Adaptive Pre-whitening Based on Parametric NMF

Alfredo Esquivel Jaramillo, Jesper Kjær Nielsen and Mads Græsbøll Christensen
Audio Analysis Lab CREATE, Aalborg University
Emails: {aeja, jkn,mgc}@create.aau.dk

***Abstract*—Several speech processing methods assume that a clean signal is observed in white Gaussian noise (WGN). An argument against those methods is that the WGN assumption is not valid in many real acoustic scenarios. To take into account the coloured nature of the noise, a pre-whitening filter which renders the background noise closer to white can be applied. This paper introduces an adaptive pre-whitener based on a supervised non-negative matrix factorization (NMF), in which a pre-trained dictionary includes parametrized spectral information about the noise and speech sources in the form of autoregressive (AR) coefficients. Results show that the noise can get closer to white, in comparison to pre-whiteners based on conventional noise power spectral density (PSD) estimates such as minimum statistics and MMSE. A better pitch estimation accuracy can be achieved as well. Speech enhancement based on the WGN assumption shows a similar performance to the conventional enhancement which makes use of the background noise PSD estimate, which reveals that the proposed pre-whitener can preserve the signal of interest.**

***Index Terms*—pre-whitening, NMF, spectral flatness, pitch estimation, speech enhancement**

I. Introduction

The presence of additive noise is inevitable in many acoustic scenarios. Although the noise characteristics can be explicitly taken into account for estimating the parameters of a signal of interest (as in [1], [2]), many methods rely on a white Gaussian noise (WGN) condition (see, e.g. [3]–[5]), since this is convenient from a mathematical point of view. This WGN assumption can be quite unrealistic, as real noise types are typically coloured. Applying methods based on the WGN assumption in real noise scenarios can degrade their performance. One example is when sub-harmonic errors appear when estimating the fundamental frequency (a.k.a. pitch) of voiced speech segments [6], [7] from estimators which assume WGN. A pre-processor which renders the coloured noise closer to white, namely a pre-whitener, can alleviate this problem. Applying pre-whitening using a linear filter is advantageous compared to a general linear transformation with, e.g., the Cholesky factor [4], since the effect of linear filtering can be modeled by only changing the sinusoidal amplitudes and phases [6], [7]. Unlike general linear transformations, linear filtering thus enables us to use many existing model-based estimators based on a WGN assumption. A linear FIR filter with response

$$A(\omega) = 1 + \sum_{i=1}^{P} a_z(i) e^{-j\omega i} \tag{1}$$

can be used to whiten the noise if the coloured noise is modeled as an autoregressive process AR(P) resultant by passing white Gaussian excitation noise with variance σ_e^2 through an IIR filter with response $H(\omega) = 1/A(\omega)$. Here, P denotes the linear prediction order, and $\{a_z(i)\}_{i=1}^{P}$ are known as the prediction coefficients. The filter in (1) is referred as the LPC pre-whitening filter, and it corresponds to a FIR filter with coefficients $\{1, a_z(1), ..., a_z(P)\}$, which in practice are found from the estimated second-order noise statistics, namely the noise PSD (power spectral density). The influence of filtering-based pre-whitening schemes based on well-known noise PSD estimates, such as minimum statistics (MS) [8], improved minima controlled recursive average (IMCRA) [9], and minimum mean squared error (MMSE) [10]), on the pitch estimation performance, was studied in [6]. Although these schemes will help, for example, in reducing the sub-harmonic errors of the pitch estimates, it was found that the performance is far from that of the oracle pre-whitener. Consequently, we believe that performance improvements are possible if a more accurate noise PSD is estimated.

Including prior spectral information about typical speech and noise spectral shapes has been shown to be beneficial for the noise PSD estimation accuracy [11], specially under non-stationary noise conditions. In a similar way, we here investigate if an adaptive pre-whitener (i.e., an FIR filter whose parameters change every time frame) based on offline trained speech and noise spectral envelopes can render the noise closer to white, and thereby improve the estimation accuracy of a maximum likelihood (ML) pitch estimator [12], [13]. Specifically, a sum of AR processes model [14] is considered, which was motivated by the source/filter speech production model. In this model, the likelihood maximization corresponds to a parametric non-negative matrix factorization (NMF) [15] of the observed periodogram matrix into a dictionary matrix of pre-trained spectral envelopes, parametrized by AR coefficients, and a matrix of activation coefficients, with the Itakura-Saito (IS) divergence as the optimization criterion.

The rest of the paper is organized as follows. In Section II, the problem is formulated. In Section III, we detail how to estimate the noise PSD using a parametric NMF approach, and we give a summary of the pre-whitening process. Next, in section IV, we compare the noise flatness from the new pre-whitener to others based on conventional noise PSD estimators, and we also evaluate its influence on pitch estimation and on speech enhancement. Finally, section V concludes the presented work.

II. Problem formulation

In this work, we assume that coloured noise $z(n)$ is added to a clean speech signal of interest $s(n)$, i.e.,

The work of A. Esquivel Jaramillo was supported by CONACYT.

$$x(n) = s(n) + z(n), \tag{2}$$

where $x(n)$ is the observed noisy signal. For the purpose of pre-whitening $z(n)$ with an LPC pre-whitener, i.e., rendering coloured noise white, the prediction coefficients $\{a_z(i)\}_{i=1}^{P}$ in (1) have to be estimated. Given the noise PSD $\phi_z(k),\ k = 1, ...,\ K$, the noise autocovariance sequence is obtained from the Wiener-Khintchine theorem as [13]

$$r_z(n) = \frac{1}{K}\sum_{k=0}^{K-1} \phi_z(k) \exp\left(j\frac{2\pi}{K}nk\right), 0 \leq n \leq P, \tag{3}$$

where k denotes the frequency bin and K is the number of frequency bins. Then, the Levinson-Durbin recursion [16] is used to compute the WGN excitation variance σ_e^2 and the P noise prediction coefficients $\{a_z(i)\}_{i=1}^{P}$, which forms the LPC pre-whitening filter in (1).

In practice, the noise PSD $\phi_z(k)$ is estimated for every frame from the noisy signal periodogram $\phi(k)$. This can be done for example, with one of the well-known noise tracking methods, such as MS [8] or MMSE based on speech presence probabilities [10]. However, as was seen in [6], LPC pre-whitening performance based on these noise PSD estimates is still far from the oracle one in, e.g., non-stationary noise. An MMSE-based noise PSD estimate can be obtained as [10]

$$\phi_z(k) = \left(\frac{1}{1+\xi(k)}\right)\phi(k) + \left(\frac{\xi(k)}{1+\xi(k)}\right)\lambda_z^2(k), \tag{4}$$

where $\xi(k) = \lambda_S^2(k)/\lambda_Z^2(k)$ is known as the *a priori* SNR, with $\lambda_S^2(k)$ and $\lambda_Z^2(k)$ being the PSDs of $s(n)$ and $z(n)$ respectively, at frequency bin k. For the proposed pre-whitener, we still use (4). However, we obtain an estimate of $\xi(k)$ from a parametric NMF derived from the sum of AR processes model introduced in [14], and explained in the next section. Because of the Kolmogorov-Szego theorem [16], even if the sum of two or more AR processes is not theoretically AR, in order to apply an LPC pre-whitener, an AR approximation of the PSD is possible if a large prediction order P is used. [1]

III. Noise PSD estimate based on Parametric NMF

In [14], the sum of AR processes model was introduced in an NMF context. There, a noisy signal frame $\mathbf{x} = [x(0), ..., x(K-1)]^T$ is represented as a sum of $U = U_s + U_z$ AR processes $\mathbf{t}_u$, i.e.,

$$\mathbf{x} = \sum_{u=1}^{U} \mathbf{t}_u = \sum_{u=1}^{U_s} \mathbf{t}_u + \sum_{u=U_s+1}^{U} \mathbf{t}_u, \tag{5}$$

where U_s is the number of AR processes corresponding to the speech signal, U_z is the number of AR noise processes, $(\cdot)^T$ denotes transpose, and K is the segment length in samples, corresponding also to the number of frequency bins. Each one of these AR processes is expressed as a multivariate Gaussian $\mathbf{t}_u \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{Q}_u)$. Here, $\mathbf{Q}_u$ is the gain normalized covariance matrix, which can asymptotically be approximated as $\mathbf{Q}_u = K^{-1}\mathbf{F}\mathbf{D}_u\mathbf{F}^H$ [17], where $\mathbf{F} = \{\exp(j2\pi nk/K)\}, n, k = 0, 1, ..., K-1$ and

$$\mathbf{D}_u = \left(\mathbf{\Lambda}_u^H \mathbf{\Lambda}_u\right)^{-1}, \quad \mathbf{\Lambda}_u = \mathrm{diag}\left(\mathbf{F}^H \left[\mathbf{a}_u^T\ \mathbf{0}\right]^T\right), \tag{6}$$

where $\mathbf{a}_u$ is the AR coefficients vector of the u^{th} spectral basis. The different pre-trained basis, i.e., spectral envelopes, are contained in a dictionary matrix $\mathbf{D} \in \mathbb{R}_{\geq 0}^{K\times U}$. In order to maximize the likelihood as a function of U excitation variances and U AR spectral envelopes, the $U \times 1$ vector of activation coefficients $\boldsymbol{\sigma} = \left[\sigma_1^2\ ...\ \sigma_U^2\right]^T$ is estimated online as

$$\hat{\boldsymbol{\sigma}} = \underset{\boldsymbol{\sigma}\geq 0}{\arg\max}\ p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) = \underset{\boldsymbol{\sigma}\geq 0}{\arg\max}\ \mathcal{N}\left(0, \sum_{u=1}^{U} \sigma_u^2 \mathbf{Q}_u\right). \tag{7}$$

This vector corresponds to the excitation variances of each one of the trained *a priori* AR processes. The log-likelihood can be computed and simplified as (see [14] for further details)

$$\ln p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{K-1}\left(\frac{\phi(k)}{\sum_{u=1}^{U}\hat{\phi}_u(k)} + \ln\sum_{u=1}^{U}\hat{\phi}_u(k)\right) \tag{8}$$

.The summation over U spectral basis in (8) is the parametrized representation of the PSD per frequency bin k, and is expressed as $\sum_{u=1}^{U}\hat{\phi}_u(k) = \mathbf{d}_k^T\boldsymbol{\sigma}$, where $\mathbf{d}_k = [d_1(k)\ ...\ d_U(k)]^T$ is the k^{th} row of $\mathbf{D}$. Therefore, the likelihood maximization is equivalent to the minimization of the IS divergence between the observed periodogram $\boldsymbol{\phi} = [\phi(1)\ ...\ \phi(K)]^T$ and the parametrized PSD $\mathbf{D}\boldsymbol{\sigma}$ where $\mathbf{D} = [\mathbf{d}_1\ ...\ \mathbf{d}_K]^T$, under the constraint $\phi(k) > 0\ \forall k$, i.e.,

$$\hat{\boldsymbol{\sigma}} = \underset{\boldsymbol{\sigma}\geq 0}{\arg\min}\ d_{IS}\left(\boldsymbol{\phi}|\mathbf{D}\boldsymbol{\sigma}\right). \tag{9}$$

Each one of this set of activation coefficients can be iteratively estimated by means of a multiplicative update (MU) rule

$$\hat{\boldsymbol{\sigma}} \leftarrow \hat{\boldsymbol{\sigma}} \odot \left\{\mathbf{D}^T\left(\mathbf{D}\hat{\boldsymbol{\sigma}}\right)^{[-2]} \odot \phi\right\} \oslash \left\{\mathbf{D}^T\left(\mathbf{D}\hat{\boldsymbol{\sigma}}\right)^{[-1]}\right\}, \tag{10}$$

where $\odot$ and $\oslash$ are element-wise product and division, respectively. The exponentiation is also an element-wise operation.

The observed periodogram matrix $\mathbf{\Phi} \in \mathbb{R}_{\geq 0}^{K\times R}$ can be expressed as $\mathbf{\Phi} \approx \mathbf{D}\mathbf{\Sigma}$, where R is the number of frames and $\mathbf{\Sigma} \in \mathbb{R}_{\geq 0}^{U\times R}$ is the activation matrix which contains in each one of its columns the activation coefficients for a single frame. Therefore, this corresponds to a supervised NMF where $\mathbf{D}$ contains the gain-normalized (i.e., unitary variance) parametrized AR spectral envelopes [13] in each one of its columns as $\widetilde{\mathbf{d}}_u = \left[\widetilde{d}_u(0)\ ...\ \widetilde{d}_u(k)\ ...\ \widetilde{d}_u(K-1)\right]^T$, where each frequency-bin element is given by

$$\widetilde{d}_u(k) = \frac{1}{\left|1 + \sum_{i=1}^{P'} a_u(i)\exp\left(-\frac{2\pi jik}{K}\right)\right|^2}, \tag{11}$$

where $\{a_u(i)\}_{i=1}^{P'}$ are the P' AR coefficients of the u^{th} spectral basis. The first U_s columns of $\mathbf{D}$ correspond to AR speech spectral envelopes and the last U_z ones to AR noise spectral envelopes, i.e., $\mathbf{D} = [\mathbf{D}_s\ \mathbf{D}_z]$.

[1] The value of P will be limited by the available data [16] (usually $P < K/3$), where a low P could result on a very smooth spectrum, while a P too large could result on spurious peaks.

Finally, after estimating $\mathbf{\Sigma}$, in order to estimate the noise PSD $\phi_z(k)$ as in (4), estimates $\hat{\lambda}_S$ and $\hat{\lambda}_Z$ can be obtained as $\hat{\lambda}_S^2(k) = [\mathbf{D}_s\mathbf{\Sigma}_s]_{(k+1),i}$ and $\hat{\lambda}_Z^2(k) = [\mathbf{D}_z\mathbf{\Sigma}_z]_{(k+1),i}$, where $\mathbf{\Sigma}_s$ corresponds to the first U_s rows of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_z$ to the last U_z ones. Then, an estimate $\hat{\xi}(k) = \hat{\lambda}_S^2(k)/\hat{\lambda}_Z^2(k)$ is found.

For a more robust adaptive pre-whitener, which takes into account noise types or samples which may not be well represented in the pre-trained spectral basis, we also append as a last column in $\mathbf{D}$ a spectral envelope corresponding to the MMSE noise PSD based pre-whitener $\{a_{mmse}(i)\}_{i=1}^{P'}$ [10]

$$\widetilde{d}_{mmse}(k) = \frac{1}{\left|1 + \sum_{i=1}^{P} a_{mmse}(i) \exp\left(-\frac{2\pi jik}{K}\right)\right|^2}. \quad (12)$$

A summary of the pre-whitening process is given in Table I.

Table I: Summary of the proposed pre-whitening scheme.

1) Train speech and noise codebooks on LSF coefficients, convert them to $\{a_u(i)\}_{i=1}^{P'}$ coefficients and build $\mathbf{D} = [\mathbf{D}_S \ \mathbf{D}_Z]$ whose columns are given by (11).
2) For every frame, estimate $\phi(k) = |X(k)|^2/N, \ k = 1, ..., K$.
3) Add spectral envelope from MMSE PSD estimator to $\mathbf{D}$.
    a) Estimate the MMSE noise PSD estimate from [10].
    b) Estimate $r_{mmse}(n)$ from (3)
    c) Estimate $\{a_{mmse}(i)\}_{i=1}^{P'}$ from Levinson-Durbin recursion and form spectral envelope as (12), for each frame.
4) Find $\hat{\boldsymbol{\sigma}}_{est}$ per frame, and therefore $\mathbf{\Sigma}$.
    a) Initialize $\hat{\boldsymbol{\sigma}}_{est}$ with random positive numbers.
    b) Compute $\hat{\boldsymbol{\sigma}}_{est}$ with the MU rule in (10) for 40 iterations.
5) Compute $\hat{\lambda}_S^2(k) = [\mathbf{D}_s\mathbf{\Sigma}_s]_{(k+1),i}$, $\hat{\lambda}_Z^2(k) = [\mathbf{D}_z\mathbf{\Sigma}_z]_{(k+1),i}$.
6) Compute $\hat{\xi}(k) = \hat{\lambda}_S^2(k)/\hat{\lambda}_Z^2(k)$.
7) Compute pre-whitening filter based on estimated noise PSD.
    a) Estimate noise PSD $\phi_z(k)$ per frame as in (4).
    b) Estimate noise covariance from (3).
    c) Estimate noise prediction coefficients from Levinson Durbin recursion which form filter in (1).

IV. Experimental evaluation

In this section, we quantify how well the described pre-whitener works in terms of the spectral flatness measure (SFM), how it improves pitch estimation performance and how well it works for speech enhancement. For these purposes, a general speech codebook was trained from approximately 54 minutes of sentences from 4 different speakers of the CMU Arctic database [18], resampled from 16 to 8 kHz. The offline training of the codebooks was done using a standard vector quantization technique from speech coding [19] on the line spectral frequency (LSF) coefficients. The parameters for both the training and for the NMF based pre-whitening (LPC Par-NMF) are summarized in Table II. The noise codebook was trained on noise samples from the Aurora database [20] of restaurant, street, car and airport noise types. Excerpts from the Keele database [21], resampled to 8 kHz, with added babble or exhibition noise from the Aurora database, were used for the evaluation. It is important to note that these noise types were not included in the training stage, and also that the testing speech involves other speakers (i.e., of another database) different from those of the training stage. LPC pre-whiteners based on other noise PSD estimates (e.g., MS, MMSE, IMCRA), as well as the oracle (AR parameters directly computed from the noise signal), with the same frame duration and overlap as in Table II, were also applied to compare their performance to our proposed pre-whitener.

Table II: NMF Pre-whitener parameters

Parameters	Value
sampling frequency(Hz)	8000
frame duration	32 ms
frame overlap	50%
speech order P'	14

Parameters	Value
noise order P'	14
U_s	32
U_z	14
MU iterations	40

A. Spectral flatness measure (SFM)

To demonstrate how well the described pre-whitener renders noise closer to white, the whiteness of the noise is quantified in terms of the SFM, defined as [13], [16] the ratio of the geometric mean to the arithmetic mean of the pre-whitened noise PSD ϕ_{zw}, i.e., SFM $= \frac{\left(\sqrt[K]{\prod_{k=0}^{K-1}\phi_{zw}(k)}\right)}{\left(\frac{1}{K}\sum_{k=0}^{K-1}\phi_{zw}(k)\right)}$. The SFM is bounded between 0 (more coloured noise) and 1 (perfect white noise). Babble and exhibition noise types were added at SNRs from -10 to 10 dB. Before pre-whitening, the mean SFM of babble noise was 0.07 and for exhibition noise it was 0.30 at all SNRs. Results of the pre-whitened noise SFM are shown in Fig. 1 for two LPC pre-whitening orders ($P = 20$ and $P = 30$). It is observed that the highest SFM (closest to the oracle pre-whitener) can be achieved with the NMF based pre-whitening scheme for babble noise at all SNRs, while for exhibition this happens for SNRs below 5dB, since at greater SNRs a similar SFM to pre-whiteners based on MS and MMSE is observed. It is also noted that using a higher LPC pre-whitening order implies a higher SFM, i.e., the noise gets closer to white.

B. Pitch estimation

We now consider the task of estimating the pitch ω_0 of a periodic signal buried in additive coloured noise. Voiced speech segments can be modeled as a periodic signal $s(n)$ consisting of L harmonics whose frequencies are an integer multiple of ω_0, having a real amplitude C_l and phase $\psi_l \in [0, 2\pi)$. When such signal segments are contaminated by uncorrelated additive coloured gaussian noise $z(n)$, the signal model becomes $x(n) = \sum_{l=1}^{L} C_l \cos(n\omega_0 l + \psi_l) + z(n)$. In particular for speech, this model is valid for short time segments (∼20-30 ms) where the speech is considered as stationary.

When K noisy samples are stacked in a vector as $\mathbf{x} = [x(0) \ ... \ x(K-1)]^T$, the signal model becomes $\mathbf{x} = \mathbf{s} + \mathbf{z} = \mathbf{Bc} + \mathbf{z}$, where $\mathbf{c} = \frac{1}{2}[C_1 e^{j\psi_1} \ C_1 e^{-j\psi_1} \ ... \ C_L e^{j\psi_L} \ C_L e^{-j\psi_L}]$, $\mathbf{B} = [\mathbf{b}(\omega_0) \ \mathbf{b}^*(\omega_0) \ ... \ \mathbf{b}(\omega_0 L) \ \mathbf{b}^*(\omega_0 L)]$ and $\mathbf{b}(\omega_0 l) = [1 \ e^{-j\omega_0 l} \ ... \ e^{-j\omega_0 l(K-1)}]^T$. If $\mathbf{z} = [z(0) \ z(1) \ ... \ z(K-1)]^T$ is WGN, the ML pitch estimate $\hat{\omega}_0$ is [4], [13]

$$\hat{\omega}_0 = \underset{\omega_0}{\arg\max} \ \mathbf{x}^T \mathbf{\Pi}_B \mathbf{x}, \quad (13)$$

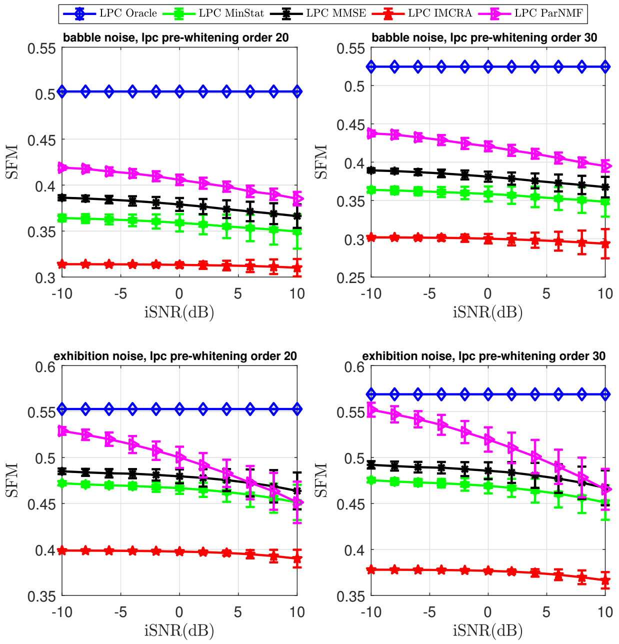

Figure 1: mean and 95% confidence interval of the SFM as a function of SNR.

where $\mathbf{\Pi}_B = \mathbf{B}(\mathbf{B}^H\mathbf{B})^{-1}\mathbf{B}^H$ with $(\cdot)^H$ denoting the hermitian transpose. As we are here concerned with pitch estimation in coloured noise, an LPC pre-whitener can be applied to the noisy vector $\mathbf{x}$ since asymptotically this only modifies the complex amplitude vector $\mathbf{c}$ [6] and not ω_0. Solving (13) in a fast way is described in [12].

In the tested Keele database excerpts, the pitches which were manually annotated are considered here as the ground truth [21]. In order to match the available ground truth, segments of duration 30 ms and an overlap of 20 ms between them were used for the pitch estimation setup. Babble and exhibition noise were added to the testing sentences at SNRs from -4 to 10 dB. After pre-whitening the noisy signals, the pitch was estimated in an interval $[60, 380]$ Hz, with a maximum possible of 15 harmonics. The evaluation was done in terms of gross error rates (GER), which is the proportion of frames where both the ground truth and the pitch estimator result in the presence of a pitch (i.e., $\hat{L} > 0$), where the relative error of the estimated pitch is larger than a certain percentage [22]. Here we use 10%. An LPC pre-whitening order $P = 30$ is used for both scenarios, since from the SFM experiment we saw that with a higher P the noise can get closer to white. As a reference, the pitch was also estimated without any pre-whitening (WGN assumption). The results are depicted in the first row of Fig. 2.

We also conducted an experiment with a specific speaker of the CMU Arctic database, for which a codebook was trained on 24 minutes of speech material (with the same parameters as the general speech codebook), and then we evaluated the pitch estimates on 40 sentences from the same speaker, not included in the training. The evaluation was also done with 30 ms segments, with an overlap of 20 ms between them. For this case, the ground truth was obtained by estimating pitches from the clean speech segments using (13). We also evaluated the pitch estimation performance on 40 sentences from same speakers of the general speech codebook, which were not used for the training. Results for the specific speaker are depicted in the second row of Fig. 2, and for general speakers in the last row of Fig. 2.

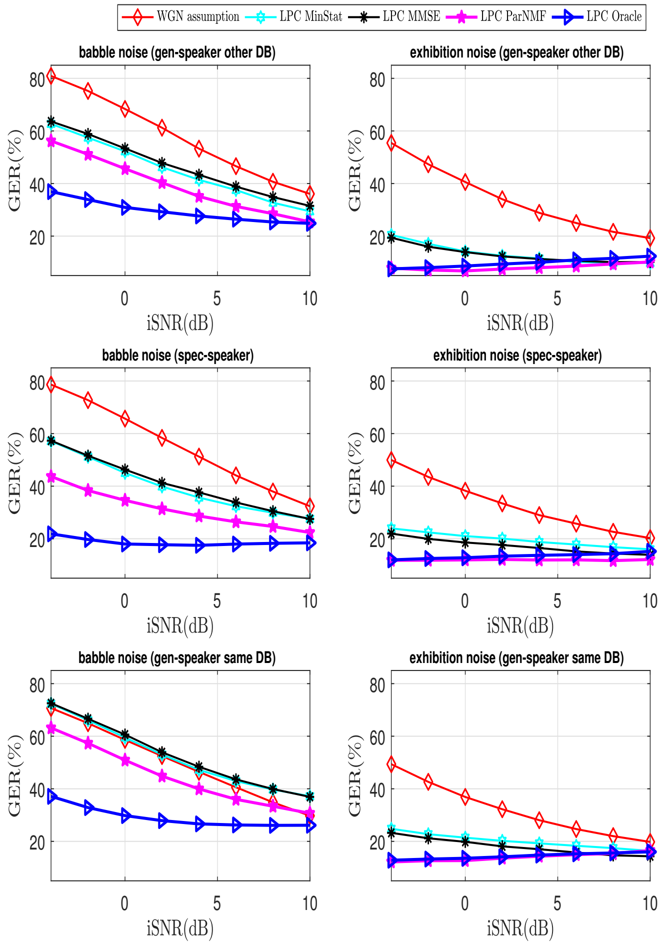

Figure 2: Gross error rate (GER) as a function of SNR.

It is seen that the suggested pre-whitener helps better in reducing the GER of the pitch estimates, in comparison to others based on well-known noise PSD estimates (MS and MMSE), since for both noise types, the parametric NMF pre-whitener performance is the closest to the oracle one. In fact, for exhibition noise type the performance gets very similar to the oracle pre-whitening, implying that a more accurate noise PSD could be captured. We speculate that this is due to that the exhibition noise is more stationary.

C. Speech enhancement

Finally, we verify that using the proposed pre-whitener as a pre-processor will not ruin the signal. The approach is to do speech enhancement on the pre-whitened noisy signal from a WGN assumption (i.e., with the WGN variance as the

Table III: Results of segSNR improvement in babble noise

Enhanc. method w/ OM-LSA	**segSNR improvement (in dB)**			
	-2dB	***1dB***	***4dB***	***7dB***
MS pre-wh	2.74±0.20	2.61±0.20	2.43±0.23	2.17±0.28
MMSE pre-wh	3.15±0.22	2.94±0.24	2.63±0.30	2.30±0.40
ParNMF pre-wh	3.75±0.29	3.28±0.25	2.71±0.33	2.07±0.51
Conv.(no pre-wh)	3.41±0.25	3.06±0.26	2.63±0.36	2.19±0.52

Table IV: Results of PESQ in babble noise

Enhanc. method w/ OM-LSA	**PESQ**			
	-2dB	***1dB***	***4dB***	***7dB***
Noisy Speech	1.63±0.15	1.77±0.12	1.96±0.11	2.16±0.10
MS pre-wh	1.71±0.08	1.93±0.08	2.16±0.07	2.39±0.07
MMSE pre-wh	1.72±0.08	1.94±0.07	2.17±0.06	2.40±0.05
ParNMF pre-wh	1.74±0.10	1.97±0.07	2.21±0.05	2.41±0.04
Conv.(no pre-wh)	1.73±0.09	1.95±0.08	2.18±0.06	2.41±0.05

single noise parameter), and then undoing the pre-whitening by applying the inverse of the pre-whitening filter. It is important to note that we do not encourage to pre-whiten a noisy signal before enhancing it in a real setup, it only serves as a mean of verification of the presented pre-whitener. We use the optimally modified log-spectral amplitude estimator (OM-LSA) [23] algorithm for this enhancement task. The WGN variance is also calculated when one computes the noise prediction coefficients from the Levinson-Durbin recursion, as explained in Sec. II. In order that this WGN variance does not change abruptly, a recursive smoothing with a smoothing factor of 0.88 is used after computing the noise PSD from (4).

The evaluation is done under babble noise conditions, and again a pre-whitening order $P = 30$ is used, including also pre-whitening based on MS and MMSE. Noisy speech is also enhanced without applying a pre-whitener, i.e., by using a conventional noise PSD estimate [10] directly with OM-LSA. Segmental SNR improvement and PESQ are reported in Tables III and IV, where 95% confidence intervals are seen for each value. In general, the performance from the proposed pre-whitener is better in comparison to the other pre-whiteners since it results in a higher average segSNR improvement (below 7dB) and higher average PESQ. Similar results to the conventional enhancement method are seen by using the presented pre-whitener, which indicates that a signal can be recovered even if it was pre-whitened for another purpose.

V. Conclusions

In this work, we proposed a new adaptive NMF based pre-whitener with pre-trained spectral envelopes parametrized with AR coefficients. The proposed pre-whitener achieves a higher spectral flatness in comparison to pre-whiteners based on classical noise PSD estimators, and therefore reduces considerably the pitch errors. Speech enhancement results based on the WGN assumption show that the pre-whitener can preserve the signal of interest. A fundamental question is why one would pre-whiten the signal instead of just enhancing it, so further research in answering this question should be conducted.

References

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

[2] M. G. Christensen and S. H. Jensen, "Variable order harmonic sinusoidal parameter estimation for speech and audio signals," in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, Oct 2006, pp. 1126–1130.

[3] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795–805, April 1991.

[4] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2009.

[5] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2042–2056, Oct 2013.

[6] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "A study on how pre-whitening influences fundamental frequency estimation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6495–6499.

[7] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for nonstationary voiced speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2354–2367, Dec 2016.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[9] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[10] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[11] J.K. Nielsen, M.S. Kavalekalam, M.G. Christensen, and J.B. Boldt, "Model-based noise psd estimation from speech in non-stationary noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[12] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Processing*, vol. 135, no. Supplement C, pp. 188 – 197, 2017.

[13] P. Stoica and R. Moses, "Spectral analysis of signals," Pearson/Prentice Hall, 2005.

[14] M. S. Kavalekalam, J. K. Nielsen, L. Shi, M. G. Christensen, and J. Boldt, "Online parametric nmf for speech enhancement," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 2320–2324.

[15] C. Févotte E. Vincent, A. Ozerov, "Single-channel audio source separation with nmf: Divergences, constraints and algorithms. audio source separation.," *Springer*, pp. 1–24, 2018.

[16] S. M. Kay, *Modern spectral estimation*, Pearson Education, 1988.

[17] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and trends in communications and information theory, Vol. 2, Iss. 3, p.155-239*, vol. 2, pp. 155–239, 2005.

[18] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[19] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, January 1980.

[20] H. Hirsch and D. Pearce, "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)*, 2000.

[21] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH*, 1995.

[22] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *INTERSPEECH*, 2011.

[23] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, April 2002.