A STUDY ON HOW PRE-WHITENING INFLUENCES FUNDAMENTAL FREQUENCY ESTIMATION

Alfredo Esquivel Jaramillo, Jesper Kjær Nielsen, Mads Græsbøll Christensen

Audio Analysis Lab, CREATE, Aalborg University, Denmark
{aeja, jkn,mgc}@create.aau.dk

ABSTRACT

This paper deals with the influence of pre-whitening for the task of fundamental frequency estimation in noisy conditions. Parametric fundamental frequency estimators commonly assume that the noise is white and Gaussian and, therefore, they are only statistically efficient under those conditions. The noise is coloured in many practical applications and this will often result in problems of misidentifying an integer divisor or multiple of the true fundamental frequency (i.e., octave errors). The purpose of this paper is to see if pre-whitening can reduce this problem, based on noise statistics obtained from existing noise PSD estimation algorithms. For this purpose, different noise types and prediction orders of LPC pre-whitening are considered. The results show that pre-whitening improves significantly the estimation accuracy of an NLS pitch estimator when the noise is fairly stationary. For nonstationary noise, the improvements are modest at best, but we hypothesize that this is due to the noise PSD estimation performance rather than the LPC pre-whitening principle.

***Index Terms*—** fundamental frequency, pre-whitening, spectral flatness measure, noise PSD estimation, gross error rate.

1. INTRODUCTION

The lowest rate at which a periodic signal repeats itself is known as the fundamental frequency. Fundamental frequency estimation is of particular interest in speech applications such as speech enhancement [1], diagnosing illnesses [2], speech decomposition [3, 4] and automatic speech recognition [5]. For example, the speech recordings obtained for the purpose of pathological voice analysis may be corrupted by background noise, and this could affect a proper diagnosis [6]. Fundamental frequency estimators can be grouped as non-parametric and parametric. The non-parametric estimators (e.g. YIN [7]), although fast and conceptually simple, have poor time-frequency resolution and poor noise robustness [8]. A signal model which takes into account the noise presence can be used to derive a parametric estimator [9], based on statistical assumptions. Recently, a fast algorithm which considerably reduces the computational complexity of a nonlinear least squares (NLS) estimator has been proposed [8, 10]. This NLS fundamental frequency estimator is only statistically efficient under a white Gaussian noise (WGN) condition. However, in most real acoustic scenarios the noise is coloured such as car noise and street noise. Estimating the fundamental frequency with a WGN assumption sometimes results in misidentifying a multiple or divisor of the true value (i.e., octave errors). Therefore, a pre-whitening scheme should be applied to the noisy signals, which renders the coloured noise closer to WGN.

The pre-whitening of noisy speech can be done either via the Cholesky factorization [9] or with a FIR filter, for example one based on linear prediction [11]. By applying the Cholesky factor, the signal model needs to be modified as in [12]. Therefore, since the structure of the problem is altered, the fast NLS method cannot be directly applied. A pre-whitening FIR filter which changes the coloured noise into white noise, can preserve the model as only the amplitudes and phases are altered [13]. We focus on this principle in this paper. Therefore, information on the noise spectrum, i.e., noise statistics, is needed. For example, in [11, 14, 15], the noise statistics and the AR parameters of the coloured noise are only estimated during speech-absence periods, assuming that the noise is stationary. Those can be obtained from a voice activity detector (VAD). However, some noise types such as babble and restaurant noise may be non-stationary, so their noise characteristics are time-varying. This issue has been addressed in some noise power spectral density (PSD) estimation algorithms, such as minimum statistics (MS) [16], improved minima controlled recursive averaging (IMCRA) [17], and minimum mean squared error (MMSE) based estimation [18]. This paper intends to extend the work in [13] on pre-whitening. In order to study the effectiveness of these noise PSD estimation algorithms when applying pre-whitening for the purpose of fundamental frequency estimation, the evaluation will be done for both male and female speech, as well as considering different types of real-life noise.

The rest of the paper is structured as follows. Section 2 details the signal model, the fundamental frequency estimator that assumes WGN and details on the pre-whitening schemes. Section 3 explains the experimental setup and the results in terms of spectrograms, gross error rates and spectral flatness measure. Finally, section 4 concludes the work.

2. SIGNAL MODEL AND PRE-WHITENING

We present the signal model, the fundamental frequency estimator, and the pre-whitening schemes in this section. For voiced speech segments, the signal $s(n)$ is modelled by L harmonic components whose frequencies are an integer multiple of the fundamental frequency ω_0, having real amplitude $A_l > 0$ and phase $\psi_l \in [0, 2\pi)$. The signal is buried in additive (white or coloured) Gaussian noise $e(n)$, which is uncorrelated with $s(n)$. For $n = 0, 1, ..., N-1$ (where the clean signal is considered being stationary), the signal model is given as

$$x(n) = s(n) + e(n) = \sum_{l=1}^{L} A_l \cos(n\omega_0 l + \psi_l) + e(n). \qquad (1)$$

By using the Euler's identity, the model can be expressed as

This work is funded by CONACYT, under grant 418437.

$$x(n) = \sum_{l=1}^{L} \left(a_l z^l(n) + a_l^* z^{-l}(n) \right) + e(n), \qquad (2)$$

where $a_l = \frac{A_l}{2} e^{j\psi_l}$, $z(n) = e^{j\omega_0 n}$, and $*$ denotes complex conjugation. For a frame of length N, (2) can be written in vector form as

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}, \qquad (3)$$

where $\mathbf{x} = [x(n)\ x(n+1)\ ...\ x(n+N-1)]^T$ and $\mathbf{e}$ is defined in the same form, $\mathbf{Z} = [\mathbf{z}(1)\ \mathbf{z}(-1)\ ...\ \mathbf{z}(L)\ \mathbf{z}(-L)]$ with $\mathbf{z}(l) = \left[(z(1))^l\ ...\ (z(N))^l\right]^T$, $\mathbf{a} = [a_1\ a_1^*\ ...\ a_L\ a_L^*]$ and $(\cdot)^T$ denotes transpose. With the WGN assumption, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$, σ^2 being the noise variance and $\mathbf{I}_N$ the $N \times N$ identity matrix, the maximum likelihood estimator of ω_0 is found by first replacing the amplitudes in (3) by their least-squares estimates, $\hat{\mathbf{a}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}$, and then by minimizing the residual power $\|\mathbf{x} - \mathbf{Z}\hat{\mathbf{a}}\|_2^2$, i.e.,

$$\hat{\omega}_0 = \arg\min_{\omega_0} \|\mathbf{x} - \mathbf{Z}\hat{\mathbf{a}}\|_2^2 = \arg\min_{\omega_0} \|\mathbf{x} - \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}\|_2^2. \quad (4)$$

Here $(\cdot)^H$ denotes hermitian-transposition. This nonlinear least squares (NLS) minimization problem can be solved in a fast way by exploiting the matrix structure (for further details, see [8]). However, this is only statistically efficient with the WGN assumption. In real scenarios, the noise is usually coloured, i.e., $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_e)$, where $\mathbf{Q}_e$ is the noise covariance matrix. A matrix $\mathbf{L}$ can be used to transform the observed signal as $\mathbf{L}^H\mathbf{x} = \mathbf{L}^H\mathbf{Z}\mathbf{a} + \mathbf{L}^H\mathbf{e}$ such that $\mathbf{v} = \mathbf{L}^H\mathbf{e}$ now is distributed as $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, i.e., the noise is now WGN. The required matrix $\mathbf{L}$ must be the Cholesky factor of $\mathbf{Q}_e^{-1}$, i.e., $\mathbf{L}\mathbf{L}^H = \mathbf{Q}_e^{-1}$. However, the harmonic part is also affected and therefore, the structure of the matrices involved in the fast computation of the cost function of (4). Another approach to pre-whiten the noisy signal (i.e., that renders coloured noise white) is by applying a filter.

To apply a filter that pre-whitens the noisy signal, the coloured noise can be seen as the output of a filter $H(\omega)$ excited with WGN. When the coloured noise is the output of an all-pole (IIR) filter $H(\omega) = \frac{1}{B(\omega)}$, where $B(\omega) = 1 + \sum_{p=1}^{P} b_p e^{-j\omega p}$, the process is said to be autoregressive (AR). Here, P denotes the prediction order and $b_1, ..., b_P$ are the linear prediction coefficients (LPC). In this sense, the inverse FIR filter $B(\omega)$, can be used to recover the white Gaussian samples given the samples of the AR process and the LPC AR coefficients. Applying this filter (b_n in the time domain) to the noisy signal preserves the signal model for the harmonic model part in (2), since

$$b_n * s(n) = b_n * \sum_{l=-L, l\neq 0}^{L} a_l e^{jn\omega_0 l} = \sum_{l=-L, l\neq 0}^{L} \tilde{a}_l e^{jn\omega_0 l}, \qquad (5)$$

where $\tilde{a}_l = a_l \sum_{p=0}^{P} b_p e^{-j\omega_0 p}$, $b_0 = 1$, so only the complex amplitudes are affected and the fundamental frequency remains unchanged. An estimate of b_p, $p = 1, ...P$ can be obtained from the Levinson-Durbin recursion of order P [19] after the noise statistics are estimated. Given $\mathbf{x}$, some noise tracking algorithms such as MS, IMCRA, and MMSE can be used to estimate the noise PSD, defined as [20]

$$\phi_e(\omega) = \lim_{N \to \infty} \frac{1}{N} \mathbb{E}\left[|E(\omega)|^2 \,|\mathbf{x} \right] \qquad (6)$$

where $E(\omega) = \mathbf{f}^H(\omega)\mathbf{e}$ is the DFT of the noise with $\mathbf{f}(\omega) = \left\{e^{jn\omega}\right\}_{n=0}^{N-1}$, and $\mathbb{E}$ denotes the statistical expectation operator. The inverse DTFT of the noise PSD allows us to recover the noise covariance sequence via [20]

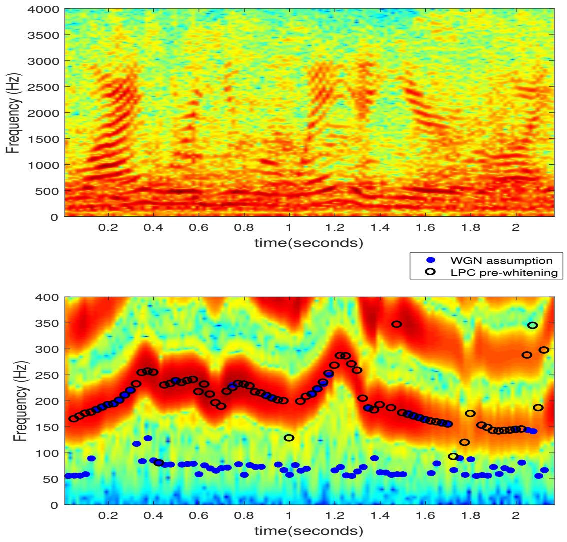

Fig. 1: Spectrogram of a female speech signal contaminated by babble noise at SNR = 3dB (top), and estimated fundamental frequency estimates imposed on the clean signal spectrogram (bottom).

$$r_e(n) = \int_{-\pi}^{\pi} \phi_e(\omega) e^{jn\omega} \frac{d\omega}{2\pi}. \qquad (7)$$

From this estimated covariance, the LPC parameters can be found from the Levinson-Durbin recursion, which form the b_n pre-whitening FIR filter of order P. We refer to this as the LPC pre-whitener.

Another possibility [13] is to derive a FIR filter directly from the N frequency coefficients of the noise PSD $\phi_e(\omega)$. Since $\phi_e(\omega) = \sigma^2 |H(\omega)|^2 = \frac{\sigma^2}{|B(\omega)|^2}$, and assuming a white Gaussian unit variance $\sigma^2 = 1$, the frequency response of the pre-whitening filter is obtained as $B(\omega) = \frac{1}{\sqrt{\phi_e(\omega)}}$, for N frequency points. An FIR filter of order N is found via the inverse DTFT, i.e. $b_n = \int_{-\pi}^{\pi} B(\omega) e^{jn\omega} \frac{d\omega}{2\pi}$, $n = 0, 1, ...N-1$. We refer to this as the FIR pre-whitener.

3. EXPERIMENTAL EVALUATIONS

In this section, we evaluate the influence of the LPC and FIR pre-whitening filters on the fundamental frequency estimation performance, and how well they render the coloured noise closer to white.

We start by demonstrating how pre-whitening can lead to better fundamental frequency estimates. For this, we consider the voiced female speech sentence "Why were you away a year, Roy?", sampled at 8 kHz, with added babble noise from the AURORA database [21] at an SNR of 3 dB. The fundamental frequency is estimated using the NLS estimator every 25 ms from the interval [55 Hz, 370 Hz]: first from WGN assumption and then, after applying an LPC-prewhitener where the LPC coefficients are directly obtained from the noise signal using $P = 7$. The results are depicted in Fig.1. As observed, the fundamental frequency estimates obtained after pre-whitening result in fewer errors compared to the case with no pre-whitening (WGN assumption).

We now consider the speech signals from the Keele reference database [22], which consists of five male and five female speech recordings, where the fundamental frequency is annotated from

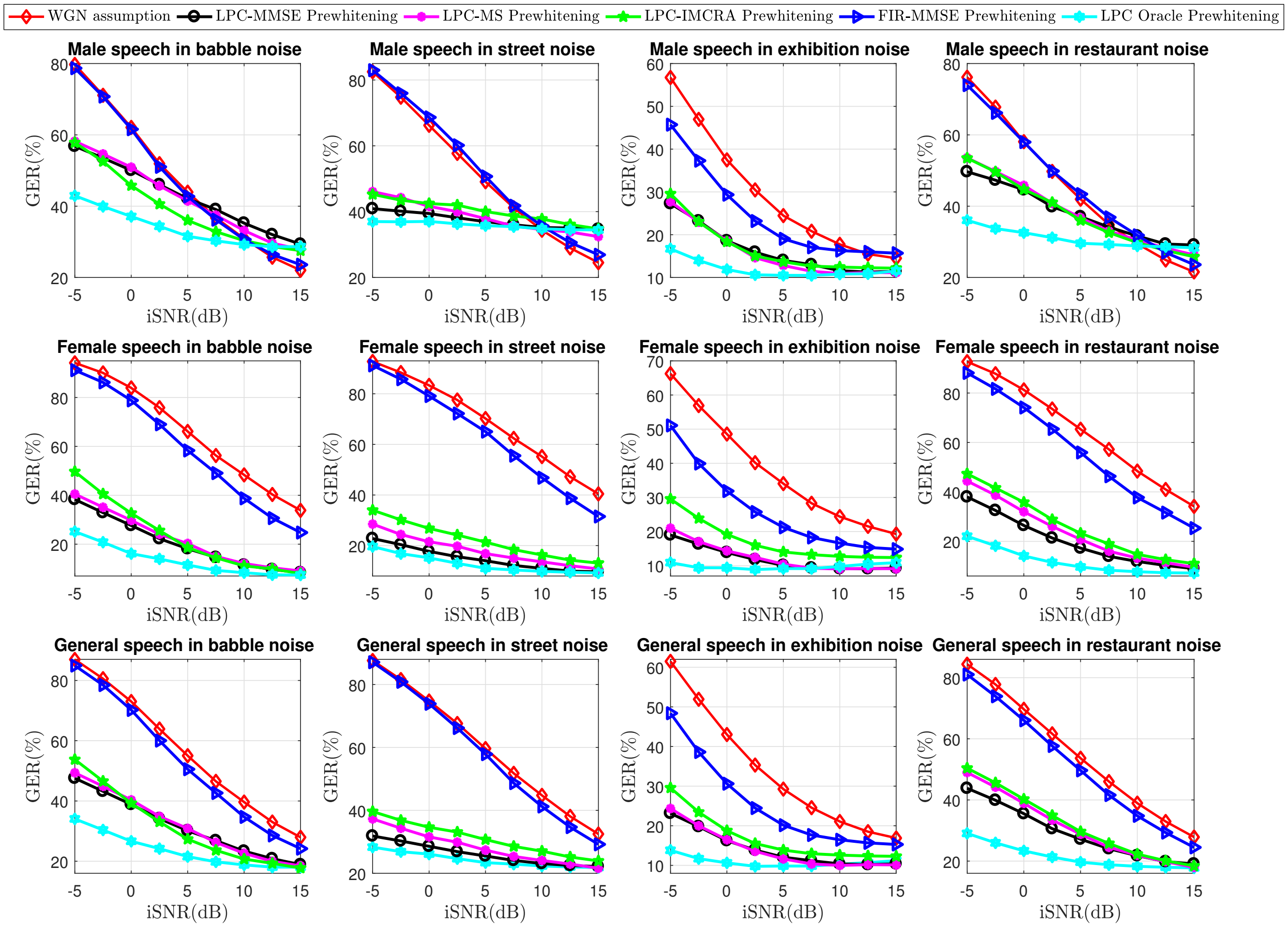

Fig. 2: Gross error rate (GER) as a function of the iSNR for male, female and general speech on different types of real noise.

laryngograph measurements at a frame rate of 10 ms. The signals are resampled from 20 kHz to 8 kHz. The evaluation was done on the first 80,000 samples (10 s) of each speech file. It is important to notice that the annotated fundamental frequencies do not necessarily correspond to the ground truth, but they also correspond to an estimate which was obtained using an autocorrelation method [23].

For evaluating the fundamental frequency estimation accuracy, only the voiced speech frames with periodicity in both the laryngograph signal and on the speech data were considered (refer to [22] for further description). The assessment was done in terms of the gross error rate (GER), which is defined as the percent of voiced frames whose estimated fundamental frequency deviates more than a certain percentage from the ground truth [24]. We here use 10%. The segment length was set to be $N = 240$ (corresponding to 30 ms), and the fundamental frequency was searched using the NLS estimator in an interval [55 Hz, 370 Hz][1], with a maximum possible of $L = 15$ harmonics. In order to have the same frame rate as the ground truth, the shift between frames was set to $N = 80$ (i.e., 10 ms). The evaluation was done with four noise types: street, babble, exhibition and restaurant, which are obtained from the AURORA database [21]. The iSNR is varied from -5 to 15 dB. Three different LPC pre-whiteners were used, according to three noise PSD estimates: MMSE [18], MS [16], and IMCRA [17], so the comparison will allow us to determine which one of them helps better for the task of fundamental frequency estimation. For the FIR pre-whitener, only the MMSE noise PSD estimate is presented, since similar results were observed with respect to the other noise PSD estimators. In order to get an insight in to what is the best performance that can be achieved, the results also include the case where an LPC oracle pre-whitener is used, i.e., where the LPC parameters were computed directly from the noise signal. The order of the LPC pre-whiteners was set to $P = 7$, as this seemed to work well (see also the explanation for the next experiment). The results are displayed in Fig.2, the results are shown separately for male and female speech, and also for general speech. In general, the GER from the LPC oracle pre-whitener is lower for female than for male speech, since most of the power of the coloured noise is in the lower frequencies which coincide with the range of fundamental frequencies of male speech.

The performance from the LPC pre-whitener based on MMSE noise PSD estimation is mostly the closest to the LPC oracle pre-whitener, followed by the one based on MS. For the case of male speech above an iSNR of 10 dB, it seems that it is better to assume WGN or to do FIR pre-whitening to estimate the fundamental frequency (except in the exhibition noise case). Otherwise, in most cases, the benefit of LPC pre-whitening is clear, as the GER resulting from WGN assumption and from FIR pre-whitening is higher. The performance of LPC pre-whitening from noise PSD MMSE es-

[1] The lowest fundamental frequency in an evaluated segment of the Keele database is 57 Hz.

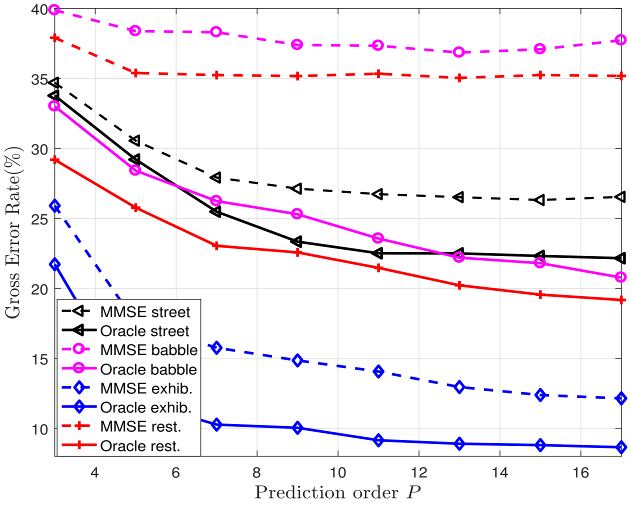

Fig. 3: Gross error rate (GER) as a function of the prediction order P at iSNR = 0 dB, for general speech.

timates is very close to the oracle for the street noise case, while for the other noise types (babble, exhibition and restaurant) there is still room for improvement for attaining lower GERs (closer to the oracle performance).

In the next experiment, we investigate the influence of the prediction order P for LPC pre-whitening. We used the same setup from previous experiment. Since from it, lower GERs were seen from the MMSE noise PSD estimate, and due to the lack of space, we only show the curves corresponding to the pre-whitener from the MMSE noise PSD tracker and compare them to those obtained from LPC oracle pre-whitening. The results are shown in Fig. 3 for an iSNR = 0 dB for the general speech case. The GERs corresponding to the WGN assumption and the FIR pre-whitening can be seen for comparison purposes from Fig. 2 at 0 dB. From the oracle pre-whitening curves, the best possible performance was obtained for the exhibition noise, followed by restaurant and with street and babble noise having the highest GER depending on which P is used. However, by increasing P the GER slightly reduced or kept nearly constant. By applying LPC pre-whitening based on the MMSE noise PSD estimate, the GER also slightly decreased or remained nearly constant as P increased. The lowest GER is also seen for the exhibition noise, but the next lower GER is for street and not for restaurant noise, as opposed to the oracle pre-whitener case. The differences between the GER from oracle and estimated LPC pre-whitener are larger for restaurant (between 8.5 and 16 %, increasing with P) and babble noise (between 6.5 and 17 %, increasing with P) than for street (between 1 and 4.5 %) and exhibition (between 3.5 and 5.5 %) noise types. We speculate that this is due to that street and exhibition are more stationary than restaurant and babble noise types, whose statistics may be more difficult to estimate. Larger differences occuring when P is high, for the babble and restaurant noise types, implies that even if a better noise PSD spectrum could be captured (since a lower GER could be achieved), the conventional noise PSD estimators do not react quickly to nonstationary noise conditions and, therefore, the estimated noise PSD spectrum does not correctly fit the true one. This suggests a future improvement of pre-whitening, for example based on codebook based approach [25, 26], which can better encompass the noise characteristics. Based on this, we did not select a very high value of P for the previous experiment.

Table 1: Comparison of SFM at iSNR = 0 dB for general speech.

		SFM (Spectral Flatness Measure)				
		FIR	LPC1	LPC2	LPC3	LPCO
Street (0.04)	$P = 7$	0.13	0.45	0.44	0.34	**0.50**
	$P = 14$	0.13	0.46	0.45	0.35	**0.53**
Babble (0.07)	$P = 7$	0.17	0.40	0.39	0.37	**0.47**
	$P = 14$	0.17	0.41	0.39	0.36	**0.51**
Exhib. (0.29)	$P = 7$	0.43	0.45	0.45	0.43	**0.48**
	$P = 14$	0.43	0.48	0.47	0.43	**0.53**
Rest. (0.08)	$P = 7$	0.20	0.42	0.40	0.38	**0.49**
	$P = 14$	0.20	0.43	0.40	0.35	**0.52**

A measure of the correlation structure of the noise, and therefore its color degree, is given by the spectral flatness measure (SFM). Therefore, the pre-whitening schemes can be compared in terms of this SFM, which is defined as

$$\mathrm{SFM} = \frac{\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \phi(\omega) d\omega\right)}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \phi(\omega) d\omega} \qquad (8)$$

which is interpreted as the ratio between the geometric mean and the arithmetic mean of the power spectrum $\phi(\omega)$ [19]. The larger this value, the flatter the noise becomes. This quantity is bounded between 0 and 1, where SFM $\rightarrow$ 0 means that the noise is more coloured and SFM $\rightarrow$ 1 implies white noise.

The mean SFM was calculated at an iSNR = 0 dB for the different noise types, for two prediction orders $P = 7$ and $P = 14$. The SFM values after pre-whitening are similar to other iSNRs, as was also evaluated in [13], so only the results at 0 dB are shown in Table 1. The SFM for each noise type before pre-whitening is shown in brackets. The table reports the SFM of the noise after pre-whitening the noisy signal with the FIR method using MMSE noise PSD estimate, and also with the LPC pre-whitening with the noise trackers MMSE, MS and IMCRA (LPC1, LPC2 and LPC3, respectively). The last column, LPCO, corresponds to the SFM obtained by using the LPC oracle pre-whitener, i.e., the highest possible SFM with a specific P. For MMSE and MS LPC pre-whiteners, the SFM increases as P increases, something that not always happens by using IMCRA. The closest SFM to the oracle SFM can be obtained from the LPC MMSE pre-whitener. The difference between them is larger for $P = 14$ than for $P = 7$. The SFM obtained from FIR pre-whitening is much lower compared to LPC pre-whitening in most cases, except for exhibition noise, in which the value is very near to the one attained from the LPC pre-whitening. Larger differences between the SFM from oracle and noise trackers are seen for more nonstationary noise types, i.e., restaurant and babble.

4. CONCLUSIONS

In this paper, we evaluated the influence of pre-whitening filters based on noise PSD estimation methods for fundamental frequency estimation. We also evaluated how well the LPC and FIR pre-whiteners can distribute the noise power across the entire frequency range in terms of the SFM measure. The LPC pre-whitening based on MMSE results in lower GER of the fundamental frequency estimates and highest SFM compared to the LPC pre-whitening based on the other noise PSD estimates. Moreover, a better improvement is still possible to be achieved, specially in the case of nonstationary noise types.

5. REFERENCES

[1] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1948–1963, Sept 2012.

[2] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, March 2006.

[3] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, Oct 2001.

[4] A. Esquivel, J. K. Nielsen, and M. G. Christensen, "On optimal filtering for speech decomposition," in *26th European Signal Processing Conference (EUSIPCO)*, May 2018.

[5] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2494–2498.

[6] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "A supervised approach to global signal-to-noise ratio estimation for whispered and pathological voices," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 296–300.

[7] A. D. Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[8] 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.

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

[10] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast and statistically efficient fundamental frequency estimation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 86–90.

[11] Z. Goh, K. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 510–524, Sept 1999.

[12] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 092953–, 2007.

[13] 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.

[14] J. Huang and Y. Zhao, "An energy-constrained signal subspace method for speech enhancement and recognition in white and colored noises," vol. 26, pp. 165–181, 1998.

[15] J. Huang and Y. Zhao, "An energy-constrained signal subspace method for speech enhancement and recognition in colored noise," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, May 1998, vol. 1, pp. 377–380 vol.1.

[16] 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.

[17] 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.

[18] 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.

[19] P. P. Vaidyanathan, *The Theory of Linear Prediction*, Morgan & Claypool, 2007.

[20] P. Stoica, *Introduction to spectral analysis*, Prentice Hall, 1997.

[21] 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.

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

[23] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[24] F. Flego and M. Omologo, "Robust f0 estimation based on a multi-microphone periodicity function for distant-talking speech," in *2006 14th European Signal Processing Conference*, Sept 2006, pp. 1–4.

[25] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, Feb 2007.

[26] 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.