_{t|t} \) ← \( \gamma_{t|-1} \), \( r_{t|t-1} \), \( s_{t|t} \) from steps 10, 13 |
| 12: | \( \beta_{t|t}, \beta_{t|t} \) ← \( \beta_{t|-1}, \beta_{t|-1}, \epsilon_{t|-1} \) from steps 2, 10 and input (d) |
| 13: | \( r_{t|-1} \), \( \gamma_{t|t-1} \), \( \beta_{t|t-1} \), \( \delta_{t|t-1} \) ← \( r_{t|t-1} \), \( \gamma_{t|t}, \beta_{t|t} \) |

For clarity, from this point onwards in Sec. III.C, the time subscript is included only if it differs from \( t|t-1 \). Table I shows the time subscripts. We also denote \( n_{t|-1} \) by \( n_{t|-1} \).

The KF algorithm computes the total disturbance, \( z_t \), from (4) in step 6 using similar equations to step 5. A two-dimensional Gaussian distribution is used for \( p(r, n) \) and independence is assumed between \( r \) and \( n \). Hence, \( p(r, n) = p(r)p(n) \). The phase-sensitive KF algorithm assumes that the phase difference, \( \zeta \), between the two disturbance sources, \( r \) and \( n \), is uniformly distributed and independent of their magnitudes. From (3), we write \( e^{2r} = e^{2r} + e^{2n} + 2\cos(\zeta)e^{r+n} \) and thus \( z = 0.5\log(e^{2r} + e^{2n} + 2\cos(\zeta)e^{r+n}) \). Next, using \( p(\zeta) = \frac{1}{2\pi} \), the first two moments of \( z \) are given by

\[
\mathbb{E}\{ z \} = \int_{\zeta=0}^{\pi} \frac{1}{2\pi} \int_{r,n} 0.5\log(e^{2r} + e^{2n} + 2\cos(\zeta)e^{r+n})
\]

\[
\times p(r)p(n) d\zeta d\zeta,
\]

\[
\mathbb{E}\{ z^2 \} = \int_{\zeta=0}^{\pi} \frac{1}{2\pi} \int_{r,n} (0.5\log(e^{2r} + e^{2n} + 2\cos(\zeta)e^{r+n})^2 \times p(r)p(n) d\zeta d\zeta.
\]

where \( K_{(r,n)} \) sigma points are used to evaluate the inner integral over \((r, n)\) and \( K_\zeta \) the outer integral over \( \zeta \).

The KF algorithm performs noise suppression with steps 6 and 7. Step 7 decomposes \( y_t \) into \( s_t \) and \( z_t \), as shown in the signal model in Fig. 2, estimating both \( s_{t|t} \) and \( z_{t|t} \).

Step 7 performs the first signal decomposition, \( y_t \) into \( s_t \) and \( z_t \), when propagating backwards through the signal model in...
Fig. 2. Step 7 applies the observation constraint, $y_t$, according to \( y_t \). As in [32, 35], the variables are first transformed according to \((s,\lambda) \rightarrow (u,\sqrt{\lambda})\) where \( u = z - s \) and \( \lambda = \angle Z - \angle Z \). This variable transformation is performed to allow the imposition of the scalar KF observation, $y_t$.

The noisy reverberant log-amplitude spectrum, $y$, is given by

\[
y(t) = s(t) + 0.5 \log(1 + \exp(2(z - s))) + 2 \cos(\lambda) \exp(z - s).
\]

The KF update step assumes that $\angle Z$ is uniformly distributed, $\angle Z \sim U(-\pi, \pi)$. Therefore, $\lambda \sim U(-\pi, \pi)$. The first two moments of the posteriors of $s_t$ and $z_t$ are computed using

\[
E \{ s_{m_1}^{m_2} \mid y_0, y_1, \ldots, y_T \} = \int_{u=-\infty}^{\pi} \int_{\lambda=-\infty}^{\pi} s_{m_1}^{m_2} p(u, \lambda | y_t) du d\lambda
\]

\[
\times \frac{1}{\Delta} \int_{u=-\infty}^{\pi} \lambda \int_{\lambda=-\infty}^{\pi} s_{m_1}^{m_2} p(s, z) du d\lambda
\]

\[
\times \int_{u=0}^{\infty} \int_{u=\infty}^{\pi} s_{m_1}^{m_2} p(s) du d\lambda
\]

where the Jacobian determinant is $\Delta = -1$ and the moment indexes, $m_1$ and $m_2$, are integers, $0 \leq m_1, m_2 \leq 2$. We denote the variables for two moment indexes by $m_1$ and $m_2$.

The first two moments of the posterior distributions of $s_t$ and $z_t$ are estimated. In [15], the priors of the speech and of the noise and late reverberation are assumed to be independent, i.e.

\[
p(s, z) = p(s)p(z).
\]

In addition, in [15], $K_\lambda$ weighted sigma points are used to evaluate the outer integral over $\lambda$ [33].

Step 8 performs the second signal decomposition, \( z_t \) into \( r_t \) and \( n_t \), when propagating backwards through the proposed signal model in Fig. 2. Step 8 decomposes $z_{t-1}$ into $r_t$ and $n_t$, according to \( \delta \), estimating both $r_t$ and $n_t$. Step 8 performs an integral over $z_t$ where the integrand is similar to step 7 and [15] and to the KF update step in [32, 35]. Instead of a scalar observation, as in step 7, the observation in step 8 is a distribution; step 8 performs an outer integral over the observation distribution and the integrand is similar to step 7. Step 8 uses $E \{ r^m \} = \sum_{z} \int \{ r^m | z \} dz$ where $m \in N_0$ and $0 \leq m \leq 2$. Step 7 computes a two-dimensional integral over the variables of $(z - s)$ and of the phase difference between $S$ and $Z$, using $y_t$ that reduces the probability space from three to two dimensions. In step 8, the KF algorithm calculates a three-dimensional integral over: (a) \((r - n)\), (b) the phase difference between $R$ and $N$, i.e. $\epsilon = \angle Z - \angle N$, and (c) the posterior of $z_t$. Assuming $\square N \sim U(-\pi, \pi)$, step 8 computes

\[
E \{ r_{m_1}^{m_2} \mid y_0, y_1, \ldots, y_T \} \propto \int_{z_1=-\infty}^{\infty} p_{z_1 \mid t}(z_1)
\]

\[
\times \int_{u=0}^{\infty} \int_{u=\infty}^{\pi} r_{m_1}^{m_2} p(r) p(n) du \, dr \, dz_t
\]

where $u' = (r - n)$ and $\epsilon \sim U(-\pi, \pi)$, In [16], $K$ weighted sigma points are used to evaluate the integral over $\epsilon$ and $K_{z_1 \mid t}$ sigma points are used to evaluate the integral over $z_t$ [37].

In Table II the signals in square brackets are calculated but are not used in the KF recursion. In step 8, the posterior of $n_t$ is estimated using [16] but is not used in the KF recursion.

Step 9 determines a preliminary estimate of the posterior distribution of new reverberation component, $\epsilon_t$ in [5]. This preliminary estimate, denoted $\epsilon_{t,0}$, combines an updated estimate of the previous frame’s speech, $s_{t-1}$ with the prior estimate of the reverberation parameter, $\beta_t$. In step 9, two random variables in the log-amplitude spectral domain are added; the means and variances of the two Gaussian variables are: $\epsilon_{t} = \beta_{t-1} + s_{t-1}$ and $\Sigma_{t}^{\epsilon} = \Sigma_{t-1}^{s} + \Sigma_{t-1}^{\beta}$. In this addition, $\beta_t$ and $s_{t-1}$ are assumed to be independent.

Step 10 decomposes the reverberation, $r_t$, into a new reverberation component, $\epsilon_t$ and an old decaying reverberation component, $\delta_t$, using (1), (4) and (5). Step 10 uses the prior distributions for these two components from steps 3 and 9. Step 10 performs the same operation as step 8. In analogy to step 7, the variable transformations in steps 8 and 10 are \((r, n, t) \rightarrow (u', z)\) and \((\delta, \epsilon, \xi) \rightarrow (u', \epsilon, \xi)\), respectively. In step 10, the KF algorithm performs an integral over $\tau$ where the integrand is similar to the KF update step in [32]. Step 10 estimates the posterior distributions of $\delta$ and $\epsilon$; it estimates both $\delta_{t|t}$ and $\epsilon_{t|t}$ using $\delta_{t|t-1}$ and $\epsilon_{t|t-1}$. Step 10 computes

\[
E \{ \delta_{t}^{m_1} \epsilon_{t}^{m_2} \mid y_0, y_1, \ldots, y_T \} \propto \int_{r_t=\infty}^{\epsilon_{t}} \int_{\epsilon_{t}=\infty}^{\delta_{t}} \int_{\xi=\infty}^{\epsilon_{t}} \delta_{t}^{m_1} \epsilon_{t}^{m_2} p(\delta) p_{\epsilon_{t}}(\epsilon) d\delta' d\epsilon' d\xi
\]

where $u' = (\delta - \epsilon)$ and $\xi \sim U(-\pi, \pi)$ is the phase difference of the STFT-domain signals of $\delta$ and $\epsilon$. Using [17], we decompose $r_t$ into old reverberation and new reverberation, estimating $\delta_{t|t}$ and $\epsilon_{t|t}$. In [17], $K_\tau$ sigma points are used to evaluate the integral over $\xi$ and $K_{\epsilon_{t|t}}$ the integral over $r_t$.

Steps 10-12 perform the final signal decomposition, $r_t$ into $\gamma_t$ and $\beta_t$, when propagating backwards through the signal model in Fig. 2. In steps 11 and 12, the KF algorithm computes the first two moments of the posterior distributions of $\gamma_t$ and $\beta_t$. In step 11, the algorithm performs an integral over $\delta_t$ using weighted sigma points where the integrand models the addition of two random variables in the log-amplitude spectral domain, $\gamma_t$ and $r_{t-1}$. Likewise, in step 12, the algorithm performs an integral over $\epsilon_t$ using sigma points where the integrand models the addition of two random variables in the log-amplitude spectral domain, $\beta_t$ and $s_{t-1}$. Steps 11 and 12 perform a linear KF update and impose a straight line observation constraint because the signal model is that $\gamma_t$ and $r_{t-1}$ are additive and that $\beta_t$ and $s_{t-1}$ are additive, according to (3) and (6).

In step 11, the KF algorithm decomposes the old reverberation component, $\delta_t$, according to \( \delta_t \) into the sum of a reverberation parameter, $\gamma_t$, and the previous frame’s reverberation, $r_{t-1}$. We define $x_1 \sim N(x_1; \mu_1, S_1)$ where $x_1 = (\gamma_t, r_{t-1}, w_t)^T \in \mathbb{R}^3$, $\mu_t$ is zero-mean Gaussian with variance $\Sigma_1$, $\mu_1 = (\gamma_t, r_{t-1}, w_{t-1}, \theta_t)^T \in \mathbb{R}^3$ and $\theta_t = diag((\gamma_t, r_{t-1}, w_{t-1}, \theta_t)^T) \in \mathbb{R}^{4 \times 4}$ where $diag(J)$ is a diagonal matrix with the elements of $J$ on the main diagonal of the matrix. If $x_1 \sim N(x; \mu_1, S_1)$, then

\[
(x_1 | g) = \delta_{t|t} \sim N(x_1; (I - H_{1}g^T)\mu_1 + H_1 \delta_{t|t},
\]

\[
(1 - H_{1}g^T)S_1
\]

where $H_1 = S_1 g (g^T S_1 g)^{-1}$ and $g = (1, 1, -1)^T \in \mathbb{R}^3$. We compute $(I - H_{1}g^T)^{-1} \mu_1 + H_1 \delta_{t|t}$,
and $\Sigma_{t+1|t}$ equal to the first element of this variance. Likewise, for step 12, we define $x_{t+1} \sim N(x_{t+1}; \mu_2, S_2)$ where $x_{t+1} = (\beta_t, s_{t-1}, q_t)^T \in \mathbb{R}^3$, $q_t$ is zero-mean Gaussian with variance $\Sigma_{t}^{(q)}$, $\mu_2 = (\beta_{t-1}, s_{t-1}, 0)^T \in \mathbb{R}^3$ and, moreover, $S_2 = \text{diag}((\Sigma_{t+1|t}, \Sigma_{t+1|t}, \Sigma_{t+1|t})^T) \in \mathbb{R}^{3x3}$. Next, we use

$$(x_{t+1})^T H_{t+1} = (I - H_{t+1} g_{t+1}^T) \mu_2 + H_{t+1} \epsilon_{t+1},$$

where $H_{t+1} = S_2 g_{t+1}(t S_2 g_{t+1})^{-1}$. The KF algorithm computes $(I - H_{t+1} g_{t+1}^T) \mu_2 + H_{t+1} \epsilon_{t+1} \in \mathbb{R}^3$ and $(I - H_{t+1} g_{t+1}^T) S_{2} \in \mathbb{R}^{3x3}$ and sets $\beta_{t+1}$ equal to the first element of this computed mean and $\Sigma_{t+1|t}$ equal to the first element of this computed variance.

In steps 11 and 12, the presented KF algorithm computes $\gamma_{t|t}$, $\Sigma_{t|t}$ and $\beta_{t|t}$, $\Sigma_{t|t}$, respectively. Steps 11 and 12 use $p(t_r, r_{t-1}|t_{t-1})$ and $p(\beta, s_{t-1}|t_{t-1})$, respectively.

Step 13 applies a one-frame delay, i.e., $z^{-1}$, to continue the KF recursion as shown in Fig. 3. In summary, the 13 steps have the four main operations of: (a) step 5, (b) step 7, (c) step 8, and (d) step 11. The operations performed in the other steps are either simple or identical to these operations.

According to the 13 steps in Table 1 the $\gamma_t$ and $\beta_t$ reverberation parameters are affected by the observed $y_t$ through a series of operations that estimate the first two moments of the posterior distributions. The estimate of $\gamma_t$ is affected by the previous estimate of $\beta_t$ because of the sequence of operations: $\beta_{t-1|t-1} \Rightarrow \beta_{t|t-1} \Rightarrow \delta_{t|t} \Rightarrow \gamma_{t|t}$. The estimate of $\beta_t$ depends on the previous estimate of $\gamma_t$ because of the sequence of operations: $\gamma_{t-1|t-1} \Rightarrow \beta_{t-1|t-1} \Rightarrow \delta_{t|t} \Rightarrow \beta_{t|t}$. The proposed 13 steps do not include the speech KF prediction step, which is shown in Fig. 3, that is used to calculate $s_{t-1|t}$ that is needed in steps 9 and 12. After step 7, according to Fig. 3 and Sec. III.B, re-correlation of the speech KF state is performed with $B_t^{-1}$. Using $\delta_{t|t}$, from the re-correlation operation, $s_{t-1|t}$ is obtained. In steps 9 and 12, $s_{t-1|t}$ is used as a better estimate of $s_{t-1}$ than $s_{t-1|t-1}$.

In step 9, we note that two sub-indexes of $t|t-1$ and $t-1|t$ give rise to a sub-index of $t|t$ if $\epsilon_{t|t} = \epsilon_{t|t-1} + \epsilon_{t-1|t}$. In step 9, we introduce the notation $\epsilon_{t|t}$ to avoid using the same symbol for two different posterior distributions, $\epsilon_t$ and $\epsilon_t$.}

D. The Priors for the Reverberation Parameters

This section describes the priors for the $\gamma_t$ and $\beta_t$ reverberation parameters, which are based on [39] [40] and [41]. The priors for $\gamma_t$ and $\beta_t$ are imposed using Gaussian-Gaussian multiplication: $\gamma_t$ is modelled with a Gaussian distribution and its internal prior is also a Gaussian. Likewise, $\beta_t$ is modelled with a Gaussian and its internal prior is also a Gaussian.

The priors for $\gamma_t$ and $\beta_t$ are estimated from spectral log-amplitude observations in the free decay region (FDR), which is comprised of $M_0$ consecutive frames with decreasing $L$. We define the look-ahead factor $C$ and the frame index $l = t - M_t + C + 1, t - M_t + C + 2, ..., t + C$. The least squares (LS) fit to the FDR is found using $r_l = \theta_x l + \theta_y$ where $x_l$ is the time index in seconds and depends on $L$. The parameters of the straight line are the slope, $\theta_1$, and the y-intercept, $\theta_2$. For clarity, the frame subscript $l$ is omitted from $\theta_1$ and $\theta_2$. We define $\theta = (\theta_1, \theta_2)^T \in \mathbb{R}^2$. Its Gaussian distribution is

$$N(\theta; \mu_\theta, \Sigma_\theta) \sim |\Sigma_\theta|^{-0.5} \times \exp\left(-0.5(\theta - \mu_\theta)^T \Sigma_\theta^{-1}(\theta - \mu_\theta)\right)$$

where $\mu_\theta \in \mathbb{R}^2$ and $\Sigma_\theta \in \mathbb{R}^{2x2}$. The log-likelihood of $\theta$ is given by $l(\theta) = c_0 - 0.5(\theta - \mu_\theta)^T \Sigma_\theta^{-1}(\theta - \mu_\theta) = c_2 - 0.5 \Sigma^{-1}_{x, x_{t+1}} \Sigma_{x, y} + \Sigma^{-1}_{x, z} (\theta - r) (\theta - r)^T \Sigma^{-1}_{r, r}$. We define the vectors $x \in \mathbb{R}^M$ and $r \in \mathbb{R}^M$ from $x_l$ and $r_l$, respectively. We define $\Sigma_{x, r} \in \mathbb{R}^{M \times M}$ as the covariance matrix of $r$. The regression coefficients as a Gaussian distribution are $N(\theta; \mu_\theta, \Sigma_\theta)$ where $\mu_\theta = A^{-1} b$ and $\Sigma_\theta = A^{-1}$. We denote the noise floor in the power spectral domain by $|N|^2$. In a FDR, at high RNRs and when $|N|^2$ is low, we have $r_l \approx z_l$ because $z \approx r + 0.5 \log(1 + \text{RNR}^{-1} + 2 \cos(\zeta)) \text{RNR}^{-0.5}$, which was also used in [13] and [14], where RNR = $\exp(2(r - n))$.

To solve the problem of finding $r_l$ and $\Sigma_{r(t)}$ given the observed $z_l$, we first employ a minimum MSE (MMSE) algorithm, such as the traditional Log-MMSE estimator [43], to remove the noise and estimate the signal’s spectral amplitude, as in the end of Sec. 3.1 in [41], and then perform a KF update step and use $z_l$ as the KF observation. For the linear KF [44], when the observation noise covariance matrix tends to zero at high RNRs, then the mean of the KF state goes to the KF observation and the variance of the KF state goes to zero. Therefore, a solution to finding $r_l$ and $\Sigma_{r(t)}$ is to set $r_l \approx z_l$ and $\Sigma_{r(t)}$ equal to a small value, such as 1 dB².

The $T_{\theta_1}$ can be estimated from $\theta_1$ using $T_{\theta_1} = \frac{3 \log(10)}{\theta_1}$, where $\theta_1$ is computed from FDR data in the log-magnitude...
spectral domain \[40\]. The KF algorithm, which models \( \gamma_t \) with a Gaussian distribution, uses a Gaussian prior for \( \gamma \) with mean \( L_t \gamma_0 \) and variance \( L^2 \Sigma \theta_1 \), because \( \gamma = L \theta_1 \). Using \( \gamma = L \theta_1 \), \( \theta_1 \) is directly mapped to \( \gamma_t \), without going via the \( T_{60} \).

The algorithm estimates a prior for \( \beta \), rather than first estimating the \( T_{60} \) and the DRR. Figure 5 depicts \( y_t, s_t \) and \( r_t \) over time, at 1 kHz, when the \( T_{60} \) is 0.8 s, the DRR is -2.52 dB, the noise type is white and the SNR is 20 dB. In Fig. 5, \( s_t \) and \( r_t \) are the true speech and reverberation powers; the latter is computed by convolving the RIR without its first 30 ms with the speech signal \[2\]. If we choose an appropriate FDR, then we can see that [1] is related to the FDR and the first frame of the FDR to the DRR. The \( \beta \) reverberation parameter is computed from the difference between the log-amplitude of the first frame and the \( \gamma \)-intercept of the LS fit, i.e. \( \theta_2 \). For the LS fit to the FDR, we choose to not use the first frame of the FDR. For the subtraction of two independent random variables, their means subtract and their variances add; \( \theta_2 \) has a mean and a variance and the log-amplitude of the first frame of the FDR has a mean and a fixed small variance. The prior for \( \beta \) is given by

\[
\beta = r_t - M_t + C_1 = \theta_2.
\]

We denote the mean of the internally computed Gaussian \( \gamma_t \) prior by \( \gamma_{t,t-1}'' \) and its variance by \( \Sigma_{t,t-1}'' \). We use

\[
\gamma_{t,t-1}'' = \gamma_{t,t-1}'' + \gamma_{t,t-1}'', \quad \Sigma_{t,t-1}'' = \Sigma_{t,t-1}''.
\]

Likewise, we denote the mean of the Gaussian \( \beta_t \) prior by \( \beta_{t,t-1}'' \) and its variance by \( \Sigma_{t,t-1}'' \). For \( \beta_t \), we use

\[
\beta_{t,t-1}'' = \beta_{t,t-1}'' + \beta_{t,t-1}'', \quad \Sigma_{t,t-1}'' = \Sigma_{t,t-1}''.
\]

Assuming that the RIR is known, it is straightforward to compute the \( T_{60} \); compute the energy decay curve, plot it on a log scale and estimate the \( T_{60} \) from its slope. On the contrary, estimating the \( T_{60} \) blindly is not trivial. The KF algorithm estimates the \( T_{60} \) by applying internally estimated priors using the decay rate of the LS fit to the FDR [40], so that the KF does not diverge and does not treat reverberation as speech.

E. The Peripheral Blocks of the Algorithm

This section describes the unshaded peripheral blocks of the KF algorithm in Fig. 3. The algorithm uses pre-cleaning before performing speech AR modelling, as in [33] and [43]. Pre-cleaning has also been used in [46] [47], in [11] [44] and in [10] [48]. The “Speech pre-cleaning” block in Fig. 5 affects the “Speech KF prediction” block and not the observation of the non-linear “KF Update” block, i.e. \( y_t \), used in step 7.

For the “Noise power estimation” block, an external noise power estimator, such as [49], is used. External noise estimation and log-normal noise power modelling, as in [50] [51], are used for \( r_{t-1} \), which is then used in step 6 of Table I. In summary, the KF algorithm tracks speech and reverberation in the spectral log-magnitude domain along with \( \gamma_t \) and \( \beta_t \), as described in Fig. 3 and in the signal model in Fig. 2.

IV. IMPLEMENTATION, TESTING AND VALIDATION

We use acoustic frames of length 32 ms and an acoustic frame increment of \( L = 8 \) ms in (2). We also use modulation frames of 64 ms and a modulation frame increment of 8 ms [33] [34]. In Fig. 3 for speech amplitude spectrum pre-cleaning, we use the Log-MMSE estimator [43] followed by the WPE dereverberation algorithm [52] [53]. The dimension of the speech KF state, \( s_t \), in [6] is \( p = 2 \). In the “Noise power estimation” block of Fig. 3 we use external noise power estimation from [49] using the implementation in [54].

The outer integrals in Secs. III.B and III.C are performed using sigma points, as in [33]. The numbers of sigma points used in [11] [17] are \( K_{(\delta,\epsilon)} = K_{(\epsilon,\gamma)} = K_{z_{(\gamma)}} = 3 \) and \( K_{(\delta,\epsilon)} = K_{(\gamma,\delta)} = K_{(\epsilon,\delta)} = 6 \). For the latter cases, for \( \xi \), the sigma points are at \( \xi = (1: K_{(\xi)} - 0.5) \times \frac{\pi}{4} \) and the weights are all equal to \( \frac{1}{N_{\xi}} \) [33] [32]. With this choice of sigma points for \( \xi \), the integral will be exact for an integrand that is a sum of terms of the form \( \cos(n\xi) \) for \( 0 \leq n \leq 2K_{(\xi)} + 1 \).

In Sec. III.D, for the \( \gamma_t \) and \( \beta_t \) priors, the look-ahead factor is \( C = 3 \) frames. For the FDR that is comprised of frames with decreasing energy, \( M_t \) is computed in every frame. Figures 6(a)-(c) illustrate \( s_t, z_t \) and \( r_t \) at 1 kHz over time, as in Fig. 3 in [5]. Figure 6(a) shows the predicted and true speech power, i.e. \( s_{t|t} \) and the true \( s_t \), and Fig. 6(b) the predicted and true noise and reverberation power, i.e. \( z_{t|t} \) and the true \( z_t \).
Table II

| Index | $T_{60}$ (s) | DRR (dB) | Room (m x m x m) | $a$ | $b$ |
|-------|-------------|----------|-----------------|-----|-----|
| A     | 0.18        | 8.45     | 5 x 4 x 4       | 0.54| 0.07|
| B     | 0.25        | 7.78     | 5 x 4 x 4       | 0.56| 0.11|
| C     | 0.33        | 7.13     | 5 x 4 x 4       | 0.72| 0.14|
| D     | 0.40        | 6.95     | 5 x 4 x 4       | 0.76| 0.16|
| E     | 0.47        | 6.25     | 5 x 4 x 4       | 0.79| 0.20|
| F     | 0.54        | 5.74     | 5 x 4 x 4       | 0.81| 0.23|
| G     | 0.61        | 5.14     | 5 x 4 x 4       | 0.86| 0.25|
| H     | 0.64        | 4.73     | 5 x 4 x 4       | 0.84| 0.26|
| I     | 0.68        | 4.25     | 5 x 4 x 4       | 0.85| 0.27|
| J     | 0.21        | 3.07     | 10 x 7 x 3      | 0.59| 0.06|
| K     | 0.31        | 2.74     | 10 x 7 x 3      | 0.70| 0.16|
| L     | 0.40        | 2.17     | 10 x 7 x 3      | 0.76| 0.23|
| M     | 0.50        | 1.78     | 10 x 7 x 3      | 0.80| 0.19|
| N     | 0.59        | 1.73     | 10 x 7 x 3      | 0.83| 0.20|
| O     | 0.64        | 0.95     | 10 x 7 x 3      | 0.84| 0.20|
| P     | 0.69        | -1.12    | 10 x 7 x 3      | 0.85| 0.19|
| Q     | 0.71        | -1.68    | 10 x 7 x 3      | 0.86| 0.21|
| R     | 0.76        | -2.01    | 10 x 7 x 3      | 0.88| 0.22|
| S     | 0.85        | -2.69    | 10 x 7 x 3      | 0.88| 0.19|
| T     | 0.97        | -2.95    | 10 x 7 x 3      | 0.87| 0.21|
| U     | 1.01        | -3.11    | 10 x 7 x 3      | 0.90| 0.20|
| V     | 1.05        | -3.33    | 10 x 7 x 3      | 0.90| 0.22|

The correlation coefficient between the true $z_t$ and $z_{|t|}$ is 0.89. Figure 6(c) shows the predicted and true reverberation power, i.e. $r_{|t|}$ and the true $r_t$. The correlation coefficient between the true $r_t$ and $r_{|t|}$ is 0.77 ($< 0.89$). The noise type is white, the SNR is 20 dB, $T_{60} = 0.61$ s and DRR = $-1.74$ dB.

The ordering of the graphs in Figs. 6(a)-(c) matches the ordering of the signal decompositions in the KF algorithm, in Table I. The noisy reverberant observation is first decomposed into $s_t$ and $z_t$ with step 7 in Table I. Figs. 6(a) and (b) illustrate $s_{|t|}$ and $z_{|t|}$, respectively, over time. Then, $s_t$ is decomposed into $r_t$ and $r_{|t|}$ with step 8; Fig. 6(c) depicts $r_{|t|}$ over time.

The $T_{60}$ and DRR estimates should converge to their true constant values when the talker and the microphone are not moving and the frequency variations in the reflection coefficients are not modelled [55–58]. Internally estimated priors for $\gamma_t$ and $\beta_t$ make the reverberation parameters converge over time. Figure 6(d) shows the predicted and true $T_{60}$ and DRR reverberation parameters over time. The dashed lines are the signals’ standard deviations, computed from $\gamma_t$ and $\beta_t$.

V. EVALUATION OF THE ALGORITHM

The proposed KF algorithm is evaluated in terms of the perceptual evaluation of speech quality (PESQ) [57], the cepstral distance (CD) spectral divergence metric [58], the reverberation decay tail (RDT) dereverberation metric [59] and the STOI speech intelligibility metric [60–61]. The ideal values of CD and RDT, which have been also used in [13], are zero. For evaluation, the TIMIT database [52], sampled at 16 kHz, and the RSG-10 noise database [63] are used. From the TIMIT database, 52 clean speech utterances are chosen. We use artificially-created reverberation with the image method [55] using the implementation in [64]. The wall reflection coefficient is used to vary the $T_{60}$ and hence also the DRR.

The KF algorithm is evaluated with noisy reverberant speech signals at various SNRs, from 5 to 20 dB, using random noise segments and the noise types of white, babble and factory. The KF algorithm is compared to the SPENDRED algorithm [13–19], which jointly performs blind denoising and dereverberation, and to algorithms that sequentially perform denoising and dereverberation, specifically to the Log-MMSE estimator [43] followed by the WPE algorithm [53] [60] and to the Log-MMSE estimator followed by an inverse-filter (IF) dereverberation method that is based on $T_{60}$ and DRR.
estimates [41]. For the IF, we blindly estimate the $T_{60}$ and the DRR for the entire speech utterance using the algorithm in [41] [65], whose implementation was generously provided by the author. As found in the Acoustic Characterisation of Environments (ACE) challenge [11], the $T_{60}$ estimator in [41] had the best performance of the examined $T_{60}$ estimators.

For evaluation, the examined acoustic conditions are shown in Table II. Two different rooms with dimensions $5 \times 4 \times 4$ m and $10 \times 7 \times 3$ m are used. The distance between the microphone and the talker is 1.5 m. Table II is sorted with respect to the $T_{60}$ and the DRR from A to I and from J to V. We plot the improvement in PESQ, $\Delta$PESQ, against the index of the acoustic environment to evaluate the KF algorithm. The top axis shows the raw PESQ of the noisy reverberant speech and is monotonic from A to I and from J to V. Similar graphs are also plotted for the other metrics. The ordering of the legends matches that of the algorithms at low $T_{60}$ values.

Table II also shows the $a$ and $b$ reverberation values of the RIRs and we note that one of the baselines that we use, i.e. the SPENDRED algorithm [18] [19], assumes that $b \leq 1$. This is valid in Table II because $b < 1$ and its range is small.

A. Overall Performance Against the $T_{60}$ and the DRR

This section investigates the overall performance of the KF algorithm against the $T_{60}$ and the DRR. Figure 7 examines: (a) the $\Delta$PESQ, where higher scores signify better speech quality, (b) the $\Delta$CD, where lower values signify better speech quality, (c) the $\Delta$RDT, where lower values signify better dereverberation, and (d) the $\Delta$STOI, where higher scores signify better intelligibility. Figure 7 first presents the results that are related to speech quality, in (a) and (b). Then, it shows the dereverberation results, in (c), and the speech intelligibility results, in (d). The graphs are against the index of the acoustic environment and, hence, against the $T_{60}$ and the DRR. The average over the noise types of white, babble and factory at 20 dB SNR is shown. We note that noisy reverberant speech with $T_{60} \approx 0.6$ s has a raw PESQ score of 2 in Fig. 7(a).

In Fig. 7(a), the top axis is monotonic from A to I and from J to V and there is no transition from index I to index J.

From Fig. 7(a), in terms of PESQ, the proposed algorithm consistently yields improved speech quality in challenging environments compared to the examined baselines. Compared to the unprocessed noisy reverberant speech, the algorithm shows improved performance for all the examined $T_{60}$ range from 0.18 to 1.05 s and DRR range from 8.43 to $-3.33$ dB. Compared to the unprocessed speech, for a $T_{60}$ of 0.3 and 1 s, the algorithm has a $\Delta$PESQ of 0.35 and 0.25, respectively, for the SNR of 20 dB averaged over the examined noises.

From Fig. 7(b), in terms of CD, the KF algorithm yields improved speech quality in acoustic environments with a $T_{60}$ from 0.18 to 1.05 s. The KF algorithm shows a deteriorating CD improvement with increasing $T_{60}$. Compared to the unprocessed signal, for a $T_{60}$ of 0.3 and 1 s, the algorithm has a CD improvement of approximately $-1.1$ and $-0.6$, respectively, for the SNR of 20 dB averaged over the tested noise types.
From Fig. 7(c), we observe that the KF algorithm yields improved speech dereverberation in challenging acoustic conditions and that the ΔRDT improves with increasing $T_{60}$ values. From the raw RDT of noisy reverberant speech in Fig. 7(c), we observe that the indexes A and J have very low reverberation. The indexes A and J have raw RDT scores of 0.4 and 0.7, respectively. The indexes G-I and T-V have high reverberation with a high raw RDT score. The proposed algorithm in these high reverberation cases achieves a large RDT improvement decreasing the RDT metric from approximately 2.6 to 0.8.

From Fig. 7(d), the KF algorithm shows marginally improved STOI performance for the examined $T_{60}$ range compared to the unprocessed speech and to the examined baselines. For a $T_{60}$ of 0.3 and 1 s, the algorithm has a ΔSTOI of 0.03 and 0.07, respectively. With increasing $T_{60}$, the ΔSTOI scores slightly increase. The baselines have negative ΔSTOI.

We use the noise types of white, babble and factory at 10 dB SNR and we obtain Fig. 8. Figure 8 shows similar graphs to Fig. 7 but for a SNR of 10 dB. For all the examined $T_{60}$ values, the KF algorithm shows a consistent improvement in the evaluation metrics of PESQ, CD, RDT and STOI.

We now use the noise types of white, babble and factory at 5 dB SNR and we obtain Fig. 9. Figure 9 shows similar graphs to Figs. 7 and 8. For all the examined SNRs, the KF algorithm shows a consistent improvement in the examined metrics, depending on the $T_{60}$. From Fig. 9(a), we observe that the ΔPESQ scores of the KF algorithm at 5 dB SNR are similar to the ΔPESQ scores at 10 dB SNR in Fig. 8(a).

The KF algorithm shows improved PESQ performance for the examined SNRs from 5 dB to 20 dB. The algorithm has better speech quality performance compared to the baselines for all the examined $T_{60}$ values. Compared to the unprocessed noisy reverberant speech, for the SNR of 5 dB, the algorithm has a ΔPESQ of about 0.45 for a $T_{60}$ of 0.6 s and a raw PESQ of 1.6. For the SNR of 20 dB, the algorithm has a ΔPESQ of approximately 0.3 for a $T_{60}$ of 0.6 s and a raw PESQ of 2. Comparing Figs. 7-9, we observe that the raw PESQ decreases with decreasing SNR while the ΔPESQ increases.

The KF algorithm yields improved dereverberation performance in terms of RDT in adverse conditions in Fig. 9(c). The algorithm’s performance improves with decreasing SNR. We observe that the indexes A-C and J-K have a high raw RDT despite that they have low reverberation at 5 dB SNR, when the effect of noise is higher than that of reverberation.

The results of the log-likelihood ratio (LLR) speech quality metric [66] [67], which is used as the main evaluation metric in the dereverberation algorithm in [68], resemble the results of CD in Fig. 10. Both CD and LLR have been used in the REVERB challenge [4]. It was found in [67] that LLR correlates well with speech quality although slightly less well than PESQ. Figure 10 examines ΔLLR when white noise is used at 20 and 10 dB SNR. From Fig. 10, we observe that decreasing the SNR from 20 to 10 dB increases the raw LLR of noisy reverberant speech and deteriorates the ΔLLR.

B. Overall Performance Against the SNR

This section investigates the overall performance of the KF algorithm against the SNR. Figure 11 presents the algorithm’s PESQ performance compared to the baselines. Figures 11(a) and 11(b) depict the PESQ improvement, ΔPESQ, against the SNR for white noise when: (a) $T_{60} = 0.40$ s and DRR = 0.17 dB, which is case L in Table I and (b) $T_{60} = 0.73$ s and DRR = -2.1 dB, which is case R in Table I. The indexes L and R were chosen because they are not extreme cases and they have different $T_{60}$ and DRR values. The ordering of the legends in Fig. 11 matches that of the algorithms at high SNRs.

Figures 11(c) and 11(d) show the ΔPESQ against the SNR for babble and factory noises, for cases L and R in Table I respectively, for the KF algorithm compared to the baselines. Babble noise is used for the solid lines, the top axis and the legends while factory noise is used for the dashed lines. The KF algorithm has a ΔPESQ that is dependent on the SNR and noise type and is, most of the times, increasing with decreasing SNRs. Compared to the unprocessed speech in Fig. 11(c), the algorithm has a ΔPESQ of about 0.35 for factory noise for SNRs from 5 to 20 dB, while decreasing for lower SNRs.

Figure 12 examines the RDT improvement, ΔRDT, against the SNR for white noise, for cases L and R in Table I in (a)
and (b). Figure 12 also depicts the $\Delta$RDT against the SNR for babble and factory noises, for cases L and R in Table II in (c) and (d). Babble noise is used for the solid lines, the top axis and the legends. Factory noise is used for the dashed lines. Figures 12(a)-(d) show that the $\Delta$RDT improves and the raw RDT of noisy reverberant speech increases with decreasing SNR. The $\Delta$RDT scores are better for stationary white noise than for non-stationary factory and babble noises.

Figure 13 shows the $\Delta$PESQ and the $\Delta$RDT against the SNR, from 5 to 20 dB, for the KF algorithm for white, factory and babble noises. Figure 13 examines the acoustic conditions than for non-stationary factory and babble noises. Figure 13 differs from Figs. 11 and 12 in presenting the indexes J to V and not only L and R. From Fig. 13(a), we observe that the $\Delta$PESQ decreases as the $T_{60}$ increases. The $\Delta$PESQ improves with decreasing SNR and the $\Delta$PESQ is higher for white noise than for factory and babble noises. In Fig. 13(b), the $\Delta$RDT improves as the $T_{60}$ increases and, moreover, the $\Delta$RDT improves with decreasing SNR.

In summary, in this evaluation section, we have tested the KF algorithm in different SNRs, noise types and acoustic conditions with different $T_{60}$ and DRR values. The algorithm is effective in enhancing distorted speech, decomposing noisy reverberant speech into speech, reverberation and noise. Regarding noise robustness, Figs. 7-13 show that the proposed KF algorithm achieves a significant performance gain over different noise types and SNRs, compared to the unprocessed noisy reverberant speech and to the examined baselines.

VI. CONCLUSION

In this paper, we present a monaural speech enhancement algorithm based on Kalman filtering in the log-magnitude spectral domain to blindly suppress noise and reverberation while accounting for inter-frame speech dynamics. The first two moments of the posterior distribution of the speech log-magnitude spectrum are estimated in noisy reverberant environments using a model that adaptively updates the $T_{60}$ and DRR reverberation parameters. The non-linear KF algorithm updates and tracks the two reverberation parameters of $\gamma_t$ and $\beta_t$ to further improve the estimation of the speech log-magnitude spectrum. In this paper, we show by means of theoretical and experimental analyses that Kalman filtering can be performed in challenging conditions by performing the proposed signal decompositions into speech, reverberation and noise, propagating backwards through the proposed signal model. Experimental results using instrumental measures show improved performance against both the $T_{60}$ and the SNR compared to the unprocessed noisy reverberant speech and to alternative competing techniques that perform blind denoising and dereverberation either in concatenation or jointly.

REFERENCES

[1] J. Eaton, N. D. Gaubitch, A. H. Moore and P. A. Naylor, “Estimation of room acoustic parameters: The ACE challenge,” IEEE Trans. on Audio, Speech and Language Process., vol. 24, no. 10, pp. 1681-1693, Oct. 2016.

[2] J. Li and L. Deng and R. Haeb-Umbach and Y. Gong, Robust automatic speech recognition: A bridge to practical applications, Ch. 9: Reverberant speech recognition. Elsevier, ISBN: 978-0-12-802398-3, 2016.

[3] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets et al., “A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research,” EURASIP J. Adv. Signal Process., vol. 7, pp. 1-119, 2010.

[4] K. Kinoshita, M. Delcroix, T. Yoshio, T. Nakatani, E. Habets et al., “The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Palz, NY, Oct. 2013.

[5] C. S. J. Doire, M. Brookes, P. A. Naylor, D. Betts, C. M. Hicks, M. A. Dmour and S. H. Jensen, “Single-channel blind estimation of reverberation parameters,” in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Sydney, Australia, pp: 35-35, April 2015.

[6] K. Lebart, J. M. Boucher, and P. N. Denbigh, “A New Method Based on Spectral Subtraction for Speech Dereverberation,” Acta Acustica, vol. 87, pp. 359–366, 2001.

[7] S. O. Sadjadi and J. H. L. Hansen, “Blind spectral weighting for robust speaker identification under reverberation mismatch,” IEEE Trans. on Audio, Speech and Language Process., vol. 22, no. 5, pp. 937-945, May 2014.

[8] I. J. Tasihev, Sound Capture and Processing. Ch. 8: De-reverberation. Wiley and Sons, ISBN: 0-470-13198-9, Aug. 2009.

[9] E. A. P. Habets, “Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement, Ch. 6: Late Reverberant Spectral Variance Estimation,” Ph.D. dissertation, Technische Universiteit Eindhoven, 2007.

[10] V. Wang and M. Brookes, “Model-based speech enhancement in the modulation domain,” IEEE Trans. on Audio, Speech and Language Process., vol. 26, no. 3, pp. 580-594, March 2018.

[11] Y. Wang and M. Brookes, “Speech enhancement using an MMSE spectral amplitude estimator based on a modulation domain Kalman filter with a Gamma prior” in Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Process., Shanghai, March 2016.

[12] S. Braun, B. Schwartz, S. Gannot and E. A. P. Habets, “Late reverberation PSD estimation for single-channel dereverberation using relative convolutive transfer functions,” in Proc. IEEE Intl. Workshop on Acoustics, Speech and Signal Process., Xi’an, China, Sept. 2016.

[13] A. Maezawa, K. Itoyama, K. Yoshii and H. G. Okuno, “Nonparametric Bayesian dereverberation of power spectrograms based on infinite-order autoregressive processes,” IEEE Trans. on Audio, Speech and Language Process., vol. 22, no. 12, pp. 1918-1930, Dec. 2014.

[14] M. Parchami, W.-P. Zhu and B. Champagne, “Model-based estimation of late reverberant spectral variance using modified weighted prediction error method,” Speech Communication, vol. 92, pp. 100-113, 2017.

[15] E. A. P. Habets, S. Gannot and I. Cohen, “Late reverberant spectral variance estimation based on a statistical model,” IEEE Signal Processing Letters, vol. 16, no. 9, pp. 770-773, Sept. 2009.

[16] J. Moorer, “About this reverberation business,” Computer Music Journal, vol. 3, no. 2, pp. 13-28, June 1979.

[17] M. Wölfel, “Enhanced speech features by single-channel joint compensation of noise and reverberation,” IEEE Trans. on Audio, Speech and Language Process., vol. 17, no. 2, pp. 312-323, Feb. 2009.

[18] C. S. J. Doire, M. Brookes, P. A. Naylor et al, “Single-channel online enhancement of speech corrupted by reverberation and noise,” IEEE Trans. on Audio, Speech and Language Process., vol. 25, no. 5, pp. 577-587, March 2017.

[19] C. S. J. Doire, “Single-channel enhancement of speech corrupted by reverberation and noise, Ch. 4: Single-channel enhancement of speech,” Ph.D. dissertation. Imperial College London, 2016.

[20] V. Leutnant, “Bayesian estimation employing a phase-sensitive observation model in the logarithmic Mel power spectral domain for the automatic recognition of noisy reverberant speech,” IEEE Trans. on Audio, Speech and Language Process., vol. 26, no. 1, pp. 95-109, Jan. 2014.

[21] V. Leutnant, “Bayesian estimation employing a phase-sensitive observation model for noise and reverberation robust automatic speech recognition, Ch. 4: Bayesian estimation of the speech feature posterior,” Ph.D. dissertation, Paderborn University, 2015.

[22] S. Braun and E. A. P. Habets, “Linear prediction based online dereverberation and noise reduction using alternating Kalman filters,” IEEE Trans. on Audio, Speech, and Language Process., vol. 26, no. 6, pp. 1119–1129, June 2018.

[23] S. Braun, “Speech dereverberation in noisy environments using time-frequency domain signal models, Ch. 4: Single-channel late reverberation PSD estimation,” Ph.D. dissertation, University of Erlangen-Nuremberg, International Audio Laboratories Erlangen, 2018. [Online]. Available: http://theses.eurasiapress.com/theses/776/speech-dereverberation-in-noisy-environments/
