Article:
Jaramillo, A.E., Nielsen, J.K. and Christensen, M.G. (2023) An adaptive autoregressive pre-whitener for speech and acoustic signals based on parametric NMF. Speech Communication, 151. pp. 9-23. ISSN: 0167-6393

https://doi.org/10.1016/j.specom.2023.04.002

Speech Communication 151 (2023) 9–23

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

An adaptive autoregressive pre-whitener for speech and acoustic signals based on parametric NMF

Alfredo Esquivel Jaramillo [a,*,1], Jesper Kjær Nielsen [b], Mads Græsbøll Christensen [b,2]

[a] The University Of Sheffield, Speech and Hearing Group, Department of Computer Science, United Kingdom
[b] Aalborg University, Audio Analysis Lab, Department of Electronic Systems, Denmark

ARTICLE INFO

Dataset link: https://github.com/alfredoej87/Autoregressive-Pre-whitening-based-on-Parametric-NMF, https://github.com/alfredoej87/iterativeF0Ar_NLS

Keywords:
Colored
Pre-whitening
Enhancement
Pitch
NMF
TOA

ABSTRACT

A common assumption in many speech and acoustic processing methods is that the noise is white and Gaussian (WGN). Although making this assumption results in simple and computationally attractive methods, the assumption is often too simple and crude in many applications. In this paper, we introduce a general purpose and online pre-whitener which can be used as a pre-processor with methods based on the WGN assumption, improving their reliability and performance in applications with colored noise. The pre-whitener is a time-varying filter whose coefficients are found using a parametric non-negative matrix factorization (NMF), based on autoregressive (AR) mixture modeling of both the noise component and the signal component constituting the noisy signal. Compared to other types of pre-whiteners, we show that the proposed pre-whitener has the best performance, especially in applications with non-stationary noise. We also perform a large number of experiments to quantify the benefits of using a pre-whitener as a pre-processor for methods based on the WGN-assumption. The applications of interest were pitch estimation and time-of-arrival (TOA) estimation, where the WGN assumption is very popular.

1. Introduction

In many speech and acoustic applications, the signal of interest is contaminated with noise. To cope with this, methods or estimators designed to extract the signal (or a quantity) of interest must be robust to the noise whose level and spectral shape are often unknown a priori. Like the signal of interest, a noise model can also be elicited from which a robust, joint estimator of the signal and noise model parameters can be derived (see examples in, e.g., Quinn, 2007; Emiya et al., 2007; Yoshii and Goto, 2012; Dou et al., 2017; Quinn et al., 2021). A Gaussian noise model is popular, but estimating its covariance or its parametrization jointly with the signal model parameters often leads to intractable estimators. Moreover, this approach is not very flexible since a new estimator has to be re-derived when the noise model changes. As an alternative to the joint approach, it is possible to keep using methods which were derived based on the simple WGN assumption, provided that a pre-whitener is used as a pre-processor so that the noise color of the pre-whitened signal is approximately white. Various acoustic and speech processing methods (Christensen, 2013; Nielsen et al., 2017; Swärd et al., 2017; Feder and Weinstein, 1988; Zou and Liu, 2020; Blanco and Nájar, 2012; Al-Aboosi and Sha'ameri, 2017; Jensen et al., 2019) have assumed that the noise is WGN to retain the mathematical simplicity of the problem and to achieve a fast implementation. However, if those WGN-based methods are applied without any pre-processing, certain problems may appear. An example of this is found in pitch estimation where a pronounced noise peak at low frequencies causes the pitch estimator to produce an estimate which is an integer fraction of the true pitch (Jaramillo et al., 2019b); an estimation error which is often referred to as the subharmonic error. To combat this, applying a pre-whitener as a pre-processor is desirable since the noise will be whitened, thereby better fulfilling the model assumptions made in the WGN-based method. As we show later, however, accurate noise statistics information is needed.

The task of applying a pre-whitening scheme has been important in several areas, such as remote sensing (Jakobsson et al., 2005), sonar (Trucco, 2001), biomedical engineering (Birch et al., 1988), speech (Zhao et al., 2003; Nørholm et al., 2016; Jaramillo et al., 2021, 2018) and acoustic array processing (Okamoto et al., 2012). Pre-whitening of the noise can be performed by, e.g., applying a general

* Corresponding author.
E-mail addresses: a.e.jaramillo@sheffield.ac.uk (A.E. Jaramillo), jesperkn.research@gmail.com (J.K. Nielsen), mgc@es.aau.dk (M.G. Christensen).

[1] The work of A. Esquivel Jaramillo was supported by the National Council of Science and Technology (CONACYT) under grant 418437. The author was affiliated with Aalborg University at the time of initial submission.

[2] EURASIP member.

https://doi.org/10.1016/j.specom.2023.04.002
Received 26 November 2022; Received in revised form 13 April 2023; Accepted 19 April 2023
Available online 12 May 2023

linear transformation. This can be a matrix such as the Cholesky factor of the inverse noise covariance matrix (Jaramillo et al., 2019b; Christensen and Jakobsson, 2009) that will decorrelate the noise samples, but will also modify the signal of interest, including the frequency content (Nørholm et al., 2016). In Nørholm et al. (2016), Jaramillo et al. (2019b) it was shown how applying the Cholesky factor as a transformation to noisy signals modifies the harmonic model structure, often used in pitch estimation, of the voiced speech parts. An alternative way to pre-whiten the noise is to apply a linear filter whose amplitude response is the inverse of the spectral shape of the noise. An example is the (autoregressive) AR pre-whitener (Kay and Salisbury, 1990) which is a (finite impulse response) FIR filter whose coefficients are the AR parameters (Stoica et al., 2005) describing the noise spectral shape. This filter corresponds to the classical prediction error filter (Therrien, 1992), and its application only modifies the sinusoidal amplitudes and phases of the desired signal and not its frequency content (asymptotically) (Jaramillo et al., 2019b). Therefore, model-based estimators assuming WGN such as the (nonlinear least squares) NLS pitch estimator (Quinn and Thomson, 1991; Christensen, 2013; Nielsen et al., 2017) can be reliably used after the signal has been pre-processed with that pre-whitening filter. It should be noted that fitting the noise (power spectral density) PSD with a smoothed spectrum, as the AR spectrum is, is preferable to directly using the estimated noise PSD coefficients to generate the FIR filter that counteracts the noise spectral shape, since this option could possibly lead to inaccurate estimates (Jaramillo et al., 2019b).

This paper proposes an adaptive pre-whitener for speech and acoustic signals, which is auto-regressive (AR) and is based on parametric NMF. Particularly, the required noise statistics used to derive the pre-whitening filter coefficients are obtained from the model introduced in Kavalekalam et al. (2018), in which both the signal and noise are modeled as a sum of time-varying AR processes. The estimation of the parameters of this model was performed using parametric NMF (Kavalekalam et al., 2018) which is a generalization of traditional NMF of superimposed Gaussian sources (Févotte et al., 2009). AR-dictionaries were pre-trained offline on typical envelopes of speech and noise sources represented by AR parameters. Given the pre-trained AR-dictionaries, the parametric NMF method continuously re-computes the activation coefficients, which are the excitation noise variances of the pre-trained AR-spectra. The proposed pre-whitener can be applied to many different problems to simplify the computations and the paper looks at some of these problems to demonstrate this. The NLS pitch estimator (Christensen, 2013; Nielsen et al., 2017) achieves a better performance when the proposed pre-whitener is used as a pre-processor as compared to a pre-whitener based on a noise PSD tracker (Martin, 2001; Gerkmann and Hendriks, 2012). The solution of the cascade of the AR pre-whitener with the NLS pitch estimator can be further refined by post-processing the initial estimates through iterative refinement, leading to an improved accuracy. Preliminary work was presented in Jaramillo et al. (2019a), Jaramillo et al. (2020). The present manuscript extends on such previous work, providing a comprehensive and in-depth experimental study of the introduced pre-whitener. Additional performance measures are assessed and in addition to the NLS pitch estimator, the influence of the pre-whitener on other methods, such as well-known speech processing algorithms (e.g., non-parametric pitch estimators), a recently introduced Bayesian parametric pitch estimator and source localization methods (TOA estimation), is exemplified. The remainder of this paper is organized as follows. Section 2 introduces the related work. We describe how to obtain the AR pre-whitening filter on a segment-by-segment basis in Section 3. In Section 4, we describe how to estimate the required noise statistics for the pre-whitening filter. The experimental setting for the applications of interest, the procedure for training the dictionaries, the performance measures and the discussion of the observed results are presented in Section 5. Section 6 concludes the work.

2. Related work

To cope with the non-stationary nature of both the signal of interest and the noise, a pre-whitener should update its parameters on a segment-by-segment basis (Christensen and Jakobsson, 2009). In the literature, as the noise statistics are unknown, the parameters of the pre-whitener are often determined only from segments in which the desired signal is absent, i.e., where only the noise is present. For example, in a sonar application to detect a low-Doppler target (Kay and Salisbury, 1990), an AR pre-whitener obtained its parameters only when the reverberation was assumed to be present, thus ignoring a more realistic scenario in which both the reverberation and the signal of interest coexist. Similarly, in Hansen and Jensen (2007), the noise statistics were only computed during silent periods obtained from a voice activity detector (VAD) (Sohn et al., 1999). Other works (Huang and Zhao, 1998) have assumed that the noise AR parameters describing the noise spectral shape are known beforehand, but this is unrealistic in non-stationary noise cases where the noise spectral shape changes quickly between segments.

For non-stationary noise, the noise statistics may change significantly during speech presence, and this will not be tracked when a VAD is used, potentially leading to a poor performance of the pre-whitener as well as the estimator assuming WGN. During speech presence, information about the noise spectrum can be tracked across time using various well-known state-of-the art methods (e.g., minimum statistics (MS) Martin, 2001 and MMSE based on speech presence probabilities (SPP) Gerkmann and Hendriks, 2012). Pre-whitening reliant on these approaches results in a good performance when the noise is stationary, but not when the noise is highly non-stationary. For pitch estimation, this was demonstrated in Jaramillo et al. (2019b).

To cope with noise statistics estimation in nonstationary noise, a model-based estimator (Nielsen et al., 2018) using *a priori* spectral information about typical speech and noise AR parameters stored in codebooks has been found to improve the noise PSD estimation accuracy compared to traditional noise trackers. As opposed to codebook-based approaches which use a log-spectral distortion approximation and noise classification (Srinivasan et al., 2006, 2007), multiplicative-update (MU) based approaches (He et al., 2017) result in more accurate excitation variance estimates, i.e., they capture better the noise spectral envelope. For the introduced pre-whitener based on parametric NMF, the minimization of the spectral distance between the periodogram and the modeled PSD leads to an MU rule of the activation coefficients. The parametric NMF method differs from unsupervised approaches such as (Févotte et al., 2009) by parametrizing the dictionary with normalized AR-envelopes. Thus, in parametric-NMF, only the activation coefficients are estimated, as the so-called spectral basis vectors are pre-trained offline and kept fixed during inference.

3. AR pre-whitener

This section describes the principle of how an AR pre-whitening filter is applied when the noise statistics are available. The next section provides details on how such noise statistics are estimated. An observed signal $x(n)$ is assumed to be formed by the mixture of a signal of interest (e.g., speech) $s(n)$ and a colored noise signal $c(n)$, i.e.,

$$x(n) = s(n) + c(n) \ . \tag{1}$$

Furthermore, we assume that $c(n)$ is well modeled as an AR process, i.e.,

$$c(n) = -\sum_{i=1}^{P} w_c(i)c(n-i) + e(n) \ . \tag{2}$$

where $e(n)$ is white Gaussian excitation noise of variance σ_e^2, $\{w_c(i)\}_{i=1}^{P}$ denote the AR parameters which describe the spectral shape of the colored noise, and P is the AR order. The generative noise model is

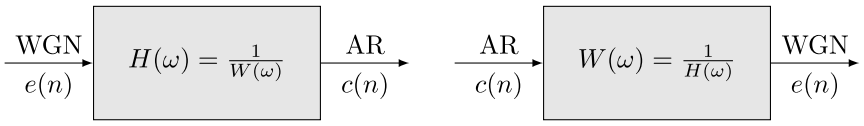

Fig. 1. Generative noise model (left) and whitening FIR filter (right).

illustrated in the left part of Fig. 1), and the model to recover the white Gaussian excitation noise samples given colored noise samples is on the right part. Such filter $W(\omega)$ is a whitening filter, and the prefix "pre" denotes that it is applied before some other method.

The parameters σ_e^2 and $\{w_c(i)\}_{i=1}^{P}$ are seldom known and they can be obtained from noise statistics, namely the noise covariance sequence $\{r_c(i)\}_{i=0}^{P}$, by solving the Yule–Walker equations (Stoica et al., 2005), using the Levinson–Durbin recursion (Stoica et al., 2005; Therrien, 1992). If instead N samples from the noise PSD $\{\Phi_c(k)\}_{k=0}^{N-1}$ is available, the noise covariance sequence can be computed as (Stoica et al., 2005)

$$r_c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \Phi_c(k) \exp\left(j\frac{2\pi}{N}nk\right), \quad 0 \le n \le P\ . \tag{3}$$

Due to the time-varying noise statistics, the AR parameters will be time-varying. In practice, we implement this by dividing the data into overlapping segments, each of length N. Given such N data samples

$$\mathbf{x}(l) = \begin{bmatrix} x(0,l) & x(1,l) & \cdots & x(N-1,l) \end{bmatrix}^T, \tag{4}$$

with $(\cdot)^T$ denoting transpose, and time-varying AR parameters $\mathbf{w}_P(l)$ in segment l, the pre-whitener is implemented in the frequency domain,[3]. That is, the discrete Fourier transform (DFT) of the pre-whitened signal is computed as

$$\hat{X}_W(k,l) = W(k,l)X(k,l) \tag{5}$$

where $W(k,l)$ and $X(k,l)$ are the kth bin of the N-length DFT of time-varying AR parameters $\mathbf{w}_P(l)$ and the data segment $\mathbf{x}(n+lM)v(n)$, respectively, where M denotes the hop size in samples between segments and $v(n)$ is the analysis window. The whitened signal in the time domain $x_W(n,l)$ is then obtained by computing an inverse DFT of $\{\hat{X}_W(k,l)\}_{k=0}^{N-1}$. As the processing is done on overlapping segments, a synthesis window $v(n)$ is applied to update the full pre-whitened signal as $x_W(n+lM) = x_W(n+lM) + v(n)x_W(n,l)$.

4. Noise PSD estimation based on parametric NMF

As mentioned above, segment-wise estimates of the noise PSD $\Phi_c(k)$ are required to compute the AR-coefficients used in the pre-whitening filter. In this section, we describe how the noise PSD is estimated in the proposed pre-whitener from a segment of data. Note that we omit the segment index l in this section to simplify the notation. To get good performance in even non-stationary noise conditions, we here propose that the noise PSD estimate is obtained by taking typical spectral shapes of speech and noise into account. For this purpose, we model the data vector in (4) as a summation of U AR processes $\{\mathbf{t}_u\}_{u=1}^{U}$ where each AR-process describe a typical spectral shape. Specifically, the data vector is modeled as

$$\mathbf{x} = \sum_{u=1}^{U}\mathbf{t}_u = \sum_{u=1}^{U_s}\mathbf{t}_u + \sum_{u=U_s+1}^{U}\mathbf{t}_u, \tag{6}$$

where the first U_s AR processes model clean signals (e.g., speech), and the last $U_c = U - U_s$ AR processes model noise signals. A stationary and stable AR process can be described as a realization from a multivariate Gaussian probability density function (pdf) (Srinivasan et al., 2007), i.e., $\mathbf{t}_u \sim \mathcal{N}(\mathbf{0}, \sigma_u^2\mathbf{R}_u(\mathbf{a}_u))$, where σ_u^2 is the excitation variance, $\mathbf{R}_u(\mathbf{a}_u)$ is its gain normalized covariance matrix, and

$$\mathbf{a}_u = \begin{bmatrix} 1 & a_u(1) & \cdots & a_u(P') \end{bmatrix}^T \tag{7}$$

is the vector containing the AR parameters of the uth spectral basis. Here P' is the AR order.

The likelihood of the observation $\mathbf{x}$ as a function of U excitation variances and U spectral shapes is given by

$$p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) \sim \mathcal{N}\left(\mathbf{0}, \sum_{u=1}^{U}\sigma_u^2\mathbf{R}_u(\mathbf{a}_u)\right) \tag{8}$$

where

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_U^2 \end{bmatrix}^T \tag{9}$$

is a $U \times 1$ vector containing the U excitation variances and is referred to as the vector of activation coefficients. The matrix $\mathbf{D}$ of dimension $N \times U$ is referred as either the spectral basis matrix or the AR dictionary, and its column vectors are the U gain normalized PSDs parametrized by the AR parameters, i.e.,

$$\mathbf{D} = \begin{bmatrix} d_1(0) & \cdots & d_u(0) & \cdots & d_U(0) \\ \vdots & & \vdots & & \vdots \\ d_1(k) & \cdots & d_u(k) & \cdots & d_U(k) \\ \vdots & & \vdots & & \vdots \\ d_1(N-1) & \cdots & d_u(N-1) & \cdots & d_U(N-1) \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{d}}_0^T \\ \vdots \\ \tilde{\mathbf{d}}_k^T \\ \vdots \\ \tilde{\mathbf{d}}_{N-1}^T \end{bmatrix} \tag{10}$$

$$= \begin{bmatrix} \mathbf{d}_1 & \cdots & \mathbf{d}_u & \cdots & \mathbf{d}_U \end{bmatrix}, \tag{11}$$

where $\tilde{\mathbf{d}}_k^T$ and $\mathbf{d}_u$ are the kth row and uth column of $\mathbf{D}$, respectively. As shown in the Appendix, the (k,u)th element of $\mathbf{D}$ is given by

$$d_u(k) = \frac{1}{\left|1 + \sum_{i=1}^{P'} a_u(i)\exp(-\frac{2\pi jik}{N})\right|^2} \tag{12}$$

which is the kth bin of the uth gain normalized PSD. The U different sets of AR parameters $\{a_u(i)\}_{i=1}^{P'}$ are obtained from a training stage which is detailed in the next section. The matrix $\mathbf{D}$ can be partitioned as $\mathbf{D} = \begin{bmatrix} \mathbf{D}_s & \mathbf{D}_c \end{bmatrix}$, where $\mathbf{D}_s$ of size $N \times U_s$ contains only the U_s signal spectral envelopes, and $\mathbf{D}_c$ of size $N \times U_c$ contains only the U_c noise spectral envelopes. The kth row of $\mathbf{D}$ can be partitioned similarly, and we write this as $\tilde{\mathbf{d}}_k^T = \begin{bmatrix} \tilde{\mathbf{d}}_{s,k}^T & \tilde{\mathbf{d}}_{c,k}^T \end{bmatrix}$.

The AR parameters describing the spectral shapes contained in $\mathbf{D}$ are obtained offline. Thus, only the activation coefficients in $\boldsymbol{\sigma}$ have to be estimated online which we do by maximizing the likelihood in (8) w.r.t. $\boldsymbol{\sigma}$, i.e.,

$$\hat{\boldsymbol{\sigma}} = \arg\max_{\boldsymbol{\sigma}\ge 0}\ p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) = \arg\max_{\boldsymbol{\sigma}\ge 0}\ \mathcal{N}\left(0, \sum_{u=1}^{U}\sigma_u^2\mathbf{R}_u(\mathbf{a}_u)\right)\ . \tag{13}$$

As shown in the Appendix, the log-likelihood function can be expanded as

$$\ln p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}D_{\mathrm{IS}}(\boldsymbol{\Phi}|\mathbf{D}\boldsymbol{\sigma}) - \frac{1}{2}\sum_{k=0}^{N-1}\ln\Phi(k) + \frac{N}{2} \tag{14}$$

where we have defined

$$\Phi(k) = \frac{1}{N}|X(k)|^2 = \frac{1}{N}\left|\sum_{n=0}^{N-1}x(n)\exp\left(-j2\pi\frac{nk}{N}\right)\right|^2 \tag{15}$$

[3] In Jaramillo et al. (2019b,a) the pre-whitening filter was applied in the time domain.

$$\boldsymbol{\Phi} = \begin{bmatrix} \Phi(0) & \cdots & \Phi(N-1) \end{bmatrix}^T \tag{16}$$

$$D_{\text{IS}}(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2) = \frac{1}{N}\sum_{k=0}^{N-1}\left(\frac{\psi_1(k)}{\psi_2(k)} - \ln\frac{\psi_1(k)}{\psi_2(k)} - 1\right). \tag{17}$$

The function $D_{\text{IS}}(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)$ is the Itakura–Saito distance (ISD) measure between two discrete spectra $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ (Itakura, 1968). As $\sum_{k=0}^{N-1}\ln\Phi(k)$ does not depend on $\boldsymbol{\sigma}$, maximizing the likelihood with respect to $\boldsymbol{\sigma}$ simply corresponds to minimizing the ISD between the periodogram $\boldsymbol{\Phi}$ and the modeled PSD $\mathbf{D}\boldsymbol{\sigma}$, under the constraint that $\Phi(k) \geq 0\ \forall k$. That is, the maximum likelihood (ML) estimate of $\boldsymbol{\sigma}$ is obtained by solving the supervised non-negative matrix factorization (NMF) problem

$$\hat{\boldsymbol{\sigma}} = \begin{bmatrix} \hat{\boldsymbol{\sigma}}_s^T & \hat{\boldsymbol{\sigma}}_c^T \end{bmatrix}^T = \arg\min_{\boldsymbol{\sigma}\geq 0} D_{\text{IS}}(\boldsymbol{\Phi}|\mathbf{D}\boldsymbol{\sigma}). \tag{18}$$

We remark here that unlike in Févotte et al. (2018), the matrix $\mathbf{D}$ is parametrized by pre-trained AR-envelopes for which reason the problem here is referred to as parametric NMF (Kavalekalam et al., 2018). The optimization problem can easily be extended to the case where V overlapping segments are available. For such case, the modeled PSD can be written as the matrix product $\mathbf{D}\boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ of dimension $U \times V$ is the activation matrix containing the activation coefficients of a segment as a column vector, i.e., $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\sigma}(1) & \cdots & \boldsymbol{\sigma}(V) \end{bmatrix}$.

Focusing on estimating $\boldsymbol{\sigma}$ for a single segment from (18), it is well-known in the NMF-literature that (18) cannot be solved analytically. The problem is reminiscent of NMF under the β divergence, with $\beta = 0$, and approaches based on maximization–minimization or heuristic algorithms based on multiplicative updates (MU) (Févotte and Idier, 2011) can be used to iteratively approach a solution. Typically, the heuristic algorithm based on MU is adopted as it requires less iterative updates for convergence. As described in Févotte and Idier (2011), under the $\beta = 0$ condition, each iterative update leads to a decrease of the objective function in (18). Specifically, the value of the variable of interest at the $(i+1)$th iteration is updated by multiplying its value at the previous ith iteration by the ratio of the negative part to the positive part of the gradient of the criterion with respect to this variable. Thus, $\boldsymbol{\sigma}$ is computed iteratively from

$$\hat{\boldsymbol{\sigma}}^{(i+1)} \leftarrow \hat{\boldsymbol{\sigma}}^{(i)} \odot \frac{\mathbf{D}^T\left(\mathbf{D}\hat{\boldsymbol{\sigma}}^{(i)}\right)^{[-2]} \odot \boldsymbol{\Phi}}{\mathbf{D}^T\left(\mathbf{D}\hat{\boldsymbol{\sigma}}^{(i)}\right)^{[-1]}}, \tag{19}$$

where $\leftarrow$ refers to an iterative overwrite, $\odot$ denotes element-wise multiplication, $(\cdot)^{[-2]}$ denotes element-wise inverse squared operator and $(\cdot)^{[-1]}$ denotes element-wise inverse operator. The division is also element-wise, and the number of iterations is denoted as I.

Having computed $\hat{\boldsymbol{\sigma}} = \begin{bmatrix} \hat{\boldsymbol{\sigma}}_s^T & \hat{\boldsymbol{\sigma}}_c^T \end{bmatrix}^T$, we can compute an SNR estimate as

$$\xi(k) = \frac{\lambda_s^2(k)}{\lambda_c^2(k)} \tag{20}$$

where

$$\lambda_s^2(k) = \tilde{\mathbf{d}}_{s,k}^T\hat{\boldsymbol{\sigma}}_s \tag{21}$$

$$\lambda_c^2(k) = \tilde{\mathbf{d}}_{c,k}^T\hat{\boldsymbol{\sigma}}_c\ . \tag{22}$$

In the noise PSD estimation literature, these quantities are often referred to as the *a priori* SNR, the prior speech PSD, and the prior noise PSD, respectively. Given values for these quantities, it can be shown that the MMSE estimator of the noise PSD is Gerkmann and Hendriks (2012)

$$\Phi_c(k) = \left(\frac{1}{1+\xi(k)}\right)^2\Phi(k) + \left(\frac{\xi(k)}{1+\xi(k)}\right)\lambda_c^2(k), \tag{23}$$

for $k = 0, \ldots, N-1$. The differences between the estimate in (23) and the MMSE-based estimate in Gerkmann and Hendriks (2012) are how the a priori SNR and the prior noise PSD are computed. While we here obtain values for these via the parametric NMF method, the approach in Gerkmann and Hendriks (2012) relies on speech presence probabilities (SPPs).

Algorithm 1 Proposed Pre-whitening for a single segment, based on parametric NMF noise PSD estimate, assuming U_s signal and U_c noise spectral envelopes whose columns given by (12) are contained on $\mathbf{D}$.

1: Obtain $\Phi(k)$, $k = 0, ..., N-1$ from (15) ▷$\mathcal{O}(N \log N)$
2: Estimate MMSE-SPP noise PSD (Gerkmann and Hendriks, 2012) and fit it to an AR spectrum. Augment $\mathbf{D}$ with envelope whose elements are given by (24) ▷$\mathcal{O}[(N+P)\log N] + \mathcal{O}(P^2)$
3: Initialize $\hat{\sigma}^{(i)}$ with random positive numbers ▷$\mathcal{O}(1)$
4: **for** `i=1:I` **do** ▷$\mathcal{O}(UNI)$
5: Compute $\hat{\sigma}^{(i)}$ using (19)
6: **end for**
7: Compute $\lambda_s^2(k)$ and $\lambda_c^2(k)$, for $k = 0, ..., N-1$, from (21) and (22) ▷$\mathcal{O}(UN)$
8: Obtain $\xi(k)$ from (20), for $k = 0, ..., N-1$ ▷$\mathcal{O}(N)$
9: Estimate $\Phi_c(k), k = 0, ..., N-1$ from (23) ▷$\mathcal{O}(N)$
10: Fit noise PSD to AR spectrum of order P. The pre-whitening filter is $W(\omega) = 1 + \sum_{i=1}^{P} w_c(i)e^{-j\omega i}$ ▷$\mathcal{O}(P\log N) + \mathcal{O}(P^2)$

To add robustness to cases where the observed noise samples are not well-represented by the pre-trained spectral envelopes in $\mathbf{D}$, it can be augmented with a single time-varying entry corresponding to the normalized AR spectral envelope that is fitted to the MMSE-SPP (Gerkmann and Hendriks, 2012) noise PSD based pre-whitener $\{w_{\text{MMSE-SPP}}(i)\}_{i=1}^{P'}$, in which each frequency-bin entry is given by

$$d_{\text{MMSE-SPP}}(k) = \frac{1}{\left|1 + \sum_{i=1}^{P'} w_{\text{MMSE-SPP}}(i)\exp\left(-\frac{2\pi jik}{N}\right)\right|^2}. \tag{24}$$

A summary on how the pre-whitening filter is updated for a single segment is outlined in Alg. 1. Note that the computational complexity of each step is given using big $\mathcal{O}$ notation. A block diagram of the pre-whitening method based on parametric NMF is shown in Fig. 2. The proposed pre-whitening method has a time complexity of $\mathcal{O}[(N+P)\log N] + O(P^2) + \mathcal{O}(NUI)$, while pre-whitening based on MMSE-SPP and MS has simply an order of $\mathcal{O}[(N+P)\log N] + \mathcal{O}(P^2)$.

5. Experimental setup and results

In this section, we present an extensive performance evaluation of the proposed pre-whitener on real signals under different colored noise scenarios. Except for the last experiment, which is concerned with time-of-arrival (TOA) estimation, we focus on speech processing problems. Specifically, the results of the following experiments are presented.

1. We seek to answer if whitening the noise using a pre-whitener is preferable to removing the noise using a speech enhancement (or noise reduction) algorithm (Huang et al., 2020). Specifically, we evaluated the accuracy of the nonlinear least squares (NLS) pitch estimator (Christensen, 2013; Nielsen et al., 2017), which is optimal under a WGN assumption, when its input speech signal has either been pre-whitened or enhanced. The comparison also included the baseline approach where no pre-processing is performed.
2. We demonstrate that the proposed pre-whitener outperforms other pre-whiteners in terms of whiteness and spectral distance to an oracle pre-whitener. The oracle pre-whitener is the pre-whitener obtained from the AR parameters computed directly from the noise signal.
3. We investigate how the pre-whitening performance depends on the AR-order and the number of spectral shapes of the pre-trained dictionaries.

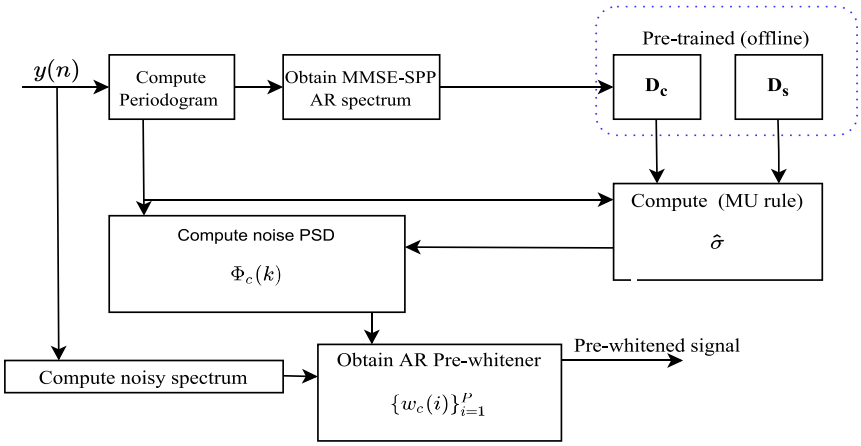

Fig. 2. The block diagram of a parametric NMF-based AR pre-whitener.

4. We aimed to verify that a better estimation accuracy of the NLS pitch estimator could be obtained when the signal was pre-processed with the proposed pre-whitener, especially in non-stationary noise conditions. The comparison also included the case in which a fixed (i.e., non-adaptive) pre-whitener is applied. Moreover, for a fairer comparison to typical non-parametric pitch estimators (e.g., RAPT Talkin, 1995), we then conducted an experiment in which we applied either speech enhancement or pre-whitening before the pitch was estimated with those classical approaches, thus allowing us to determine whether there is a greater benefit with certain types of pre-processing. The computational complexity of the different pre-processing approaches was also evaluated.
5. We applied a last stage of post-processing in order to contrast the performance to individual pitch estimators.
6. Finally, the last experiment dealt with TOA estimation, and it was assessed how much the proposed pre-whitener improved the estimation accuracy.

5.1. Codebook training

As we have alluded to earlier, the matrix **D** in , containing spectral envelopes of typical speech and noise segments, must first be obtained via a training step. The spectral envelopes are determined from autoregressive parameters which were obtained by using a standard vector quantization technique of speech coding. Specifically, the generalized Lloyd algorithm (Linde et al., 1980) was used to obtain cluster centers of line spectral frequency (LSF) coefficients of order $P' = 12$ computed from a large number of windowed data segments. The LSF parametrization was used in the clustering to ensure that the cluster centers corresponded to stable AR-processes. The obtained cluster centers were converted into AR parameters of order $P' = 12$. The collection of cluster centers converted into AR-parameters are often referred to as a codebook (Srinivasan et al., 2006, 2007). A speech codebook of U_s entries was obtained from training on 54 minutes of sentences uttered by four speakers (two male and two female) from the CMU Arctic database (Kominek and Black, 2004), which were re-sampled from 16 to 8 kHz. We note that another database was used in the evaluation. Similarly, a noise codebook of U_c entries was obtained from training samples from the NOISEX-92 database (Varga and Steeneken, 1993), and we used the noise types babble, factory, F-16 and street, all resampled to 8 kHz. The duration of segments for the training was 32 ms. The segments were windowed using a Hanning window with 50% overlap between adjacent segments. Examples of speech and noise spectral envelopes are illustrated in Rosenkranz (2010). The codebook sizes U_s and U_c are intentionally kept as variables as we evaluate the pre-whitening performance for different codebook sizes.

5.2. Performance measures

To compare the different pre-whiteners, both the spectral flatness measure (SFM) and the Itakura–Saito distance (ISD) are used. The SFM is defined as the ratio between the geometric mean and the arithmetic mean of a PSD (Madhu, 2009), i.e., as

$$\text{SFM} = \frac{\sqrt[N]{\prod_{k=0}^{N-1} \boldsymbol{\Phi}_{c,w}(k)}}{\frac{1}{N}\sum_{k=0}^{N-1} \boldsymbol{\Phi}_{c,w}(k)}. \tag{25}$$

where $\boldsymbol{\Phi}_{c,w}$ is the noise PSD of the pre-whitened noise. The SFM indicates how correlated the noise samples are and, therefore, the degree of coloring. A value of 0 means that the noise is very correlated (colored), whereas an SFM of 1 means that the noise is perfectly white (the samples are perfectly uncorrelated). Therefore, an important goal of a pre-whitener is to increase the SFM, and we can also use the SFM to quantify the performance of a pre-whitener. Another approach to quantifying pre-whitening performance is to measure a spectral distance between a pre-whitener and the oracle pre-whitener. We here measure this spectral distance using the ISD defined in (17). Note that both the SFM and ISD are computed on a segment-by-segment basis and averaged over the test set.

While the SFM and ISD can be used to evaluate a pre-whitener directly, we can also evaluate it indirectly by measuring the performance improvement of the estimator cascaded with a pre-whitener. For pitch estimation, typical performance measures are Chu and Alwan (2009):

- Gross Error Rate (*GER*): GER is defined as

$$GER = \frac{N_g}{N_{VV}} \times 100\ \% \tag{26}$$

  where N_{VV} is the number of voiced segments and N_g is the number of voiced segments in which the magnitude of the relative difference between the estimate and the ground truth is greater than a threshold. Here, we used a relative threshold of 20%. Note that only segments which are correctly classified as being voiced are included in the GER.
- Voicing Detection Error (*VDE*): VDE is defined as

$$VDE = \frac{N_{VU} + N_{UV}}{N} \times 100\ \% \tag{27}$$

  where N_{VU}, N_{UV}, and N are the number of segments misclassified as voiced, the number of segments misclassified as not-voiced (i.e., as unvoiced or pauses), and the total number of segments, respectively.
- Full Frame Error (*FFE*) (Chu and Alwan, 2009): FFE is defined as

$$FFE = \frac{N_{VU} + N_{UV} + N_g}{N} \times 100\ \% \tag{28}$$

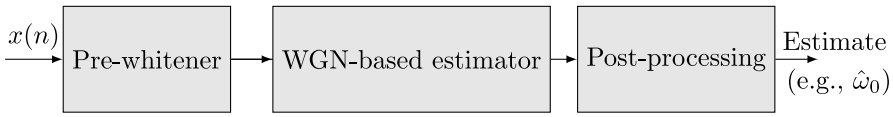

Fig. 3. Structure diagram for obtaining estimates on colored noise scenarios based on a WGN method.

where the different quantities are the same as in the GER and VDE. Note that FFE is a composite metric which is simply the sum of the VDE and the FFE when all segments are voiced and correctly classified as being voiced (i.e., when $N = N_{VV}$).

5.3. Experimental results with the Keele speech database

The set of experiments in this subsection was conducted on the Keele database (Plante et al., 1995), which consists of speech recordings of around 40 seconds from five male and five female speakers. The signals were resampled from 20 kHz to 8 kHz. Pitch estimates[4] extracted from laryngograph measurements segmented into 26.5 ms frames with 16.5 ms overlap are available in the database, and we treat these as being the ground truth estimates. We used the same segment length and overlap for the pitch estimators. Note that we in the evaluation have ignored segments for which the ground truth estimate has been labeled unreliable. These segments represents only approximately 3 % of the total number of segments.

Different noise types such as babble, factory, F-16 and street noise from the NOISEX-92 database (Varga and Steeneken, 1993) were added at different values of iSNR. The iSNR indicates the power level of the clean speech signal relative to the noise power component, i.e.,

$$\text{iSNR} = \frac{\sigma_s^2}{\sigma_c^2}, \tag{29}$$

where σ_s^2 is the variance of the speech signal, and σ_c^2 is the noise signal variance. The samples used for the testing were different from those used in the training of the noise codebooks. In addition, samples of restaurant noise from the Aurora database (Hirsch and Pearce, 2000), already sampled at 8 kHz, were used in the evaluation to assess the robustness against new encountered noise types.

5.3.1. Comparison to no pre-processing and to speech enhancement

First, the accuracy of a WGN-based method, namely the nonlinear least squares (NLS) pitch ω_0 estimator (Nielsen et al., 2017; Christensen, 2013), was assessed for the cases where the input signal to the estimator is either pre-whitened, enhanced using a speech enhancement method (an approach suggested in Huang et al., 2020), or unprocessed. Fig. 3 illustrates the case where a pre-whitener is used as a pre-processor. Note that a final post-processing step can be used to refine the initially obtained parameter estimates. Such post-processing step is ignored in all the subsections, except the last one, as we want first to verify that a pre-whitener applied as a pre-processor will result in a better accuracy of the pitch estimator. The NLS estimator of ω_0 (Christensen, 2013; Quinn and Thomson, 1991) corresponds to the ML estimator under the WGN assumption and is given by

$$\hat{\omega}_0 = \arg\max_{\omega_0} \ \underline{\mathbf{x}}^T \mathbf{Z}_L(\omega_0) \left[\mathbf{Z}_L^H(\omega_0)\mathbf{Z}_L(\omega_0)\right]^{-1} \mathbf{Z}_L^H(\omega_0)\underline{\mathbf{x}}, \tag{30}$$

where $\mathbf{Z}_L(\omega_0) = \left[\mathbf{z}(\omega_0)\ \mathbf{z}^*(\omega_0)\ \ldots\ \mathbf{z}^*(\omega_0 L)\right]$ is a Fourier matrix constructed from $2L$ complex exponential vectors $\mathbf{z}(\omega_0 l) = [1\ e^{j\omega_0 l}\ \ldots\ e^{j\omega_0 l(N-1)}]^T$. Here, $\underline{\mathbf{x}}$ is the vector used in the estimation, either the vector of noisy speech (i.e., discarding the pre-processing block), or enhanced speech or pre-whitened speech. An important feature of this problem is that it jointly estimates ω_0 and the number of real sinusoids L. The frequency of each sinusoid is an integer multiple of the fundamental ω_0, in contrast to the case of independent sinusoids, which are not harmonically related (Quinn, 2007). To obtain L, some Bayesian model comparison methods (e.g., based on maximum a posteriori) (Stoica and Selen, 2004) can be used to find the most likely model order, after estimates of ω_0 have been obtained for all candidate model orders. The model comparison is the key in reducing the sub-harmonic error problems, such as doublings or halvings. Such model comparison also includes the case $L = 0$, i.e., it is possible to do voicing detection. In all the experiments related to pitch estimation, the pitch range was [60,400] Hz, and a maximum model order of $L = 30$ harmonics is set for the model comparison.

To ensure that the differences in noise PSD estimates do not influence the result, both the applied AR pre-whitener and the speech enhancement method were based on the introduced parametric NMF (Par-NMF) noise PSD estimate. In the first experiment, $U_s = 32$ and $U_c = 16$ pre-trained spectral shapes were used, whereas we assessed the performance as a function of the number of entries in the second experiment. The iSNR was varied from −5 to 10 dB, and three Monte-Carlo simulations (MCS) were run for each noise type at each iSNR for each file from the Keele database. In each MCS, the noise samples were randomly selected, as there are more available noise samples, for each noise type, than the number of samples of each file from the Keele database. The performance measures in (26)–(28) were computed and are depicted in Fig. 4 with 95 % confidence intervals. The pre-whitening order was set to $P = 30$, and the pre-whitening filter coefficients were updated on segments of length 32 ms, with a time shift of 16 ms between them. $P = 30$ was empirically chosen after verifying better performance in (26)–(28) than when using $P = 16$ at an iSNR of −5 dB for all noise types. The pre-processing based on speech enhancement was performed with the optimally modified LSA (OM-LSA) speech estimator (Cohen, 2002). The average number of pitch errors was very high when $\underline{\mathbf{x}}$ is the unprocessed input signal, even in high SNR conditions. When $\underline{\mathbf{x}}$ is produced by the speech enhancement method, the pitch estimation accuracy improved considerably in most cases compared to when the input signal was unprocessed. When $\underline{\mathbf{x}}$ is the pre-whitened input signal, this gives the overall best performance, as noted from the non-overlapping confidence intervals between speech enhancement and pre-whitening. This is more evident at lower iSNRs, but still a considerable gap is seen at high SNRs, specially for non-stationary noise types, such as babble and restaurant noise.

5.3.2. Comparison of AR pre-whiteners

We investigated the pre-whitening performance of AR pre-whiteners based on three noise PSD estimates: MS (Martin, 2001), MMSE-SPP (Gerkmann and Hendriks, 2012), and the proposed Par-NMF based approach. Since the codebooks were trained on segments of 32 ms, overlapped by 50 %, the same segment length and overlapping percentage were used in the different pre-whiteners. The SFM of the pre-whitened noise in (25) and the ISD between the frequency responses of the oracle and the estimated pre-whiteners were evaluated.

First, we studied the performance as a function of the AR-order P. The iSNR used in this setup was 0 dB. The SFM results include four curves, as the performance from applying the oracle pre-whitener is also included while the ISD plots only involve three curves, as comparing the response of the oracle pre-whitener to itself leads to an ISD of 0. $U_s = 32$ and $U_c = 16$ pre-trained spectral shapes were used. Before pre-whitening, the average SFM values at all the iSNRs were: babble noise (0.065), factory noise (0.045), restaurant noise

[4] The pitch estimates were computed using the RAPT (Talkin, 1995) pitch estimation method and manually checked afterwards.

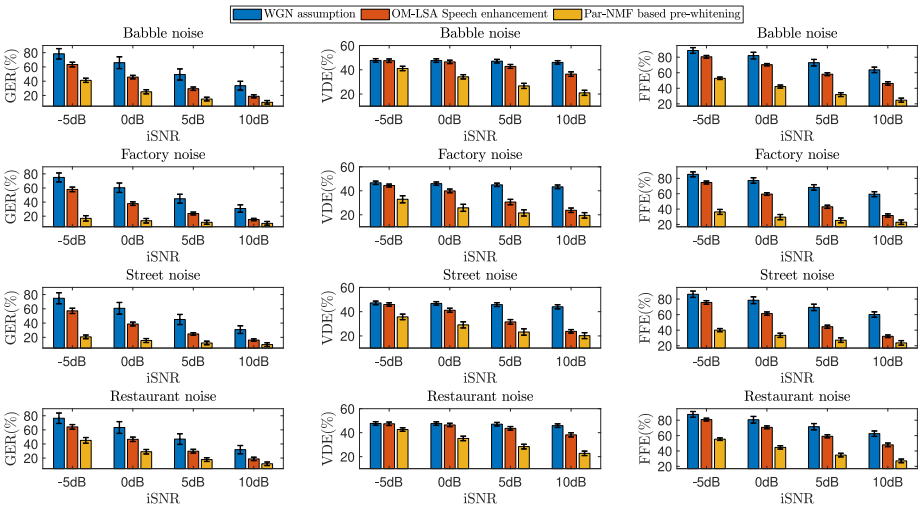

Fig. 4. The gross pitch errors, voicing detection errors and full frame errors of the estimated pitch computed by the NLS pitch estimator, for different iSNRs, with different colored noise types, after assuming that the noise is WGN, applying speech enhancement, or applying pre-whitening.

(0.129), and F-16 noise (0.115). The results are depicted in Fig. 5. For the proposed Par-NMF based pre-whitener, the SFM increased as a function of P, but at the same time, the ISD between the oracle and the estimated pre-whitener increased. Thus, even if the noise gets closer to being white by increasing P, the spectral response of the pre-whitener becomes more different from the oracle pre-whitener response, given that fitting the estimated noise PSD to an AR spectrum of a higher order P is more prone to overfitting. With a lower P, the noise PSD can be more easily fitted to a much smoother spectral shape, which may lack some detail. The performance of pre-whitening based on MMSE (Gerkmann and Hendriks, 2012) is better than the one based on MS (Martin, 2001). The SFM from Par-NMF based pre-whitening was always higher than that based on MS and MMSE for both babble and F-16 noise. For factory noise, the Par-NMF pre-whitener can achieve better performance when P is not too low. For the restaurant noise scenario, with a $P > 30$, the Par-NMF based pre-whitener had lower ISD than the MMSE based pre-whitener. Increasing P did not appear to significantly improve the noise whiteness by pre-whitening based on MS in babble, restaurant and factory noise, and neither by pre-whitening based on MMSE in babble and factory noise. In such cases, the noise PSD could be either overestimated or underestimated at some contiguous frequency bins where an important peak of the noise spectrum was present. Thus, if not enough fine details were obtained, when fitting such noise PSD to an AR spectral shape, the modeled AR PSD revealed similar smooth characteristics for different orders P of the AR model. Therefore, when applying the pre-whitener obtained from the modeled noise AR spectrum to counteract the spectral shape of the noise component, a flatter spectrum will not necessarily be engendered if the order P is increased. An increase in the SFM from the proposed pre-whitener is possible since in several cases the AR spectral shape fitted to the Par-NMF noise PSD can become sharper for a large P. This is seen, e.g., in cases where there is more than one pronounced peak in the noise spectrum. However, from the observed results in the figure, to balance the noise whiteness and the accuracy of the pre-whitener response, we found using a value of P in the interval [30,40] to be convenient. Since a slightly lower ISD was observed by using $P = 36$ in the Par-NMF as compared to the MMSE-based pre-whitener, as opposed to the $P = 30$ case (in which the ISD is quite identical) in the restaurant noise scenario, we decided to use $P = 36$ in the subsequent experiments, instead of the used value of $P = 30$ in the first experiment.

Next, we evaluated how the performance of the Par-NMF pre-whitener depends on how many speech and noise entries are used in $\mathbf{D}$. The performance was evaluated for different combinations of noise and speech AR dictionary sizes where the speech AR dictionary $\mathbf{D}_s$ could have 2^{b_s} spectral shapes for $b_s \in \{5, 6, 7, 8\}$ and the noise AR dictionary $\mathbf{D}_c$ could have 2^{b_c} spectral shapes for $b_c \in \{3, 4, 5, 6, 7, 8, 10\}$. This will allow us to compare different combinations of U_s and U_c in terms of the pre-whitener performance. Again, the iSNR was fixed at 0 dB. The results are displayed in Fig. 6. For a small number of entries in the noise codebooks, using $U_s = 32$ speech spectral envelopes leads to a higher SFM and lower ISD for babble and restaurant noise. In such cases, a larger U_s degrades the performance, potentially due to overfitting, as was similarly seen in Bao et al. (2014) in a speech enhancement framework. However, when more entries are available in the noise codebook, the performance from using $U_s = 64$ is similar to that of $U_s = 32$. To lower computational complexity, we keep with the setup of $U_s = 32$. An important observation is that speech codebooks trained from a larger or smaller dataset could slightly change the performance, and this deserves future investigation. From the current codebook configuration, the combination of $U_s = 32$ with $U_c \geq 128$ spectral shapes lead to the best performance for both babble and restaurant noise, although increasing from 256 to 1024 noise spectral envelopes, did not decrease the ISD significantly, as the confidence intervals overlapped. In the F-16 noise scenario, there is not a noticeable difference in performance in most cases, except when a very large number of $U_s = 256$ speech shapes is used. Using $U_c = 16$ entries seemed to be enough for factory and F-16 noise types, although using a higher number of entries did not degrade the performance too much. We therefore used both combinations $U_s = 32, U_c = 16$ and $U_s = 32, U_c = 256$ in the next experiments.

We then conducted an evaluation of the pre-whitening performance as a function of the iSNR, and the results are depicted in Fig. 7. For babble noise, Par-NMF based pre-whitening had the best performance, regardless how many U_c entries were used, although with $U_c = 16$, a considerable lower ISD was observed at higher iSNRs. However, $U_c = 256$ allowed for a slightly better SFM. For restaurant noise, a similar SFM was achieved for Par-NMF based pre-whitening using $U_c = 16$ entries and MMSE pre-whitening, with a slightly lower ISD for the Par-NMF. For this noise type, not included in the training step, $U_c = 256$ entries lead to a much better performance. For factory and F-16 noise, using a lower number of U_c entries is more convenient. In these cases, the benefit of Par-NMF pre-whitening (with $U_c = 16$) was seen at lower iSNRs, because at higher ones, the ISD from MMSE or MS based pre-whitening became lower, although the noise whiteness from the three approaches were similar. By increasing the iSNR, the ISD between the oracle pre-whitener and the estimated pre-whitener increases, and also the average SFM of the different pre-whiteners differs more from the one that can be obtained from the oracle pre-whitener. This is expected since the noise PSD estimates are typically less accurate in

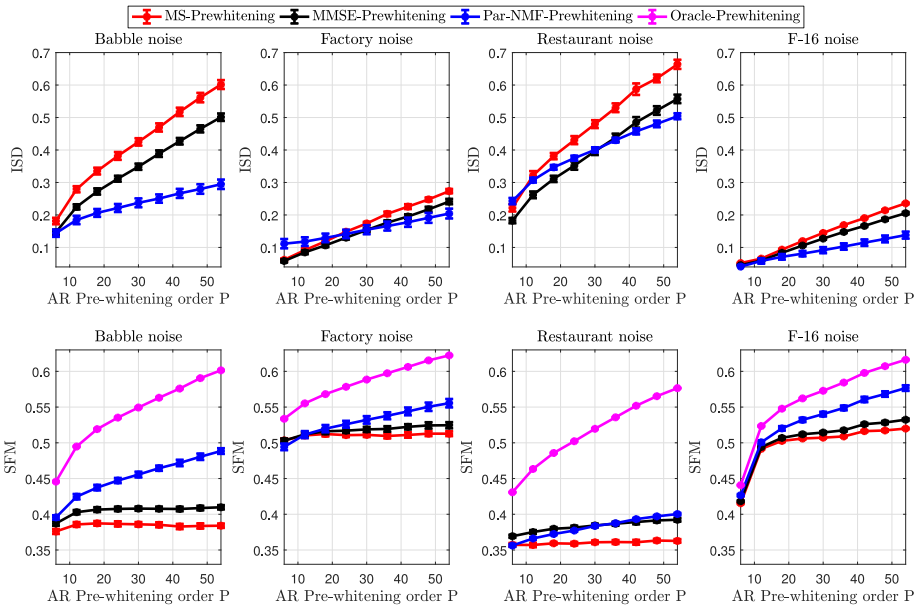

Fig. 5. The ISD between the oracle and estimated pre-whiteners and the spectral flatness measure of the pre-whitened noise as a function of the AR pre-whitening order, at iSNR = 0 dB, under different colored noise scenarios. A lower ISD is preferred, and a higher SFM is desirable. Results are reported in 95% confidence intervals.

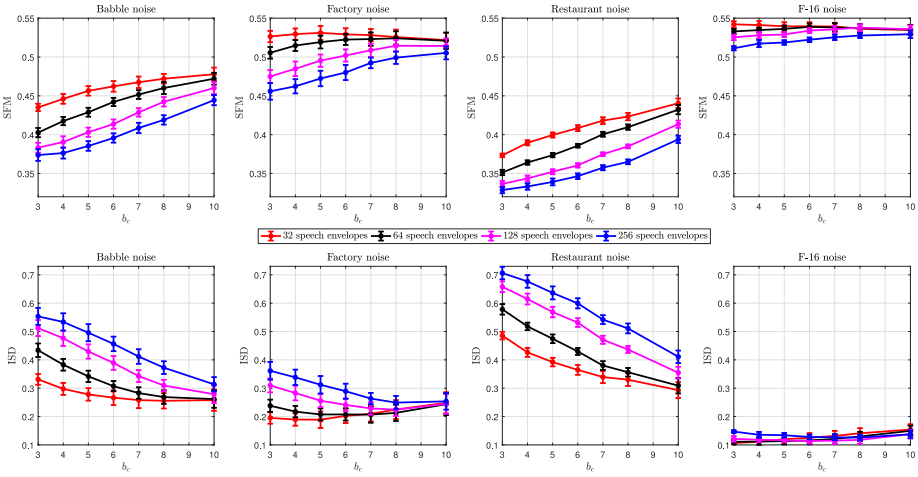

Fig. 6. The spectral flatness measure of the pre-whitened noise and the ISD between the oracle and estimated Par-NMF based pre-whitener as a function of b_c, for different number 2^{b_s} speech spectral envelopes, at iSNR = 0 dB, under different colored noise scenarios. A lower ISD is preferred, and a higher SFM is desirable.

high iSNR conditions as reported in the literature (Gerkmann and Hendriks, 2012). Again, in most cases, MS based pre-whitening was outperformed by either MMSE-SPP or Par-NMF based pre-whitening.

5.3.3. Evaluation of pitch estimation accuracy with pre-whitening

In the next experiment, we compared four different pre-whiteners by evaluating the performance improvement of the NLS pitch estimator when its input signal is the pre-whitened signal. The four pre-whiteners are based on noise PSD estimates obtained by MS, MMSE, the proposed Par-NMF (with both $U_c = 16$ and $U_c = 256$ entries), and a fixed noise PSD computed from the long-term averaged spectrum of the samples of the noise of interest used in the codebooks training. That is, a fixed pre-whitening filter is applied to verify that an adaptive pre-whitener based on the local characteristics of speech and noise signals should be preferred as a pre-processor. Typical shapes of the long-term averaged spectrum of some noise types can be seen in Hirsch and Pearce (2000), Hilkhuysen et al. (2014). Only in the restaurant noise case, which was not included in the training, the samples used for the testing were used to determine the long-term average spectrum. The post-processing block in Fig. 3 is still not used. Babble, factory, street, and restaurant noise were added at different iSNRs from −5 to 10 dB, and three MCS were run for each file at each iSNR. The performance measures in (26)–(28) were computed after estimating the pitch. The results are depicted in Fig. 8 with 95 % confidence intervals. Clearly, using a fixed pre-whitener resulted in poorer pitch estimates and voicing detections than using the time-varying pre-whiteners. For babble and restaurant noise, the best accuracy of the NLS pitch estimator was achieved when the cascading was done with the Par-NMF based pre-whitener, because the confidence intervals of the FFE were clearly separated from those of MMSE-SPP or MS based pre-whitening. When using $U_c = 256$ entries, a slightly better performance is seen than when using $U_c = 16$ entries. For street noise, using $U_c = 256$ entries has a positive effect in reducing the GER, although it will not benefit the VDE. In terms of FFE, for factory and street noise, which are more stationary noise types, the accuracy after pre-whitening based on the three approaches is very similar, thus indicating that the proposed Par-NMF based pre-whitener is of greater benefit in non-stationary noise scenarios.

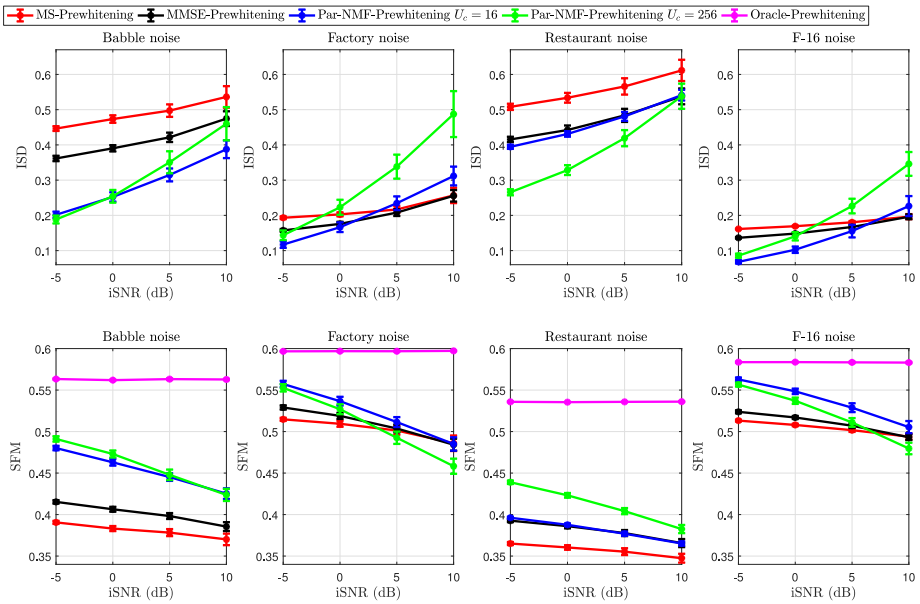

Fig. 7. The ISD between the oracle and estimated pre-whiteners, and the spectral flatness measure of the pre-whitened noise as a function of the iSNR, for an AR pre-whitening order P=36. A lower ISD is preferred, and a higher SFM is desirable.

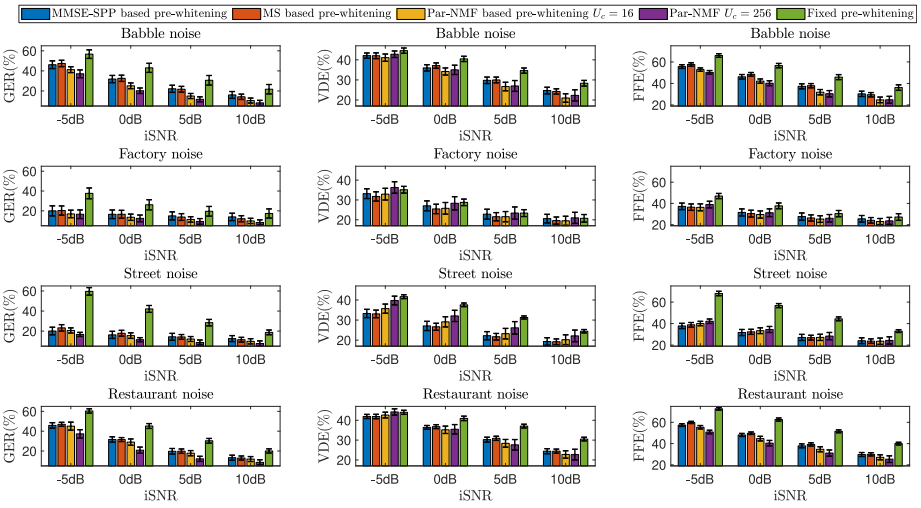

Fig. 8. Estimation accuracy of the NLS pitch estimator under different noise conditions, after application of AR pre-whitening based on different noise PSD estimates, and also after applying a fixed pre-whitening filter.

We then investigated if various non-parametric pitch estimators, which are not derived under a WGN assumption, improved their accuracy by using either the proposed pre-whitener or an enhancement system as a pre-processor. Particularly, the Cepstrum-based method (Noll, 1967), RAPT (Talkin, 1995), SHRP (Sun, 2002), and SWIPE' (Camacho and Harris, 2008), all of them with a final smoothing step, were used in the evaluation. To determine which is the best pre-processing method on average, we present the averaged performance from those four estimators in three different ways: in its naïve (out-of-the-box) form (i.e., without a pre-processor), and when either an OMLSA based enhancement system or the proposed Par-NMF pre-whitener were used as a pre-processor. The performance of the recently introduced robust Bayesian pitch tracker (Shi et al., 2019), also derived under a WGN assumption, was also evaluated. This method models the dynamic evolution of the pitch, the number of harmonics, and the voicing state by using first-order Markov processes. The already introduced NLS pitch estimates were again included, but a Ney smoothing step between consecutive independent-segment values (Ney, 1983) was applied, to be fairer in the comparison, as all the other methods had tracking or refinement capabilities. The NLS and the Bayesian pitch tracker estimates were evaluated only when the Par-NMF pre-whitener was used as a pre-processor, as we verified in Section 5.3.1 that the performance of the NLS estimator, has a better improvement from pre-whitening. An example of the estimates produced by the Bayesian pitch tracker for a female speaker in babble noise at an iSNR of 3 dB is shown in Fig. 9. The pitch was estimated after either OMLSA-based enhancement or the Par-NMF based pre-whitening or MS based pre-whitening were used as pre-processors. Clearly, the resulting estimates after enhancement showed a large number of not-voiced (e.g., silent) segments wrongly detected as voiced, and also a high number of octave errors. When the proposed pre-whitener is instead applied, the pitch contour is better captured as less octave errors and less voicing detection errors are obtained. MS-based pre-whitening appears to be effective in reducing the octave errors, however it still leads to several consecutive voiced frames incorrectly detected as unvoiced, e.g., between 4 and 6 s. In such cases, the Par-NMF based pre-whitener produces a more accurate pitch contour. A similar plot, but for car noise, is displayed in Fig. 10, so that we can verify the generalization capacity of the pre-whitener in

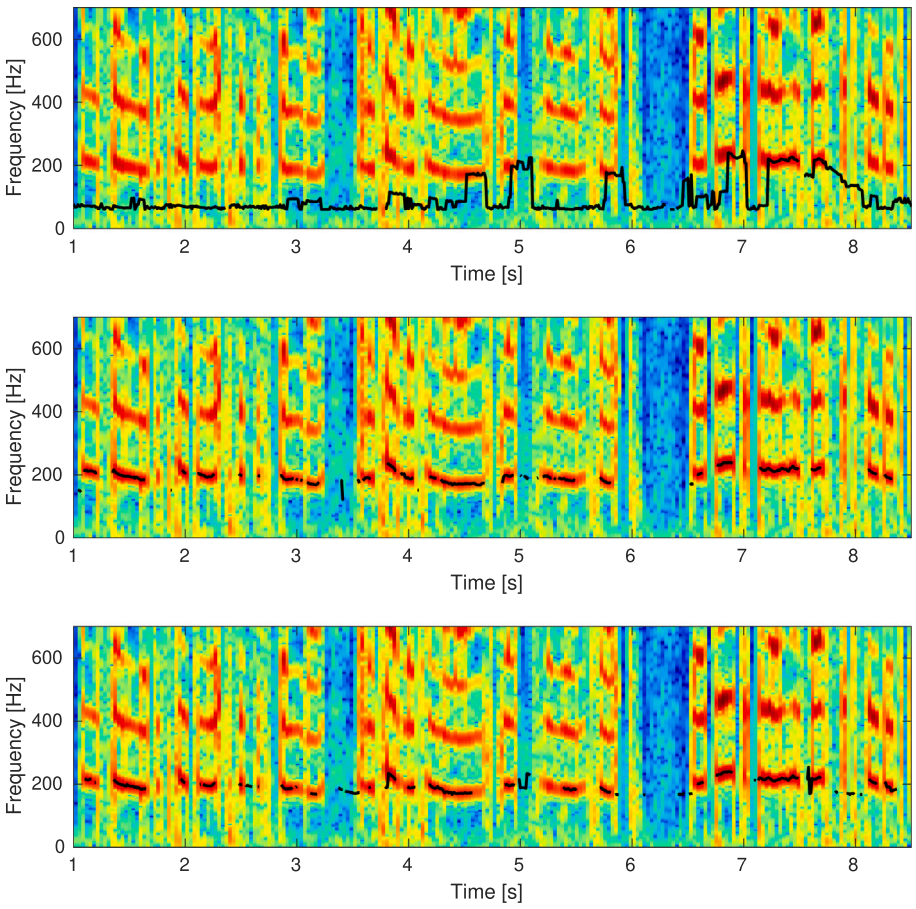

Fig. 9. Pitch estimates from the Bayesian tracker for a female excerpt in 3 dB babble noise after either enhancement (top) or pre-whitening based on Par-NMF (middle) pre-whitening based on MS (bottom) is applied as pre-processor. Note that only the spectrograms of the clean signal are shown to facilitate an easier visual evaluation of the produced pitch estimates.

noise environments of different nature than the ones considered in the training stage. According to the spectral shape visualization, in general, this noise type appears to be more stationary as compared to babble or restaurant noise. The car noise was obtained at 50 km/h from the NOISEX-92 database. The pitch contour obtained after enhancement is better captured as opposed to the babble noise case counterpart. Both pre-whitening approaches lead to an acceptable capture of the pitch contour, although at 3.5 and between 6 and 7 s, some not-voiced frames were incorrectly classified as voiced. This kind of error was less visible in the babble noise setup, possibly due to the fact that the testing samples are fairly similar to the ones used in the training (although not identical). This suggests that a more diverse codebook trained on a larger variety of noise types could improve the robustness.

The overall performance is assessed by adding either babble or factory noise at iSNRs from −5 to 11 dB. Three MCS were run for each file from the Keele database at each iSNR. The results are depicted in Fig. 11. The best performance was achieved by cascading the pre-whitener with the Bayesian pitch tracker, although in some cases under factory noise conditions, the confidence intervals overlapped with the NLS pitch estimates. The performance in the babble noise case was worse than the factory noise case, which is expected due to the fact that babble noise is a random mixture of human speech signals, making more challenging the pitch estimation task. The displayed results from the Bayesian pitch tracker correspond to using $U_c = 256$ envelopes in the babble noise case and $U_c = 16$ envelopes for the factory noise case, as these configurations produced slightly better results, but there was not a very significant difference in the performance from using the other U_c configuration. On average, under babble noise conditions, the GER from non-parametric estimators was improved by pre-processing via pre-whitening, and in the factory noise, also from pre-whitening based on $N_c = 16$ spectral shapes, for iSNRs below 7 dB. In contrast, the VDE was improved by applying enhancement below 7 dB for babble noise, and in the factory noise case, at iSNRs below 11 dB. The FFE slightly decreased when pre-whitening based on $U_c = 16$ entries was used, but only at iSNRs lower than 3 dB in babble noise. In the factory noise case, the full frame errors were reduced by applying enhancement at iSNRs below 7 dB. However, although the performance of non-parametric pitch methods was improved by either enhancing or pre-whitening, the best performance was achieved from pre-whitening followed by the Bayesian pitch tracking. Pre-whitening combined with NLS pitch estimation followed by nonlinear smoothing also resulted in less full frame errors than non-parametric pitch estimators (even if they obtained a benefit from a pre-processing step) for babble noise at iSNRs below 7 dB.

5.3.4. Evaluation of computational complexity of pre-processors

We also evaluated the computation time of the various pre-processors, including OMLSA-based enhancement based on the parametric NMF noise PSD estimate. The testing was done with one excerpt of the Keele database with a duration of 40.3 seconds. The total time for each type of pre-processing (enhancement and pre-whitening based on different noise PSD estimates) is reported in Table 1. The other approaches, especially the pre-whiteners based on MS or MMSE, are computationally faster than the pre-whitener based on parametric NMF. However, as seen from previous experiments, such an increase in computation time for the proposed pre-processing scheme resulted in an improvement in the accuracy of the WGN-based estimators, specially under non-stationary noise.

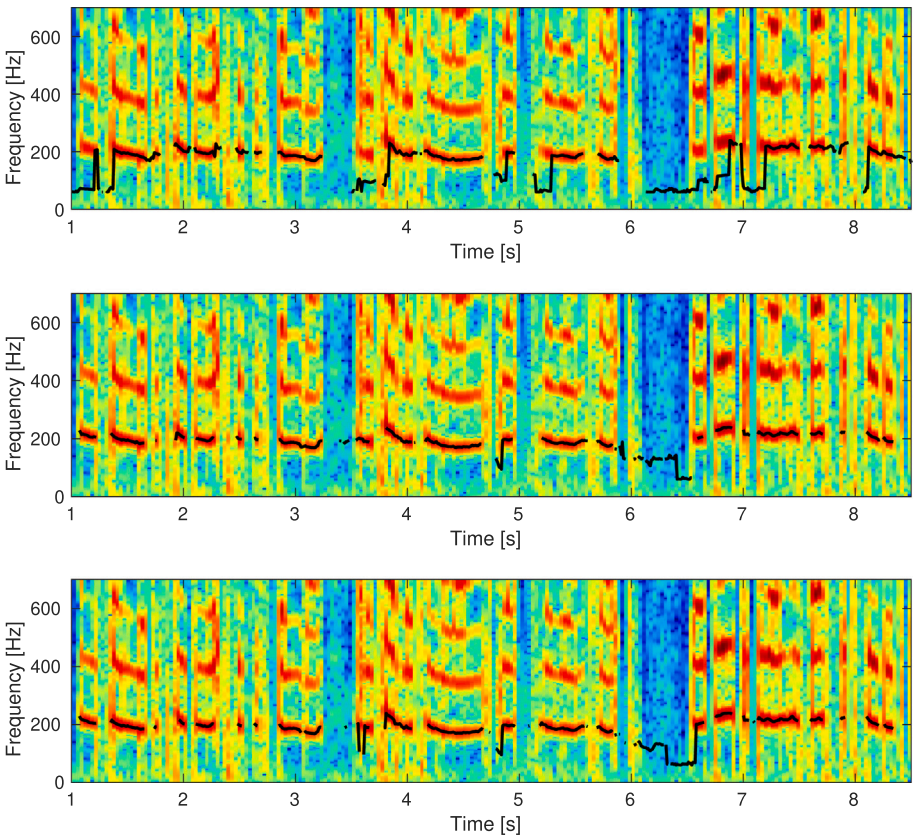

Fig. 10. Pitch estimates from the Bayesian tracker for a female excerpt in 3 dB car noise after either enhancement (top) or pre-whitening based on Par-NMF (middle) pre-whitening based on MS (bottom) is applied as pre-processor. Note that only the spectrograms of the clean signal are shown to facilitate an easier visual evaluation of the produced pitch estimates.

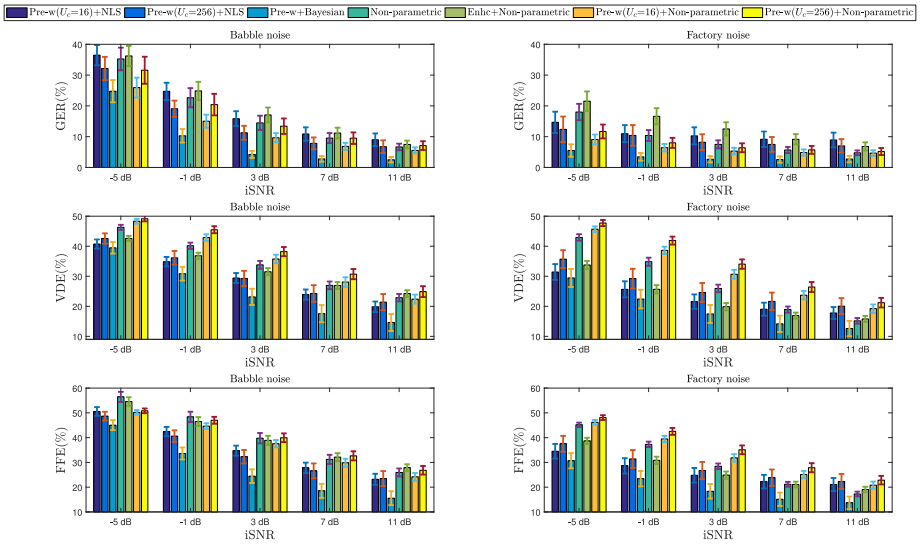

Fig. 11. Estimation accuracy of parametric pitch estimators (NLS and Bayesian) after pre-whitening, and averaged performance of four non-parametric pitch estimators (Cepstrum, SWIPE, SHRP and RAPT) in their naïve states, and when either speech enhancement or pre-whitening was previously applied, under different noise conditions.

Table 1
Computation time in [ms] for different pre-processing schemes for a single segment.

OMLSA	MS	MMSE	Par-NMF (U_c = 16)	Par-NMF (U_c = 256)
2.09	0.42	0.39	2.71	4.08

5.3.5. *Evaluation of pitch estimation accuracy including post-processing*

In the last evaluation of pitch estimation, we included the last post-processing block of Fig. 3. As previously shown experimentally, using the Par-NMF pre-whitener as a pre-processor leads to the largest improvement of the accuracy of the NLS pitch estimator. However, there might still be some segments in which the solution produces estimates resulting in either a gross error or a voicing detection error. To further reduce these errors, post-processing of the initial pitch

estimates is performed by iterating the following two steps (Jaramillo et al., 2020):

1. The harmonic amplitudes for a given estimated $\hat{\omega}_0$ and order $\hat{L}$ are $\hat{\boldsymbol{\alpha}} = \left[\mathbf{Z}_{\hat{L}}^H(\hat{\omega}_0)\mathbf{Z}_{\hat{L}}(\hat{\omega}_0)\right]^{-1}\mathbf{Z}_{\hat{L}}^H(\hat{\omega}_0)\mathbf{x}$ (Christensen, 2013), so the residual representing what is not captured by the harmonic model, including the stochastic parts of the speech, is $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{Z}_{\hat{L}}^H(\hat{\omega}_0)\hat{\boldsymbol{\alpha}}$. Thus, the AR parameters for an updated pre-whitener can be directly re-estimated from this residual using the autocorrelation method (Stoica et al., 2005).
2. The re-estimated AR parameters of the residual are directly used as the coefficients of a new pre-whitening filter, which is applied to the noisy signal. From the new pre-whitened signal, the pitch ω_0 and L are again estimated with the NLS estimator (30).

In this iterative process, the new pre-whitener is no longer computed using the parametric NMF based noise PSD estimator, as it is now instead computed from the residual. As seen below, however, the key to achieve the final best pitch estimation accuracy is having applied a better pre-whitener as a pre-processor.

The full setup in Fig. 3 was evaluated, and although this involves an even higher computational complexity than the one reported in Sec. 5.3.4, it also leads to an improved pitch estimation accuracy. Both pre-whiteners based on MMSE-SPP and on the proposed Par-NMF were applied as a pre-processor. For the Par-NMF noise PSD estimate, $U_s = 32$ and either $U_c = 16$ or $U_c = 256$ spectral shapes were considered, for factory and babble, respectively. For the iterative estimation, the iteration was performed a maximum of 10 times. Moreover, if a frame was detected as being not-voiced (i.e., $\hat{L} = 0$), the estimation was stopped for that segment. We compared to the performance of individual non-parametric estimators SWIPE', PEFAC (Gonzalez and Brookes, 2014), SHRP, and RAPT which all include a final smoothing step between consecutive estimates. Their individual performance was also assessed after pre-processing the noisy signal, being pre-whitened in the babble noise case, and enhanced using OMLSA for factory noise, according to the averaged preferred pre-preprocessing that was noted previously. The FFE for babble and factory noise are depicted in Fig. 12. The configurations including post-processing are denoted by Iter(MMSE) and Iter(Par-NMF) in which either pre-whiteners based on MMSE or Par-NMF noise PSD estimates were applied as a pre-processor. By including the post-processing step, there was a reduction between 2.2 and 4.5% compared to the single cascade of pre-whitening and NLS pitch estimation. The independent consecutive pitch estimates were not smoothed in this case, but by doing so we expect that the performance could be improved.

It is seen that only pre-processing using parametric NMF combined with the NLS pitch estimates was not enough to have better accuracy than SWIPE' in babble noise. However, by post-processing the initially obtained estimates (i.e., Iter(Par-NMF)), a better performance was achieved below 5 dB, and similar one at 5 and 10 dB. That did not occur if the initial pre-whitener based on MMSE-SPP was applied, even if the post-processing based on iterative refinement was later applied (i.e., Iter(MMSE)). Also, SWIPE' improved its accuracy when its input signal was the pre-whitened signal, achieving similar or better accuracy than the proposed method Iter(Par-NMF) at SNRs above 0 dB. Similarly, for factory noise, by applying the post-processing in the proposed method, SWIPE' was outperformed at −5 and 0 dB and RAPT was outperformed below 10 dB. When the estimates are obtained from the pre-processed (enhanced) signal, RAPT was able to achieve a similar performance to the proposed method for all the SNRs. Also, when SWIPE' was applied to the enhanced signal, it achieved a similar performance at 0 and 5 dB, and a better one at 10 dB.

5.4. Experimental results regarding TOA estimation

Finally, we evaluated the accuracy of an estimator of the time-of-arrival (TOA) of a signal emitted by a source such as a loudspeaker and received by a receiver such as a microphone. For this application, we use a model in which the received signal is modeled as

$$x(n) = g s(n-\tau) + c(n) \tag{31}$$

where $s(n)$ is a known signal emitted by the source, g is the attenuation of the signal, and τ is the time it takes for the signal to propagate from the source to the receiver (i.e., the TOA). If the noise term $c(n)$ is assumed to be WGN, and by considering vectors of N successive samples in bold notation, the ML estimator of τ and g are the solutions to

$$\{\hat{\tau}, \hat{g}\} = \min_{\tau \in T, g>0} \|\boldsymbol{x}(n) - g\boldsymbol{s}(n-\tau)\|_2^2, \tag{32}$$

over the possible set T of TOAs. If the analysis window is long relative to the size of $s(n)$, the TOA estimator can be accurately approximated as

$$\hat{\tau} = \arg\max_{\tau \in T} \ \boldsymbol{x}(n)\boldsymbol{s}(n-\tau) \tag{33}$$

which is often referred to as the matched filter (Feder and Weinstein, 1988). In practical setups, the noise is likely to be non-white. In such cases, pre-whitening should be applied as a pre-processor, but we remark that it has to be applied to both $\boldsymbol{x}(n)$ and $\boldsymbol{s}(n)$ since applying the pre-whitener to only $\boldsymbol{x}(n)$ would introduce an additional delay, resulting in a biased estimator.

We used the recorded signals from the SMARD database (Nielsen et al., 2014) at both the loudspeaker and the single microphone, both of them separated 3.13 m, with configuration number 0001. The known source signal was an artificial white noise synthetic signal, and the size of the burst was 3500 samples at a sampling frequency of 48 kHz. The rooms where the signals were recorded had a reverberation time of approximately 0.15 s. The colored noise was taken from the DREGON database (Strauss et al., 2018). Specifically, rotor noise from a drone running at 70 rounds per second was added to the signal picked up by the single microphone at different signal-to-diffuse-noise ratios (SDNR) before the TOA was estimated. 200 MCS were run at each SDNR. The rotor noise was resampled from 44.1 to 48 kHz to match the rate of the source signal. In the evaluation, we compared the performance of the matched filter with and without a pre-whitener. To pre-whiten the observation, a spectral basis matrix of four AR spectra shapes of the rotor noise was built by training a noise codebook on samples of the rotor noise. The training samples were different from those for testing. The testing samples were randomized at each MCS. An additional entry, corresponding to the known source signal, was also included as the clean signal spectral shape which was simply a flat PSD. The training was performed with an order $P' = 35$ on segments with a duration of 20 ms with an overlap of 50 % between them. This order was chosen according to our observation that the best oracle performance was obtained with a higher order, as important envelope components that might be present at medium and high frequencies were not smoothed out. From the estimated TOA, the distance between the loudspeaker and the microphone was obtained, and we computed the mean squared error (MSE) of the measured distance at each SDNR. The results are shown in Fig. 13. As seen, pre-whitening the received signal leads to a lower MSE as compared to the case where the received signal was not pre-whitened prior to the TOA estimation.

6. Conclusion

The accuracy of statistical-based estimators based on the WGN assumption in real acoustic scenarios can be considerably improved when an AR pre-whitener is applied as a pre-processor of the noisy observation. In this paper, we introduced a time-varying pre-whitener

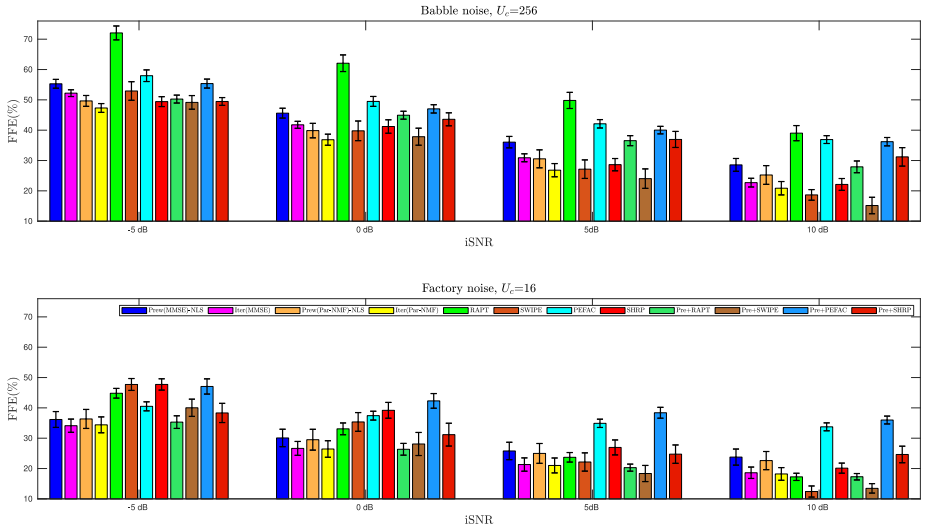

Fig. 12. Estimation accuracy of NLS pitch estimation by considering both pre-processing and post-processing, and of non-parametric pitch estimators (PEFAC, SWIPE', SHRP, and RAPT) in their naïve states and with pre-processing, under different noise conditions.

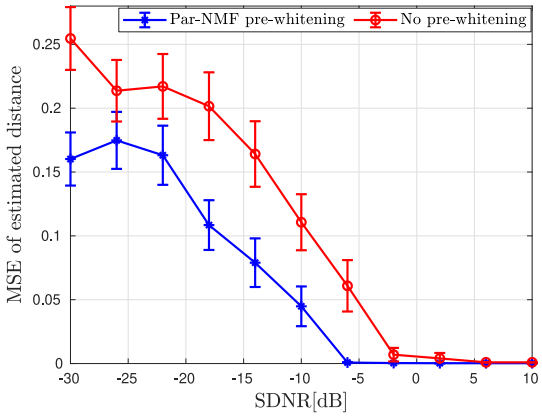

Fig. 13. MSE of single-channel TOA estimation with pre-whitening applied or not applied, versus the SDNR. Results are reported as 95% confidence intervals.

which requires the activation coefficients of pre-trained spectral shapes in the parametric NMF method. Through numerous simulations, we have shown that using an AR pre-whitener based on the parametric NMF method results in a higher noise whiteness and a more similar spectral response to that of the oracle pre-whitener compared to conventional noise PSD estimators, especially in non-stationary noise situations. Although the training stage of the spectral shapes may initially require additional effort compared to using traditional noise trackers, it offers a consistent way of including prior information about speech and noise types, resulting in a better performance of WGN-based estimators such as the NLS pitch estimator. Although well-known non-parametric pitch estimators can improve their accuracy from some pre-processing, the combination of pre-whitening with fast and efficient statistical-based WGN methods gives the best performance in terms of pitch errors and voicing detection errors, specially in scenarios of high noise levels (i.e., low SNRs). An additional improvement can be obtained by post-processing the resulting NLS pitch estimates, and this will result in a better overall accuracy than individual non-parametric pitch estimators, even if they are using a pre-processor, specially under low SNR conditions. This may require high computation time, but it allows to extract both the harmonic and autoregressive components of the speech signal, which is useful, e.g., in the speech decomposition problem (Jaramillo et al., 2021). The pre-whitener was also applied before a time-of-arrival estimation method formulated under the WGN assumption. In that case, the TOA estimation accuracy is improved by a pre-whitening step which relies on pre-trained shapes of the involved source signal and of the real noise in the recording environment, such as wind noise or drone ego-noise.

CRediT authorship contribution statement

Alfredo Esquivel Jaramillo: Conceptualization, Formal analysis, Data curation, Funding acquisition, Investigation, Writing – original draft, Writing – review & editing. **Jesper Kjær Nielsen:** Conceptualization, Formal analysis, Investigation, Project administration, Writing – original draft, Writing – review & editing. **Mads Græsbøll Christensen:** Conceptualization, Formal analysis, Investigation, Resources, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alfredo Esquivel Jaramillo reports financial support was provided by National Council of Science and Technology (CONACYT).

Data availability

Code and Keele database files are available at https://github.com/alfredoej87/Autoregressive-Pre-whitening-based-on-Parametric-NMF and https://github.com/alfredoej87/iterativeF0Ar_NLS.

Acknowledgments

This research was supported by the National Council of Science and Technology (CONACYT), Mexico under grant 418437.

Appendix

To maximize the data likelihood $p(\mathbf{x}|\boldsymbol{\sigma},\mathbf{D}) \sim \mathcal{N}\left(\mathbf{0}, \sum_{u=1}^{U} \sigma_u^2 \mathbf{R}_u(\mathbf{a}_u)\right)$, we use the well-known fact (Gray et al., 2006) that $\mathbf{R}_u(\mathbf{a}_u)$ can be approximated as circulant and therefore diagonalized by the Fourier transform if N is much larger than the AR-order P'. Thus, the approximate diagonalization of the covariance is

$$\mathbf{R}_u(\mathbf{a}_u) \approx \frac{1}{N}\mathbf{F}\mathbf{D}_u(\mathbf{a}_u)\mathbf{F}^H \tag{A.1}$$

where $\mathbf{F}$ is the DFT matrix whose entries are given by $[\mathbf{F}]_{n,l} = \exp(j2\pi nl/N)$, $n, l = 0, 1, \ldots, N-1$, and

$$\mathbf{D}_u(\mathbf{a}_u) = \left(\boldsymbol{\Lambda}_u^H(\mathbf{a}_u)\boldsymbol{\Lambda}_u(\mathbf{a}_u)\right)^{-1}, \quad \boldsymbol{\Lambda}_u(\mathbf{a}_u) = \operatorname{diag}\left(\mathbf{F}^H\left[\mathbf{a}_u^T\ \mathbf{0}^T\right]^T\right). \tag{A.2}$$

The diagonal entries of $\mathbf{D}_u(\mathbf{a}_u)$ represent the eigenvalues of $\mathbf{R}_u(\mathbf{a}_u)$, which correspond to the normalized PSD of the uth AR process.

Using the above definitions, the log-likelihood can be written as (Kavalekalam et al., 2018)

$$\ln p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln\left|\sum_{u=1}^{U}\frac{\sigma_u^2\mathbf{F}\mathbf{D}_u(\mathbf{a}_u)\mathbf{F}^H}{N}\right| - \frac{1}{2}\mathbf{x}^T\left[\sum_{u=1}^{U}\frac{\sigma_u^2\mathbf{F}\mathbf{D}_u(\mathbf{a}_u)\mathbf{F}^H}{N}\right]^{-1}\mathbf{x}, \tag{A.3}$$

which can be simplified as

$$\ln p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln\prod_{k=0}^{N-1}\sum_{u=1}^{U}\sigma_u^2 d_u(k) - \frac{1}{2N}\mathbf{x}^T\mathbf{F}\left[\sum_{u=1}^{U}\sigma_u^2\mathbf{D}_u(\mathbf{a}_u)\right]^{-1}\mathbf{F}^H\mathbf{x} \tag{A.4}$$

$$= -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{N-1}\ln\sum_{u=1}^{U}\hat{\boldsymbol{\Phi}}_u(k) - \frac{1}{2}\sum_{k=0}^{N-1}\frac{\boldsymbol{\Phi}(k)}{\sum_{u=1}^{U}\hat{\boldsymbol{\Phi}}_u(k)} \tag{A.5}$$

where $\boldsymbol{\Phi}(k)$ is the kth element of the periodogram of $\mathbf{x}$ and $\hat{\boldsymbol{\Phi}}_u(k) = \sigma_u^2 d_u(k)$. Each $d_u(k)$ is the kth diagonal element of $\mathbf{D}_u(\mathbf{a}_u)$. The summation over U spectral basis, i.e., $\sum_{u=1}^{U}\hat{\boldsymbol{\Phi}}_u(k) = \tilde{\mathbf{d}}_k^T\boldsymbol{\sigma}$ is the modeled PSD at frequency bin k. The expression in (A.5) can now be re-written into (14).

References

Al-Aboosi, Y.Y., Sha'ameri, A.Z., 2017. Improved underwater signal detection using efficient time–frequency de-noising technique and Pre-whitening filter. Appl. Acoust. 123.

Bao, F., Dou, H.-j., Jia, M.-s., Bao, C.-c., 2014. Speech enhancement based on a few shapes of speech spectrum. In: 2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP). IEEE, pp. 90–94.

Birch, G.E., Lawrence, P.D., Lind, J.C., Hare, R.D., 1988. Application of prewhitening to AR spectral estimation of EEG. IEEE Trans. Biomed. Eng. 35 (8), 640–645.

Blanco, L., Nájar, M., 2012. Sparse covariance fitting for direction of arrival estimation. EURASIP J. Adv. Signal Process. (1).

Camacho, A., Harris, J.G., 2008. A sawtooth waveform inspired pitch estimator for speech and music. J. Acoust. Soc. Am. 124 (3), 1638–1652.

Christensen, M.G., 2013. Accurate estimation of low fundamental frequencies from real-valued measurements. IEEE Trans. Audio Speech Lang. Process. 21 (10), 2042–2056.

Christensen, M.G., Jakobsson, A., 2009. Multi-Pitch Estimation. In: Synthesis Lectures on Speech and Audio Processing, Morgan & Claypool Publishers.

Chu, W., Alwan, A., 2009. Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 3969–3972.

Cohen, I., 2002. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal Process. Lett. 9 (4), 113–116.

Dou, Z., Shi, C., Lin, Y., Li, W., 2017. Modeling of non-Gaussian colored noise and application in CR multi-sensor networks. EURASIP J. Wireless Commun. Networking 2017 (1), 1–11.

Emiya, V., Badeau, R., David, B., 2007. Multipitch estimation of quasi-harmonic sounds in colored noise. In: 10th Int. Conf. on Digital Audio Effects (DAFx-07). p. 1,5.

Feder, M., Weinstein, E., 1988. Parameter estimation of superimposed signals using the EM algorithm. IEEE Trans. Acoust. Speech Signal Process. 36 (4), 477–489.

Févotte, C., Bertin, N., Durrieu, J.-L., 2009. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. Neural Comput. 21 (3), 793–830.

Févotte, C., Idier, J., 2011. Algorithms for nonnegative matrix factorization with the β-divergence. Neural Comput. 23 (9), 2421–2456.

Févotte, C., Vincent, E., Ozerov, A., 2018. Single-channel audio source separation with NMF: divergences, constraints and algorithms. In: Audio Source Separation. Springer, pp. 1–24.

Gerkmann, T., Hendriks, R.C., 2012. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang. Process. 20 (4), 1383–1393.

Gonzalez, S., Brookes, M., 2014. PEFAC-a pitch estimation algorithm robust to high levels of noise. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (2), 518–530.

Gray, R.M., et al., 2006. Toeplitz and circulant matrices: A review. Found. Trends® Commun. Inf. Theory 2 (3), 155–239.

Hansen, P.C., Jensen, S.H., 2007. Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis. EURASIP J. Adv. Signal Process. 2007, 092953.

He, Q., Bao, F., Bao, C., 2017. Multiplicative update of auto-regressive gains for codebook-based speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (3), 457–468.

Hilkhuysen, G., Gaubitch, N., Brookes, M., Huckvale, M., 2014. Effects of noise suppression on intelligibility. II: An attempt to validate physical metrics. J. Acoust. Soc. Am. 135 (1), 439–450.

Hirsch, H., Pearce, D., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000-Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop. ITRW.

Huang, Q., Bao, C., Wang, X., Xiang, Y., 2020. Speech enhancement method based on multi-band excitation model. Appl. Acoust. 163.

Huang, J., Zhao, Y., 1998. An energy-constrained signal subspace method for speech enhancement and recognition in white and colored noises. Speech Commun. 26 (3), 165–181.

Itakura, F., 1968. Analysis synthesis telephony based on the maximum likelihood method. In: The 6th International Congress on Acoustics, 1968. pp. 280–292.

Jakobsson, A., Mossberg, M., Rowe, M.D., Smith, J.A.S., 2005. Frequency-selective detection of nuclear quadrupole resonance signals. IEEE Trans. Geosci. Remote Sens. 43 (11), 2659–2665.

Jaramillo, A.E., Jakobsson, A., Nielsen, J.K., Christensen, M.G., 2020. Robust fundamental frequency estimation in coloured noise. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 741–745.

Jaramillo, A.E., Nielsen, J.K., Christensen, M.G., 2018. On optimal filtering for speech decomposition. In: 2018 26th European Signal Processing Conference. EUSIPCO, IEEE, pp. 2325–2329.

Jaramillo, A.E., Nielsen, J.K., Christensen, M.G., 2019a. Adaptive pre-whitening based on parametric NMF. In: 2019 27th European Signal Processing Conference. EUSIPCO, pp. 1–5.

Jaramillo, A.E., Nielsen, J.K., Christensen, M.G., 2019b. A study on how pre-whitening influences fundamental frequency estimation. In: IEEE, ICASSP International Conference on Acoustics, Speech and Signal Processing. pp. 6495–6499.

Jaramillo, A.E., Nielsen, J.K., Christensen, M.G., 2021. Speech decomposition based on a hybrid speech model and optimal segmentation. In: Interspeech.

Jensen, J.R., Saqib, U., Gannot, S., 2019. An EM method for multichannel TOA and DOA estimation of acoustic echoes. In: IEEE Workshop on Applications of Signal Processing To Audio and Acoustics. WASPAA, pp. 120–124.

Kavalekalam, M.S., Nielsen, J.K., Shi, L., Christensen, M.G., Boldt, J., 2018. Online parametric NMF for speech enhancement. In: 2018 26th European Signal Processing Conference. EUSIPCO, pp. 2320–2324.

Kay, S., Salisbury, J., 1990. Improved active sonar detection using autoregressive prewhiteners. J. Acoust. Soc. Am. 87 (4), 1603–1611.

Kominek, J., Black, A.W., 2004. The CMU arctic speech databases. In: Fifth ISCA Workshop on Speech Synthesis.

Linde, Y., Buzo, A., Gray, R., 1980. An algorithm for vector quantizer design. IEEE Trans. Commun. 28 (1), 84–95.

Madhu, N., 2009. Note on measures for spectral flatness. Electron. Lett. 45 (23), 1195–1196.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9 (5), 504–512.

Ney, H., 1983. A dynamic programming algorithm for nonlinear smoothing. Signal Process. 5 (2), 163–173.

Nielsen, J.K., Jensen, J.R., Jensen, S.H., Christensen, M.G., 2014. The single- and multichannel audio recordings database (SMARD). In: 14th International Workshop on Acoustic Signal Enhancement. IWAENC, pp. 40–44.

Nielsen, J.K., Jensen, T.L., Jensen, J.R., Christensen, M.G., Jensen, S.H., 2017. Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient. Signal Process. 135 (Supplement C), 188–197.

Nielsen, J.K., Kavalekalam, M.S., Christensen, M.G., Boldt, J.B., 2018. Model-based noise PSD estimation from speech in non-stationary noise. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP.

Noll, A.M., 1967. Cepstrum pitch determination. J. Acoust. Soc. Am. 41 (2), 293–309.

Nørholm, S.M., Jensen, J.R., Christensen, M.G., 2016. Instantaneous fundamental frequency estimation with optimal segmentation for nonstationary voiced speech. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (12), 2354–2367.

Okamoto, T., Iwaya, Y., Suzuki, Y., 2012. Wide-band dereverberation method based on multichannel linear prediction using prewhitening filter. Appl. Acoust. 73, 50–55.

Plante, F., Meyer, G.F., Ainsworth, W.A., 1995. A pitch extraction reference database. In: EUROSPEECH.

Quinn, B.G., 2007. Efficient estimation of the parameters in a sum of complex sinusoids in complex autoregressive noise. In: Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers. pp. 636–640.

Quinn, B.G., Nielsen, J.K., Christensen, M.G., 2021. Fast algorithms for fundamental frequency estimation in autoregressive noise. Signal Process. 180, 107860.
Quinn, B., Thomson, P., 1991. Estimating the frequency of a periodic function. Biometrika 78 (1), 65–74.
Rosenkranz, T., 2010. Noise codebook adaptation for codebook-based noise reduction. In: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC). Tel Aviv.
Shi, L., Nielsen, J.K., Jensen, J.R., Little, M.A., Christensen, M.G., 2019. Robust Bayesian pitch tracking based on the harmonic model. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (11), 1737–1751.
Sohn, J., Kim, N.S., Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Process. Lett. 6 (1), 1–3.
Srinivasan, S., Samuelsson, J., Kleijn, W.B., 2006. Codebook driven short-term predictor parameter estimation for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 14 (1), 163–176.
Srinivasan, S., Samuelsson, J., Kleijn, W.B., 2007. Codebook-based Bayesian speech enhancement for nonstationary environments. IEEE Trans. Audio Speech Lang. Process. 15 (2), 441–452.
Stoica, P., Moses, R.L., et al., 2005. Spectral analysis of signals. Pearson.
Stoica, P., Selen, Y., 2004. Model-order selection: a review of information criterion rules. IEEE Signal Process. Mag. 21 (4), 36–47.
Strauss, M., Mordel, P., Miguet, V., Deleforge, A., 2018. Dregon: Dataset and methods for uav-embedded sound source localization. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 1–8.
Sun, X., 2002. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, vol. 1, pp. I–333–I–336.
Swärd, J., Li, H., Jakobsson, A., 2017. Off-grid fundamental frequency estimation. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (2), 296–303.
Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT). Speech Coding and Synth. 495, 518.
Therrien, C.W., 1992. Discrete Random Signals and Statistical Signal Processing. Prentice Hall PTR.
Trucco, A., 2001. Experimental results on the detection of embedded objects by a prewhitening filter. IEEE J. Ocean. Eng. 26 (4), 783–794.
Varga, A., Steeneken, H.J., 1993. Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251.
Yoshii, K., Goto, M., 2012. Infinite composite autoregressive models for music signal analysis.. In: ISMIR. Citeseer, pp. 79–84.
Zhao, Y., Hu, R., Nakamura, S., 2003. Whitening processing for blind separation of speech signals. In: Proc. ICABSS. pp. 331–336.
Zou, Y., Liu, H., 2020. A simple and efficient iterative method for TOA localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 4881–4884.