Robust Fundamental Frequency Estimation in Coloured Noise

May. 6, 2020

Alfredo Esquivel Jaramillo[1], Andreas Jakobsson[2], Jesper Kjær Nielsen[1], and Mads Græsbøll Christensen[1]

aeja@create.aau.dk

[1]Audio Analysis Lab, CREATE, Aalborg University
[2]Dept. of Mathematical Statistics, Lund University

Funded by the National Council of Science and Technology of Mexico and by the Swedish Research Council

AALBORG UNIVERSITY
DENMARK

LUND UNIVERSITY

Agenda

Introduction

Signal Model

Single cascade estimation

Proposed Reestimation

Experimental Setup and Results

Conclusion

- **Fundamental frequency (f_0)** is important in several applications: music transcription, audio coding, and speech decomposition.
- Correlation-based methods (Ghahremani, 2014) (Talkin,1995) suffer from octave errors, are not robust to the noise and have time-frequency resolution problems (Nielsen, 2017)
- Parametric methods (Christensen, 2009) (Shi, 2019) also present the problem of octave errors in coloured noise conditions, as they are typically derived under a WGN assumption.
- The spectral shape of the noise needs to be considered.
- In that case, the ML solution is not longer the same as the NLS one (Parra, 2001).
- We propose 2 approaches to estimate the speech signal and noise parameters in an iterative manner to improve accuracy.

- Jointly estimating the signal and noise parameters is a nonlinear multidimensional optimization problem (Kay, 1994).
- In the case of independent sinusoids:
    1. The sinusoidal and noise AR parameters are estimated separately in two different steps. This, in only one iteration (Li, 1996).
    2. In the first step, the different sinusoids are estimated using iterative algorithms such as matching pursuit and RELAX.
    3. Despite the noise spectral shape, the one-iteration solution still results in statistically efficient estimates (Stoica, 1997).
- This is not possible for harmonically related sinusoids. Exploiting the harmonic structure (Elvander, 2020) is necessary to allow for improved estimates.
- Ignoring the noise AR structure increases the risk of selecting a wrong peak (octave errors).

Signal Model
Speech and noise model

Harmonic signal model:

$$x(n) = s(n) + e(n) = \sum_{l=-L}^{L} \alpha_l e^{j\omega_l n} + e(n), \tag{1}$$

where

- L: number of harmonics,
- $\alpha_0 = 0, \alpha_l = \alpha^*_{-l}$: complex amplitude of the l^{th} harmonic,
- $\omega_l = \omega_0 l$, i.e., sinusoids are harmonically related.

Coloured noise modelled as an AR process:

$$e(n) = -\sum_{i=1}^{P} a_i e(n-i) + w(n), \tag{2}$$

where

- $\{a_i\}_{i=1}^{P}$: noise AR parameters,
- $w(n)$ is a driving WGN process $\mathcal{N}(0, \sigma_w^2)$.

Audio Analysis Lab, CREATE, Aalborg University, Denmark

Vector model

A segment of N samples is expressed as

$$\mathbf{x} = \mathbf{s} + \mathbf{e} = \mathbf{Z}_L(\omega_0)\boldsymbol{\alpha} + \mathbf{e}, \quad (3)$$

where $\mathbf{x} = \begin{bmatrix} x(0) & \cdots & x(N-1) \end{bmatrix}^T$, and $\mathbf{e}$ is similarly defined, and

$$\mathbf{Z}_L(\omega_0) = \begin{bmatrix} \mathbf{z}(\omega_0) & \mathbf{z}^*(\omega_0) & \cdots & \mathbf{z}(\omega_0 L) & \mathbf{z}^*(\omega_0 L) \end{bmatrix}, \quad (4)$$

$$\mathbf{z}(\omega) = \begin{bmatrix} 1 \; e^{j\omega} & \cdots & e^{j\omega(N-1)} \end{bmatrix}^T, \quad (5)$$

$$\boldsymbol{\alpha} = \frac{1}{2} \begin{bmatrix} A_1 e^{j\psi_1} & \cdots & A_L e^{j\psi_L} & A_L e^{-j\psi_L} \end{bmatrix}^T. \quad (6)$$

When $L \neq 0$, the model describes a voiced speech segment, otherwise is a not-voiced one (unvoiced or speech pauses).

Single cascade estimation
Pre-whitening+NLS pitch estimates

- The signal is first pre-whitened. Otherwise, we observe a high number of gross errors (> 70%) and voicing detection errors (> 45%), especially at low SNRs (Esquivel, 2019).

Single cascade estimation
Pre-whitening+NLS pitch estimates

- The signal is first pre-whitened. Otherwise, we observe a high number of gross errors (> 70%) and voicing detection errors (> 45%), especially at low SNRs (Esquivel, 2019).
- An AR spectrum is fitted to the estimated noise PSD $\Phi_e(k)$, $k = 0, 1, ..., N-1$ (e.g. MS, MMSE-SPP or parametric NMF).

Single cascade estimation

Pre-whitening+NLS pitch estimates

- The signal is first pre-whitened. Otherwise, we observe a high number of gross errors (> 70%) and voicing detection errors (> 45%), especially at low SNRs (Esquivel, 2019).
- An AR spectrum is fitted to the estimated noise PSD $\Phi_e(k)$, $k = 0, 1, ..., N - 1$ (e.g. MS, MMSE-SPP or parametric NMF).
- The time-varying AR pre-whitener $A(\omega) = 1 + \sum_{i=1}^{P} a_i e^{-j\omega i}$ is applied (it does not modify ω_0).

Single cascade estimation

Pre-whitening+NLS pitch estimates

- The signal is first pre-whitened. Otherwise, we observe a high number of gross errors (> 70%) and voicing detection errors (> 45%), especially at low SNRs (Esquivel, 2019).
- An AR spectrum is fitted to the estimated noise PSD $\Phi_e(k)$, $k = 0, 1, ..., N-1$ (e.g. MS, MMSE-SPP or parametric NMF).
- The time-varying AR pre-whitener $A(\omega) = 1 + \sum_{i=1}^{P} a_i e^{-j\omega i}$ is applied (it does not modify ω_0).
- As the noise is now closer to white, an initial ω_0 is obtained from the nonlinear least squares (NLS) estimator

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{x}^T \mathbf{Z}_L(\omega_0) \left[\mathbf{Z}_L^H(\omega_0)\mathbf{Z}_L(\omega_0)\right]^{-1} \mathbf{Z}_L^H(\omega_0)\mathbf{x}, \qquad (7)$$

which can be solved in a rapid order-recursive manner (Nielsen, 2016). L is selected from model selection criteria (e.g BIC).

Proposed Reestimation
First approach: Approximate Maximum Likelihood (ML)

- Using the harmonic structure, for a given $\hat{\omega}_0$, the LS amplitudes estimates are

$$\hat{\boldsymbol{\alpha}} = [\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{Z}_L(\hat{\omega}_0)]^{-1}\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{x}. \quad (8)$$

Proposed Reestimation
First approach: Approximate Maximum Likelihood (ML)

- Using the harmonic structure, for a given $\hat{\omega}_0$, the LS amplitudes estimates are

$$\hat{\boldsymbol{\alpha}} = [\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{Z}_L(\hat{\omega}_0)]^{-1}\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{x}. \quad (8)$$

- The residual noise may be estimated as $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{Z}_L(\hat{\omega}_0)\hat{\boldsymbol{\alpha}}$.

Proposed Reestimation
First approach: Approximate Maximum Likelihood (ML)

- Using the harmonic structure, for a given $\hat{\omega}_0$, the LS amplitudes estimates are

$$\hat{\boldsymbol{\alpha}} = [\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{Z}_L(\hat{\omega}_0)]^{-1}\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{x}. \tag{8}$$

- The residual noise may be estimated as $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{Z}_L(\hat{\omega}_0)\hat{\boldsymbol{\alpha}}$.
- Reestimate $\{\hat{a}_i\}_{i=1}^{P}$ using the AR modeling *autocorrelation method* $\rightarrow$ form a new pre-whitened signal vector $\rightarrow$ new $\hat{\hat{\omega}}_0$ and $\hat{L}$ (rapid implementation (Nielsen, 2017)).

Proposed Reestimation

First approach: Approximate Maximum Likelihood (ML)

- Using the harmonic structure, for a given $\hat{\omega}_0$, the LS amplitudes estimates are

$$\hat{\boldsymbol{\alpha}} = [\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{Z}_L(\hat{\omega}_0)]^{-1}\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{x}. \quad (8)$$

- The residual noise may be estimated as $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{Z}_L(\hat{\omega}_0)\hat{\boldsymbol{\alpha}}$.
- Reestimate $\{\hat{a}_i\}_{i=1}^P$ using the AR modeling *autocorrelation method* $\rightarrow$ form a new pre-whitened signal vector $\rightarrow$ new $\hat{\hat{\omega}}_0$ and $\hat{L}$ (rapid implementation (Nielsen, 2017)).
- This is repeated until convergence, when either the NLS cost function (7) between 2 consecutive iterations is below a threshold value or a maximum number of iterations is reached.

Proposed Reestimation

First approach: Approximate Maximum Likelihood (ML)

- Using the harmonic structure, for a given $\hat{\omega}_0$, the LS amplitudes estimates are

$$\hat{\boldsymbol{\alpha}} = [\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{Z}_L(\hat{\omega}_0)]^{-1}\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{x}. \tag{8}$$

- The residual noise may be estimated as $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{Z}_L(\hat{\omega}_0)\hat{\boldsymbol{\alpha}}$.
- Reestimate $\{\hat{a}_i\}_{i=1}^{P}$ using the AR modeling *autocorrelation method* $\rightarrow$ form a new pre-whitened signal vector $\rightarrow$ new $\hat{\hat{\omega}}_0$ and $\hat{L}$ (rapid implementation (Nielsen, 2017)).
- This is repeated until convergence, when either the NLS cost function (7) between 2 consecutive iterations is below a threshold value or a maximum number of iterations is reached.
- When $\hat{L} = 0$, $\hat{\mathbf{e}} = \mathbf{x}$. If in the next iteration the segment is still detected as not-voiced, the process is stopped for that segment.

Proposed Reestimation

Second approach: based on LCMV filtering

- The speech vector can be estimated by linear filtering as

$$\hat{\mathbf{s}} = \mathbf{H}^T\mathbf{x} = \mathbf{H}^T\mathbf{Z}_L(\hat{\omega}_0)\boldsymbol{\alpha} + \mathbf{H}^T\mathbf{e}. \quad (9)$$

- To minimize the residual noise power and avoid distortion,

$$\min_{\mathbf{H}} \operatorname{Tr}\left\{\mathbf{H}^T\mathbf{R}_e\mathbf{H}\right\} \quad \text{s.t.} \quad \mathbf{H}^T\mathbf{Z}_L(\hat{\omega}_0) = \mathbf{Z}_L(\hat{\omega}_0). \quad (10)$$

- The solution of this optimization problem is

$$\mathbf{H}_{\text{LCMV}} = \mathbf{R}_e^{-1}\mathbf{Z}_L(\hat{\omega}_0)\left(\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{R}_e^{-1}\mathbf{Z}_L(\hat{\omega}_0)\right)^{-1}\mathbf{Z}_L^H(\hat{\omega}_0) \quad (11)$$

- From the GS formula, $\mathbf{R}_e^{-1} = \frac{1}{\sigma_w^2}\left\{\mathbf{L}_1^T\mathbf{L}_1 - \mathbf{L}_2^T\mathbf{L}_2\right\}$, where $\mathbf{L}_1$ and $\mathbf{L}_2$ are upper triangular Toeplitz matrices whose first rows are $[1 \; a_1 \; \ldots \; a_P \; 0]$ and $[0 \; a_P \; \ldots \; a_1]$, respectively.
- Similar to the approximate ML, iteration between harmonic signal and AR parameter estimates until convergence.

Experimental Setup
Speech and Noise Databases

- Keele DB has ground truth, which is an estimate from an autocorrelation method for segments of length 26.5 ms.

Experimental Setup
Speech and Noise Databases

- Keele DB has ground truth, which is an estimate from an autocorrelation method for segments of length 26.5 ms.
- Comparison to non-parametric methods: YIN, RAPT and Cepstrum based, which have a final step of refinement (e.g. median filter or dynamic programming)

Experimental Setup
Speech and Noise Databases

- Keele DB has ground truth, which is an estimate from an autocorrelation method for segments of length 26.5 ms.
- Comparison to non-parametric methods: YIN, RAPT and Cepstrum based, which have a final step of refinement (e.g. median filter or dynamic programming)
- For the proposed, search in [60,400] Hz, with a maximum of $L = 27$ harmonics.

Experimental Setup
Speech and Noise Databases

- Keele DB has ground truth, which is an estimate from an autocorrelation method for segments of length 26.5 ms.
- Comparison to non-parametric methods: YIN, RAPT and Cepstrum based, which have a final step of refinement (e.g. median filter or dynamic programming)
- For the proposed, search in [60,400] Hz, with a maximum of $L = 27$ harmonics.
- Added noise from NOISEX-92 DB: babble, factory and F-16, @iSNR of -5, 5 and 15 dB.

Experimental Setup
Speech and Noise Databases

- Keele DB has ground truth, which is an estimate from an autocorrelation method for segments of length 26.5 ms.
- Comparison to non-parametric methods: YIN, RAPT and Cepstrum based, which have a final step of refinement (e.g. median filter or dynamic programming)
- For the proposed, search in [60,400] Hz, with a maximum of $L = 27$ harmonics.
- Added noise from NOISEX-92 DB: babble, factory and F-16, @iSNR of -5, 5 and 15 dB.
- 6 Monte Carlo simulations per Keele file (10 files) at each iSNR.

Experimental Setup

Speech and Noise Databases

- Keele DB has ground truth, which is an estimate from an autocorrelation method for segments of length 26.5 ms.
- Comparison to non-parametric methods: YIN, RAPT and Cepstrum based, which have a final step of refinement (e.g. median filter or dynamic programming)
- For the proposed, search in [60,400] Hz, with a maximum of $L = 27$ harmonics.
- Added noise from NOISEX-92 DB: babble, factory and F-16, @iSNR of -5, 5 and 15 dB.
- 6 Monte Carlo simulations per Keele file (10 files) at each iSNR.
- The initial applied pre-whitener is based on a parametric NMF noise PSD Φ_e estimate (Esquivel, 2019). AR order $P = 25$.

Experimental Setup
Speech and Noise Databases

- Keele DB has ground truth, which is an estimate from an autocorrelation method for segments of length 26.5 ms.
- Comparison to non-parametric methods: YIN, RAPT and Cepstrum based, which have a final step of refinement (e.g. median filter or dynamic programming)
- For the proposed, search in [60,400] Hz, with a maximum of $L = 27$ harmonics.
- Added noise from NOISEX-92 DB: babble, factory and F-16, @iSNR of -5, 5 and 15 dB.
- 6 Monte Carlo simulations per Keele file (10 files) at each iSNR.
- The initial applied pre-whitener is based on a parametric NMF noise PSD Φ_e estimate (Esquivel, 2019). AR order $P = 25$.
- For this, a spectral basis matrix of typical speech and noise spectral envelopes was required and trained offline.

Experimental Setup
Performance Measures

$$GER = \frac{N_g}{N_{VV}} \times 100\%, \quad (12)$$

$$VDE = \frac{N_{VU} + N_{UV}}{N} \times 100\%, \quad (13)$$

$$FFE = \frac{N_{VU} + N_{UV} + N_g}{N} \times 100\%, \quad (14)$$

- N_{VV}: segments which are voiced by both the ground truth and the estimated f_0.
- N_g: total of the N_{VV} frames with gross relative difference (>20%).
- N_{VU}: segments which are voiced but misdetected as not-voiced.
- N_{UV}: segments which are not-voiced but misdetected as voiced.
- N: total number of segments of an excerpt.

Experimental Results

Example of f_0 estimates of male excerpt @10 dB in F-16 noise

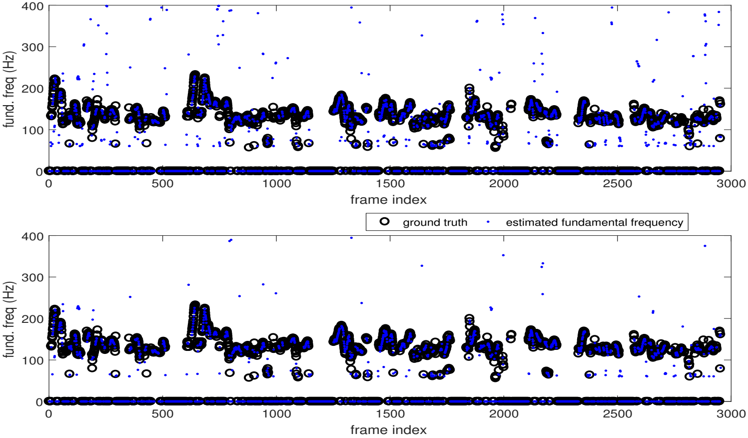

Figure: f_0 ground truth and estimates without iteration (top) and from approx-ML approach (bottom).

Experimental Results

Example of f_0 estimates of female excerpt @15 dB in factory noise

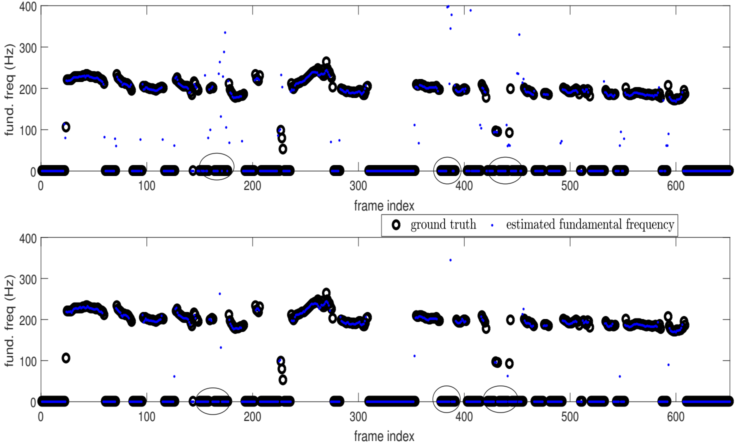

Figure: f_0 ground truth and estimates without iteration (top) and from iterative LCMV filtering approach (bottom).

Experimental Results

Evaluation of parametric and non-parametric estimators

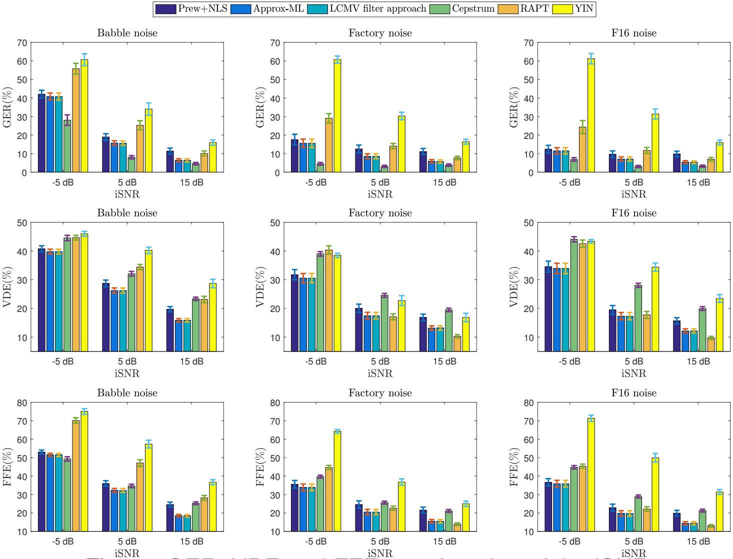

Figure: GER, VDE and FFE as a function of the iSNR.

Conclusion

- Parametric estimators, as the NLS, are only statistically efficient in WGN conditions.

Conclusion

- Parametric estimators, as the NLS, are only statistically efficient in WGN conditions.
- The NLS estimator (rapid implementation) is still useful as long as the additive noise shape is taken into account in forming a pre-whitening filter.

Conclusion

- Parametric estimators, as the NLS, are only statistically efficient in WGN conditions.
- The NLS estimator (rapid implementation) is still useful as long as the additive noise shape is taken into account in forming a pre-whitening filter.
- f_0 along with model order and amplitudes, and AR noise parameters are estimated iteratively. This reduces the risk of octave errors and voicing detection errors noticeably.

Conclusion

- Parametric estimators, as the NLS, are only statistically efficient in WGN conditions.
- The NLS estimator (rapid implementation) is still useful as long as the additive noise shape is taken into account in forming a pre-whitening filter.
- f_0 along with model order and amplitudes, and AR noise parameters are estimated iteratively. This reduces the risk of octave errors and voicing detection errors noticeably.
- The proposed approach offers better performance than state-of-the-art f_0 methods only when the reiterations are applied.

Conclusion

- Parametric estimators, as the NLS, are only statistically efficient in WGN conditions.
- The NLS estimator (rapid implementation) is still useful as long as the additive noise shape is taken into account in forming a pre-whitening filter.
- f_0 along with model order and amplitudes, and AR noise parameters are estimated iteratively. This reduces the risk of octave errors and voicing detection errors noticeably.
- The proposed approach offers better performance than state-of-the-art f_0 methods only when the reiterations are applied.
- Even if the proposed does not take the correlation of consecutive estimates into account, is more robust to the noise.