

Phase-Aware Single-Channel Speech Enhancement with Modulation-Domain Kalman Filtering

Nikolaos Dionelis, Mike Brookes, Member, IEEE

Abstract—We present a speech enhancement algorithm that performs modulation-domain Kalman filtering to track the speech phase using circular statistics, along with the log-spectra of speech and noise. In the proposed algorithm, the speech phase posterior is used to create an enhanced speech phase spectrum for the signal reconstruction of speech. The Kalman filter prediction step separately models the temporal inter-frame correlation of the speech and noise spectral log-amplitudes and of the speech phase, while the Kalman filter update step models their nonlinear relations under the assumption that speech and noise add in the complex short-time Fourier transform domain. The phase-sensitive enhancement algorithm is evaluated with speech quality and intelligibility metrics, using a variety of noise types over a range of SNRs. Instrumental measures predict that tracking the speech log-spectrum and phase with modulation-domain Kalman filtering leads to consistent improvements in speech quality, over both conventional enhancement algorithms and other algorithms that perform modulation-domain Kalman filtering.

Index Terms—Speech enhancement, speech phase

I. INTRODUCTION

SPEECH enhancement in non-stationary noise environments is a challenging research area. The modulation domain is an often-used representation in models of the human auditory system; in enhancement, the modulation domain can be used to model the temporal inter-frame correlation of frames rather than treating each frame independently, [1], [2]. A number of authors have found that the performance of a speech enhancer can be improved by using a speech model that imposes temporal structure, [3], [4], [5]. Inter-frame speech correlation modelling can be performed with a Kalman filter (KF) with a state of low dimension, as in [1] and [6]. The enhancement algorithms in Chapter 5 of [7] track the time evolution of the speech Short Time Fourier Transform (STFT) amplitude domain coefficients. In [8], speech inter-frame correlation is modeled. Considering KF algorithms, many papers, such as [9], [10] and [11], utilise the nonlinear observation model relating clean and noisy speech in the log-spectral domain.

An overview of traditional statistical-based enhancement algorithms is given in Chapter 7 of [12]. Whereas these algorithms treat each frame independently, an alternative approach performs filtering in the modulation domain to model the time correlation of speech. Inter-frame speech correlation modelling can help enhancement algorithms achieve greater noise reduction with reduced speech distortion, [13]. Modulation-domain Kalman filtering refers to sequentially updating the statistics of speech imposing temporal constraints on the estimation of the

spectrum with a KF prediction step. In [14] and [15], speech tracking is presented. A number of authors perform Kalman filtering in the amplitude spectral domain assuming additivity of speech and noise in the same domain, [1], [16], [17].

A number of enhancement algorithms have been proposed that use a KF in the time domain and many of them are based on [18]. Kalman filtering in the time domain, [19], [18], is different from modulation-domain Kalman filtering, [1], [2]. Kalman filtering in the time domain changes the spectrum, without explicitly computing it. According to [1], [6] and [20], modulation-domain Kalman filtering operates in a spectral time-frequency domain and changes the modulation spectrum, without explicitly computing the modulation spectrum.

Existing enhancement KF algorithms that work in the time-frequency domain differ in their choice of the KF state, the KF prediction and the KF update. The KF state can be in the speech amplitude spectral domain, [1], [16], the power spectral domain or the log-spectral domain, [21], [22]. Speech spectra are well modelled by Gaussians in the log-spectral domain (and not so well in other domains), mean squared errors in the log-spectral domain are a good measure to use for perceptual speech quality and the not non-negative log-spectral domain is most suitable for infinite-support Gaussian modelling.

The KF update is affected by the signal model used for the addition of speech and noise, [23]. If speech and noise are independent, then they add in the complex STFT domain, [24], [25]; it may however be analytically simpler to assume that speech and noise add either in the power spectral domain or the amplitude spectral domain, [1], [16]. These alternative models are related to assumptions about the phase factor, the cosine of the phase difference between speech and noise, [24], considered in Chapter 4 of [26]. Assuming speech and noise additivity in the power domain is equivalent to assuming that the phase factor is zero. Assuming speech and noise additivity in the amplitude domain is equivalent to assuming that the phase factor is unity. The KF algorithms in [1], [16] and [17] assume that speech and noise add in the amplitude domain that results in noise oversubtraction in the region of $\text{SNR} = 0$ dB and may sometimes be perceptually beneficial, [12].

Noise tracking using a KF of low order in a spectral time-frequency domain can be beneficial for speech enhancement and a number of authors advocate this, [27], [7]. Noise tracking is performed in [28], in [29] and in the algorithm in [20].

Enhancement algorithms can benefit either from modelling the phase factor, [21], or from estimating the speech phase itself, [30]. The two related but distinct issues here are: (a) modelling the phase factor allows a better estimate of the speech amplitude, and (b) modelling the phase itself allows a better estimate of the speech phase. Both (a) and (b) can

N. Dionelis and M. Brookes are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: nikolaos.dionelis11@imperial.ac.uk; mike.brookes@imperial.ac.uk).

improve the performance of the enhancer. The phase factor is considered in [24] and [31]; the phase factor is assumed to be zero in [11]. Estimating the speech phase and not estimating the noise phase does not affect the phase factor distribution since if the noise phase is uniformly distributed, the difference of the speech and noise phases is also uniformly distributed.

Recent research has shown the importance of estimating the speech phase: approximating the speech phase by the noisy phase can degrade the performance of enhancement algorithms. The estimation of the speech phase has been addressed in [32], [30] and [33]. Circular statistics take phase periodicity explicitly into account so that speech phase tracking does not fail when the discontinuity between $-\pi$ and π is crossed, [34]. The speech phase is not perceptually irrelevant, [35], [36], and in low SNRs, the ear is sensitive to the phase. In [30] and [37], several speech phase estimation algorithms are evaluated.

In this paper, we perform phase-sensitive speech enhancement using modulation-domain Kalman filtering in the log-spectral domain, performing speech phase tracking with circular probability distributions and a nonlinear KF update step. In this paper, speech and noise are taken to be additive in the complex STFT domain, as in [21] and [22]. Modulation-domain Kalman filtering is performed using speech phase inter-frame correlation modelling and a KF update that uses local priors for the speech phase. In the KF update step, the posterior distributions of the speech phase and of the speech and noise log-spectra are computed using a specific four-dimensional distribution parameterization. A particular feature of this work is that the log-amplitude spectrum is used along with a nonlinear KF update step that computes the first two moments of the posterior distributions of the speech phase and of the speech and noise log-spectra. Speech enhancement is performed in the spectral log-amplitude time-frequency domain using a KF to model inter-frame correlations. The reasons for choosing the log-spectral time-frequency domain are related to the observation that good statistical models of speech and noise exist in the log-spectral domain. In this work, as in [21] and [22], regarding speech and noise estimation, modulation-domain Kalman filtering refers to imposing temporal inter-frame constraints in the log-spectral domain.

In this paper, the temporal dynamics of the speech log-spectrum and phase are modelled using the KF prediction step. In the KF update, the estimated dynamics of the speech log-spectrum and phase are combined with the observed noisy log-spectrum and phase to obtain a minimum mean squared error (MMSE) estimate of the speech log-spectrum and phase. As a main contribution, we advance modulation-domain Kalman filtering to include speech phase tracking, along with joint speech and noise spectral log-amplitudes tracking. Speech enhancement using modulation-domain Kalman filtering in the spectral log-amplitude and phase domains can now be performed and understood with this paper. The ordinary technique in order to incorporate the speech phase is to track the complex STFT domain of speech or the real and imaginary parts of the STFT domain of speech, [15], [38], or the log real and log imaginary parts of the complex STFT domain of speech.

This paper presents a Kalman filtering framework to perform speech phase tracking. We create a KF prediction step

that separately models the inter-frame relations of the speech log-spectrum, the speech phase and the noise log-spectrum. In addition, we formulate a phase-sensitive KF update step that models the nonlinear relations between the speech phase and the speech and noise spectral log-amplitudes. The KF update step computes the first two moments of the posterior using both the phase difference between noise and speech and the phase difference between noisy speech and speech.

The main difference between the proposed KF-based enhancement algorithm and the algorithms in [21] and [22] is that speech phase tracking is now performed along with joint speech and noise log-spectra tracking. The KF update step in this paper is different from the KF update in [21] and takes into account both the phase factor and the speech phase.

The proposed algorithm differs from the modulation-domain Kalman filtering algorithm in [1] in using the log-amplitude spectrum as the KF state, tracking the speech phase and using a phase-aware KF update step based on the additivity of speech and noise in the complex STFT domain. Furthermore, the proposed algorithm differs from the modulation-domain Kalman filtering algorithms in [16], [17], [6] and [20] in tracking and estimating the speech phase. Finally, the proposed algorithm differs from the phase-sensitive enhancement algorithms in [39], [40] and [41] in estimating the first and second moments of the distributions of the speech phase and of the speech and noise log-spectra and propagating them in the KF loop using a nonlinear KF update. The speech spectral amplitude and phase are estimated in [41] using the local SNR and the phase difference between noisy speech and speech. The enhancement algorithm in [42] estimates the speech phase using a Gaussian mixture model for speech in the log-spectral domain.

The structure of this paper is as follows. Section II presents the signal model, Sec. III describes the phase-sensitive enhancement algorithm that tracks the speech phase and Secs. IV and V present the implementation and evaluation of the algorithm, respectively. Conclusions are drawn in Sec. VI.

II. SIGNAL MODEL AND NOTATION

In the time domain, the noisy speech signal, $\tilde{y}(l)$, is given by $\tilde{y}(l) = \tilde{s}(l) + \tilde{n}(l)$, where $\tilde{s}(l)$ and $\tilde{n}(l)$ are the clean speech and additive noise, respectively, and l is the discrete time index. Applying the STFT, we obtain $Y_t(k) = S_t(k) + N_t(k)$ where t indexes the time-frame and k the frequency bin. Within the KF algorithm, frequency bins are processed independently and for clarity, in the remainder of this paper, the frequency bin index, k , is omitted and the time-frame index, t , is included only in equations that involve multiple time-frames.

In the complex STFT domain, we have: $Y = S + N \Leftrightarrow |Y|e^{j\theta} = |S|e^{j\phi} + |N|e^{j\psi}$ where the noisy speech phase is θ , the speech phase is ϕ and the noise phase is ψ . The STFT spectral log-amplitudes of the noisy speech, speech and noise are denoted by $y = \log |Y|$, $s = \log |S|$ and $n = \log |N|$.

Substituting for $|Y|$, $|S|$ and $|N|$ in the STFT coefficients gives $e^{y+j\theta} = e^{s+j\phi} + e^{n+j\psi}$. We define the relative phases of the noisy speech and speech and of the noise and speech by $\delta = \theta - \phi$ and $\gamma = \psi - \phi$, respectively. Figure 1 depicts the complex STFT phasors. Substituting for $(\theta - \phi)$ and $(\psi - \phi)$

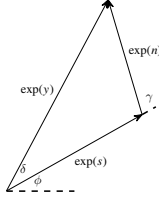


Figure 1. Plot of the STFT phasors in the complex plane.

gives $e^{y+j\delta} = e^s + e^{n+j\gamma}$. If δ and γ are chosen to lie in the range $[-\pi, \pi)$, then, since e^s is real and positive, δ and γ will have the same sign, i.e. $\text{sgn}(\delta) = \text{sgn}(\gamma)$. If S and N are independent, then γ is uniformly distributed, $\gamma \sim U(-\pi, \pi)$.

In this paper, $s_{t|t-1} = \mathbb{E}\{s_t | \mathcal{Y}_{t-1}\}$ denotes the mean of a random variable, s_t , given all the noisy complex STFT coefficients up to frame $(t-1)$. The semicolon, $;$, denotes vertical concatenation of vectors, $(\mathbf{x}_1; \mathbf{x}_2) = (\mathbf{x}_1^T \mathbf{x}_2^T)^T$.

A. Circular Distributions and Moments

The distribution of a phase, ϕ , is commonly modelled using either the von Mises (vM) distribution or the Wrapped Normal (WN) distribution which are both defined by two parameters. In both cases, the distribution can be uniquely defined by its complex-valued first circular moment, $\mathbb{E}\{\exp(j\phi)\}$, [34]. The vM distribution and its first circular moment are given by

$$p_{\text{vM}}(\phi; \mu^{(\phi)}, \kappa^{(\phi)}) = \frac{\exp(\kappa^{(\phi)} \cos(\phi - \mu^{(\phi)}))}{2\pi I_0(\kappa^{(\phi)})} \quad (1)$$

$$\mathbb{E}\{\exp(j\phi)\} = \exp(j\mu^{(\phi)}) \frac{I_1(\kappa^{(\phi)})}{I_0(\kappa^{(\phi)})} \quad (2)$$

where $\mu^{(\phi)}$ and $\kappa^{(\phi)}$ are the mean and concentration of the vM distribution and $I_0(x)$ and $I_1(x)$ are the Bessel functions of order zero and one, [34]. In (2), the first circular moment of the vM distribution, $\mathbb{E}\{\exp(j\phi)\}$, is $\exp(j\mu^{(\phi)})F(\kappa^{(\phi)})$, where the function $F(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}$ is the ratio of Bessel functions. We denote the vM distribution with mean $\mu_t^{(\phi)}$ and concentration $\kappa_t^{(\phi)}$ by $\text{vM}(\mu_t^{(\phi)}, \kappa_t^{(\phi)})$. For a given first circular moment, $m_1 = |m_1|e^{j\angle m_1}$, the vM distribution with this m_1 has density $\text{vM}(\angle m_1, F^{-1}(|m_1|))$ where $F^{-1}(\cdot)$ is the inverse of $F(\cdot)$.

In this paper, as in [39] and in contrast to [43], we choose to use the circular vM distribution and not the WN distribution. In particular, we model the distribution of the speech phase, ϕ , with a vM distribution because its probability density function, $p_{\text{vM}}(\phi; \mu^{(\phi)}, \kappa^{(\phi)})$, in (1), does not involve a summation.

III. DESCRIPTION OF THE ENHANCEMENT ALGORITHM

A. Overview of the Algorithm

The flowchart diagram of the algorithm is shown in Fig. 2. The core of the algorithm is the KF enclosed in the figure by a dotted line. The KF tracks the posterior distributions of the speech phase and of the log-spectra of both the speech and the noise. Rather than tracking the speech phase, ϕ , directly, it is more convenient to track the unit-magnitude complex quantity, $\exp(j\phi)$, since this eliminates the discontinuity at $\pm\pi$.

The algorithm's first step is to transform the noisy speech signal into the time-frequency domain using the STFT. The

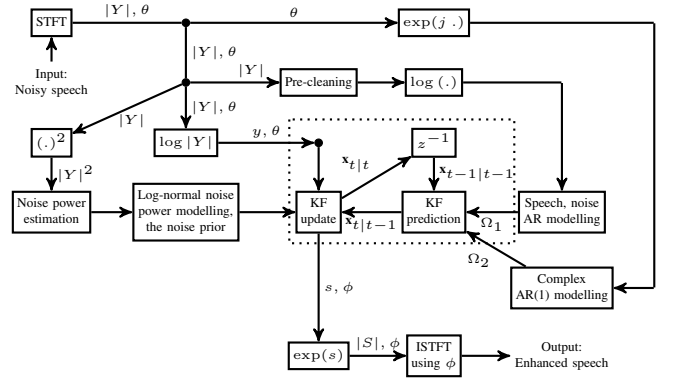


Figure 2. The flowchart of the enhancement algorithm is shown. The blocks in the dotted rectangle constitute the KF that tracks the speech and noise log-spectra, (s, n) , and the speech phase, ϕ . In the dotted rectangle, the term z^{-1} refers to one-frame delay. As described in Sec. III.B, the parameter sets Ω_1 and Ω_2 are $\{\mathbf{a}_t^{(s)}, \mathbf{e}_t^{(s)}, \zeta_t^{(s)}, \mathbf{a}_t^{(n)}, \mathbf{e}_t^{(n)}, \zeta_t^{(n)}\}$ and $\{A_t^{(p)}, \epsilon_t^{(p)}\}$, respectively.

algorithm considers the spectral amplitude and phase of the noisy speech separately. The noisy spectral amplitude is used in three ways: (a) it is converted to the log-spectral domain and used as the KF observation together with the noisy phase, θ , (b) it is pre-cleaned, using a conventional speech enhancer, and is used for speech autoregressive (AR) modelling of order p in the log-spectral domain, and (c) it is used for noise AR modelling of order q in the log-spectral domain. The noisy phase, θ , is represented as a unit magnitude phasor, $\exp(j\theta)$, which is used for phase complex AR modelling of order 1.

For the prediction step of the KF, autoregressive AR models are estimated for $\exp(j\phi)$ and for the log-spectra of speech and noise. To estimate the speech log-spectrum AR model, the noisy speech spectrum is pre-cleaned using a conventional speech enhancer, converted to the log-spectral domain and then divided into overlapping modulation frames followed by AR analysis. To estimate the AR model for $\exp(j\phi)$, the noisy phase is converted to the complex domain and is divided into overlapping modulation frames followed by AR analysis.

The linear KF prediction step (described in Sec. III.B) uses the results of the AR analysis to predict the KF state in the current time-frame, t , from past observations, from $\mathcal{Y}_{t-1}$.

The KF update (described in Secs. III.C and III.D) is non-linear and combines the output of the KF prediction step, the prior, with the observed noisy speech log-spectrum and phase while incorporating an additional prior distribution for the noise log-spectrum (Sec. III.F). This noise prior distribution is determined using an independent noise power estimator.

The algorithm's final step is to extract the estimated speech log-spectrum and phase of the current frame from the KF state vector and to combine these to generate the corresponding STFT coefficient. The inverse STFT (ISTFT) is then used to reconstruct the enhanced speech signal in the time domain.

B. The KF State and the KF Prediction Step

In Fig. 2, within the dotted rectangle, the KF state vector, $\mathbf{x}_t$, has dimension $(p + q + 1)$ and may be partitioned as

$$\mathbf{x}_t = \begin{pmatrix} \mathbf{x}_t^{(s)}; \mathbf{x}_t^{(n)}; x_t^{(p)} \end{pmatrix} \quad (3)$$

where

$$\mathbf{x}_t^{(s)} = (s_t \ s_{t-1} \ \dots \ s_{t-p+1})^T \in \mathbb{R}^p \quad (4)$$

$$\mathbf{x}_t^{(n)} = (n_t \ n_{t-1} \ \dots \ n_{t-q+1})^T \in \mathbb{R}^q \quad (5)$$

$$x_t^{(p)} = \exp(j\phi_t) \in \mathbb{C} \quad (6)$$

where the speech KF state is $\mathbf{x}_t^{(s)}$, the noise KF state is $\mathbf{x}_t^{(n)}$ and the phase KF state is $x_t^{(p)}$. For $\mathbf{x}_t^{(s)}$, the signal model is

$$\mathbf{x}_{t|t-1}^{(s)} = \mathbf{A}_t^{(s)} (\mathbf{x}_{t-1|t-1}^{(s)} - \zeta_t^{(s)} \mathbf{1}_p) + \zeta_t^{(s)} \mathbf{1}_p + \mathbf{w}_t^{(s)} \quad (7)$$

where $\mathbf{1}_p$ denotes the unity vector, $\mathbf{1}_p \in \mathbb{R}^p$. In (7), $\zeta_t^{(s)} \in \mathbb{R}$ is the AR mean, $\mathbf{w}_t^{(s)} \in \mathbb{R}^p$ with only its first element being non-zero is zero-mean with covariance matrix $\mathbf{Q}_t^{(s)}$ and

$$\mathbf{A}_t^{(s)} = \begin{pmatrix} -\mathbf{a}_t^{(s)T} & \\ \mathbf{I}_{(p-1)} & \mathbf{0}_{(p-1) \times 1} \end{pmatrix} \in \mathbb{R}^{p \times p} \quad (8)$$

$$\mathbf{Q}_t^{(s)} = \begin{pmatrix} \epsilon_t^{(s)} & \mathbf{0}_{(p-1) \times 1} \\ \mathbf{0}_{(p-1) \times 1} & \mathbf{0}_{(p-1) \times (p-1)} \end{pmatrix} \in \mathbb{R}^{p \times p} \quad (9)$$

where $\mathbf{I}_{(p-1)}$ denotes the identity matrix, $\mathbf{I}_p \in \mathbb{R}^{p \times p}$, and $\mathbf{0}_{(p-1) \times 1}$ denotes the zero matrix, $\mathbf{0}_{p \times p} \in \mathbb{R}^{p \times p}$. In (7), the speech KF transition matrix is $\mathbf{A}_t^{(s)}$, the speech KF transition noise covariance matrix is $\mathbf{Q}_t^{(s)}$ and the speech KF transition noise is $\mathbf{w}_t^{(s)}$. The speech KF transition noise is zero-mean Gaussian with covariance matrix $\mathbf{Q}_t^{(s)}$. The vector of speech AR coefficients is $\mathbf{a}_t^{(s)} \in \mathbb{R}^p$ and the variance of the transition noise is $\epsilon_t^{(s)}$. Also, the prior of $\mathbf{x}_t^{(s)}$ is denoted by $\mathbf{x}_{t|t-1}^{(s)}$.

In the linear KF prediction equation, (7), $\mathbf{A}_t^{(s)}$, $\zeta_t^{(s)}$ and $\mathbf{Q}_t^{(s)}$ are estimated from AR(p) modelling, AR modelling of order p , as described in Sec III.G. The AR mean, $\zeta_t^{(s)}$, is the mean speech log-amplitude estimated as a speech AR parameter.

The speech KF state, $\mathbf{x}_t^{(s)}$, consists of a speech KF state mean, $\boldsymbol{\mu}_t^{(s)} \in \mathbb{R}^p$, and a speech KF state covariance matrix, $\mathbf{P}_t^{(s)} \in \mathbb{R}^{p \times p}$. Both $\boldsymbol{\mu}_t^{(s)}$ and $\mathbf{P}_t^{(s)}$ are updated using

$$\boldsymbol{\mu}_{t|t-1}^{(s)} = \mathbf{A}_t^{(s)} (\boldsymbol{\mu}_{t-1|t-1}^{(s)} - \zeta_t^{(s)} \mathbf{1}_p) + \zeta_t^{(s)} \mathbf{1}_p \quad (10)$$

$$\mathbf{P}_{t|t-1}^{(s)} = \mathbf{A}_t^{(s)} \mathbf{P}_{t-1|t-1}^{(s)} \mathbf{A}_t^{(s)T} + \mathbf{Q}_t^{(s)}. \quad (11)$$

For the noise KF state, $\mathbf{x}_t^{(n)}$, the inter-frame correlation between adjacent noise frames is modelled in the log-spectral domain using a noise KF prediction step that is based on noise AR(q) modelling described in Sec III.G. Corresponding to (10) and (11), the noise KF prediction step equations are

$$\boldsymbol{\mu}_{t|t-1}^{(n)} = \mathbf{A}_t^{(n)} (\boldsymbol{\mu}_{t-1|t-1}^{(n)} - \zeta_t^{(n)} \mathbf{1}_q) + \zeta_t^{(n)} \mathbf{1}_q \quad (12)$$

$$\mathbf{P}_{t|t-1}^{(n)} = \mathbf{A}_t^{(n)} \mathbf{P}_{t-1|t-1}^{(n)} \mathbf{A}_t^{(n)T} + \mathbf{Q}_t^{(n)} \quad (13)$$

where the quantities with the superscript (n) refer to the noise KF state and are analogous to the speech quantities.

We now define the parameter set Ω_1 , which is shown in Fig. 2, by $\Omega_1 = \{\mathbf{a}_t^{(s)}, \epsilon_t^{(s)}, \zeta_t^{(s)}, \mathbf{a}_t^{(n)}, \epsilon_t^{(n)}, \zeta_t^{(n)}\}$, based on (7)-(13).

The proposed KF-based algorithm uses a joint speech and noise KF state with a full covariance matrix and the joint speech and noise KF state is $\mathbf{x}_t^{(j)} = (\mathbf{x}_t^{(s)}; \mathbf{x}_t^{(n)}) \in \mathbb{R}^{p+q}$. The superscript (j) denotes the joint KF state. For the joint KF state, $\mathbf{x}_t^{(j)}$, the joint speech and noise signal model is

$$\mathbf{x}_{t|t-1}^{(j)} = \mathbf{A}_t^{(j)} (\mathbf{x}_{t-1|t-1}^{(j)} - \boldsymbol{\zeta}_t^{(j)}) + \mathbf{w}_t^{(j)} + \boldsymbol{\zeta}_t^{(j)} \quad (14)$$

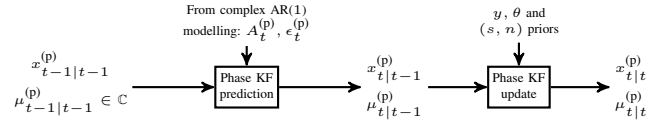


Figure 3. The speech phase KF is shown. The inputs/outputs match the inputs/outputs of the dotted rectangle in Fig. 2 for the speech phase KF.

where

$$\mathbf{A}_t^{(j)} = \begin{pmatrix} \mathbf{A}_t^{(s)} & \mathbf{0}_{p \times q} \\ \mathbf{0}_{q \times p} & \mathbf{A}_t^{(n)} \end{pmatrix} \in \mathbb{R}^{(p+q) \times (p+q)} \quad (15)$$

$$\boldsymbol{\zeta}_t^{(j)} = (\zeta_t^{(s)} \mathbf{1}_p; \zeta_t^{(n)} \mathbf{1}_q) \in \mathbb{R}^{(p+q)} \quad (16)$$

$$\mathbf{Q}_t^{(j)} = \begin{pmatrix} \mathbf{Q}_t^{(s)} & \mathbf{0}_{p \times q} \\ \mathbf{0}_{q \times p} & \mathbf{Q}_t^{(n)} \end{pmatrix} \in \mathbb{R}^{(p+q) \times (p+q)}. \quad (17)$$

The KF prediction step equations for the mean, $\boldsymbol{\mu}_t^{(j)}$, and covariance matrix, $\mathbf{P}_t^{(j)}$, are therefore

$$\boldsymbol{\mu}_{t|t-1}^{(j)} = \mathbf{A}_t^{(j)} (\boldsymbol{\mu}_{t-1|t-1}^{(j)} - \boldsymbol{\zeta}_t^{(j)}) + \boldsymbol{\zeta}_t^{(j)} \quad (18)$$

$$\mathbf{P}_{t|t-1}^{(j)} = \mathbf{A}_t^{(j)} \mathbf{P}_{t-1|t-1}^{(j)} \mathbf{A}_t^{(j)T} + \mathbf{Q}_t^{(j)} \quad (19)$$

where

$$\boldsymbol{\mu}_{t|t}^{(j)} = (\boldsymbol{\mu}_{t|t}^{(s)}; \boldsymbol{\mu}_{t|t}^{(n)}) \in \mathbb{R}^{(p+q)} \quad (20)$$

$$\mathbf{P}_{t|t}^{(j)} = \begin{pmatrix} \mathbf{P}_{t|t}^{(s)} & \mathbf{Z}_{t|t}^T \\ \mathbf{Z}_{t|t} & \mathbf{P}_{t|t}^{(n)} \end{pmatrix} \in \mathbb{R}^{(p+q) \times (p+q)} \quad (21)$$

where $\mathbf{P}_{t|t}^{(j)}$ is decomposed as shown in (21) and where the off-diagonal blocks, $\mathbf{Z}_{t|t}$, are, in general, non-zero.

Although S and N are independent, their posteriors given Y are dependent and the KF algorithm tracks the correlation between speech and noise in the log-spectral domain, [21].

For the speech phase KF state, $x_t^{(p)}$, $\exp(j\phi_t)$ is utilised. The speech phase KF state variance, $P_t^{(p)} \in \mathbb{R}$, is related to the speech phase KF state mean, $\mu_t^{(p)} \in \mathbb{C}$, with

$$P_t^{(p)} = 1 - |\mu_t^{(p)}|^2. \quad (22)$$

The prior distribution of the phase is fully defined by the complex mean, $\mu_{t|t-1}^{(p)} = \mathbb{E}\{\exp(j\phi_t) | \mathcal{Y}_{t-1}\}$. Using $\exp(j\phi_t)$ as the speech phase KF state leads to $\mathbb{E}\{\exp(j\phi) \exp(-j\phi)\} = 1$ and, in turn, this leads to (22). According to (22), $\mu_t^{(p)}$ uniquely defines $P_t^{(p)}$. Therefore, the KF algorithm tracks the speech phase KF state mean, $\mu_t^{(p)}$, and does not utilise $P_t^{(p)}$.

The KF state, $\mathbf{x}_t$, is defined in (3). The KF state mean is

$$\boldsymbol{\mu}_{t|t-1} = (\boldsymbol{\mu}_{t|t-1}^{(j)}; \mu_{t|t-1}^{(p)}). \quad (23)$$

It should be noted that $\boldsymbol{\mu}_t$ and $\mathbf{P}_t^{(j)}$ do not have the same dimensions because $\boldsymbol{\mu}_t$ includes both $\boldsymbol{\mu}_{t|t-1}^{(j)}$ and $\mu_{t|t-1}^{(p)}$.

Figure 3 shows how the speech phase KF state, $x_t^{(p)}$, evolves over time. The three main points of Fig. 3 are: (a) only $\mu_t^{(p)}$ is used to track $x_t^{(p)}$, (b) $x_{t|t-1}^{(p)}$ is computed from $x_{t-1|t-1}^{(p)}$ and from the two parameters of complex AR(1) modelling, and (c) the speech phase posterior, $x_{t|t}^{(p)}$, given the noisy y and θ is calculated using the log-spectrum local priors of (s, n) from the KF prediction step. Regarding (b), the two parameters of

the complex AR modelling are the AR coefficient, $A_t^{(p)} \in \mathbb{C}$, and the speech phase KF transition noise variance, $\epsilon_t^{(p)} \in \mathbb{R}$.

For the speech phase KF prediction step, the KF uses

$$\mu_{t|t-1}^{(p)} = A_t^{(p)} \mu_{t-1|t-1}^{(p)} \exp(-0.5\epsilon_t^{(p)}) \quad (24)$$

where the term $\exp(-0.5\epsilon_t^{(p)})$ approximates the Bessel ratio factor in (2) and decreases the absolute value of $A_t^{(p)} \mu_{t-1|t-1}^{(p)}$ to account for the transition noise variance, $\epsilon_t^{(p)}$. Equation (24) shows how $\mu_t^{(p)}$ changes in the speech phase KF prediction step and how $\mu_{t|t-1}^{(p)}$ is obtained using $A_t^{(p)}$ and $\epsilon_t^{(p)}$, which reduces the absolute value of the complex first circular moment considering the uncertainty in the complex AR prediction.

We now define the parameter set Ω_2 , shown in Fig. 2, by $\Omega_2 = \{A_t^{(p)}, \epsilon_t^{(p)}\}$, according to equation (24) and Fig. 3.

The phase KF update is speech-log-spectrum-sensitive, as seen in Fig. 3. Likewise, the speech log-spectrum KF update is speech-phase-sensitive. The modulation-domain Kalman filtering algorithm does not track the correlation between ϕ and (s, n) , assuming that $p(s, n, \phi) = p(s, n) p(\phi)$, but tracks μ_t and $\mathbf{P}_t^{(j)}$ and, according to Secs. III.C and III.D, performs phase-aware spectral log-amplitude estimation. Correlation between the concentration parameter of the vM distribution of the speech phase and the speech log-spectrum exists in voiced frames, compared to unvoiced/silent frames, and this is why phase-aware log-spectrum estimation is performed. In voiced frames, complex AR modelling for the speech phase and the speech phase KF prediction step are most beneficial.

C. The Decorrelation and Recorrelation Steps

The algorithm performs the update $\mathbf{x}_{t|t} = (\mathbf{x}_{t|t}^{(j)}; x_{t|t}^{(p)})$. The Kalman filtering algorithm first estimates $x_{t|t}^{(p)}$ using

$$\mu_{t|t}^{(p)} = \mathbb{E} \left\{ x_t^{(p)} | \mathcal{Y}_t \right\} \quad (25)$$

and, then, computes $\mathbf{x}_{t|t}^{(j)}$.

To perform the KF update, to compute $\mathbf{x}_{t|t}^{(j)}$, the KF state vector, $\mathbf{x}_{t|t-1}$, is first transformed as in Sec. 5.3.2.3 of [7] so that the elements corresponding to the current frame, $(s_t; n_t)$, are uncorrelated with the other elements of the KF state. This decorrelation allows the mean and the covariance matrix of the current-frame elements, $(s_t; n_t)$, to be updated as described in Sec. III.D without affecting the mean and covariance matrix of the other state vector elements. Finally, the transformation is inverted in order to restore the original KF state vector.

The KF state, $\mathbf{x}_t$, is of dimension $(p + q + 1)$, where $p \geq 1$ and $q \geq 1$, and the KF update step operates on the current-time elements, $(s_t; n_t; \exp(j\phi_t))$. The KF update estimates the posterior of $(s_t; n_t; \exp(j\phi_t))$ using the observed y_t, θ_t .

In the KF prediction step presented in Sec. III.B, the priors, $\mu_{t|t-1}^{(j)}$ and $\mathbf{P}_{t|t-1}^{(j)}$, are calculated in (18) and (19). In the nonlinear KF update step in Sec. III.D, the posteriors, $\check{\mu}_{t|t}^{(j)}$ and $\check{\mathbf{P}}_{t|t}^{(j)}$, are found. This section shows how the decorrelation and recorrelation steps are used to calculate $\mu_{t|t}^{(j)}$ and $\mathbf{P}_{t|t}^{(j)}$.

Before performing the nonlinear KF update step, the elements of the current time-frame in the KF state are decorrelated with respect to the other elements of the KF state.

The decorrelation operation is a two step procedure. Before actually decorrelating via a decorrelation matrix, $\mathbf{C}$, the current time-frame elements are first rearranged/swapped via a permutation matrix, $\mathbf{B}$. The speech and noise state vector, $\mathbf{x}^{(j)}$, is multiplied by the matrix $\mathbf{D} = \mathbf{C}\mathbf{B}$ where the permutation matrix, $\mathbf{B}$, swaps elements 2 and $(p + 1)$ of the vector $\mathbf{x}^{(j)}$ so that the first two elements relate to the current state. The permutation matrix, $\mathbf{B} \in \mathbb{R}^{(p+q) \times (p+q)}$, is given by

$$\mathbf{B} = [e_1 \ e_{p+1} \ e_3 \ e_4 \ \dots \ e_p \ e_2 \ e_{p+2} \ \dots \ e_{p+q}]^T \quad (26)$$

where e_i is the i -th column of the identity matrix.

The transformation matrix, $\mathbf{D}$, is chosen to decorrelate the current frame elements, s_t and n_t , from the rest of the state vector. The multiple-element decorrelation operation preserves the inter-frame correlation that is created from the KF prediction step so that the algorithm performs the KF update step with the variables in the current time-frame.

The covariance matrix, $\mathbf{B}\mathbf{P}_{t|t-1}^{(j)}\mathbf{B}^T$, is first decomposed as

$$\mathbf{B}\mathbf{P}_{t|t-1}^{(j)}\mathbf{B}^T = \begin{pmatrix} \mathbf{P}_A & \mathbf{P}_B^T \\ \mathbf{P}_B & \mathbf{P}_C \end{pmatrix} \quad (27)$$

where $\mathbf{P}_{t|t-1}^{(j)}$ is defined in (19) and where the dimensions of the sub-matrices are: $\mathbf{P}_A \in \mathbb{R}^{2 \times 2}$, $\mathbf{P}_B \in \mathbb{R}^{(p+q-2) \times 2}$ and $\mathbf{P}_C \in \mathbb{R}^{(p+q-2) \times (p+q-2)}$. The subscript $t|t-1$ of $\mathbf{P}_A$, $\mathbf{P}_B$ and $\mathbf{P}_C$ is omitted for clarity in (27)-(31).

We note that a sufficient condition for the sub-matrix $\mathbf{P}_A$ to be non-singular is that $\epsilon_t^{(s)}$ and $\epsilon_t^{(n)}$ are non-zero.

Then, the decorrelation matrix, $\mathbf{C}$, is

$$\mathbf{C} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_{2 \times (p+q-2)} \\ -\mathbf{P}_B\mathbf{P}_A^{-1} & \mathbf{I}_{(p+q-2)} \end{pmatrix}. \quad (28)$$

Using a breve diacritic to indicate the transformed domain, the transformed mean and covariance matrix are

$$\check{\mu}_{t|t-1}^{(j)} = \mathbf{D} \mu_{t|t-1}^{(j)} = (s_{t|t-1}; n_{t|t-1}; \check{\mu}_{C,t|t-1}^{(j)}) \quad (29)$$

$$\begin{aligned} \check{\mathbf{P}}_{t|t-1}^{(j)} &= \mathbf{D} \mathbf{P}_{t|t-1}^{(j)} \mathbf{D}^T \\ &= \begin{pmatrix} \mathbf{P}_A & \mathbf{0}_{2 \times (p+q-2)} \\ \mathbf{0}_{(p+q-2) \times 2} & \mathbf{P}_C \end{pmatrix} \end{aligned} \quad (30)$$

where

$$\check{\mathbf{P}}_C = \mathbf{P}_C - \mathbf{P}_B\mathbf{P}_A^{-1}\mathbf{P}_B^T. \quad (31)$$

From this point onwards, we include the subscript $t|t-1$ in the sub-matrices of $\mathbf{P}_A$, $\mathbf{P}_B$, $\mathbf{P}_C$ and $\check{\mathbf{P}}_C$.

The decorrelation step is the linear transformation in equations (29) and (30). In (29), the vector $\check{\mu}_{C,t|t-1}^{(j)} \in \mathbb{R}^{(p+q-2)}$ contains the elements that will not change in the transformed KF state mean after the KF update step in Sec. III.D.

In (27) and (30), we observe that $\mathbf{P}_A$ is preserved after the multiple-element decorrelation step. This is important since $\mathbf{P}_A$ is updated in the KF update in Sec. III.D. After the KF update in Sec. III.D, the inverse transformation is applied with

$$\mu_{t|t}^{(j)} = \mathbf{D}^{-1} \check{\mu}_{t|t}^{(j)} \quad (32)$$

$$\mathbf{P}_{t|t}^{(j)} = \mathbf{D}^{-1} \check{\mathbf{P}}_{t|t}^{(j)} \mathbf{D}^{-T}. \quad (33)$$

Equations (32) and (33) constitute the recorrelation step.

In (32) and (33), the posteriors $\check{\mu}_{t|t}^{(j)}$ and $\check{\mathbf{P}}_{t|t}^{(j)}$ are computed using the equations in (34) and (36). Now, $\check{\mu}_{t|t}^{(j)}$ is defined by

$$\check{\mu}_{t|t}^{(j)} = \left(s_{t|t}; n_{t|t}; \check{\mu}_{C,t|t}^{(j)} \right) \quad (34)$$

$$\text{where } \check{\mu}_{C,t|t}^{(j)} = \check{\mu}_{C,t|t-1}^{(j)} \in \mathbb{R}^{(p+q-2)} \quad (35)$$

where $\check{\mu}_{C,t|t-1}^{(j)}$ is defined in (29) and where $s_{t|t} = \mathbb{E}\{s_t | \mathcal{Y}_t\}$ is defined in Sec. II. According to (35), $\check{\mu}_{C,t|t-1}^{(j)}$ is not altered.

The posterior covariance matrix, $\check{\mathbf{P}}_{t|t}^{(j)}$, is obtained by replacing $\mathbf{P}_{A,t|t-1}$ in (30) by $\check{\mathbf{P}}_{A,t|t}$ to give

$$\check{\mathbf{P}}_{t|t}^{(j)} = \begin{pmatrix} \check{\mathbf{P}}_{A,t|t} & \mathbf{0}_{2 \times (p+q-2)} \\ \mathbf{0}_{(p+q-2) \times 2} & \check{\mathbf{P}}_{C,t|t} \end{pmatrix} \quad (36)$$

$$\text{where } \check{\mathbf{P}}_{C,t|t} = \check{\mathbf{P}}_{C,t|t-1} \in \mathbb{R}^{(p+q-2) \times (p+q-2)} \quad (37)$$

$$\check{\mathbf{P}}_{A,t|t} \in \mathbb{R}^{2 \times 2}$$

$$\check{\mathbf{P}}_{A,t|t} = \begin{pmatrix} \mathbb{E}\{s_t^2 | \mathcal{Y}_t\} & \mathbb{E}\{s_t n_t | \mathcal{Y}_t\} \\ \mathbb{E}\{s_t n_t | \mathcal{Y}_t\} & \mathbb{E}\{n_t^2 | \mathcal{Y}_t\} \end{pmatrix} - (s_{t|t}; n_{t|t}) (s_{t|t}; n_{t|t})^T \quad (38)$$

where $\check{\mathbf{P}}_{A,t|t}$ is obtained from the KF update step in Sec. III.D. According to (37), we note that $\check{\mathbf{P}}_{C,t|t-1}$ is not altered. In Sec. III.D, $\mathbb{E}\{(s_t; n_t; x_t^{(p)}) | \mathcal{Y}_t\}$ is calculated along with the covariance matrix of the first two elements of this vector.

D. The Phase-Sensitive KF Update Step

This section describes the evaluation of the quantities needed to perform the phase-sensitive KF update step, namely the posterior distribution parameters $(s_{t|t}; n_{t|t}; \mu_{t|t}^{(p)})$ in (34) and (25) together with the speech and noise covariance matrix $\check{\mathbf{P}}_{A,t|t}$ in (36). For clarity, in this section, the time-frame subscript, t , is omitted from all the random variables.

The prior distribution, $p(s, n, \phi, \psi)$, may be obtained from the KF prediction step outputs, $\mu_{t|t-1}^{(j)}$ and $\mathbf{P}_{t|t-1}^{(j)}$ in (18) and (19), and from the KF prediction step output, $\mu_{t|t-1}^{(p)}$ in (24), assuming that ψ is uniformly distributed, $\psi \sim U(-\pi, \pi)$. We assume that $p(s, n, \phi, \psi) = p(s, n) p(\phi) p(\psi)$ can be decomposed as the product of three independent distributions over the domain $-\infty < s, n < \infty$ and $-\pi < \phi, \psi \leq \pi$.

To calculate the posterior distribution, we need to obtain the conditional distribution of (s, n, ϕ, ψ) subject to the observation constraint, $(y, \theta) = (y_t, \theta_t)$. Applying the complex observation constraint reduces the dimension of the distribution from four to two. To impose this constraint, we make a transformation of variables from (s, n, ϕ, ψ) to (u, γ, y, θ) . Since the transformed parameterization includes y and θ as explicit variables, it becomes straightforward to impose the observation constraint. The remaining free parameters, $u = n - s$ and $\gamma = \psi - \phi$, were chosen because they are linear functions of the original parameters and because their domains of validity do not depend on the values taken by y_t and θ_t . This invertible transformation is given by

$$u = n - s \quad (39)$$

$$y = 0.5 (s + n + \log(2 \cosh(u) + 2 \cos(\gamma))) \quad (40)$$

$$\gamma = \psi - \phi \quad (41)$$

$$\theta = \phi + \delta$$

$$\begin{aligned} &= \phi + \text{sgn}(\gamma) \cos^{-1} \left(\frac{e^{2y} + e^{2s} - e^{2n}}{2e^{s+y}} \right) \\ &= \phi + \text{sgn}(\gamma) \cos^{-1} (\cosh(y - s) - 0.5e^{2n-s-y}). \end{aligned} \quad (42)$$

When $\cos^{-1}(\cdot)$ is taken to be within the range $[0, \pi]$, (39)-(42) are over the domain $-\infty < u, y < \infty$ and $-\pi < \gamma \leq \pi$.

To impose the observation constraint, we use the relationship $e^{y+j\delta} = e^s + e^{n+j\gamma}$ illustrated by the STFT phasor diagram in Fig. 1. Equation (40) can be derived by multiplying $e^{y+j\delta} = e^s + e^{n+j\gamma}$ by its complex conjugate to obtain $e^{2y} = e^{s+n} (e^{s-n} + e^{n-s} + 2 \cos(\gamma))$ and then taking the log of both sides. Equation (42) can be derived by writing $e^{y+j\theta} - e^{s+j\phi} = e^{n+j\psi}$ and multiplying this equation by its complex conjugate to obtain $e^{2y} + e^{2s} - 2e^{s+y} \cos(\delta) = e^{2n}$.

The inverse transformation of (39)-(42) is given by

$$v = s + n = 2y - \log(2 \cosh(u) + 2 \cos(\gamma)) \quad (43)$$

$$s = 0.5(v - u) \quad (44)$$

$$n = 0.5(v + u) \quad (45)$$

$$\delta = \text{sgn}(\gamma) \cos^{-1} (\cosh(y - s) - 0.5e^{2n-s-y}) \quad (46)$$

$$\phi = \theta - \delta \quad (47)$$

$$\psi = \gamma + \phi. \quad (48)$$

Using (39)-(48), the Jacobian matrix, $\frac{\partial(u, y, \gamma, \theta)}{\partial(s, n, \phi, \psi)}$, may be computed and its determinant is $\Delta = 1$.

To calculate the posterior mean and covariance matrices, $\check{\mu}_{t|t}^{(j)}$ and $\check{\mathbf{P}}_{t|t}^{(j)}$, that are required in the recorrelation step in equations (32) and (33), we compute $(s_{t|t}; n_{t|t})$ and $\check{\mathbf{P}}_{A,t|t}$ in (34) and (38), respectively. To calculate the posterior mean, $\mu_{t|t}^{(p)}$ in (25), together with $(s_{t|t}; n_{t|t})$ and $\check{\mathbf{P}}_{A,t|t}$, we apply the observation constraint to form the conditional distribution $p(u, \gamma | y, \theta)$ and then compute expectations of the form

$$\begin{aligned} \mathbb{E}\{f(u, \gamma, y_t, \theta_t) | y_t, \theta_t\} &= \mathbb{E}\{f(\cdot) | y_t, \theta_t\} \\ &= \int_{\gamma} \int_u f(\cdot) p(u, \gamma | y_t, \theta_t) du d\gamma \end{aligned} \quad (49)$$

where $f(\cdot) = f(u, \gamma, y_t, \theta_t)$ is an arbitrary function and

$$\begin{aligned} \mathbb{E}\{f(\cdot) | y_t, \theta_t\} &= \int_{\gamma} \int_u f(\cdot) \frac{p(u, \gamma, y_t, \theta_t)}{p(y_t, \theta_t)} du d\gamma \\ &\propto \int_{\gamma} \int_u f(\cdot) p(u, \gamma, y_t, \theta_t) du d\gamma \\ &= \int_{\gamma} \int_u f(\cdot) |\Delta|^{-1} p(s, n, \phi, \psi) du d\gamma \end{aligned} \quad (50)$$

where, in (50), (s, n, ϕ, ψ) are obtained from $(u, \gamma, y_t, \theta_t)$ using (43)-(48), $p(u, \gamma, y_t, \theta_t) = |\Delta|^{-1} p(s, n, \phi, \psi)$.

The first two moments of the posterior given (y_t, θ_t) are computed using (50). Substituting $\Delta = 1$, we obtain

$$\begin{aligned} \mathbb{E}\{f(\cdot) | y_t, \theta_t\} &\propto \int_{\gamma} \int_u f(\cdot) p(s, n, \phi, \psi) du d\gamma \\ &= \int_{\gamma} \int_u f(\cdot) p(s, n) p(\phi) p(\psi) du d\gamma. \end{aligned} \quad (51)$$

In (51), $p(s, n, \phi, \psi) = p(s, n) p(\phi) p(\psi)$ is assumed. The prior phases are assumed to be independent of the prior log-spectra. As noted in Sec. II, $\psi \sim U(-\pi, \pi)$ and hence

$$\mathbb{E}\{f(\cdot) | y_t, \theta_t\} \propto \int_{\gamma} \int_u f(\cdot) p(s, n) p(\phi) du d\gamma. \quad (52)$$

In (52), the integration variables are u and γ and we note that, given the KF observation (y_t, θ_t) , the variables s , n and ϕ are functions of u and γ , according to equations (43)-(48).

Using $0 \leq a, b \leq 2$, where a and b are integers, we now compute the first two moments of the posterior of (s, n) as

$$\mathbb{E}\{s^a n^b | y_t, \theta_t\} \propto \int_{\gamma} \int_u s^a n^b p(s, n) p(\phi) du d\gamma. \quad (53)$$

In (53), $p(\phi)$ is used along with the mapping from $\mathbb{E}\{x_t^{(p)}\}$ to $p(\phi)$ using the vM distribution presented in Sec. II.A.

Similarly, using (52), the posterior mean of $x_t^{(p)}$ is given by

$$\mathbb{E}\{x_t^{(p)} | y_t, \theta_t\} \propto \int_{\gamma} \int_u \exp(j\phi) p(s, n) p(\phi) du d\gamma. \quad (54)$$

Equations (53) and (54) are used to compute the first two moments of the posterior distributions of the speech phase and of the speech and noise log-spectra. For (53) and (54), as described in more detail in Sec. IV.A, we use uniformly weighted sigma points for the outer integration over γ and numerical integration for the inner integration over u .

Equations (53) and (54) provide the link between Sec. III.C and Sec. III.D. With (54), we compute $\mu_{t|t}^{(p)}$ in (25) and with (53), we calculate $(s_{t|t}, n_{t|t})$ in (34). Finally, with (53), we find $\mathbb{E}\{s_t^2 | \mathcal{Y}_t\}$, $\mathbb{E}\{s_t n_t | \mathcal{Y}_t\}$ and $\mathbb{E}\{n_t^2 | y_t, \mathcal{Y}_t\}$ in (38).

E. Discussion of the KF Update Step

The nonlinear KF update in Sec. III.D computes the first two moments of the posterior of the speech and noise log-spectra and of the speech phase. The KF update in (39)-(54) uses a local prior for the speech phase from the speech phase KF prediction step, as shown in Fig. 3. Using such a speech phase local prior, based on inter-frame speech phase modelling, is different from using the pitch and intra-frame phase correlation modelling, [30], [36]. The KF update computes the speech phase posterior using the speech phase local prior together with the speech and noise log-spectra local priors.

The two related but distinct issues here are: (a) how the speech phase prior, $p(\phi_t | \mathcal{Y}_{t-1})$, is obtained and (b) how the speech phase prior is applied to obtain the speech estimate. Regarding (a), inter-frame speech phase modelling is performed using equation (24). Regarding (b), equation (53) uses the speech phase prior to compute the clean speech estimate.

With (54), the complex-valued first circular moment for the speech phase posterior is estimated. As noted in Sec. II.A, the vM distribution can be computed from the first circular moment with moment matching, [34]. For the phase posterior, the speech phase concentration, $\kappa_{t|t}^{(\phi)}$, of the vM distribution can be calculated. In the algorithm, correlation between $\kappa_{t|t}^{(\phi)}$ and the estimated speech log-spectrum is apparent in voiced frames. For voiced frames, $\kappa_{t|t}^{(\phi)}$ is large, which means that the posterior variance is small, and the estimated clean speech spectral log-amplitude, $s_{t|t}$, is high. Conversely, in silence or unvoiced frames, $\kappa_{t|t}^{(\phi)}$ is small and $s_{t|t}$ is relatively low.

Figure 4 shows the absolute value of the complex-valued first circular moment, $|\mu_{t|t}^{(p)}| = F(\kappa_{t|t}^{(\phi)})$, when white noise at the SNR of 10 dB is used. According to Fig. 4, $|\mu_{t|t}^{(p)}|$ seems to pick out the harmonics of voiced speech shown in Fig. 5. Figures 4

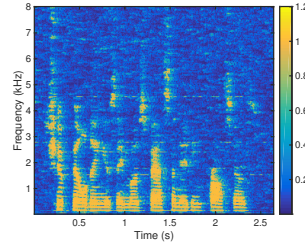


Figure 4. Plot of $|\mu_{t|t}^{(p)}| = F(\kappa_{t|t}^{(\phi)})$ when speech with white noise at the SNR of 10 dB is used.

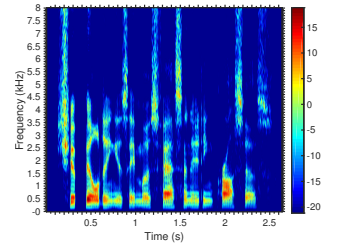


Figure 5. The spectrogram, power per decade, of the clean speech signal that is used with noise in Fig. 4.

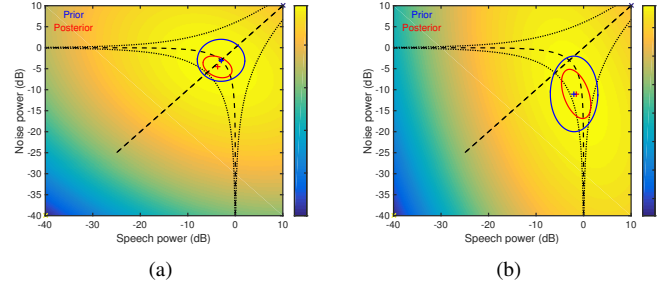


Figure 6. Plots of the prior and posterior distributions. The horizontal and vertical axes show s and n , respectively. The plots show the constraint curves for $\alpha = \{-1, 0, 1\}$ that are respectively the outer pair of dotted curves, the dashed central curve and the dotted lower left curve. The $\alpha = \{1, -1\}$ curves are (55), (56). The plots also show the 0-dB-SNR $u = 0$ line, which is the dashed diagonal line. The ellipses are the covariance matrices of the prior (outer, blue) and posterior (inner, red) distributions with the means indicated by crosses. The background shading shows the log probability density function of the posterior distribution. Here: (a) the mean of the KF prior is on the 0-dB-SNR $u = 0$ line, and (b) the mean of the KF prior has a positive SNR.

and 5 depict the relation between the speech spectrum and the concentration parameter of the vM distribution of the speech phase that characterizes voiced frames, [30], [33].

According to Fig. 4, at high frequencies, $\kappa_{t|t}^{(\phi)}$ will be very low and the algorithm will make the speech phase equal to the noisy phase to avoid the introduction of artefacts. The speech phase, ϕ , is tracked in all frequency bins using (24) and (54), in contrast to the algorithm in [36] where it is necessary to choose an upper bound, such as 4 kHz, as a frequency threshold to avoid introducing spurious artefacts in high frequency regions.

Figure 6 illustrates the means and the covariance matrices of the KF prior and posterior distributions when the proposed nonlinear KF update step in equations (39)-(54) is used. The ellipses are the covariance matrices of the prior (outer, blue) and posterior (inner, red) distributions. The background shading shows the log probability density function of the posterior distribution. In Fig. 6, $\text{vM}(\mu_{t|t-1}^{(\phi)}, \kappa_{t|t-1}^{(\phi)})$, $\mu_{t|t-1}^{(\phi)} = 0.66$, $\kappa_{t|t-1}^{(\phi)} = 3.01$, $y_t = 0$ and $\theta_t = 0$ are used.

We now define the phase factor, α , by $\alpha = \cos(\gamma)$, [21]. In Fig. 6, the mean of the speech and noise KF posterior lies within the curvy triangle that is mathematically defined by

$$\alpha = 1, \quad \cosh(u) = 0.5 \exp(-v) \quad (55)$$

$$\alpha = -1, \quad |\sinh(u)| = 0.5 \exp(-v). \quad (56)$$

The curvy triangle defines the observation constraint region in the (s, n) plane, [21]. The curvy triangle in Fig. 6 and in (55) and (56) is related to the different values that α can take,

[44]. In Fig. 6(a), the mean of the KF prior is in the curvy triangle and on the $u = 0$ line, which is for SNR = 0 dB. In Fig. 6(b), the mean of the KF prior is in the curvy triangle and off the $u = 0$ line. The mean of the KF posterior lies within the curvy triangle, within the observation constraint region.

In Secs. III.B-III.D, the KF algorithm that performs speech log-spectrum and phase tracking with inter-frame modelling has been described. Speech phase tracking with (24) and (54), taking into account the speech log-spectrum as explained in the end of Sec. III.B and in Fig. 3, differs from estimating ϕ using the post-processing technique in [45], as in [46].

F. The Noise Prior

The noise prior, which is out of the KF loop according to Fig. 2, constitutes a mechanism that helps the KF algorithm distinguish between speech and noise in order to solve the otherwise ill-conditioned problem of simultaneously tracking the speech and noise spectral log-amplitudes.

The KF-based local prior is given by

$$p(s_t, n_t, \phi_t | \mathcal{Y}_{t-1}) = p(s_t, n_t | \mathcal{Y}_{t-1}) p(\phi_t | \mathcal{Y}_{t-1}). \quad (57)$$

The next step is to consider the noise prior and include it in the proposed model. Using the noise prior, (57) becomes

$$\begin{aligned} & p(s_t, n_t, \phi_t | \mathcal{Y}_{t-1}, \mathcal{G}^{(n)}) \\ &= p(s_t, n_t | \mathcal{Y}_{t-1}, \mathcal{G}^{(n)}) p(\phi_t | \mathcal{Y}_{t-1}) \\ &\propto p(s_t, n_t | \mathcal{Y}_{t-1}) p(n_t | \mathcal{G}^{(n)}) p(\phi_t | \mathcal{Y}_{t-1}) \end{aligned} \quad (58)$$

where $\mathcal{G}^{(n)}$ is the knowledge from the external noise prior and where we assume that the information in $\mathcal{G}^{(n)}$ and $\mathcal{Y}_{t-1}$ is independent. In (58), $p(s_t, n_t | \mathcal{Y}_{t-1}, \mathcal{G}^{(n)})$ is given by the multiplication of the correlated KF-based local prior for speech and noise, $p(s_t, n_t | \mathcal{Y}_{t-1})$, by the noise prior, $p(n_t | \mathcal{G}^{(n)})$. When the noise KF-based local prior has a small variance, then the algorithm weights the noise KF-based local prior more than $p(n_t | \mathcal{G}^{(n)})$. We note that the product of two Gaussian distributions is itself a Gaussian distribution. The noise prior is assumed to be a Gaussian distribution because, according to Fig. 2, log-normal noise power modelling, [47], [48], is used to obtain the noise prior. Log-normal noise power modelling has previously been utilised in [47] and in [49], [50].

G. The Processing Outside the KF Loop

We now describe the processing that is carried outside the KF loop, outside the dotted rectangle in Fig. 2. The speech pre-cleaning in the “pre-cleaning” block in Fig. 2 applies a conventional enhancement algorithm, such as the Log-MMSE, [51], to the noisy speech, as in [1], [6], and [20]. Speech pre-cleaning has also been used in [52]. The speech is then divided into overlapping modulation frames in the log-amplitude-spectral domain and AR(p) modelling is performed to estimate the coefficients $\mathbf{a}_t^{(s)}$, $\zeta_t^{(s)}$ and $\epsilon_t^{(s)}$ in (7)-(9). AR modelling is performed using the covariance method, estimating both the speech AR coefficients and the speech AR mean. Estimating the speech AR mean, $\zeta_t^{(s)}$, is beneficial because we operate in the log-spectral domain and so $\zeta_t^{(s)}$ represents a positive scale factor in the amplitude spectral domain. The AR coefficients

need to be estimated from noisy speech and their estimate is inevitably biased, [53]. The model misspecification error is reduced by speech pre-cleaning, according to [1] and [7].

As indicated in Fig. 2, the proposed algorithm also performs noise AR modelling. For the same reasons as for speech, the algorithm performs noise pre-cleaning. The noise pre-cleaning in the “pre-cleaning” block in Fig. 2 applies a voice activity detection (VAD), [12], that is based on an one-pole smoother, as in [7], and on the estimated SNR from the KF state. The VAD computes $\hat{n}_t = \lambda_t \hat{n}_{t-1} + (1 - \lambda_t) y_t$, where $\hat{n}$ denotes the noise estimate and λ_t is obtained from applying a sigmoid function to the estimated SNR, $\eta_t = \exp(2s_t - 2n_t)$, using $\lambda_t = (1 - \exp(-\eta_t))^{-1}$. The estimation of the noise parameters $\mathbf{a}_t^{(n)}$, $\zeta_t^{(n)}$ and $\epsilon_t^{(n)}$ in (12) and (13) are created from noise AR(q) modelling in the same way as for speech.

IV. IMPLEMENTATION

For the implementation of the modulation-domain Kalman filtering algorithm, we use acoustic frames of length 32 ms, an acoustic frame time increment of 8 ms, modulation frames of 64 ms and a modulation frame time increment of 8 ms. For the noise prior, we utilise external noise estimation based on [54]. For speech amplitude spectrum pre-cleaning, we use the Log-MMSE estimator, [51]. The KF state dimensions for speech and noise are respectively $p = 2$ and $q = 2$. For the mapping between the parameters of the circular vM distribution and its complex-valued first circular moment, we use [55].

The algorithm processes each frequency bin independently, as discussed in Sec. II. Table I presents the pseudocode for the algorithm’s recursion for processing a single frequency bin. The pseudocode connects the different parts of the proposed KF algorithm. In Table I, step 7 continues the recursion.

In step 1 of Table I, in the KF prediction step, the speech and noise AR modelling parameters are obtained after pre-cleaning. Referring to Fig. 2, speech pre-cleaning affects $\mathbf{a}_t^{(s)}$, $\epsilon_t^{(s)}$ and $\zeta_t^{(s)}$ from the Ω_1 set and noise pre-cleaning affects $\mathbf{a}_t^{(n)}$, $\epsilon_t^{(n)}$ and $\zeta_t^{(n)}$ from Ω_1 . Pre-cleaning does not affect the KF observation, (y_t, θ_t) . We have evaluated the effect of omitting pre-cleaning for the speech and/or the noise but found that, in these cases, the output of the KF frequently diverged.

A. Evaluation of Integrals

We now present the integration details of the KF update in Sec. III.D. Using (53) and (54), the KF algorithm computes the first two moments of the posterior of the speech and noise log-spectra and the first circular moment of the posterior of the speech phase, respectively. The inner integration over u is performed by numerical integration, using the MATLAB function `integral` with default parameters. The outer integration over γ is performed with weighted sigma points, using the Unscented transform, [56], [23], using $G = 6$ sigma points for approximating integration with summation, [21]. We use uniformly weighted sigma points for the outer integration over γ at $\gamma = \frac{\pi}{2} + \frac{\pi}{0.5G}z$, where $z = \{0, 1, 2, \dots, G-1\}$, and hence at $\gamma \in \{\pm \frac{\pi}{6}, \pm \frac{\pi}{2}, \pm \frac{5\pi}{6}\}$. This choice of sigma points and weights ensures that the integral is exact for harmonic functions up to order 5. This choice of sigma points includes sigma points at the values of γ corresponding to $\alpha = 0$.

Table I

The algorithm's recursion for processing each frequency bin independently.

Inputs: (a) Noisy speech in the log-spectral and phase domains of one frequency bin of length M frames, and (b) the current KF state mean and covariance matrix.	
Recursive Processing:	
1:	KF prediction step: Calculate $\boldsymbol{\mu}_{t t-1}^{(i)}$ and $\mathbf{P}_{t t-1}^{(i)}$ using (18) and (19) for the speech and noise KF state. Compute $\mu_{t t-1}^{(p)}$ using (24) for $x_t^{(p)}$. The KF prediction step propagates the first two moments of the speech and noise log-spectra and the first circular moment of ϕ .
2:	Decorrelation step: Calculate $\check{\boldsymbol{\mu}}_{t t-1}^{(i)}$ and $\check{\mathbf{P}}_{t t-1}^{(i)}$ using equations (29) and (30).
3:	Multiply by the non-KF-based noise prior using (58).
4:	Perform the KF update step using the current-frame complex-valued noisy observation, y_t and θ_t . Impose the complex KF observation constraint and update the KF state mean and covariance matrix. Use equations (53) and (54) to compute the first two moments of the posterior of the current-frame speech phase and speech and noise log-spectra, as presented in Sec. III.D. Use sigma points for the integration over γ , as noted in Sec. IV.A.
5:	Recorrelation step: Compute $\boldsymbol{\mu}_{t t}^{(i)}$ and $\mathbf{P}_{t t}^{(i)}$ using equations (32) and (33).
6:	Store the current-frame estimated speech log-spectrum and phase from the updated KF state mean. According to Fig. 2, the current-frame estimated speech log-spectrum and phase are used for ISTFT signal reconstruction.
7:	Repeat steps 1-6 with the noisy speech in the log-spectral and phase domains of length $(M - 1)$ frames and with the updated KF state mean and covariance matrix.

V. EVALUATION

In this section, we evaluate the proposed modulation-domain Kalman filtering algorithm. We use a subset of the core test set of the TIMIT database, [57], sampled at 16 kHz, to evaluate the algorithm in different noise types. We use 110 utterances from the TIMIT core test containing different sentence texts. We also use noise recordings from the RSG-10 database, [58], at SNRs from 0 to 30 dB. Randomly selected noise segments are used in each test. Noisy speech is created at specific SNRs using [59], using the active speech level from [60].

The KF algorithm is evaluated in terms of speech quality with the perceptual evaluation of speech quality (PESQ), [61], metric. While PESQ has been initially developed for assessing the perceived quality of coded speech, it also shows good correlation with speech quality in the speech enhancement context, [12]. The proposed KF-based algorithm is also evaluated with the segmental SNR (SegSNR), cepstrum distance (CD), [62], overall speech quality (OVRL), [63], and short-time objective intelligibility (STOI), [64], [65], metrics.

The structure of the evaluation section is as follows. Section V.A presents the baselines that perform and that do not perform modulation-domain Kalman filtering. Section V.B evaluates the proposed KF algorithm in terms of speech quality and compares it with the non-KF and the KF baselines. Section V.C evaluates the KF algorithm using the STOI intelligibility metric. Section V.D examines computational complexity.

A. The Non-KF Baselines and the KF Baselines

The proposed algorithm is compared with two non-KF-based approaches to speech enhancement: (a) the traditional technique of Log-MMSE, [51], with the MMSE noise estimator, [54], and (b) the optimally modified log-spectral amplitude (OMLSA) estimator, [66], with the improved minima controlled recursive averaging (IMCRA) noise estimator, [67]. The (a) and (b) algorithms constitute the non-KF baselines of this paper. The proposed algorithm is also compared with the two KF baselines presented in [21]. The first KF baseline performs speech log-spectrum tracking; the second KF baseline performs joint speech and noise log-spectra tracking, [21].

We denote the proposed algorithm by SNPT that refers to speech, noise and phase tracking. We denote the algorithm in [21] by ST that refers to speech tracking, leaving out both noise and phase tracking. Correspondingly, leaving out phase tracking, we denote the KF algorithm in [21] by SNT that refers to speech and noise tracking. Hence, we compare SNPT with ST and SNT and, furthermore, we compare the proposed SNPT algorithm with the STA algorithm, which, similarly to [1], performs speech tracking using a KF state in the amplitude spectral domain and an update that assumes speech and noise additivity in the amplitude spectral domain, [1], [16], [17].

The proposed SNPT algorithm and the ST and SNT baselines, which have also been evaluated in [21] and [22], use a log-spectrum KF state and assume that speech and noise add in the complex STFT domain. SNPT, ST and SNT assume that $\gamma \sim U(-\pi, \pi)$ and that the speech and noise phases are independent. Considering the number of tracked quantities, SNPT tracks the most quantities. Considering the choice of signal model, STA assumes $\alpha = 1$ and a simpler signal model than SNPT. The algorithm in [1] differs from STA in using a linear KF update step and the algorithms in [16] and [17] in using a KF as a post-processor that follows a conventional enhancer. STA, ST and SNT have lower computational complexity than SNPT because they leave out speech phase tracking.

B. Speech Quality Experimental Results

In this section, the SNPT algorithm is evaluated with speech quality metrics and compared to KF and non-KF baselines. To study the robustness of SNPT to noise, coloured boxplots of the differential (Δ) scores are used. For each result, a positive score indicates a higher metric score and better performance compared to the unprocessed noisy speech. The boxplots in Figs. 7-11 show the median, the inter-quartile range and the 5% and 95% points of the distribution of a speech quality metric. The upper horizontal axis shows the (raw) evaluation metric of noisy speech for the corresponding noise types.

The SNPT algorithm is evaluated in terms of PESQ improvement, ΔPESQ , compared to unprocessed speech, in Figs. 7 and 8. In this section, the proposed SNPT algorithm is also evaluated in terms of SegSNR improvement, ΔSegSNR , in Fig. 9, in terms of CD improvement, $-\Delta\text{CD}$, in Fig. 10 and in terms of OVRL improvement, ΔOVRL , in Fig. 11.

Figure 7 shows that for each of the examined noise types, SNPT achieves a higher ΔPESQ metric score compared to the KF-baselines of ST and SNT, an even higher ΔPESQ score

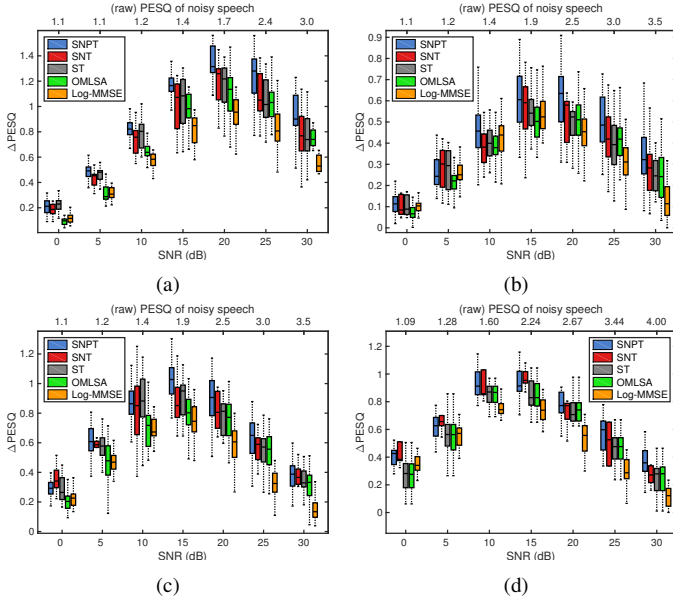


Figure 7. Boxplots of the Δ PESQ for speech with: (a) white noise, (b) babble noise, (c) F16 noise, and (d) factory noise.

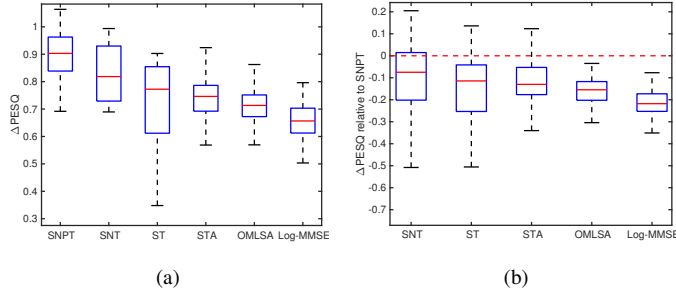


Figure 8. Boxplots using four noise types (white, babble, F16 and factory) with the average SNR for each noise type chosen to give a mean PESQ score of 2.0 for the noisy speech. (a) Boxplot of the average of the Δ PESQ scores for the six competing algorithms. (b) Boxplot of the average of the difference in the Δ PESQ scores between the baselines and the SNPT algorithm.

compared to the OMLSA non-KF baseline and a further higher Δ PESQ score compared to the Log-MMSE baseline.

Figure 7 shows that all algorithms give the greatest Δ PESQ at the SNR corresponding to a raw PESQ of about 2.0. For positive SNRs, SNPT is consistently better than the baselines for all the examined noise types, presenting greater differences compared to the non-KF algorithms. The benefits of SNPT are most apparent in stationary noise, in white and F16 noise, followed by babble and then by factory noise. SNT and ST are in babble noise only slightly better than OMLSA indicating that most of the improvement comes from tracking the speech phase. In Fig. 7, Log-MMSE was consistently the worst of the tested algorithms. The median of SNPT is approximately 0.2 higher than the median of the two non-KF algorithms.

To examine the performance over a range of noise types, we evaluate the algorithms in Fig. 8 using four noise types with the average SNR for each noise type chosen to give a mean PESQ of 2.0 for the noisy speech. We use the noise types: white, babble, F16 and factory. Combining different noises into a single graph, Fig. 8(a) shows the average of the Δ PESQ scores for each algorithm. Figure 8(b) shows the average of

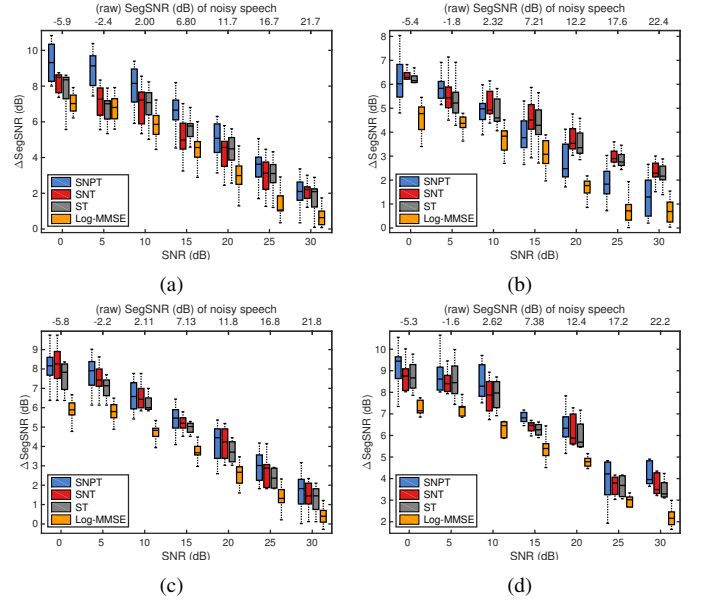


Figure 9. Boxplots of the Δ SegSNR for speech with: (a) white noise, (b) babble noise, (c) F16 noise, and (d) factory noise.

the difference in the Δ PESQ scores between the competing algorithms and SNPT and therefore excludes the effects of test material variability that are common to all algorithms.

In Fig. 8, the proposed SNPT algorithm is compared to the STA KF baseline, as presented in Sec. V.A, using the Δ PESQ speech quality metric. Figures 8(a) and 8(b) compare SNPT with the STA, ST and SNT KF baselines. The presented SNPT algorithm achieves a consistent improvement in PESQ compared to the KF baselines of STA, ST and SNT. Furthermore, Figs. 8(a) and 8(b) show that the presented SNPT algorithm has a higher PESQ improvement score compared to both the non-KF and the KF baselines. Considering the median values in Fig. 8(b), the SNPT algorithm shows a higher PESQ improvement score of approximately 0.09 compared to the KF baselines of ST and SNT, and of approximately 0.18 compared to the non-KF baselines of OMLSA and Log-MMSE.

In Fig. 8(a), the SNPT algorithm has an improvement in median PESQ values of approximately 0.9 compared to the unprocessed noisy speech. Considering Fig. 8(b), for noisy speech signals with a PESQ score of 2.0, using the paired-sample t-test with Bonferroni correction, the difference in performance between the SNPT algorithm and each of the other algorithms was found to be significant with $P < 0.01$.

Figure 9 shows that for each of the tested noise types, SNPT achieves a SegSNR improvement, of approximately 2 dB on average, compared to Log-MMSE. All algorithms give the greatest Δ SegSNR at low SNRs with improvements of up to 9 dB. For white noise, SNPT gives a greater improvement in SegSNR than the other algorithms, especially at low SNRs. For other noise types, however, there is little difference between SNPT, SNT and ST. Compared to the unprocessed noisy speech, the proposed SNPT algorithm consistently improves the SegSNR metric and enhances the noisy speech signal.

Figure 10 shows the decrease in CD, $-\Delta$ CD, for each of the algorithms. Higher values of $-\Delta$ CD correspond to

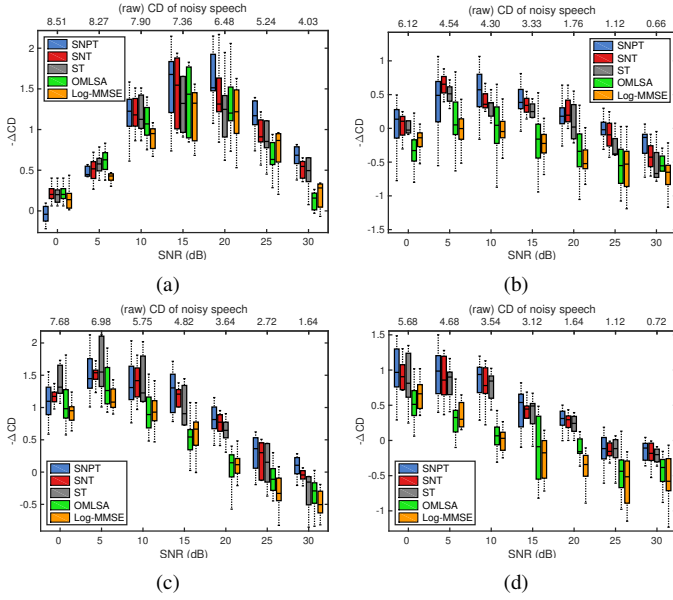


Figure 10. Boxplots of the $-\Delta\text{CD}$ for speech with: (a) white noise, (b) babble noise, (c) F16 noise, and (d) factory noise.

improved speech quality. For most positive SNRs and for each of the examined noise types, SNPT achieves a higher $-\Delta\text{CD}$ compared to the non-KF baselines. The $-\Delta\text{CD}$ of SNPT are higher for factory noise and for babble noise with an average improvement of about 0.5, followed by F16 noise and white noise. Comparing SNPT with the non-KF baselines, OMLSA and Log-MMSE, we observe a high $-\Delta\text{CD}$ improvement; comparing SNPT with the KF baselines, ST and SNT, we observe a smaller $-\Delta\text{CD}$ improvement. For the examined noise types, the SNPT, SNT and ST algorithms were consistently better than OMLSA and Log-MMSE in terms of $-\Delta\text{CD}$.

According to Fig. 11, the ΔOVRL results are similar to PESQ presenting the highest ΔOVRL gains for all algorithms at around 20 dB SNR. The ΔOVRL results in Fig. 11 are broadly consistent with the ΔPESQ results in Fig. 7, but with less difference between the SNPT, SNT and ST KF algorithms. SNPT achieves a higher OVRL score, compared to the non-KF baselines. Compared to the unprocessed noisy speech, SNPT consistently improves the OVRL metric, with maximum ΔOVRL scores of approximately 1 in the middle SNRs.

According to our results, the performance of SNPT is mainly because of phase-sensitive log-spectrum estimation, as explained in the end of Sec. III.B. Both phase estimation and phase-sensitive log-spectrum estimation improve speech quality. From Figs. 7-11 and the analysis in Sec. V.B, SNPT shows consistent improvement compared to the non-KF and the KF baselines using PESQ, SegSNR, CD and OVRL.

C. Speech Intelligibility Metric Experimental Results

The SNPT algorithm is evaluated using the STOI metric that performs correlation in the modulation domain. Figure 12 shows the ΔSTOI of SNPT, SNT, ST, OMLSA and Log-MMSE for the noise types of white, babble, F16 and factory. In Fig. 12, the raw STOI scores were ≥ 0.66 . In [64], this score was found to correspond to intelligibility $> 86\%$ on two

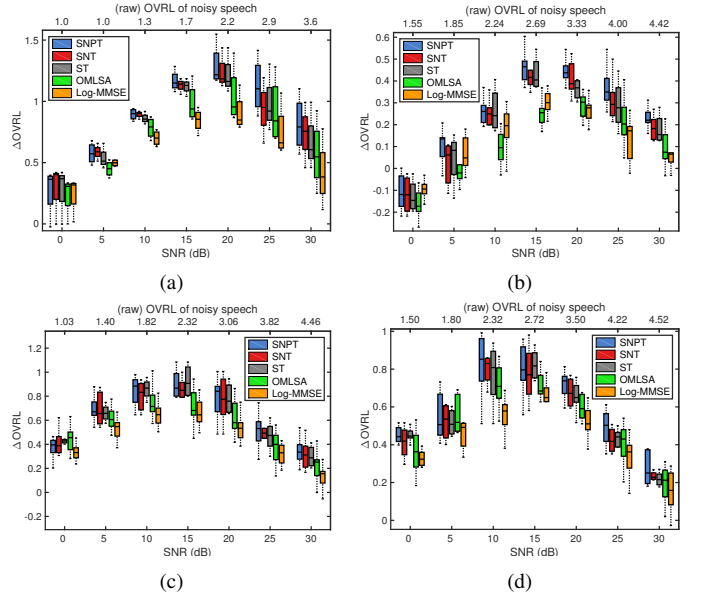


Figure 11. Boxplots of the ΔOVRL for speech with: (a) white noise, (b) babble noise, (c) F16 noise, and (d) factory noise.

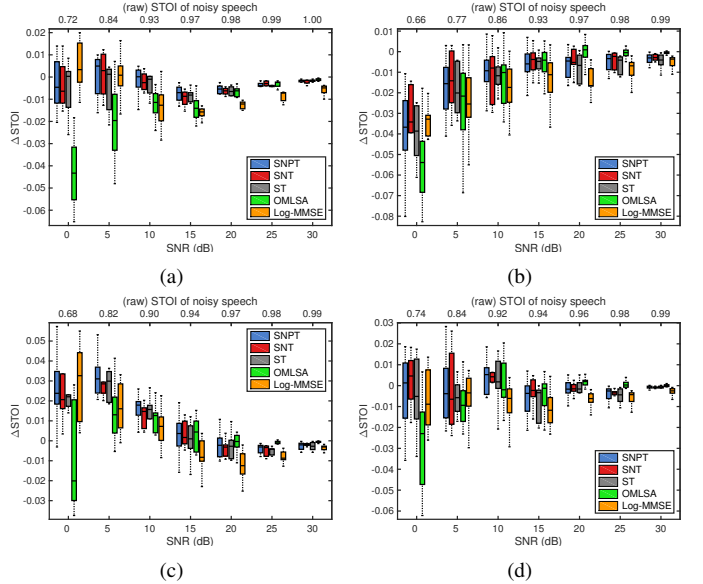


Figure 12. Boxplots of the ΔSTOI for speech with: (a) white noise, (b) babble noise, (c) F16 noise, and (d) factory noise.

different databases. The ΔSTOI values were very small for all algorithms and noise types. OMLSA was noticeably the worst at low SNRs. For SNPT, for $0 \leq \text{SNR} \leq 15$ dB, marginal ΔSTOI improvements are obtained only for F16 noise.

D. Computational Complexity

In this section, we provide a brief analysis on the computational requirements of the proposed SNPT algorithm using the parameters in Sec. IV. The real-time factor, R , as defined in [68] is examined. Using MATLAB on a laptop equipped with an Intel Core i5 processor and using parallel computation with two cores for processing frequency bins in parallel, $R = 31$ for the presented SNPT algorithm and $R = 17$ for the ST and SNT KF baselines. The computational complexity of the

SNPT algorithm is high because the speech phase is tracked, along with the speech and noise spectral log-amplitudes, for all frequencies using (24) and (54), because the KF update step in Sec. III.D is a joint speech log-amplitude and phase estimator and, in high frequencies, estimating the amplitude is important for improving speech quality, [12]. In [36], the upper bound of 4 kHz is chosen for perceptual reasons. In the SNPT algorithm, the KF observation is the noisy phase and the KF will take care of the high frequencies on its own. If needed, the KF will make the speech phase equal to the noisy phase to avoid the introduction of spurious artefacts.

For a perceptual comparison, the reader is referred to [69] where some recordings processed by the proposed modulation-domain Kalman filtering enhancement algorithm are available.

VI. CONCLUSION

In this paper, we present a phase-sensitive enhancement algorithm that tracks the time evolution of the speech phase using circular statistics, along with the log-spectra of speech and noise. We create a KF prediction step that separately models the inter-frame relations of the speech log-spectrum, the speech phase and the noise log-spectrum. In addition, we create a phase-sensitive KF update step that models the non-linear relations between the speech log-spectrum, the speech phase and the noise log-spectrum and computes their posterior distributions. In the nonlinear KF update step, considering the magnitude and phase of the speech and noise in each frequency bin separately, we start with a four-dimensional probability space and we impose a complex observation constraint that reduces the four-dimensional space to a two-dimensional space, using the phase difference between speech and noise and the phase difference between speech and noisy speech. Equations (39)-(54) describe the proposed KF update that is different from the normal linear KF update and is based on the decorrelation and recorelation steps in Sec. III.C. Instrumental measures predict a consistent moderate speech quality improvement, compared to both non-KF and KF baselines, when using the SNPT modulation-domain Kalman filtering algorithm for a range of noise types, for positive SNRs.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful, detailed and constructive suggestions.

REFERENCES

- [1] S. So and K. K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Communication*, vol. 53, no. 6, pp. 818-829, July 2011.
- [2] S. So, K. K. Wójcicki, and K. K. Paliwal, "Single-channel speech enhancement using Kalman filtering in the modulation domain," in *Proc. Conf. Int. Speech Communication Association, Makuhari*, Sept. 2010.
- [3] D. Liang, M. D. Hoffman and G. J. Mysore, "Speech dereverberation using a learned speech model," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process., Brisbane*, April 2015.
- [4] N. Boulanger-Lewandowski, G. J. Mysore and M. Hoffman, "Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process., Florence*, May 2014.
- [5] I. Andrianakis and P. R. White, "On the application of Markov Random Fields to speech enhancement," in *Proc. Int. Conf. Signal Process., pp. 198-201, Cirencester*, Dec. 2006.
- [6] 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 Int. Conf. Acoustics, Speech and Signal Process., Shanghai*, March 2016.
- [7] Y. Wang, "Speech enhancement in the modulation domain, Ch. 5: Model-based speech enhancement in the modulation domain," Ph.D. dissertation, Imperial College London, 2015.
- [8] N. Mohammadiha, P. Smaragdis and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in NMF," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process., Vancouver*, May 2013.
- [9] J. Li, L. Deng, D. Yu, Y. Gong and A. Acero, "High-performance HMM adaption with joint compensation of additive and convolutive distortions via vector Taylor series," in *Proc. IEEE Work. Automatic Speech Recognition and Understanding*, pp. 65-70, Kyoto, Dec. 2007.
- [10] J. Li, L. Deng, D. Yu, Y. Gong and A. Acero, "A unified framework of HMM adaption with joint compensation of additive and convolutive distortions," *Computer Speech and Language*, vol. 23, no. 3, pp. 389-405, July 2009.
- [11] Y. Gong, "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," *IEEE Trans. on Speech Audio Process.*, vol. 13, no. 5, pp. 975-983, 2005.
- [12] P. C. Loizou, *Speech Enhancement: Theory and Practice, Part II: Algorithms, Part III: Evaluation*. Taylor & Francis, 2013.
- [13] J. Benesty and J. Chen, *A Conceptual Framework for Noise Reduction, Ch. 4: Single-Channel Noise Reduction in the STFT Domain with Interframe Correlation*. Springer, 2015.
- [14] T. Esch and P. Vary, "Model-based speech enhancement using SNR dependent MMSE estimation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, pp. 4652-4655, Prague, May 2011.
- [15] T. Esch and P. Vary, "Speech enhancement using a modified Kalman filter based on complex linear prediction and supergaussian priors," in *Proc. IEEE Int. Conf. Audio and Speech Signal Process.*, pp. 4877-4880, Las Vegas, April 2008.
- [16] Y. Wang and M. Brookes, "Speech enhancement using a modulation domain Kalman filter post-processor with a Gaussian mixture noise model," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, pp. 7024-7028, Florence, May 2014.
- [17] Y. Wang and M. Brookes, "Speech enhancement using a robust Kalman filter post-processor in the modulation domain," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, Vancouver, May 2013.
- [18] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, vol. 12, 1987, pp. 177-180.
- [19] S. So and K. K. Paliwal, "Suppressing the influence of additive noise on the Kalman gain for low residual noise speech enhancement," *Speech Communication*, vol. 53, pp. 355-378, 2010.
- [20] Y. 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.
- [21] N. Dionelis and M. Brookes, "Modulation-domain speech enhancement using a Kalman filter with a Bayesian update of speech and noise in the log-spectral domain," in *Proc. IEEE Int. Work. Hands-free Speech Communication and Microphone Arrays, San Francisco*, March 2017.
- [22] N. Dionelis and M. Brookes, "Speech enhancement using modulation-domain Kalman filtering with active speech level normalized log-spectrum global priors," in *Proc. European Signal Process. Conf., Kos*, Aug. 2017.
- [23] J. Li and L. Deng and R. Haeb-Umbach and Y. Gong, *Robust automatic speech recognition: A bridge to practical applications, Ch. 3: Background of robust speech recognition*, ISBN: 978-0-12-802398-3. Elsevier, 2016.
- [24] L. Deng, J. Droppo, and A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. on Speech and Audio Process.*, vol. 12, no. 2, pp. 133-143, April 2004.
- [25] L. Deng and J. Droppo and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. on Speech and Audio Process.*, vol. 12, no. 3, pp. 218-233, May 2004.
- [26] 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.
- [27] T. Esch and P. Vary, "Exploiting temporal correlation of speech and noise magnitudes using a modified Kalman filter for speech enhancement,"

- in *Proc. Conf. Voice Communication (SprachKommunikation)*, pp. 1-4, Aachen, Oct. 2008.
- [28] B. Raj, R. Singh and R. Stern, "On tracking noise with linear dynamical system models," in *Proc. Conf. Int. Speech Communication Association*, pp. 965-968, Lisbon, Oct. 2004.
- [29] Matthias Wolfel, "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.
- [30] P. Mowlaee, R. Saeidi and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1-29, July 2016.
- [31] V. Leutnant and R. Haeb-Umbach, "An analytic derivation of a phase-sensitive observation model for noise-robust speech recognition," in *Proc. Conf. Int. Speech Communication Association*, pp. 2395-2398, Brighton, Sept. 2009.
- [32] J. Kulmer and P. Mowlaee, "Harmonic phase estimation in single-channel speech enhancement using von Mises distribution and prior SNR," in *Proc. IEEE Int. Conf. Audio and Speech Signal Process.*, Brisbane, April 2015.
- [33] P. Mowlaee, J. Kulmer, J. Stahl and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*, Ch. 4: *Phase Processing for Single-Channel Speech Enhancement*, ISBN: 978-1-119-23881-2. John Wiley & Sons, Dec. 2016.
- [34] G. Kurz, I. Gilitschenski and U. D. Hanebeck, "Recursive Bayesian filtering in circular state spaces," *IEEE A&E Systems Magazine*, vol. 31, no. 3, pp. 70-87, DOI: No. 1109/MAES.2016.150083, March 2016.
- [35] K. Paliwal, K. Wojcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465-494, 2011.
- [36] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 22, no. 12, pp. 1931-1940, Dec. 2014.
- [37] T. Gerkmann, M. Krawczyk-Becker and J. Le Roux, "Phase processing for single-channel speech enhancement, History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55-66, DOI: 10.1109/MSP.2014.2369251, March 2015.
- [38] T. Esch, "Model-based speech enhancement exploiting temporal and spectral dependencies, Ch. 3: Speech enhancement incorporating temporal correlation," Ph.D. dissertation, Aachen University, 2012.
- [39] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Trans. on Signal Processing*, vol. 62, no. 16, pp. 4199-4208, Aug. 2014.
- [40] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 129-132, Febr. 2013.
- [41] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *Signal Process. Lett.*, vol. 20, no. 12, pp. 1235-1239, Dec. 2013.
- [42] X. Wang and C. Bao, "Speech enhancement method with geometric phase estimation by incorporating MIXMAX model," in *Proc. IEEE Signal and Information Process. Association Annual Summit and Conf., Asia-Pacific, Jeju, South Korea*, Dec. 2016.
- [43] Y. Agiomyriannakis and Y. Stylianou, "Wrapped Gaussian mixture models for modeling and high-rate quantization of phase data of speech," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 17, no. 4, pp. 775-786, 2009.
- [44] J. Le Roux, E. Vincent and H. Erdogan, "Learning-based approaches to speech enhancement and separation, Section 2: The pre-deep-learning era, Slides: 1-13, and Section 3: Deep learning for single channel separation, Slides: 20-26," 2016, [Online] Available: <https://www.merl.com/publications/docs/TR2016-113.pdf>.
- [45] D. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236-243, Apr. 1984.
- [46] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. on Audio, Speech and Language Process.*, vol. 23, no. 6, pp. 982-992, March 2015.
- [47] M. Gales and S. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Communication*, vol. 12, no. 3, pp. 231-239, July 1993.
- [48] R. F. Astudillo, "Integration of short-time Fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition, Ch. 5: Fourier domain uncertainty models," Ph.D. dissertation, Technical University of Berlin, 2010.
- [49] C. S. J. Doire, M. Brookes, P. A. Naylor et al, "Single-channel online enhancement of speech corrupted by reverberation and noise," *IEEE/ACM Trans. on Audio, Speech and Language Process.*, vol. 25, no. 3, pp. 572-587, March 2017.
- [50] 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.
- [51] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Process.*, vol. 33, no. 2, pp. 443-445, April 1985.
- [52] R. Chen, C.-F. Chan and H. C. So, "Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1324-1336, May 2012.
- [53] M. Parchami, W.-P. Zhu and B. Champagne, "Speech dereverberation using weighted prediction error with correlated inter-frame speech components," *Speech Communication*, vol. 87, pp. 49-57, 2017.
- [54] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1383-1393, 2012.
- [55] G. Kurz, I. Gilitschenski, F. Pfaff and L. Drude, "libDirectional," 2015. [Online]. Available: <https://github.com/libDirectional>
- [56] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401-422, 2004.
- [57] J. Garofolo, L. Lamel, W. Fisher et al, "TIMIT acoustic-phonetic continuous speech corpus," *Corpus LDC93S1, Linguistic Data Consortium, Philadelphia*, 1993.
- [58] H. Steeneken and F. Geurtsen, "Description of the RSG-10 noise database," *TNO Institute for perception*, 1988.
- [59] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," Imperial College, Software Library, 2018. [Online]. Available: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>
- [60] ITU-T, "Objective Measurement of Active Speech Level," ITU-T Recommendation P.56, Mar. 1993.
- [61] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Rec P.862, Feb. 2001.
- [62] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low bit-rate speech coding systems," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 262-273, 1988.
- [63] Y. Hu and P. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Conf. Int. Speech Communication Association*, Pittsburgh, Sept. 2006.
- [64] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 19, no. 7, pp. 2125-2136, Sept. 2011.
- [65] C. Taal, "MATLAB code for algorithms," Software Package, 2011-2015, [Online]. Available: <http://www.ceestaal.nl/matlab-code/>.
- [66] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403-2418, 2001.
- [67] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, pp. 12-15, 2002.
- [68] N. D. Gaubitch, H. W. Lollmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Proc. IEEE Int. Work. Acoustic Signal Enhancement*, Sept. 2012.
- [69] N. Dionelis and M. Brookes, "Phase-aware single-channel speech enhancement using modulation-domain Kalman filtering," 2017, [Online]. Available: <https://www.commsp.ee.ic.ac.uk/~sap/phase-aware-speech-enhancement-using-modulation-domain-kalman-filtering/>.



Nikolaos Dionelis is a postgraduate researcher in the Speech and Audio Processing Group, which is part of the Communications and Signal Processing Group, in the Department of Electrical and Electronic Engineering at Imperial College London, London, U.K. He has been awarded with the Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Award and he is working towards his Ph.D. degree in the Speech and Audio Processing Group at Imperial College London since 2015. He received his M.Eng. degree with First Class Honours in Electrical and Electronic Engineering from Imperial College London in 2015. He completed the International Baccalaureate (IB) Diploma Programme at the Hellenic-American Educational Foundation (HAEF) Athens College - Psychico College, Athens, Greece in 2011. His current research interests include single-channel speech enhancement, speech denoising, speech separation, speech quality and intelligibility improvement, modulation-domain Kalman filtering in noisy and reverberant environments, blind speech dereverberation, speech and audio signal processing, machine learning and signal processing using statistical and probabilistic methods.



Mike Brookes (M'88) is a Reader (Associate Professor) in Signal Processing in the Department of Electrical and Electronic Engineering at Imperial College London. After graduating in Mathematics from Cambridge University in 1972, he worked at the Massachusetts Institute of Technology and, briefly, the University of Hawaii before returning to the UK and joining Imperial College in 1977. Within the area of speech processing, he has concentrated on the modelling and analysis of speech signals, the extraction of features for speech and speaker recognition and on the enhancement of poor quality speech signals. He is the primary author of the VOICEBOX speech processing toolbox for MATLAB. Between 2007 and 2012 he was the Director of the Home Office sponsored Centre for Law Enforcement Audio Research (CLEAR) which investigated techniques for processing heavily corrupted speech signals. He is currently principal investigator of the E-LOBES project that seeks to develop environment-aware enhancement algorithms for binaural hearing aids.