Pitch Estimation and Tracking with Harmonic Emphasis On The Acoustic Spectrum

Karimian-Azari, Sam; Mohammadiha, Nasser; Jensen, Jesper Rindom; Christensen, Mads Græsbøll

Published in:
I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2015.7178788

Publication date:
2015

Document Version
Early version, also known as pre-print

Pitch Estimation and Tracking with Harmonic Emphasis on the Acoustic Spectrum

April 23, 2015

Sam Karimian-Azari[1], Nasser Mohammadiha[2], Jesper R. Jensen[1], and Mads G. Christensen[1]

ska@create.aau.dk

[1] Audio Analysis Lab, AD:MT, Aalborg University

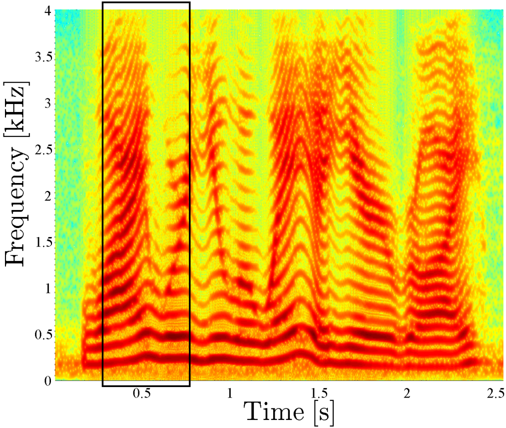
[2] University of Oldenburg

AALBORG UNIVERSITY DENMARK

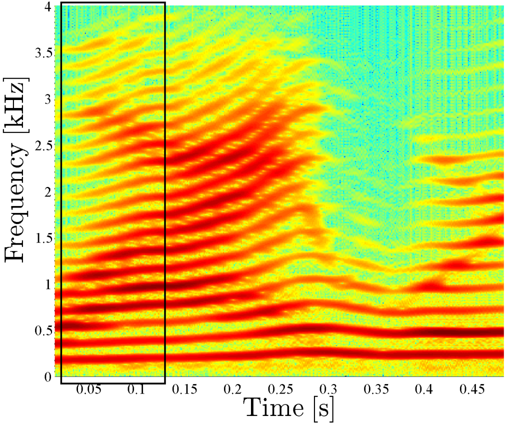
CARL VON OSSIETZKY universität OLDENBURG

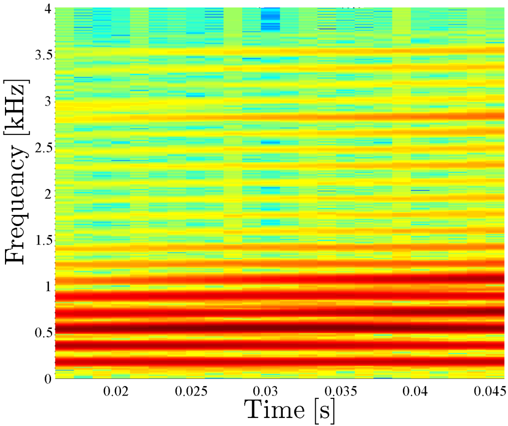
Agenda

- Introduction
  - Noisy Harmonic Signal Approximation
  - ML Pitch Estimate from UFE
- Bayesian Methods
  - Motivation
  - HMM
  - Kalman Filter
- Numerical Results
- Conclusion



Frequency [kHz]
4
3.5
3
2.5
2
1.5
1
0.5
0
0.5
1
1.5
2
2.5
Time [s]


Frequency [kHz]
Time [s]
0
0.5
1
1.5
2
2.5
3
3.5
4
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45



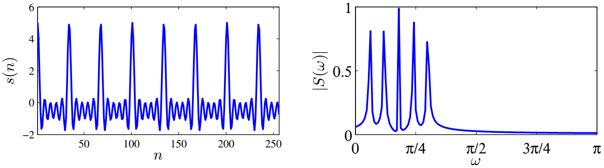
Frequency [kHz]
4
3.5
3
2.5
2
1.5
1
0.5
0
0.02
0.025
0.03
0.035
0.04
0.045
Time [s]

Harmonic Signal Model:

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

where $\omega_l(n) = l\omega_0(n)$ for $l = 1, \ldots, L(n)$,

$L(n)$: number of sinusoids
α_l : real magnitudes
ω_0 : fundamental frequency
φ_l : phases of harmonics



Signal Model
Additive noise

The observed signal can be written as a sum of a desired signal $s(n)$ and a noise signal $v(n)$, i.e.,

$$x(n) = s(n) + v(n) = \sum_{l=1}^{L} \alpha_l \, e^{j\,(\omega_l n + \varphi_l)} + v(n). \quad (2)$$

Signal Model
Phase noise

At a high narrowband SNR, the harmonic frequency ω_l is perturbed with a real-valued phase-noise [S.Tretter 1985], which has a normal distribution with zero mean and the variance

$$\mathrm{E}\{\Delta\omega_l^2(n)\} = \frac{\sigma^2}{2\alpha_l^2} \tag{3}$$

We can approximate $x(n) = \sum_{l=1}^{L} \alpha_l \, e^{j(\omega_l n + \varphi_l)} + v(n)$ like

$$x(n) \approx \sum_{l=1}^{L} \alpha_l \, e^{j(\omega_l n + \Delta\omega_l(n) + \varphi_l)} \tag{4}$$

Signal Model

Unconstrained frequency estimates (UFE)

Unconstrained frequency estimates (UFE) of the constrained frequencies:

$$\hat{\boldsymbol{\Omega}}(n) = \left[\, \hat{\omega}_1(n),\, \hat{\omega}_2(n),\, \ldots,\, \hat{\omega}_L(n) \,\right]^T \quad (5)$$

$$= \mathbf{d}_L(n)\, \omega_0(n) + \Delta\boldsymbol{\Omega}(n), \quad (6)$$

where

$$\mathbf{d}_L(n) = \left[\, 1, 2, \ldots, L(n) \,\right]^T \quad (7)$$

$$\Delta\boldsymbol{\Omega}(n), = \left[\, \Delta\omega_1(n),\, \Delta\omega_2(n),\, \ldots,\, \Delta\omega_L(n) \,\right]^T, \quad (8)$$

and

$$\boldsymbol{R}_{\Delta\Omega}(n) = \mathrm{E}\{\Delta\boldsymbol{\Omega}(n)\Delta\boldsymbol{\Omega}^T(n)\} \quad (9)$$

$$= \frac{\sigma^2}{2}\, \mathrm{diag}\left\{ \frac{1}{\alpha_1^2}, \frac{1}{\alpha_2^2}, \ldots, \frac{1}{\alpha_L^2} \right\}.$$

Max. Likelihood (ML) Pitch Estimator

For the time-frame $\mathbf{x}(n) = \left[\, x(n), x(n-1), \ldots, x(n-M-1) \,\right]^T$, the PDF of the UFE is

$$P(\hat{\boldsymbol{\Omega}}(n)|\omega_0(n)) \sim \mathcal{N}(\mathbf{d}_L(n)\,\omega_0(n), \boldsymbol{R}_{\Delta\Omega}(n)). \tag{10}$$

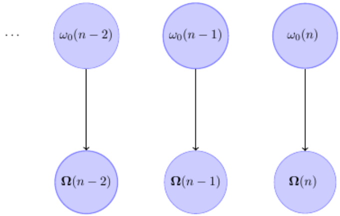
The ML pitch estimator:

$$\hat{\omega}_0(n) = \arg\max_{\omega_0(n)} \log P(\hat{\boldsymbol{\Omega}}(n)|\omega_0(n)) \tag{11}$$

$$= \left[\, \mathbf{d}_L^T(n)\boldsymbol{R}_{\Delta\Omega}^{-1}(n)\mathbf{d}_L(n) \,\right]^{-1} \mathbf{d}_L^T(n)\boldsymbol{R}_{\Delta\Omega}^{-1}(n)\hat{\boldsymbol{\Omega}}(n) \tag{12}$$

Bayesian Pitch Estimator
Motivation

- The ML Estimators are statistically efficient, e.g., the non-linear least-squares (NLS), and the weighted least squares (WLS) [H.Li, et al. 2000], but the minimum variance is limited by the number of samples.
- Consecutive pitch values are estimated independently

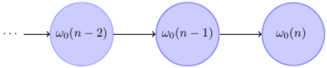


Bayesian Pitch Estimator
Motivation

- Pitch values are usually correlated in a sequence, i.e.,

$$P(\omega_0(n)|\omega_0(n-1), \omega_0(n-2), \cdots), \qquad (13)$$

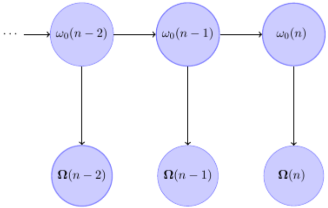
  that motivate Bayesian methods to minimize an error incorporating prior distributions.
- State-of-the-art methods mostly track pitch estimates in a sequential process without concerning noise statistics.



1- Jointly estimate and track pitch incorporating both the harmonic constraints and noise characteristics.
2- Estimate the state $\omega_0(n)$ through a series of noisy observations:

$$P(\omega_0(n)|\hat{\boldsymbol{\Omega}}(n), \hat{\boldsymbol{\Omega}}(n-1), \cdots) \quad (14)$$

3- Recursively update the prior distribution of the pitch value.



Bayesian Pitch Estimator

Discrete state-space (HMM)

$\omega_0(n)$: Discrete random variable (Hidden states)

$P(\omega_0(n)|\omega_0(n-1))$: Transition probability in a 1st-order Markov model,

$$\text{i.e.,} \sum_{\omega_0(n)} P(\omega_0(n)|\omega_0(n-1)) = 1$$

$$\begin{aligned}\hat{\omega}_0(n) &= \arg\max_{\omega_0(n)} \log P(\omega_0(n)|\hat{\mathbf{\Omega}}(n), \hat{\mathbf{\Omega}}(n-1), \cdots) \\ &= \arg\max_{\omega_0(n)} \log P(\hat{\mathbf{\Omega}}(n)|\omega_0(n)) + \log P(\omega_0(n)|\hat{\mathbf{\Omega}}(n-1), \cdots).\end{aligned} \tag{15}$$

The priori distribution is defined recursively like

$$\begin{aligned}P(\omega_0(n)|\hat{\mathbf{\Omega}}(n-1), \hat{\mathbf{\Omega}}(n-2), \cdots) &= \\ \sum_{\omega_0(n-1)} P(\omega_0(n)|\omega_0(n-1)) P(\omega_0(n-1)|\hat{\mathbf{\Omega}}(n-1), \cdots),\end{aligned} \tag{16}$$

where $P(\omega_0(n-1)|\hat{\mathbf{\Omega}}(n-1), \cdots)$ is the past estimate.

Bayesian Pitch Estimator
state-space representation of the pitch continuity

Continuous state-space:

$$\omega_0(n) = \omega_0(n-1) + \delta(n)$$
$$\hat{\boldsymbol{\Omega}}(n) = \mathbf{d}_L(n)\,\omega_0(n) + \Delta\boldsymbol{\Omega}(n),$$

where $\delta(n) \sim \mathcal{N}(0, \sigma_t^2)$ and $\Delta\boldsymbol{\Omega}(n) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R}_{\Delta\Omega}(n))$ are the state evolution and observation noise, respectively.

First, a pitch estimate is predicted using the past estimates as

$$\hat{\omega}_0(n|n-1) = \hat{\omega}_0(n-1|n-1) \tag{17}$$

with the variance

$$\sigma_{\mathrm{K}}^2(n|n-1) = \sigma_{\mathrm{K}}^2(n-1|n-1) + \sigma_t^2. \tag{18}$$

Second, the pitch estimate is updated with the error of

$$\mathbf{e}(n) = \hat{\boldsymbol{\Omega}}(n) - \mathbf{d}_L(n)\,\hat{\omega}_0(n|n-1). \tag{19}$$

Then, the predicted estimate is updated:

$$\hat{\omega}_0(n|n) = \hat{\omega}_0(n|n-1) + \mathbf{h}_{\mathrm{K}}(n)\mathbf{e}(n) \tag{20}$$

$$\mathbf{h}_{\mathrm{K}}(n) = \sigma_{\mathrm{K}}^2(n|n-1)\mathbf{d}_L^T(n)\left[\boldsymbol{\Pi}_L(n)\sigma_{\mathrm{K}}^2(n|n-1) + \boldsymbol{R}_{\Delta\Omega}(n)\right]^{-1}, \tag{21}$$

where $\boldsymbol{\Pi}_L(n) = \mathbf{d}_L(n)\mathbf{d}_L^T(n)$, and update

$$\sigma_{\mathrm{K}}^2(n|n) = \left[1 - \mathbf{h}_{\mathrm{K}}(n)\mathbf{d}_L(n)\right]\sigma_{\mathrm{K}}^2(n|n-1). \tag{22}$$

Covariance Matrix
Estimation

The ML estimator of the covariance matrix among N estimates:

$$\mathbf{R}_{\Delta\Omega}(n) = \mathrm{E}\{\Delta\boldsymbol{\Omega}(n)\Delta\boldsymbol{\Omega}^T(n)\} = \frac{1}{N}\sum_{i=n-N+1}^{n} \Delta\boldsymbol{\Omega}(i)\Delta\boldsymbol{\Omega}^T(i), \quad (23)$$

where $\Delta\boldsymbol{\Omega}(n) = \hat{\boldsymbol{\Omega}}(n) - \hat{\boldsymbol{\mu}}(n)$, and $\boldsymbol{\mu}(n) = \mathrm{E}\{\hat{\boldsymbol{\Omega}}(n)\}$.

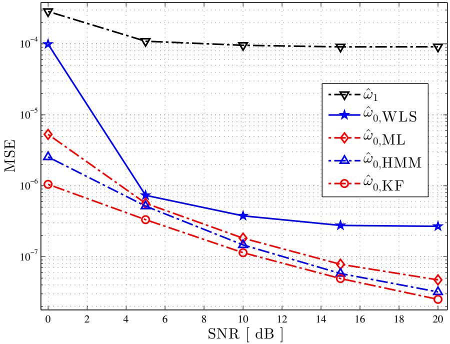
Exponential moving average:

$$\hat{\boldsymbol{\mu}}(n) = \lambda\,\hat{\boldsymbol{\Omega}}(n) + (1-\lambda)\,\hat{\boldsymbol{\mu}}(n-1) \quad (24)$$

The forgetting factor $0 < \lambda < 1$ recursively updates the time-varying mean value.

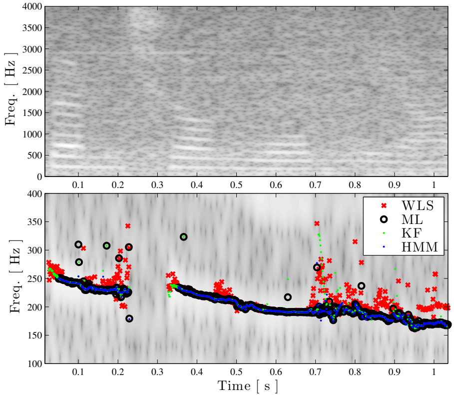
Numerical Results
Synthetic signal

A linear chirp signal ($r = 100$ Hz/s) with $L = 5$ harmonics, random phases, and identical amplitudes during 0.1 s.



$M = 80$, $\omega_0(1) = 400\pi/f_s$, $fs = 8.0$ kHz, $\sigma_t = \sqrt{2\pi r / f_s^2}$, and for the HMM-based pitch estimator, the frequency range $\omega \in [150, 280] \times (2\pi/f_s)$ was discretized into $N_d = 1000$ samples.

Speech signal + Car noise at SNR= 5 dB.



The MAP order estimation [Djuric 1998], $M = 240$, $\lambda = 0.9$, and $N = 150$.

Conclusion

- For pitch estimation, we have formulated the ML estimate from the UFE.
- For pitch estimation and tracking, we have proposed HMM- and KF-based methods.
- Experimental results showed that both HMM- and KF-based methods outperform the corresponding ML pitch estimators.
- The KF-based method statistically performs better than the HMM-based method, while the it tracks pitch changes more accurate than the KF-based method.

Thank you!

AALBORG UNIVERSITY
DENMARK

CARL
VON
OSSIETZKY
universität
OLDENBURG