A Study on how Pre-Whitening Influences Fundamental Frequency Estimation

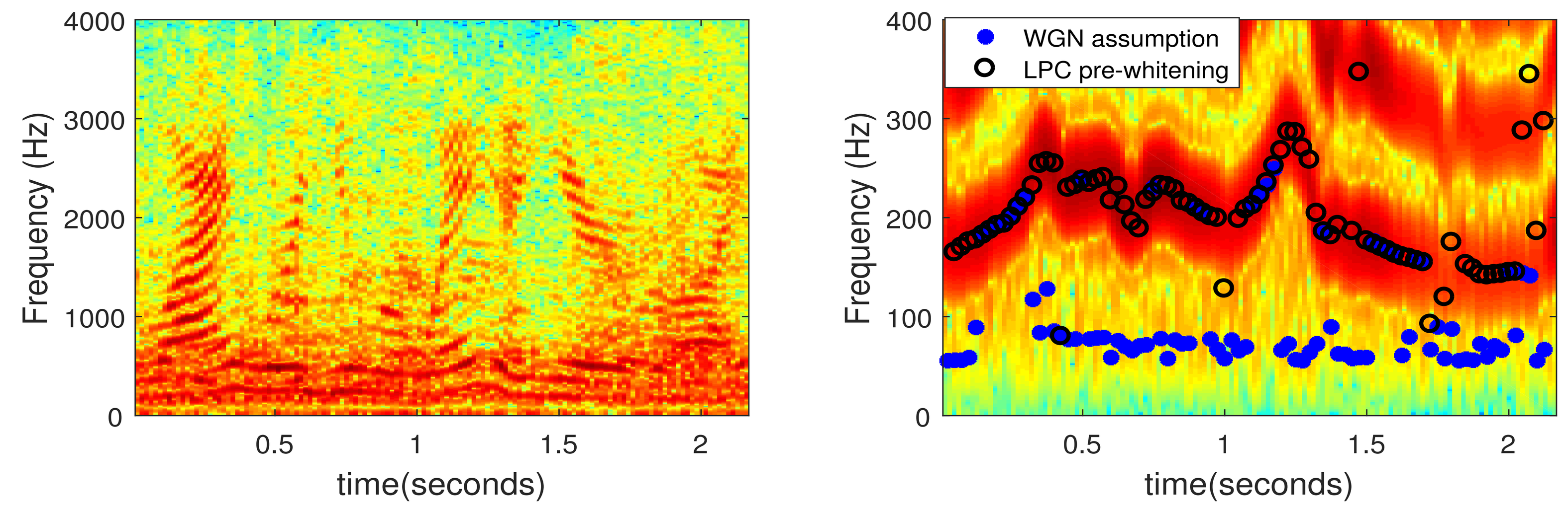
AUDIO ANALYSIS LAB

Alfredo Esquivel Jaramillo, Jesper Kjær Nielsen, and Mads Græsbøll Christensen
{aeja, jkn, mgc}@create.aau.dk

AALBORG UNIVERSITY DENMARK

Introduction

- Parametric fundamental frequency estimators (e.g. NLS), as opposed to non-parametric ones (e.g. YIN), are robust to the noise.
- NLS is statistically efficient under WGN conditions. However, noise is typically coloured.
- Assuming WGN in real scenarios can result in subharmonic fundamental frequency (a.k.a pitch) errors.
- A pre-whitening scheme which renders the coloured noise closer to WGN should be applied.
- Pre-whitening framework based on linear filtering, featuring
  - Noise statistics estimation.
  - Filtering based directly on the estimated noise PSD (FIR pre-whitening) vs. a smoother estimated noise AR spectrum (LPC pre-whitening).
  - How pitch estimation accuracy is improved.
- Example of female voiced speech sentence "Why were you away a year, Roy?", with added babble noise at SNR = 3 dB



Signal Model and NLS Pitch Estimator

- Harmonic signal model for voiced speech segments

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

L is the number of sinusoids whose frequencies are an integer multiple of ω_0, having a real amplitude $A_l > 0$ and phase $\psi_l \in [0, 2\pi)$, and $e(n)$ is the additive gaussian noise.
- For a noisy vector of N noisy samples, $\mathbf{x} = \mathbf{Za} + \mathbf{e}$,

$$\mathbf{Z} = [\mathbf{z}(\omega_0)\ \mathbf{z}^*(\omega_0)\ \dots\ \mathbf{z}(\omega_0 L)\ \mathbf{z}^*(\omega_0 L)], \quad (2)$$

$$\mathbf{z}(\omega_0 l) = [1\ e^{-j\omega_0 l}\ \dots\ e^{-j\omega_0 l(N-1)}]^T, \quad (3)$$

$$\mathbf{a} = \frac{1}{2}\left[A_1 e^{j\psi_1}\ \dots\ A_L e^{j\psi_L}\ A_L e^{-j\psi_L}\right]. \quad (4)$$

- If $\mathbf{e}$ is WGN, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$, the NLS pitch estimator is [1]

$$\hat{\omega}_0 = \arg\min_{\omega_0} \|\mathbf{x} - \mathbf{Z}\hat{\mathbf{a}}\|_2^2 = \arg\min_{\omega_0} \|\mathbf{x} - \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}\|_2^2. \quad (5)$$

- A fast way of solving (5) is described in [2]. In real scenarios, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_e)$. One could apply the Cholesky factor $\mathbf{L}^H\mathbf{x} = \mathbf{L}^H\mathbf{Za} + \mathbf{L}^H\mathbf{e}$ so that $\mathbf{v} = \mathbf{L}^H\mathbf{e}$ is now WGN, but this modifies the estimator in (5).

This work is funded by CONACYT, under grant 418437

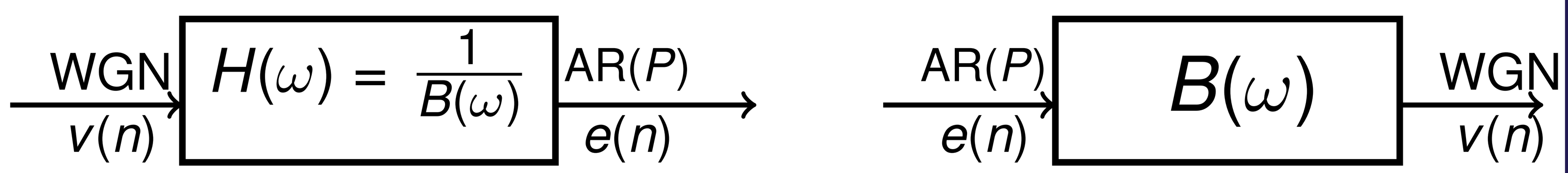
FIR Pre-whitening

- Given $\mathbf{x}$, the noise PSD is estimated using one of the noise trackers such as Minimum Statistics (MS) or MMSE [3]

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

- 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 pre-whitening filter frequency response is obtained as $B(\omega) = \frac{1}{\sqrt{\phi_e(\omega)}}$, for N frequency points.
- An FIR filter is found as $b_n = \int_{-\pi}^{\pi} B(\omega)e^{jn\omega}\frac{d\omega}{2\pi}$, $n = 0, \dots N-1$.

LPC Pre-whitening



- where $B(\omega) = 1 + \sum_{p=1}^{P} b_p e^{-j\omega p}$, and $\{b_p\}_{p=1}^{P}$ are the linear prediction coefficients (LPC) coefficients.
- It only modifies the amplitudes of the harmonic signal model and not ω_0, since

$$b_n * s(n) = b_n * \sum_{l=-L, l\neq 0}^{L} a_l e^{jn\omega_0 l} = \sum_{l=-L, l\neq 0}^{L} \tilde{a}_l e^{jn\omega_0 l}, \quad (7)$$

- In practice, the noise PSD is estimated as in (6) (e.g., using MMSE or MS) and the $\{b_p\}_{p=1}^{P}$ pre-whitening filter is obtained from the Levinson-Durbin recursion after estimating the noise covariance

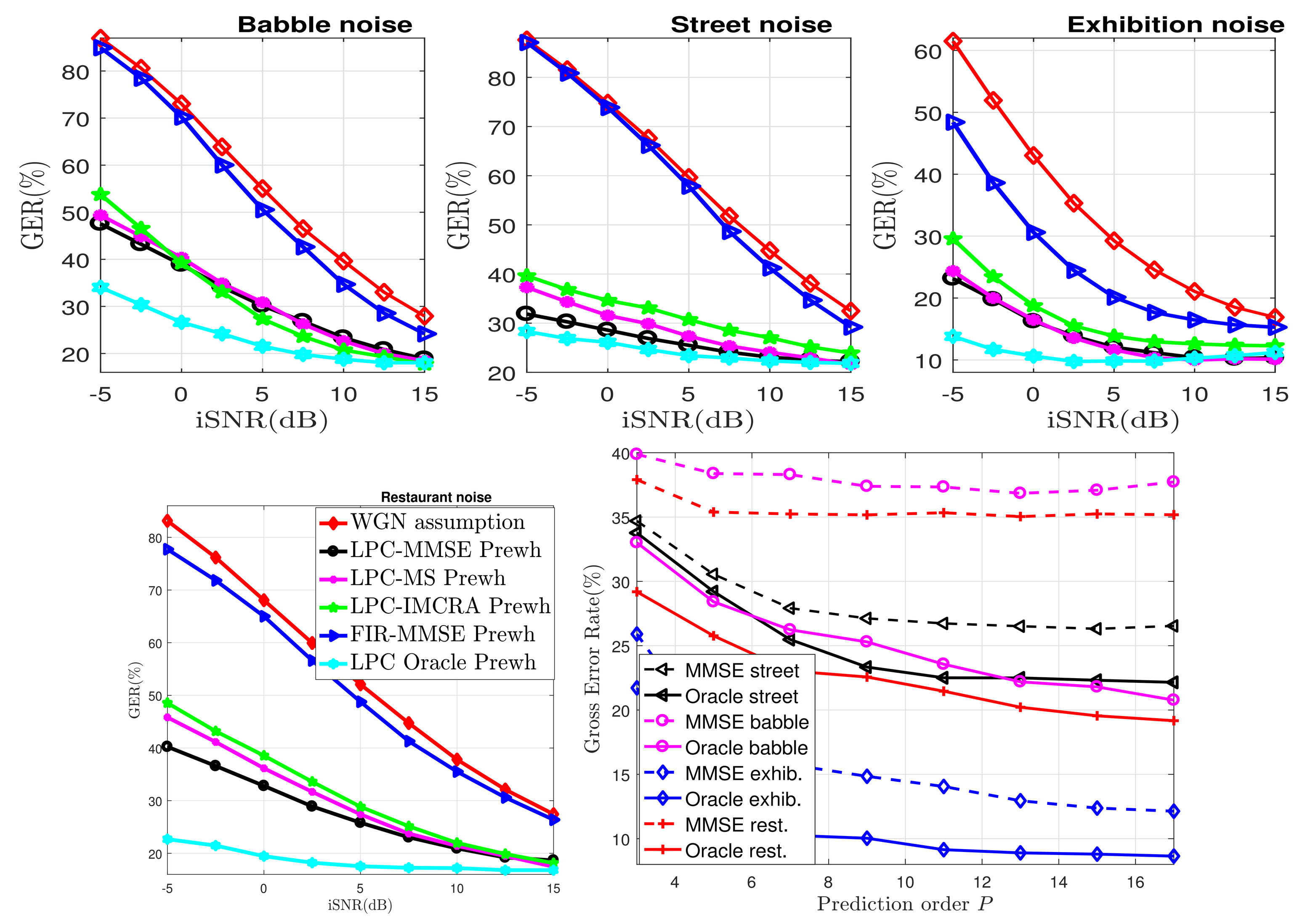
$$r_e(n) = \int_{-\pi}^{\pi} \phi_e(\omega)e^{jn\omega}\frac{d\omega}{2\pi} \quad (8)$$

Experimental Setup

- Keele DB signals at 8 kHz. Annotated ground truth from laryngograph signals and RAPT.
- Added noise from Aurora DB. Evaluation on voiced speech segments in terms of gross error rate (GER).
- NLS pitch estimation parameters: $N = 240$, interval $[55 - 370]$ Hz, maximum of $L = 15$ harmonics.
- LPC pre-whitening based on MMSE(1), MS(2) and IMCRA(3).
- FIR pre-whitening based on MMSE.
- Spectral Flatness Measure (SFM) assesses how closer or further to white noise is the noise spectrum before and after pre-whitening (bounded between 0 and 1)

$$\text{SFM} = \frac{\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\phi(\omega)d\omega\right)}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \phi(\omega)d\omega} \quad (9)$$

Experimental Results



		mean SFM at 0 dB				
		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**

Conclusion

- GER and SFM closer to the Oracle pre-whitening by using LPC pre-whitening based on MMSE.
- Fitting the estimated noise PSD with a smooth AR spectrum is desirable as we avoid inverting small PSD values caused by inaccurate estimates. Therefore FIR pre-whitening is not very helpful.
- There seems to be room for improvement, specially under non-stationary noise types (e.g., babble and restaurant).

Today at 13PM DEMO: Real-time Bayesian Pitch Tracking!

References

[1] P. Stoica, “Spectral analysis of Signals,” Prentice Hall, 2005.

[2] J. K. Nielsen, T. L. Jensen, et al. “Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient,” *Signal Processing*, vol. 135, Sup. C, pp. 188–197, 2017

[3] T. Gerkmann and R. C. Hendriks, “Unbiased mmse-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Speech and Audio Proc.*, vol. 20, no. 4, pp. 1383–1393, 2012