# Parametric Pitch Estimators for Music Signals

SMC 2012 Tutorial

Mads Græsbøll Christensen

Dept. of Architecture, Design & Media Technology
Aalborg University
mgc@create.aau.dk
http://imi.aau.dk/∼mgc

July 11, 2012

# Outline

Introduction

Statistical Methods

Filtering Methods

Subspace Methods

Amplitude Estimation

Discussion

# Introduction

What is pitch?

- *that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale* (ASA)
- *that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high* (ANSI)

Those *auditory sensations* are caused by signals generated by physical processes.

Those physical processes can (most) often be characterized by a fundamental frequency.

The properties of the observed signal determines what can and cannot be done.

# Fourier Series

A function $f(t)$ that repeats over $T$, i.e., $f(t) = f(t + T)$, is said to be periodic. It has a Fourier series (under certain conditions)

$$f(t) = \sum_{l=-\infty}^{\infty} c_l e^{j2\pi l/Tt}, \quad c_l = \frac{1}{T}\int_{-T/2}^{T/2} f(t)e^{-j\pi l/Tt}dt. \qquad (1)$$

It has fundamental frequency $\omega_0 = 2\pi/T$ and harmonics (overtones) having frequencies $\omega_0 l$ and complex amplitudes $\{c_l\}$.

Essentially states that if a signal is periodic, it has a Fourier series. If the function $f(t)$ is band-limited, it has a finite Fourier series.

Of limited use to us since the integration range (and thus the fundamental frequency) must be known. Also, our signals are not perfectly periodic and are noisy.

# An Example

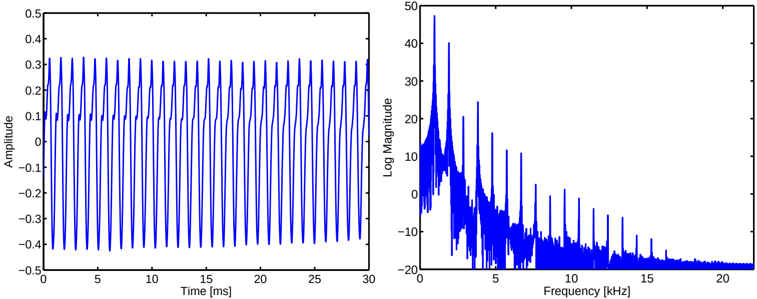


Figure: A quasi-periodic musical signal: a trumpet tone.

# Another Example

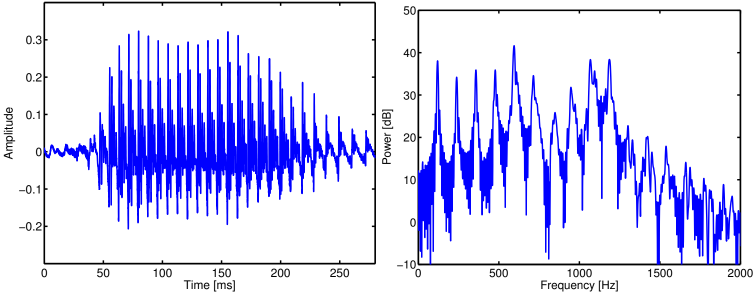


Figure: An approximately periodic speech signal and its spectrum.

We are here concerned with finding the fundamental frequency of periodic signals, i.e., the physical attribute of sounds.

Why is this a difficult problem? Because it is a non-convex, nonlinear problem. Sometimes it is not even a well-defined problem.

Pitch is not necessarily the same as the *perceived* pitch (but often is).

We will here use fundamental frequency and pitch synonymously and use *perceived* pitch when referring to the auditory sensation phenomenon.

The study of pitch perception is an entire field of its own.

Methods for determining the physical attribute pitch can be applied to a wide range of problems and signals.

# Scope

This tutorial covers:

- Methods rooted in estimation theory.
- Based on parametric models of the signal of interest.
- Analysis of pitch estimation as a mathematical problem.
- Models at signal level and on a segment-by-segment basis.

Why parametric methods?

- They lead to robust, tractable methods whose properties can be analyzed and understood.
- A full parametrization of the signal of interest is obtained.
- Back to basics... how can we hope to solve complicated problems if we cannot solve the simple ones?
- Basically a bunch of tools are out there. Why not use them?

Some questions:

- Under which conditions can a method be expected to work?
- How does performance depend on various conditions, like noise color and variance, or the number of observations?
- Is the method optimal? (and what does optimal mean?)
- Does the method work for low pitches?

Only possible to answer if assumptions are made explicit.

# Music Applications

Separation A parameterization of a signal into components allows for a natural separation of sources if the signal components have a close relation to the sources.

Enhancement Using parametric models, the enhancement problem is almost trivially solved–it is a matter of finding good estimates.

Compression Parametric models also form a natural basis for compression (e.g., HILN, SSC).

Modification It is possible to perform many kinds of otherwise complicated signal modification based on parametric models (e.g., time-stretching, pitch-shifting, morphing).

# Music Applications

Transcription Automatic transcription of music is a direct application of pitch estimators.

Tuning Very accurate real-time pitch estimators may be desirable for tuning of musical instruments.

Classification Pitch is a commonly used feature in many music information retrieval (MIR) tasks.

Can also be applied to other areas, like certain problems in RADAR, SONOR, speech analysis (prosody analysis, diagnosis of illnesses) or analysis of biological signals (e.g., ECG, bird songs).

# Signal Model

First, we introduce a source defined for $n = 0, \ldots, N-1$ as

$$x_k(n) = \sum_{l=1}^{L_k} a_{k,l} e^{j\omega_k l n} + e_k(n) = \sum_{l=1}^{L_k} a_{k,l} e^{j\psi_{k,l} n} + e_k(n) \tag{2}$$

where

$\omega_k$ is the fundamental frequency
$\psi_{k,l} = \omega_k l$ is the frequency of the $l$th harmonic.
$a_{k,l} = A_{k,l} e^{j\phi_{k,l}}$ is the complex amplitude.
$e_k(n)$ is the observation noise.

All the unknown real parameters are organized in a vector defined as

$$\boldsymbol{\theta}_k = [\ \omega_k\ A_{k,1}\ \phi_{k,1}\ \cdots A_{k,L_k}\ \phi_{k,L_k}\ ]^T . \tag{3}$$

In many cases, the observed signal consists of many such signals, i.e.,

$$x(n) = \sum_{k=1}^{K} x_k(n) = \sum_{k=1}^{K} \sum_{l=1}^{L_k} a_{k,l} e^{j\omega_k l n} + e(n). \tag{4}$$

For this model, the full parameter vector is

$$\boldsymbol{\theta} = \left[ \begin{array}{ccc} \boldsymbol{\theta}_1^T & \cdots & \boldsymbol{\theta}_K^T \end{array} \right]^T . \tag{5}$$

Estimation problems:

- Find $\omega_k$ from $x_k(n)$–a nonlinear problem.
- Find $\{\omega_k\}$ from $x(n)$ which is a multidimensional nonlinear problem.
- Find $\{a_{k,l}\}$ given $\omega_k$ which is a linear problem.
- Find the statistics of $e(n)$.

The noise term includes all stochastic signal components (i.e., including stuff like bow-noise, pick noise, etc.)!

We define vectors from $M$ consecutive samples of the observed signal as (with $M \leq N$)

$$\mathbf{x}(n) = [\ x(n)\ \cdots\ x(n+M-1)\ ]^T, \tag{6}$$

and similarly for $\mathbf{x}_k(n)$. Note that when $M = N$ we simply write $\mathbf{x}_k(n) = \mathbf{x}_k$. The signal model can be written into matrix-vector form as

$$\mathbf{x}(n) = \sum_{k=1}^{K} \mathbf{Z}_k \begin{bmatrix} e^{j\omega_k 1 n} & & 0 \\ & \ddots & \\ 0 & & e^{j\omega_k L_k n} \end{bmatrix} \mathbf{a}_k + \mathbf{e}(n) \tag{7}$$

$$= \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{a}_k(n) + \mathbf{e}(n), \tag{8}$$

or as $\mathbf{x}(n) = \sum_{k=1}^{K} \mathbf{Z}_k(n) \mathbf{a}_k + \mathbf{e}(n)$ where

$$\mathbf{Z}_k = [\ \mathbf{z}(\omega_k)\ \cdots\ \mathbf{z}(\omega_k L_k)\ ], \tag{9}$$

with $\mathbf{z}(\omega) = [\ 1\ e^{j\omega}\ \cdots\ e^{j\omega(M-1)}\ ]^T$, and $\mathbf{a}_k = [\ a_{k,1}\ \cdots\ a_{k,L_k}\ ]^T$.

The covariance matrix of $\mathbf{x}_k(n)$ can be written as (assuming independence)

$$\mathbf{R} = \sum_{k=1}^{K} \mathbf{R}_k = \sum_{k=1}^{K} \mathrm{E}\left\{\mathbf{x}_k(n)\mathbf{x}_k^H(n)\right\}. \tag{10}$$

The covariance matrix for a single source is then given by

$$\begin{aligned}\mathbf{R}_k &= \mathbf{Z}_k \mathrm{E}\left\{\mathbf{a}_k(n)\mathbf{a}_k^H(n)\right\}\mathbf{Z}_k^H + \mathrm{E}\left\{\mathbf{e}_k(n)\mathbf{e}_k^H(n)\right\} & (11)\\ &= \mathbf{Z}_k\mathbf{P}_k\mathbf{Z}_k^H + \mathbf{Q}_k, & (12)\end{aligned}$$

which is called the covariance matrix model. Note that often we will assume $\mathbf{Q} = \sigma^2\mathbf{I}$.

The matrix $\mathbf{P}_k$ is the covariance matrix for the amplitudes, which can be shown to be (under certain conditions)

$$\mathbf{P}_k \approx \mathrm{diag}\left(\left[\begin{array}{ccc} A_{k,1}^2 & \cdots & A_{k,L_k}^2 \end{array}\right]\right). \tag{13}$$

For multiple sources, we get

$$\mathbf{R} = \sum_{k=1}^{K} \mathbf{Z}_k\mathbf{P}_k\mathbf{Z}_k^H + \mathbf{Q}_k. \tag{14}$$

# Comments on Nonlinear Optimization

The estimators are usually stated as the solution to an optimization problem.

Closed-form solutions to non-linear, non-convex optimization problems rarely exist.

Hence, we must resort to numerical and often iterative optimization methods.

In practice, this is carried out by grid-searches and subsequent gradient-based optimization (Hessian matrices usually too complex).

Note that it is important to treat the fundamental frequency as a continuous parameter!

So, given a cost function $J(\cdot)$ first evaluate candidate fundamental frequencies $\omega_k$ on a grid $\Omega$ as

$$\hat{\omega}_k = \arg \min_{\omega_k \in \Omega} J(\omega_k). \tag{15}$$

The grid $\Omega$ must be sufficiently dense or we may miss the minumum!

Use this estimate as an initial estimate $\hat{\omega}_k^{(i)}$ in the following manner:

$$\hat{\omega}_k^{(i+1)} = \hat{\omega}_k^{(i)} - \alpha \nabla J(\omega_k^{(i)}). \tag{16}$$

Find the step size $\alpha$ using so-called line search:

$$\hat{\alpha} = \arg \min_{\alpha} J\left(\hat{\omega}_k^{(i)} - \alpha \nabla J(\omega_k^{(i)})\right), \tag{17}$$

which can be done in a number of ways. Usually, only a few iterations are required.

# On The Complex Signal Model

There are a number of reasons we use the complex signal model:

- Simpler math
- Faster algorithms

Real signals can be mapped to (almost) equivalent complex signals:

- Using the Hilbert transform to calculate the discrete-time analytic signal.
- Those do not hold for very low and high frequencies (relative to $N$).
- It is, in most cases, possible to account for real signals in estimators, but it is often not worth the trouble.

# Issues

The following issues occur when decomposing speech and audio signals using the signal model:

- Non-stationarity
- Noise characteristics
- Overlapping Harmonics
- Order estimation/model selection
- Inharmonicity

The one thing we want to avoid is multiple-dimensional nonlinear optimization!

# On Modified Signal Models

A myriad of modified signal models exist, including:

- AM-FM models allowing for various kinds of modulation.
- Polynomial phase and amplitude models.
- Other parametric modulation models.
- Inharmonicity models.
- Perturbed (uncertainty) models.

Sometimes easy to incorporate in estimators, sometimes difficult, depending on the type of estimator.

Prior knowledge (like amplitude smoothness, small perturbations, $\omega_k$ distribution) can be incorporated using *priors*.

# Parameter Estimation Bounds

An estimator is said to be unbiased if an estimate $\hat{\theta}_i$ of $\theta_i$ of the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^P$ whose expected value is the true parameter, i.e.,

$$\mathrm{E}\left\{\hat{\theta}_i\right\} = \theta_i \ \forall \theta_i, \tag{18}$$

and the difference $\theta_i - \mathrm{E}\left\{\hat{\theta}_i\right\}$, if any, is referred to as the bias. The Cramér-Rao lower bound (CRLB) of the parameter is given by (under so-called regularity conditions)

$$\mathrm{var}(\hat{\theta}_i) \geq \left[\mathbf{I}(\boldsymbol{\theta})\right]_{ii}^{-1}, \tag{19}$$

with

$$\left[\mathbf{I}(\boldsymbol{\theta})\right]_{il} = -\mathrm{E}\left\{\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i \partial\theta_l}\right\}, \tag{20}$$

where $\ln p(\mathbf{x};\boldsymbol{\theta})$ is the log-likelihood function of the observed signal $\mathbf{x} \in \mathbb{C}^N$.

The following asymptotic bounds can be established for the pitch estimation problem for white Gaussian noise:

$$\begin{aligned}\mathrm{var}(\hat{\omega}_k) &\geq \frac{6\sigma^2}{N(N^2-1)\sum_{l=1}^{L_k} A_{k,l}^2 l^2} && (21)\\ \mathrm{var}(\hat{A}_{k,l}) &\geq \frac{\sigma^2}{2N} && (22)\\ \mathrm{var}(\hat{\phi}_{k,l}) &\geq \frac{\sigma^2}{2N}\left(\frac{1}{A_{k,l}^2} + \frac{3l^2(N-1)^2}{\sum_{m=1}^{L_k} A_{k,m} m^2 (N^2-1)}\right). && (23)\end{aligned}$$

These depend on the following quantity:

$$PSNR_k = 10\log_{10}\frac{\sum_{l=1}^{L} A_{k,l}^2 l^2}{\sigma^2}\ [\mathrm{dB}]. \qquad (24)$$

For colored noise, the squared amplitudes should be weighted by the reciprocal of the noise psd.

Such bounds are useful for a number of reasons:

- An estimator attaining the bound is optimal.
- The bounds tell us how performance can be expected to depend on various quantities.
- The bounds can be used as benchmarks in simulations.
- Provides us with "rules of thumb" (e.g., include as many harmonics as possible, less noise should result in increasing performance, same for more samples).

Caveat emptor: it does not accurately predict the performance of non-linear estimators under adverse conditions (thresholding behavior).

It is also possible to calculate it exact CRLBs, where no asymptotic approximations are used. These predict more complicated phenomena.

An estimator attaining the bound is said to be *efficient*. A more fundamental property is *consistency*.

# Evaluation of Estimators

Basically two questions need to be answered: 1) how does an estimator perform given that the model is true? 2) is the model true?

Monte Carlo Repeated experiment with parameters and/or noise being randomized in each run.

Synthetic Signals Makes it possible to measure the performance of estimators.

MIDI Signals Same as above, but may still ultimately be model-based.

Audio Databases Real signal allows us to answer the second question. But how do we measure performance? Speech/audio databases are okay, but contain subjective aspects–the pitch may be not well-defined in a particular segment. Or we are trying to solve an ill-posed problem.

It is of course possible to violate model assumptions (whereby robustness is revealed) with synthetic signals.

Basically check whether the estimator is efficient or at least consistent.

A good measure is the root mean square estimation error (RMSE):

$$RMSE = \sqrt{\frac{1}{SK}\sum_{k=1}^{K}\sum_{s=1}^{S}\left(\hat{\omega}_k^{(s)} - \omega_k\right)^2}, \tag{25}$$

where $\omega_k$ and $\hat{\omega}_k^{(s)}$ are the true fundamental frequency and the estimate for source $k$ in Monte Carlo iteration $s$

The RMSE can be used to bound the probability of errors using the Chebyshev inequality.

Some annotated speech and audio databases: Keele Pitch Database, RWC Music Database, MAPS database, IOWA Musical Instrument Samples.

Relevant MIREX tasks: Audio melody extraction, chord detection, multi-pitch estimation & tracking, score following (training/tweaking sets available).

# Statistical Methods

Statistical methods are based on statistical models of the observed signal with the observation pdf being characterized by a number of parameters.

Maximum likelihood (ML) estimation is perhaps the most commonly used of all types of estimators.

Often based on a deterministic plus stochastic signal model, where the parameters of interest are considered deterministic but unknown and the observation noise is the stochastic part.

ML is statistically efficient for a sufficiently high number of samples and can be computationally demanding for nonlinear problems like ours.

Often approximate methods can be derived from explicit assumptions.

# Maximum Likelihood Method

For multi-variate Gaussian signals, the likelihood function of the observed signal sub-vector $\mathbf{x}_k(n)$ can be written as

$$p(\mathbf{x}_k(n); \boldsymbol{\theta}_k) = \frac{1}{\pi^M \det(\mathbf{Q}_k)} e^{-\mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}. \tag{26}$$

Assuming that the deterministic part is stationary and the noise is i.i.d., the likelihood of $\{\mathbf{x}_k(n)\}_{n=0}^{G-1}$ can be written as

$$p(\{\mathbf{x}_k(n)\}; \boldsymbol{\theta}_k) = \prod_{n=0}^{G-1} p(\mathbf{x}_k(n); \boldsymbol{\theta}_k) = \frac{1}{\pi^{MG} \det(\mathbf{Q}_k)^G} e^{-\sum_{n=0}^{G-1} \mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}.$$

The log-likelihood function is $\mathcal{L}(\boldsymbol{\theta}_k) = \ln p(\{\mathbf{x}_k(n)\}; \boldsymbol{\theta}_k)$ and the maximum likelihood estimator is

$$\hat{\boldsymbol{\theta}}_k = \arg\max \mathcal{L}(\boldsymbol{\theta}_k). \tag{27}$$

For white Gaussian noise, i.e., $\mathbf{Q}_k = \sigma^2 \mathbf{I}$, and setting $M = N$ the log-likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}_k) = -N \ln \pi - N \ln \sigma_k^2 - \frac{1}{\sigma_k^2} \|\mathbf{e}_k\|_2^2, \tag{28}$$

with $\mathbf{e}_k = \mathbf{x}_k - \mathbf{Z}_k \mathbf{a}_k$. Given $\omega_k$ and $L_k$, we can substitute the amplitudes by their LS estimates, and the ML noise variance estimate is then

$$\hat{\sigma}_k^2 = \frac{1}{N} \|\mathbf{x}_k - \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^H \mathbf{x}_k\|_2^2. \tag{29}$$

This leads to the following estimator:

$$\hat{\omega}_k = \arg\max_{\omega_k} \mathcal{L}(\omega_k) = \arg\max_{\omega_k} \mathbf{x}_k^H \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^H \mathbf{x}_k. \tag{30}$$

The above a nonlinear function and is termed the nonlinear least-squares (NLS) method.

The projection matrix $\mathbf{\Pi}_{Z_k}$ can be approximated as

$$\lim_{M\to\infty} M\mathbf{\Pi}_{Z_k} = \lim_{M\to\infty} M\mathbf{Z}_k \left(\mathbf{Z}_k^H\mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^H = \mathbf{Z}_k\mathbf{Z}_k^H. \tag{31}$$

Using this, the noise variance estimate can be simplified, i.e.,

$$\hat{\sigma}_k^2 \approx \frac{1}{N}\|\mathbf{x}_k - \frac{1}{N}\mathbf{Z}_k\mathbf{Z}_k^H\mathbf{x}_k\|_2^2. \tag{32}$$

Writing out the log-likelihood function, we get

$$\begin{aligned} \hat{\omega}_k &= \arg\max_{\omega_k} \mathcal{L}(\omega_k) = \arg\max_{\omega_k} -N\ln\pi - N\ln\hat{\sigma}_k^2 - N && (33)\\ &= \arg\max_{\omega_k} \mathbf{x}_k^H\mathbf{Z}_k\mathbf{Z}_k^H\mathbf{x}_k = \arg\max_{\omega_k} \|\mathbf{Z}_k^H\mathbf{x}_k\|_2^2 && (34)\end{aligned}$$

where $\|\mathbf{Z}_k^H\mathbf{x}_k\|_2^2 = \sum_{l=1}^{L_k} |\sum_{n=0}^{N-1} x_k(n)e^{-j\omega_k ln}|^2 \triangleq \sum_{l=1}^{L_k} |X_k(\omega_k l)|^2$, i.e., this can be computed using an FFT (i.e., *harmonic summation*)!

This is known as the approximate NLS method (ANLS). Note that these estimators are known to be robust to colored noise.

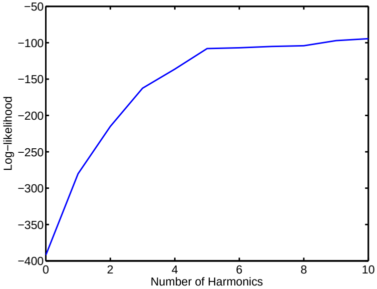


Figure: Log-likelihood function for a synthetic periodic signal (with $L_k = 5$).

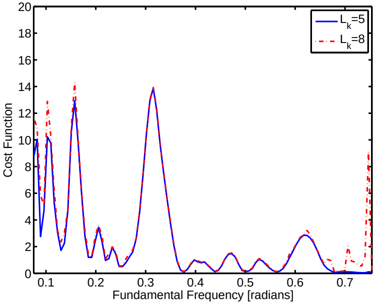


Figure: Cost function for a synthetic signal $\omega_k = 0.3142$.

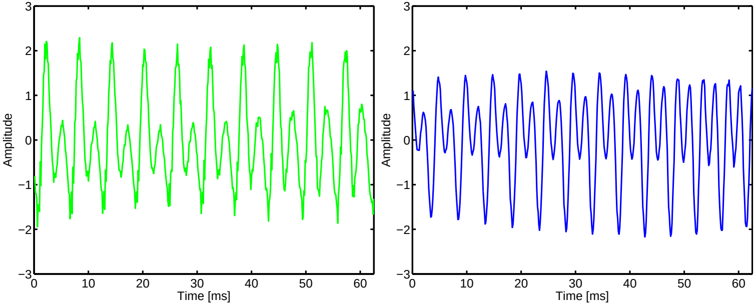


Figure: Speech signals with pitches 165 and 205 Hz.

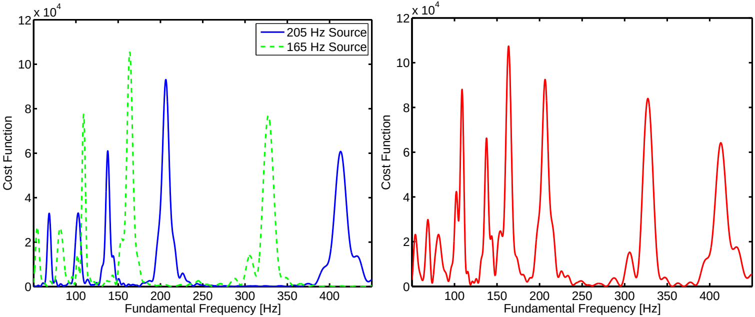


Figure: Approximate maximum likelihood cost function for the two speech signals (left) and their mixture (right).

# Inharmonicity

To incorporate the inharmonicity model, we only have to replace the frequencies $\omega_k l$ by $\psi_{k,l} = \omega_k l \sqrt{1 + B_k l^2}$, i.e., the estimator becomes

$$(\hat{\omega}_k, \hat{B}_k) = \arg \max_{\omega_k, B_k} \sum_{l=1}^{L_k} |X_k(\psi_{k,l})|^2 \tag{35}$$

$$= \arg \max_{\omega_k, B_k} \sum_{l=1}^{L_k} |X_k(\omega_k l \sqrt{1 + B_k l^2})|^2, \tag{36}$$

which means that we, in principle, have to sweep over combinations of the two nonlinear parameters to obtain the estimates.

Similarly, in the exact method, $\mathbf{Z}_k$ would then be a function of both $\omega_k$ and $B_k$.

# Model and Order Selection

For the parametric methods to work, the model order $L$ must be known!

To determine the model order (or choose between different models), one can use a number of different methods.

The MAP method penalizes nonlinear and linear parameters differently and is well-suited for our purposes.

First, we introduce $\mathbb{Z}_q = \{0, 1, \ldots, q-1\}$ which is the candidate model index set with $\mathcal{M}_m, m \in \mathbb{Z}_q$ being the candidate models.

The principle of MAP-based model selection is to choose the model that maximizes the a posteriori probability, i.e.,

$$\widehat{\mathcal{M}}_k = \arg \max_{\mathcal{M}_m, m \in \mathbb{Z}_q} p(\mathcal{M}_m | \mathbf{x}_k) = \arg \max_{\mathcal{M}_m, m \in \mathbb{Z}_q} \frac{p(\mathbf{x}_k | \mathcal{M}_m) p(\mathcal{M}_m)}{p(\mathbf{x}_k)}. \quad (37)$$

Assuming that all the models are equally probable, i.e.,

$$p(\mathcal{M}_m) = \frac{1}{q} \tag{38}$$

and noting that $p(\mathbf{x}_k)$ is constant once $\mathbf{x}_k$ has been observed, the MAP model selection criterion reduces to

$$\widehat{\mathcal{M}}_k = \arg \max_{\mathcal{M}_m, m \in \mathbb{Z}_q} p(\mathbf{x}_k | \mathcal{M}_m), \tag{39}$$

which is the likelihood function when seen as a function of $\mathcal{M}_m$.

Since the various models also depend on a number of unknown parameters, we will integrate those out as

$$p(\mathbf{x}|\mathcal{M}_m) = \int_{\mathbf{\Theta}_k} p(\mathbf{x}_k | \boldsymbol{\theta}_k, \mathcal{M}_m) p(\boldsymbol{\theta}_k | \mathcal{M}_m) d\boldsymbol{\theta}_k. \tag{40}$$

We will use the method of Laplace integration. Assuming that the likelihood function is highly peaked, we can write

$$\int_{\boldsymbol{\Theta}_k} p(\mathbf{x}_k|\boldsymbol{\theta}_k, \mathcal{M}_m) p(\boldsymbol{\theta}_k|\mathcal{M}_m) d\boldsymbol{\theta}_k$$
$$= \pi^{D_k/2} \det\left(\widehat{\mathbf{H}}_k\right)^{-1/2} p(\mathbf{x}_k|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_m) p(\hat{\boldsymbol{\theta}}_k|\mathcal{M}_m), \qquad (41)$$

where $D_k$ is the number of parameters and

$$\widehat{\mathbf{H}}_k = - \left. \frac{\partial^2 \ln p(\mathbf{x}_k|\boldsymbol{\theta}_k, \mathcal{M}_m)}{\partial\boldsymbol{\theta}_k \partial\boldsymbol{\theta}_k^T} \right|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k} \qquad (42)$$

is the Hessian of the log-likelihood function evaluated at $\hat{\boldsymbol{\theta}}_k$ (also known as the observed information matrix).

Taking the logarithm and ignoring constant terms (and $\frac{D_k}{2} \ln \pi$), we get

$$\widehat{\mathcal{M}}_k = \arg \min_{\mathcal{M}_m, m \in \mathbb{Z}_q} - \underbrace{\ln p(\mathbf{x}_k | \hat{\boldsymbol{\theta}}_k, \mathcal{M}_m)}_{\text{log-likelihood}} + \underbrace{\frac{1}{2} \ln \det \left( \widehat{\mathbf{H}}_k \right)}_{\text{penalty}}, \qquad (43)$$

which can be used directly for selecting between various models and orders.

The Hessian matrix is related to the Fischer information matrix, only it is evaluated in $\hat{\boldsymbol{\theta}}_k$. We introduce the normalization matrix

$$\mathbf{K}_N = \begin{bmatrix} N^{-3/2} & \mathbf{0} \\ \mathbf{O} & N^{-1/2}\mathbf{I} \end{bmatrix} \qquad (44)$$

where $\mathbf{I}$ is an $2L_k \times 2L_k$ identity matrix.

Using this normalization matrix, we can write the determinant of the Hessian in as

$$\det\left(\widehat{\mathbf{H}}_k\right) = \det\left(\mathbf{K}_N^{-2}\right)\det\left(\mathbf{K}_N\widehat{\mathbf{H}}_k\mathbf{K}_N\right). \tag{45}$$

And, finally, by observing that $\mathbf{K}_N\widehat{\mathbf{H}}_k\mathbf{K}_N = \mathcal{O}(1)$ and taking the logarithm, we obtain

$$\begin{aligned} \ln\det\left(\widehat{\mathbf{H}}_k\right) &= \ln\det\left(\mathbf{K}_N^{-2}\right) + \ln\det\left(\mathbf{K}_N\widehat{\mathbf{H}}_k\mathbf{K}_N\right) && (46)\\ &= \ln\det\left(\mathbf{K}_N^{-2}\right) + \mathcal{O}(1) && (47)\\ &= 3\ln N + 2L_k\ln N + \mathcal{O}(1). && (48) \end{aligned}$$

When the additive noise is a white complex Gaussian process, the log-likelihood function is $N\ln\sigma_k^2$, where $\sigma_k^2$ then is replaced by an estimate $\hat{\sigma}_k^2(L_k)$.

# Model Selection Rules

This all leads to the following rule (for $L_k \geq 1$):

$$\hat{L}_k = \arg\min_{L_k} N \log \hat{\sigma}_k^2(L_k) + L_k \log N + \frac{3}{2} \log N. \tag{49}$$

No harmonics are present if

$$N \log \hat{\sigma}_k(0)^2 < N \log \hat{\sigma}_k^2(\hat{L}_k) + \hat{L}_k \log N + \frac{3}{2} \log N. \tag{50}$$

Comments:

- Accurate order estimation is critical to the pitch estimation problem but also a very difficult problem.
- Statistical order estimation methods (MDL, MAP, AIC) are based on asymptotic approximations and are often arbitrary and suboptimal.
- Colored noise may cause problems for order estimation, more so than for fundamental frequency estimation!

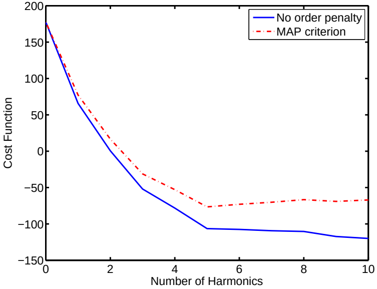


Figure: MAP model selection criterion and log-likelihood term for a synthetic signal with $L_k = 5$.

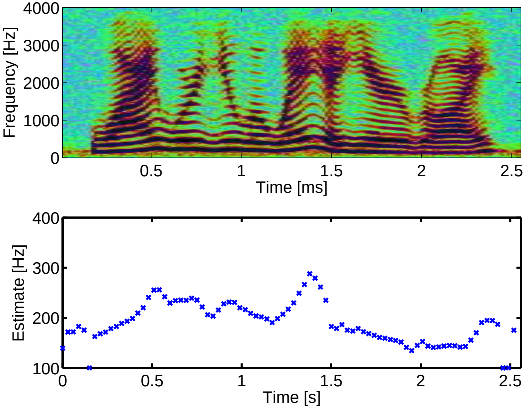


Figure: Voiced speech signal spectrogram (top) and pitch estimates (bottom).

# Expectation Maximization Algorithm

We write the signal model as a sum of $K$ sources in white additive Gaussian noise, i.e.,

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{x}_k \tag{51}$$

where the individual sources are given by $\mathbf{x}_k = \mathbf{Z}_k \mathbf{a}_k + \beta_k \mathbf{e}$, and

- the noise source is decomposed into $\mathbf{e}_k = \beta_k \mathbf{e}$ where $\beta_k \geq 0$ is chosen so that $\sum_{k=1}^{K} \beta_k = 1$.
- the set $\{\mathbf{x}_k\}$ is referred to as the complete data which is unobservable and the observed data is $\mathbf{x}$.
- $\mathbf{x}$ and $\{\mathbf{x}_k\}$ are assumed to be jointly Gaussian.
- the observations are assumed to be white Gaussian.

The problem is then to estimate the complete data set or its parameters. By stacking the complete data in a vector $\mathbf{y}$ as

$$\mathbf{y} = \left[\ \mathbf{x}_1^T\ \mathbf{x}_2^T\ \ \ldots\ \ \mathbf{x}_K^T \right]^T, \tag{52}$$

we can now write the incomplete data as

$$\mathbf{x} = \mathbf{H}\mathbf{y}, \tag{53}$$

where $\mathbf{H} = [\ \mathbf{I} \cdots \mathbf{I}\ ]$. In each iteration, where $(i)$ denotes the iteration number, the EM algorithm consists of two steps, the E-step, i.e.,

$$U(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}) = \int \ln p(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) p(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}^{(i)}) d\mathbf{y}, \tag{54}$$

and the M-step, i.e.,

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} U(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}). \tag{55}$$

Define an estimate of the $k$th source at iteration $(i)$ as

$$\hat{\mathbf{x}}_k^{(i)} = \mathbf{Z}_k^{(i)} \hat{\mathbf{a}}_k^{(i)} + \beta_k \left( \mathbf{x} - \sum_{k=1}^{K} \mathbf{Z}_k^{(i)} \hat{\mathbf{a}}_k^{(i)} \right), \tag{56}$$

where $\mathbf{Z}_k^{(i)}$ is constructed from $\hat{\omega}_k^{(i)}$. The problem of estimating the fundamental frequencies then becomes

$$\hat{\omega}_k^{(i+1)} = \arg\max_{\omega_k^{(i+1)}} \hat{\mathbf{x}}_k^{(i)H} \mathbf{Z}_k \left( \mathbf{Z}_k^H \mathbf{Z}_k \right)^{-1} \mathbf{Z}_k^H \hat{\mathbf{x}}_k^{(i)} \tag{57}$$

and the amplitudes can be found given $\hat{\omega}_k^{(i+1)}$ as

$$\hat{\mathbf{a}}_k^{(i+1)} = \left( \mathbf{Z}_k^{(i+1)H} \mathbf{Z}_k^{(i+1)} \right)^{-1} \mathbf{Z}_k^{(i+1)H} \hat{\mathbf{x}}_k^{(i)}. \tag{58}$$

This process is then repeated until convergence for $i = 0, \ldots, I-1$.

This is the same as the single-pitch ML method, but applied to the source estimates.

Comments:

- The EM algorithm leads to an implicit source separation.
- It is quite complicated to use and, especially, to initialize.
- Many different variations of these ideas can be (and have been) devised. For example, the harmonic matching pursuit.
- The model order estimation problem can cause difficulties.

# Harmonic Fitting

Idea: Estimate the unconstrained frequencies $\{\psi_{k,l}\}$ and fit the fundamental frequency to those (aka EXIP). Define

$$\boldsymbol{\theta}'_k = [\ A_{k,1}\ \phi_{k,1}\ \psi_{k,1}\ \cdots A_{k,L_k}\ \phi_{k,L_k}\ \psi_{k,L_k}\ ]^T \tag{59}$$

and

$$\boldsymbol{\eta}'_k = [\ \omega_k\ A_{k,1}\ \phi_{k,1}\ \cdots A_{k,L_k}\ \phi_{k,L_k}\ ]^T . \tag{60}$$

The basic idea of the method is that there exists a so-called selection matrix $\mathbf{S}' \in \mathbb{Z}^{3L_k \times (2L_k+1)}$ that relates the vectors as

$$\boldsymbol{\theta}'_k = \mathbf{S}'\boldsymbol{\eta}'_k. \tag{61}$$

We can now find an estimate of $\boldsymbol{\eta}'_k$ from estimates $\hat{\boldsymbol{\theta}}'_k$ as

$$\hat{\boldsymbol{\eta}}'_k = \arg\min_{\boldsymbol{\eta}'_k} \left\| \mathbf{W}'^{\frac{1}{2}} \left( \hat{\boldsymbol{\theta}}'_k - \mathbf{S}'\boldsymbol{\eta}'_k \right) \right\|_2^2 . \tag{62}$$

How to choose $\mathbf{W}'$?

If a maximum likelihood estimator is used for $\boldsymbol{\theta}'_k$ then the estimates will asymptotically be distributed according the CRLB!

Hence, we may choose $\mathbf{W}' = \mathbf{I}(\hat{\boldsymbol{\theta}}'_k)$, which is the FIM matrix. Therefore, $\mathbf{W}'$ becomes block diagonal for large $N$, i.e.,

$$\mathbf{W}' = \begin{bmatrix} \mathbf{W}'_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{W}'_{L_k} \end{bmatrix}, \tag{63}$$

where the individual sub-matrices contain the inverse of the CRLB matrix for the individual sinusoids of the unconstrained model, i.e.,

$$\mathbf{W}'_l = \frac{1}{\sigma_k^2} \begin{bmatrix} 2N & 0 & 0 \\ 0 & 2N\hat{A}_{k,l}^2 & N^2\hat{A}_{k,l}^2 \\ 0 & N^2\hat{A}_{k,l}^2 & \frac{2}{3}N^3\hat{A}_{k,l}^2 \end{bmatrix}. \tag{64}$$

The weighting does not lead to refined estimates of the amplitudes. Consequently, we define $\boldsymbol{\theta}_k \in \mathbb{R}^{2L_k \times 1}$ and $\boldsymbol{\eta}_k \in \mathbb{R}^{L_k+1 \times 1}$ like $\boldsymbol{\theta}'_k$ and $\boldsymbol{\eta}'_k$ but without the amplitudes. Now we may rewrite (61) as

$$\boldsymbol{\theta}_k = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 2 & 0 & 0 & \cdots & 0 \\ & \vdots & & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ L_k & 0 & 0 & \cdots & 0 \end{bmatrix} \boldsymbol{\eta}_k \triangleq \mathbf{S}\boldsymbol{\eta}_k. \tag{65}$$

As before, we can state our estimator as the minimizer of the norm of the error between the left and the right side of this expression, i.e.,

$$\hat{\boldsymbol{\eta}}_k = \arg\min_{\boldsymbol{\eta}_k} \left\| \mathbf{W}^{\frac{1}{2}} \left( \hat{\boldsymbol{\theta}}_k - \mathbf{S}\boldsymbol{\eta}_k \right) \right\|_2^2. \tag{66}$$

$\mathbf{W}$ is a now block diagonal with sub-matrices $\mathbf{W}_l$ defined as

$$\mathbf{W}_l = \frac{1}{\sigma_k^2} \left[ \begin{array}{cc} 2N\hat{A}_{k,l}^2 & N^2\hat{A}_{k,l}^2 \\ N^2\hat{A}_{k,l}^2 & \frac{2}{3}N^3\hat{A}_{k,l}^2 \end{array} \right], \tag{67}$$

and the cost function is

$$J = \left\| \mathbf{W}^{\frac{1}{2}} \left( \hat{\boldsymbol{\theta}}_k - \mathbf{S}\boldsymbol{\eta}_k \right) \right\|_2^2 = \frac{1}{\sigma_k^2} \sum_{l=1}^{L_k} \hat{A}_{k,l}^2 ([2N(\hat{\phi}_{k,l} - \phi_{k,l}) + N^2(\hat{\psi}_{k,l} - l\omega_k)]$$

$$\times (\hat{\phi}_{k,l} - \phi_{k,l}) + \left[ N^2(\hat{\phi}_{k,l} - \phi_{k,l}) + \frac{2}{3}N^3(\hat{\psi}_{k,l} - l\omega_k) \right] (\hat{\psi}_{k,l} - l\omega_k)).$$

Substituting the phases $\{\phi_{k,l}\}$ by estimates and solving for $\omega_k$, we get

$$\hat{\omega}_k = \frac{\sum_{l=1}^{L_k} l\hat{A}_{k,l}^2 \hat{\psi}_{k,l}}{\sum_{l=1}^{L_k} l^2 \hat{A}_{k,l}^2}. \tag{68}$$

Which is essentially a closed-form fundamental frequency estimator!

# Experimental Details

- RMSE as a function of various conditions.
- Percentage of correctly estimated model orders.
- $\omega_1 = 0.6364$ and $\omega_2 = 0.1580$, three harmonics, unit amplitudes.
- 100 Monte Carlo iterations for each point when RMSE, 1000 for orders.
- When estimating $\omega_k$, the model order is assumed known (and vice versa).
- White Gaussian noise.
- First a coarse estimate is found using grid search, then Gradient/Newton methods are used for refinement.

# Experimental Results

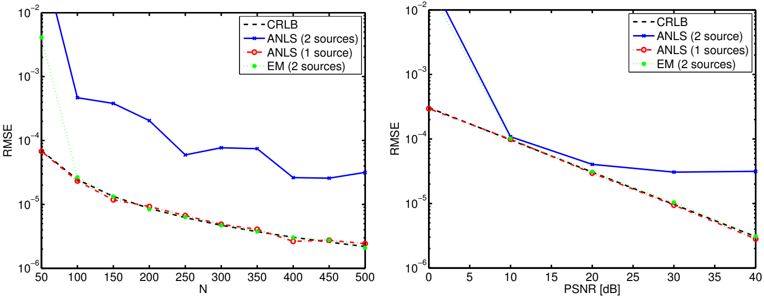


Figure: RMSE as a function of $N$ (with $PSNR = 40$ dB) and $PSNR$ (with $N = 400$).

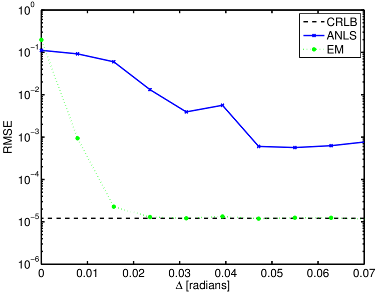


Figure: RMSE as a function of the difference between two fundamental frequencies for $N = 160$ and $PSNR = 40$ dB.

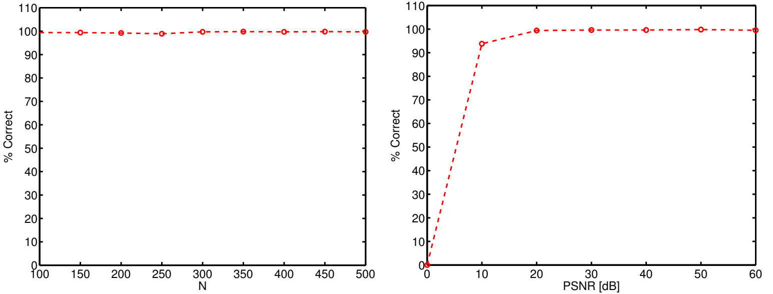


Figure: Percentage of correctly estimated model orders as a function of $N$ (with $PSNR = 40$ dB) and $PSNR$ (with $N = 500$).

# Discussion

- For the single-pitch case, the NLS methods performs extremely well, being statistically efficient, even asymptotically for colored noise.
- Associated problems of model and order selection can be solved consistently within the framework.
- Somewhat problematic for the multi-pitch case, requiring multidimensional nonlinear optimization.
- The EM algorithm and similar methodologies provide only a partial solution.
- The harmonic fitting approach is very sensitive to spurious estimates and its generalization to multiple pitches is not straightforward.

# Filtering Methods

Intuitive idea: filter the observed signal with filters having unit gain for the candidate harmonics while suppressing everything else.

Can be used for estimating parameters, extracting periodic signals, and separation of periodic signals.

One can use classical IIR/FIR filters or adaptive optimal filters.

The history of comb filters goes far back.

As we shall seen, there also exists some connections between statistical methods and filtering methods.

# Comb Filtering

Mathematically, we may express periodicity as $x(n) \approx x(n-D)$ where $D$ is the pitch period. It follows that a measure of the periodicity can be obtained using a metric on $e(n)$ defined as

$$e(n) = x(n) - \alpha x(n-D). \tag{69}$$

Taking the z-transform of this expression we get

$$\begin{aligned} E(z) &= X(z) - \alpha X(z) z^{-D} \qquad (70) \\ &= X(z)(1 - \alpha z^{-D}). \qquad (71) \end{aligned}$$

The transfer function $H(z)$ of the filter that operates on $x(n)$ can be seen to be

$$H(z) = \frac{E(z)}{X(z)} = (1 - \alpha z^{-D}). \tag{72}$$

This mathematical structure is known as a comb filter.

A more efficient alternative is notch filters which are filters that cancel out signal components at certain frequencies. These have the following form:

$$H(z) = \frac{1+\beta_1 z^{-1}}{1+\rho\beta_1 z^{-1}} = \frac{P(z)}{P(\rho^{-1}z)}. \tag{73}$$

Using $L_k$ such notch filters having notches at frequencies $\{\psi_i\}$, we obtain

$$P(z) = \prod_{i=1}^{L_k}(1-e^{j\psi_i}z^{-1}) = 1+\beta_1 z^{-1}+\ldots+\beta_{L_k}z^{-L_k}, \tag{74}$$

which has zeros on the unit circle at the desired frequencies. This polynomial defines the numerator, while the denominator is

$$P(\rho^{-1}z) = \prod_{i=1}^{L_k}(1-\rho e^{j\psi_i}z^{-1}) = 1+\rho\beta_1 z^{-1}+\ldots+\rho^{L_k}\beta_M z^{-L_k}. \tag{75}$$

Combining these, one obtains the following comb filter:

$$H(z) = \frac{P(z)}{P(\rho^{-1}z)} = \frac{1 + \beta_1 z^{-1} + \beta_2 z^{-2} + \ldots + \beta_{L_k} z^{-L_k}}{1 + \rho\beta_1 z^{-1} + \rho^2 \beta_2 z^{-2} + \ldots + \rho^{L_k} \beta_{L_k} z^{-L_k}}. \quad (76)$$

By applying this filter for various candidate fundamental frequencies to our observed signal $x(n)$, we can obtain the filtered signal $e(n)$ where the harmonics have been suppressed:

$$e(n) = x(n) + \beta_1 x(n-1) + \beta_2 x(n-2) + \ldots + \beta_{L_k} x(n - L_k) \quad (77)$$

$$- \rho\beta_1 e(n-1) - \rho^2 \beta_2 e(n-2) - \ldots - \rho^{L_k} \beta_{L_k} e(n - L_k). \quad (78)$$

Finally, we can use this signal for finding the fundamental frequency:

$$\hat{\omega}_k = \arg\min_{\omega} J \quad \text{with} \quad J = \sum_{n=1}^{N} |e(n)|^2. \quad (79)$$

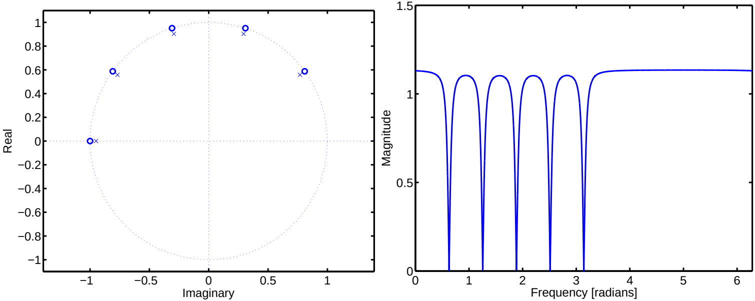


Figure: Z-plane representation of the zeros (circles) and poles (x-mark) and frequency response.

# Classical Methods

Returning to the original comb filter

$$e(n) = x(n) - \alpha x(n-D), \tag{80}$$

we see that it can be thought of as a prediction problem with unknowns $\alpha$ and $D$. We could determine these as

$$\{\hat{\alpha}, \hat{D}\} = \arg\min_{\alpha,D} \mathrm{E}\left\{|e(n)|^2\right\}, \tag{81}$$

which is also what long-term predictors in speech coders do. Assuming $\alpha = 1$, we obtain

$$\begin{aligned}\mathrm{E}\left\{|e(n)|^2\right\} &= \mathrm{E}\left\{(x^*(n) - x^*(n-D))\left((x(n) - x(n-D))\right\}\right. & (82)\\ &= \mathrm{E}\left\{|x(n)|^2\right\} + \mathrm{E}\left\{|x(n-D)|^2\right\} \\ &- \mathrm{E}\left\{x^*(n)x(n-D)\right\} - \mathrm{E}\left\{x^*(n-D)x(n)\right\}. & (83)\end{aligned}$$

Assuming that the signal is stationary, we have that $\mathrm{E}\left\{|x(n)|^2\right\} = \mathrm{E}\left\{|x(n-D)|^2\right\} = \sigma^2$.

Furtermore, we have that $r(D) = \mathrm{E}\left\{x^*(n)x(n-D)\right\}$. This leads to the following estimator:

$$\hat{D} = \arg\max_{D} 2\,\mathrm{Re}(r(D)), \tag{84}$$

which is the well-known auto-correlation method (complex version). One can generalize this principle as follows (with $p \geq 1$):

$$\hat{D} = \arg\min_{\alpha,D} \mathrm{E}\left\{|x(n) - x(n-D)|^p\right\}. \tag{85}$$

Comments:

(i) For $p = 2$, we obtain the autocorrelation method (corresponding to $e(n)$ being Gaussian).

(ii) For $p = 1$, we obtain the average magnitude difference function (AMDF) method (corresponding to $e(n)$ being Laplacian).

(iii) Restricted to integer samples (fractional delays require more work).

(iv) Non-unique estimates (and summation limits considerations).

# FIR Methods

We construct a vector from $M$ time-reversed samples of the observed signal, i.e.,

$$\mathbf{x}(n) = [\ x(n)\ x(n-1)\ \cdots\ x(n-M+1)\ ]^T, \tag{86}$$

with $M \leq N$ and with $(\cdot)^T$ denoting the transpose. Next, introducing the output signal $y_{k,l}(n)$ of the $l$th filter for the $k$th source having coefficients $h_{k,l}(n)$ as

$$y_{k,l}(n) = \sum_{m=0}^{M-1} h_{k,l}(m)x(n-m) = \mathbf{h}_{k,l}^H \mathbf{x}(n), \tag{87}$$

$\mathbf{h}_{k,l}$ being a vector containing the impulse response of the $l$th filter, i.e.,

$$\mathbf{h}_{k,l} = \left[\ h_{k,l}^*(0)\ \cdots\ h_{k,l}^*(M-1)\ \right]^H. \tag{88}$$

The output power of the $l$th filter can be written as

$$\mathrm{E}\left\{|y_{k,l}(n)|^2\right\} = \mathrm{E}\left\{\mathbf{h}_{k,l}^H \mathbf{x}(n)\mathbf{x}^H(n)\mathbf{h}_{k,l}\right\} = \mathbf{h}_{k,l}^H \mathbf{R}\mathbf{h}_{k,l}. \tag{89}$$

The total output power of all the filters is

$$\sum_{l=1}^{L_k} \mathrm{E}\left\{|y_{k,l}(n)|^2\right\} = \sum_{l=1}^{L_k} \mathbf{h}_{k,l}^H \mathbf{R}\mathbf{h}_{k,l}. \tag{90}$$

Defining a matrix $\mathbf{H}_k$ consisting of the filters $\{\mathbf{h}_{k,l}\}$ as

$$\mathbf{H}_k = \left[\ \mathbf{h}_{k,1}\ \cdots\ \mathbf{h}_{k,L_k}\ \right], \tag{91}$$

we can write the total output power as a sum of the power of the subband signals, i.e.,

$$\sum_{l=1}^{L_k} \mathrm{E}\left\{|y_{k,l}(n)|^2\right\} = \mathrm{Tr}\left[\mathbf{H}_k^H \mathbf{R}\mathbf{H}_k\right]. \tag{92}$$

# Interpretation of the NLS Method

Suppose we construct the filters from complex sinusoids as

$$\mathbf{h}_{k,l} = \left[ \begin{array}{ccc} e^{-j\omega_k l 0} & \cdots & e^{-j\omega_k l(M-1)} \end{array} \right]^T, \tag{93}$$

The matrix $\mathbf{H}_k$ is identical to the Vandermonde matrix $\mathbf{Z}_k$ except that it is time-reversed, i.e.,

$$\mathbf{Z}_k = [ \begin{array}{ccc} \mathbf{z}(\omega_k) & \cdots & \mathbf{z}(\omega_k L_k) \end{array} ], \tag{94}$$

with $\mathbf{z}(\omega) = [ \begin{array}{cccc} 1 & e^{-j\omega} & \cdots & e^{-j\omega(M-1)} \end{array} ]^T$. Then, we may write the total output power of the filterbank as

$$\begin{aligned} \mathrm{Tr}\left[\mathbf{H}_k^H \mathbf{R} \mathbf{H}_k\right] &= \mathrm{Tr}\left[\mathbf{Z}_k^H \mathbf{R} \mathbf{Z}_k\right] && (95) \\ &= \mathrm{E}\left\{\|\mathbf{Z}_k^H \mathbf{x}(n)\|_2^2\right\}. && (96) \end{aligned}$$

Except for the expectation, this is the FFT method introduced earlier.

# Optimal filterbank

Idea: find a set of filters that pass power undistorted at specific frequencies while minimizing the output power:

$$\min_{\mathbf{H}_k} \operatorname{Tr}\left[\mathbf{H}_k^H \mathbf{R} \mathbf{H}_k\right] \quad \text{s.t.} \quad \mathbf{H}_k^H \mathbf{Z}_k = \mathbf{I}, \tag{97}$$

where $\mathbf{I}$ is the $L_k \times L_k$ identity matrix. As before, the matrix $\mathbf{Z}_k \in \mathbb{C}^{M \times L_k}$ is constructed from $L_k$ complex sinusoidal vectors.

Using the method of Lagrange multipliers, the filterbank matrix $\mathbf{H}_k$ can be shown to be

$$\mathbf{H}_k = \mathbf{R}^{-1}\mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{R}^{-1} \mathbf{Z}_k\right)^{-1}. \tag{98}$$

This data and frequency dependent filter bank can then be used to estimate the fundamental frequencies as

$$\hat{\omega}_k = \arg\max_{\omega_k} \operatorname{Tr}\left[\left(\mathbf{Z}_k^H \mathbf{R}^{-1} \mathbf{Z}_k\right)^{-1}\right]. \tag{99}$$

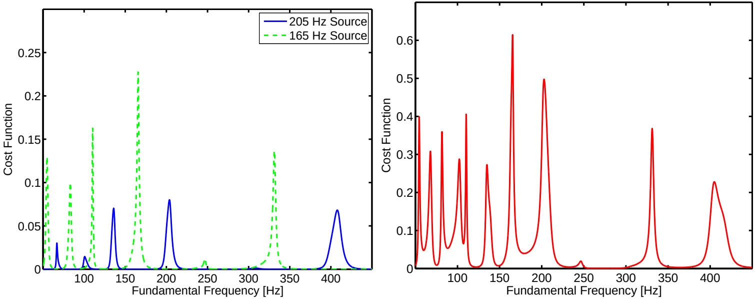


Figure: Cost function based on optimal filtering for the two speech signals (left) and their mixture (right).

# Optimal Filter

Suppose that we instead design a single filter for the $k$th source, $\mathbf{h}_k$. This filter design problem can be stated as

$$\min_{\mathbf{h}_k} \mathbf{h}_k^H \mathbf{R} \mathbf{h}_k \quad \text{s.t.} \quad \mathbf{h}_k^H \mathbf{z}(\omega_k l) = 1, \quad \text{for} \quad l = 1, \ldots, L_k. \tag{100}$$

This has the solution

$$\mathbf{h}_k = \mathbf{R}^{-1} \mathbf{Z}_k \left( \mathbf{Z}_k^H \mathbf{R}^{-1} \mathbf{Z}_k \right)^{-1} \mathbf{1} \quad \text{and} \quad \mathbf{h}_k^H \mathbf{R} \mathbf{h}_k = \mathbf{1}^H \left( \mathbf{Z}_k^H \mathbf{R}^{-1} \mathbf{Z}_k \right)^{-1} \mathbf{1}.$$

As before, we readily obtain an estimate of the fundamental frequency as

$$\hat{\omega}_k = \arg\max_{\omega_k} \mathbf{1}^H \left( \mathbf{Z}_k^H \mathbf{R}^{-1} \mathbf{Z}_k \right)^{-1} \mathbf{1}. \tag{101}$$

It is perhaps not clear how these two estimators are related.

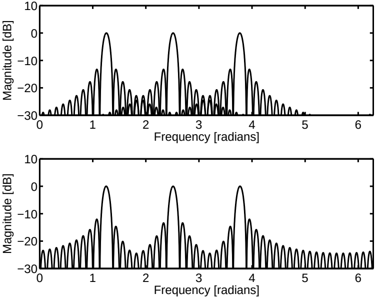


Figure: Frequency response (magnitude) of the optimal filter bank and filter for white noise.

# Asymptotic Analysis

Comparing the optimal filters, we observe that they can be related as

$$\mathbf{h}_k = \mathbf{R}^{-1}\mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{R}^{-1}\mathbf{Z}_k\right)^{-1} \mathbf{1} = \mathbf{H}_k \mathbf{1} = \sum_{l=1}^{L} \mathbf{h}_{k,l}, \tag{102}$$

but generally $\mathbf{1}^H \left(\mathbf{Z}_k^H \mathbf{R}^{-1}\mathbf{Z}_k\right)^{-1} \mathbf{1} \neq \mathrm{Tr}\left[\left(\mathbf{Z}^H \mathbf{R}^{-1}\mathbf{Z}_k\right)^{-1}\right]$.

Analyzing the asymptotic properties of the cost function, we see that

$$\lim_{M\to\infty} \frac{1}{M}\left(\mathbf{Z}_k^H \mathbf{R}\mathbf{Z}_k\right) = \mathrm{diag}\left(\left[\begin{array}{ccc} \Phi(\omega_k) & \cdots & \Phi(\omega_k L_k) \end{array}\right]\right), \tag{103}$$

with $\Phi(\omega)$ being the psd of $x(n)$. It can therefore be seen that

$$\lim_{M\to\infty} M\mathbf{1}^H \left(\mathbf{Z}_k^H \mathbf{R}^{-1}\mathbf{Z}_k\right)^{-1} \mathbf{1} = \sum_{l=1}^{L_k} \Phi(\omega_k l). \tag{104}$$

Conclusion: The methods are asymptotically equivalent and are equivalent to the NLS method too!

# Order Estimation

As stated earlier, the MAP criterion is defined for $L_k \geq 1$ as

$$\hat{L}_k = \arg\min_{L} N \log \hat{\sigma}^2(L_k) + L_k \log N + \frac{3}{2} \log N. \tag{105}$$

We now need to find the estimate $\hat{\sigma}^2(L_k)$ from the residual $\hat{e}(n) = x(n) - y_k(n)$. Additionally, $y_k(n)$ is

$$y_k(n) = \sum_{m=0}^{M-1} \sum_{l=1}^{L_k} h_{k,l}(m) x(n-m) = \sum_{m=0}^{M-1} h_k(m) x(n-m), \tag{106}$$

where $h_k(m)$ is the sum over the filters of the filterbank. We can now write (with $\mathbf{g}_k = [\ (1 - h_k(0))\ \ -h_k(1)\ \cdots\ -h_k(M-1)\ ]^H$)

$$\hat{e}(n) = x(n) - \sum_{m=0}^{M-1} h_k(m) x(n-m) \triangleq \mathbf{g}_k^H \mathbf{x}(n). \tag{107}$$

We can then estimate the noise variance as

$$\hat{\sigma}^2(L_k) = \mathrm{E}\left\{|\hat{e}(n)|^2\right\} = \mathrm{E}\left\{\mathbf{g}_k^H \mathbf{x}(n)\mathbf{x}^H(n)\mathbf{g}_k\right\} = \mathbf{g}_k^H \mathbf{R} \mathbf{g}_k. \tag{108}$$

Defining $\mathbf{g}_k = \mathbf{b}_1 - \mathbf{h}_k$, with $\mathbf{b}_1 = \left[\begin{array}{cccc} 1 & 0 & \cdots & 0 \end{array}\right]$ the variance estimate is rewritten as

$$\hat{\sigma}^2(L_k) = \mathbf{b}_1^H \mathbf{R} \mathbf{b}_1 - \mathbf{b}_1^H \mathbf{R} \mathbf{h}_k - \mathbf{h}_k^H \mathbf{R} \mathbf{b}_1 + \mathbf{h}_k^H \mathbf{R} \mathbf{h}_k. \tag{109}$$

The first term can be identified as $\mathbf{b}_1^H \mathbf{R} \mathbf{b}_1 = \mathrm{E}\left\{|x(n)|^2\right\}$ and $\mathbf{h}_k^H \mathbf{R} \mathbf{h}_k$ we know. Writing out the cross-terms yields

$$\mathbf{b}_1^H \mathbf{R} \mathbf{h}_k = \mathbf{b}_1^H \mathbf{R} \mathbf{R}^{-1} \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{R}^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{1} = \mathbf{h}_k^H \mathbf{R} \mathbf{h}_k. \tag{110}$$

Therefore, the variance estimate can be expressed as

$$\hat{\sigma}^2(L_k) = \mathrm{E}\left\{|x(n)|^2\right\} - \mathbf{1}^H \left(\mathbf{Z}_k^H \mathbf{R}^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{1}. \tag{111}$$

# Efficient Implementation

Both the filterbank method and the single filter method require the calculation of

$$\left(\mathbf{Z}_k^H \mathbf{R}^{-1} \mathbf{Z}_k\right)^{-1}. \tag{112}$$

To apply the MIL to the calculation of the cost function, we first define a matrix

$$\mathbf{Z}_k^{(L_k-1)} = [\ \mathbf{z}(\omega_k)\ \cdots\ \mathbf{z}(\omega_k(L_k-1))\ ], \tag{113}$$

and a vector $\mathbf{z}_k^{(L_k)} = \left[\ e^{-j\omega_k L_k 0}\ \cdots\ e^{-j\omega_k L_k (M-1)} \right]^T$. We can now write

$$\begin{aligned}\left(\mathbf{Z}_k^H \mathbf{R}^{-1} \mathbf{Z}_k\right)^{-1} &= \begin{bmatrix} \mathbf{Z}_k^{(L_k-1)H} \mathbf{R}^{-1} \mathbf{Z}_k^{(L_k-1)} & \mathbf{Z}_k^{(L_k-1)H} \mathbf{R}^{-1} \mathbf{z}_k^{(L_k)} \\ \mathbf{z}_k^{(L_k)H} \mathbf{R}^{-1} \mathbf{Z}_k^{(L_k-1)} & \mathbf{z}_k^{(L_k)H} \mathbf{R}^{-1} \mathbf{z}_k^{(L_k)} \end{bmatrix}^{-1} \\ &\triangleq \mathbf{\Xi}_{L_k}, \end{aligned} \tag{114}$$

where $\mathbf{\Xi}_{L_k}$ is the matrix calculated for an order $L_k$ model.

Next, define the quantities

$$\xi_{L_k} = \mathbf{z}_k^{(L_k)H} \mathbf{R}^{-1} \mathbf{z}_k^{(L_k)} \quad \text{and} \quad \boldsymbol{\eta}_{L_k} = \mathbf{Z}_k^{(L_k-1)H} \mathbf{R}^{-1} \mathbf{z}_k^{(L_k)}. \tag{115}$$

We can now express the matrix inverse using the MIL in terms of these as

$$\boldsymbol{\Xi}_{L_k} = \begin{bmatrix} \boldsymbol{\Xi}_{L_k-1} & \mathbf{0} \\ \mathbf{O} & 0 \end{bmatrix} + \begin{bmatrix} -\boldsymbol{\Xi}_{L_k-1}\boldsymbol{\eta}_{L_k} \\ 1 \end{bmatrix} \times \frac{1}{\xi_{L_k}^H - \boldsymbol{\eta}_{L_k}^H \boldsymbol{\Xi}_{L_k-1} \boldsymbol{\eta}_{L_k}} \begin{bmatrix} -\boldsymbol{\eta}_{L_k}^H \boldsymbol{\Xi}_{L_k-1} & 1 \end{bmatrix} \tag{116}$$

$$\triangleq \begin{bmatrix} \boldsymbol{\Xi}_{L_k-1} & \mathbf{0} \\ \mathbf{O} & 0 \end{bmatrix} + \frac{1}{\beta_{L_k}} \begin{bmatrix} \boldsymbol{\zeta}_{L_k} \boldsymbol{\zeta}_{L_k}^H & -\boldsymbol{\zeta}_{L_k} \\ -\boldsymbol{\zeta}_{L_k}^H & 1 \end{bmatrix}. \tag{117}$$

This shows that once $\boldsymbol{\Xi}_{L_k-1}$ is known, $\boldsymbol{\Xi}_{L_k}$ can be obtained in a simple way.

For a given $\omega_k$, we calculate the order 1 inverse matrix as

$$\boldsymbol{\Xi}_1 = \frac{1}{\xi_1}, \tag{118}$$

and then, for $l = 2, \ldots, L_k$, calculate

$$
\begin{aligned}
\boldsymbol{\kappa}_l &= \mathbf{R}^{-1}\mathbf{z}_k^{(l)} \qquad (119)\\
\xi_l &= \mathbf{z}_k^{(l)H}\boldsymbol{\kappa}_l \qquad (120)\\
\boldsymbol{\eta}_l &= \mathbf{Z}_k^{(l-1)H}\boldsymbol{\kappa}_l \qquad (121)\\
\boldsymbol{\zeta}_l &= \boldsymbol{\Xi}_{l-1}\boldsymbol{\eta}_l \qquad (122)\\
\beta_l &= \xi_l^H - \boldsymbol{\eta}_l^H\boldsymbol{\Xi}_{l-1}\boldsymbol{\eta}_l \qquad (123)\\
\boldsymbol{\Xi}_l &= \begin{bmatrix} \boldsymbol{\Xi}_{l-1} & \mathbf{0} \\ \mathbf{O} & 0 \end{bmatrix} + \frac{1}{\beta_l}\begin{bmatrix} \zeta_l\zeta_l^H & -\zeta_l \\ -\zeta_l^H & 1 \end{bmatrix}, \qquad (124)
\end{aligned}
$$

from which the fundamental frequency and the model order can be found.

# Experimental Results

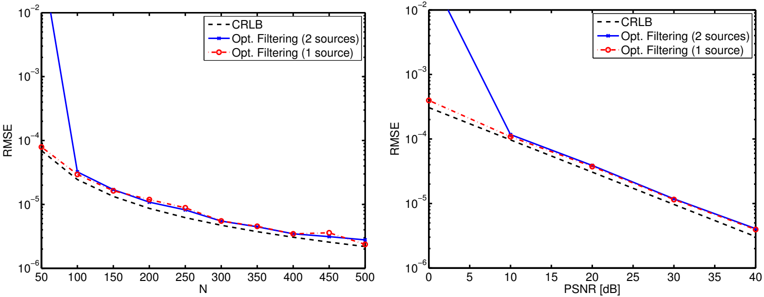


Figure: RMSE as a function of $N$ (with $PSNR = 40$ dB) and $PSNR$ (with $N = 400$).

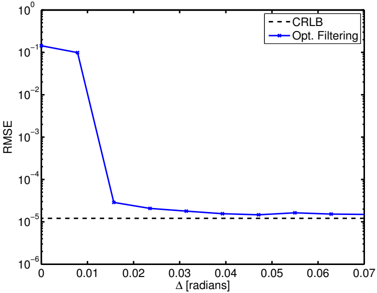


Figure: RMSE as a function of the difference between two fundamental frequencies for $N = 160$ and $PSNR = 40$ dB.

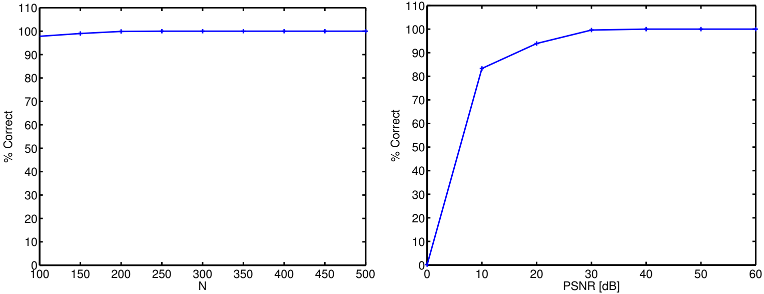


Figure: Percentage of correctly estimated model orders as a function of $N$ (with $PSNR = 40$ dB) and $PSNR$ (with $N = 500$).

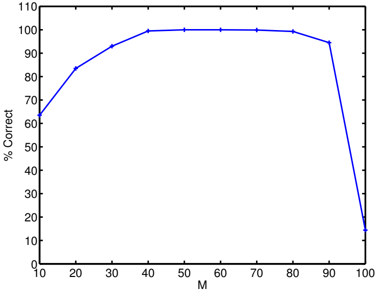


Figure: Percentage of correctly estimated model orders as a function of $M$ (with $PSNR = 40$ dB and $N = 200$).

# Discussion

- The methods based on optimal filtering form an intriguing alternative for pitch estimation.
- Especially so as they lead to a natural decoupling of the multi-pitch estimation problem.
- They can also be used directly for enhancement and separation.
- They have excellent performance under adverse conditions, like closely spaced multiple pitches.
- Robust to colored noise.
- Complexity may be prohibitive for some applications.
- Order and model selection can be performed consistently within the framework.

# Subspace Methods

In subspace methods, the full observation space is divided into signal (plus noise) and noise subspaces.

The properties of these can then be used for various estimation and identification tasks.

Some of the most elegant unconstrained frequency estimators are subspace methods, with ESPRIT, MODE, and root-MUSIC essentially being the only closed-form frequency estimators.

There has been some interesting in subspace methods in the connection with sinusoidal speech and audio modeling (unconstrained model).

# Covariance Matrix Model

Define $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_K \end{bmatrix}$ and $\mathbf{a}(n) = \begin{bmatrix} \mathbf{a}_1^T(n) & \cdots & \mathbf{a}_K^T(n) \end{bmatrix}^T$. We write the model as

$$\mathbf{x}(n) = \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{a}_k(n) + \mathbf{e}(n), \; = \mathbf{Z}\mathbf{a}(n) + \mathbf{e}(n). \tag{125}$$

As shown earlier, the covariance matrix is then ($\sum_{k=1}^{K} \mathbf{Z}_k \mathbf{P}_k \mathbf{Z}_k^H$ has rank $V = \sum_{k=1}^{K} L_k$)

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\} = \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{P}_k \mathbf{Z}_k^H + \sigma^2 \mathbf{I} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \sigma^2 \mathbf{I} \tag{126}$$

with $\mathbf{P}_k = \mathrm{diag}\left(\begin{bmatrix} |A_{k,1}|^2 & \cdots & |A_{k,L_k}|^2 \end{bmatrix}\right)$ and $\mathbf{P} = \mathrm{diag}\left(\begin{bmatrix} \mathbf{P}_1 & \cdots & \mathbf{P}_K \end{bmatrix}\right)$.

Let

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H \tag{127}$$

be the eigenvalue decomposition (EVD) of the covariance matrix. $\mathbf{U}$ contains the $M$ orthonormal eigenvectors of $\mathbf{R}$, i.e.,

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_M \end{bmatrix}, \tag{128}$$

and $\boldsymbol{\Lambda}$ is a diagonal matrix containing the corresponding (sorted) positive eigenvalues, $\lambda_k$. Let $\mathbf{S}$ be formed as

$$\mathbf{S} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_V \end{bmatrix}. \tag{129}$$

The subspace that is spanned by the columns of $\mathbf{S}$ we denote $\mathcal{R}(\mathbf{S})$, which is the same space as $\mathcal{R}(\mathbf{Z})$.

Similarly, let $\mathbf{G}$ be formed as

$$\mathbf{G} = \left[ \begin{array}{ccc} \mathbf{u}_{V+1} & \cdots & \mathbf{u}_M \end{array} \right], \tag{130}$$

where $\mathcal{R}\left(\mathbf{G}\right)$ is the so-called *noise subspace*. Using the EVD, the covariance matrix model can now be written as

$$\mathbf{U}\left(\mathbf{\Lambda} - \sigma^2 \mathbf{I}\right)\mathbf{U}^H = \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{P}_k \mathbf{Z}_k^H. \tag{131}$$

Some useful properties that can be exploited for estimation purposes can now be established. It can be seen that (MUSIC)

$$\mathbf{Z}^H \mathbf{G} = \mathbf{0} \quad \text{and} \quad \mathbf{Z}_k^H \mathbf{G} = \mathbf{0} \quad \forall k. \tag{132}$$

When only one source is present we have that $\mathbf{Z} = \mathbf{Z}_k$ and $V = L_k$.

Furthermore, we have that $\mathbf{S} = \mathbf{Z}\mathbf{B}$. Defining

$$\underline{\mathbf{S}} = [\ \mathbf{I}\ \mathbf{0}\ ]\,\mathbf{S} \quad \text{and} \quad \overline{\mathbf{S}} = [\ \mathbf{0}\ \mathbf{I}\ ]\,\mathbf{S}. \tag{133}$$

and

$$\underline{\mathbf{Z}} = [\ \mathbf{I}\ \mathbf{0}\ ]\,\mathbf{Z} \quad \text{and} \quad \overline{\mathbf{Z}} = [\ \mathbf{0}\ \mathbf{I}\ ]\,\mathbf{Z}, \tag{134}$$

we see that

$$\overline{\mathbf{Z}} = \underline{\mathbf{Z}}\mathbf{D} \quad \text{and} \quad \overline{\mathbf{S}} = \underline{\mathbf{S}}\boldsymbol{\Xi}, \tag{135}$$

with $\mathbf{D} = \mathrm{diag}\left([\ e^{j\psi_1}\ \ \cdots\ \ e^{j\psi_L}\ ]\right)$. This leads us to (ESPRIT)

$$\boldsymbol{\Xi} = \mathbf{B}^{-1}\mathbf{D}\mathbf{B}, \tag{136}$$

from which the frequencies $\{\psi_l\}$ can be found.

# Pre-whitening

The question is how to deal with colored noise. The classical approach is to either do a) sub-band processing or b) pre-whitening. When the noise is not white, the covariance matrix model is

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \mathbf{Q}, \tag{137}$$

where $\mathbf{Q} = \mathrm{E}\{\mathbf{e}(n)\mathbf{e}^H(n)\}$. Since covariance matrices are symmetric and positive definite, so are their inverses and the Cholesky factorization of $\mathbf{Q}^{-1}$ is

$$\mathbf{Q}^{-1} = \mathbf{L}\mathbf{L}^H, \tag{138}$$

where $\mathbf{L}$ is an $M \times M$ lower triangular matrix. By multiplying the observed signal vectors by this matrix, we get

$$\mathrm{E}\left\{\mathbf{L}^H\mathbf{x}(n)\mathbf{x}^H(n)\mathbf{L}\right\} = \mathbf{L}^H\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{L} + \mathbf{I}. \tag{139}$$

A simple implementation of this is via a pre-whitening filter.

# Rank Estimation

As we have seen, the likelihood function of the observed signal can for the Gaussian case be expressed as:

$$p(\{\mathbf{x}(n)\};\zeta) = \prod_{n=0}^{G-1} p(\mathbf{x}(n);\zeta) \tag{140}$$

$$= \frac{1}{\pi^{GM}\det(\mathbf{R})^{G}} e^{-\sum_{n=0}^{G} \mathbf{x}^H(n)\mathbf{R}^{-1}\mathbf{x}(n)}. \tag{141}$$

By taking the logarithm, we obtain the log-likelihood function

$$\mathcal{L}(\zeta) = \ln p(\{\mathbf{x}(n)\};\zeta) \tag{142}$$

$$= -GM\ln\pi - G\ln\det(\mathbf{R}) - \sum_{n=0}^{G-1} \mathbf{x}^H(n)\mathbf{R}^{-1}\mathbf{x}(n). \tag{143}$$

As it turns out, this can be expressed as

$$\mathcal{L}(\zeta) = -GM \ln \pi - G \ln \prod_{v=1}^{M} \hat{\lambda}_v - G(M - L') \ln \frac{\frac{1}{M-L'} \sum_{v=L'+1}^{M} \hat{\lambda}_v}{\prod_{v=L'+1}^{M} \hat{\lambda}_v^{1/(M-L')}} - GM.$$

Using the AIC ($\nu = 2$) or MDL ($\nu = \frac{1}{2} \ln N$), we obtain the following cost function to be minimized for determining the rank of the signal subspace:

$$J(L') = -\mathcal{L}(\zeta) + (L'(2M - L') + 1)\nu. \tag{144}$$

The signal subspace dimension is identical to the number of harmonics for the single-pitch case and the total number of harmonics for multi-pitch signals.

Unfortunately, the criterion performs poorly in practice, especially when the noise is colored.

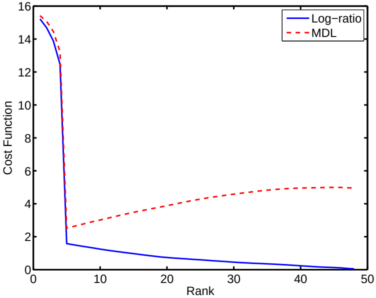


Figure: Log-ratio between the geometric and arithmetic means and the MDL cost function.

# Angles Between Subspaces

The principal angles $\{\theta_k\}$ between the two subspaces $\mathcal{Z} = \mathcal{R}(\mathbf{Z})$ and $\mathcal{G} = \mathcal{R}(\mathbf{G})$ are defined recursively for $k = 1, \ldots, K$ as

$$\cos(\theta_k) = \max_{\mathbf{u} \in \mathcal{Z}} \max_{\mathbf{v} \in \mathcal{G}} \frac{\mathbf{u}^H \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2} = \mathbf{u}_k^H \mathbf{v}_k, \qquad (145)$$

where $K$ is the minimal dimension of the two subspaces, i.e., $K = \min\{V, M - V\}$ and $\mathbf{u}^H \mathbf{u}_i = 0$ and $\mathbf{v}^H \mathbf{v}_i = 0$ for $i = 1, \ldots, k-1$.

For subspace $\mathcal{G}$, the projection matrix is

$$\mathbf{\Pi}_G = \mathbf{G}\left(\mathbf{G}^H \mathbf{G}\right)^{-1} \mathbf{G}^H = \mathbf{G}\mathbf{G}^H, \qquad (146)$$

while for subspace $\mathcal{Z}$, the projection matrix is

$$\mathbf{\Pi}_Z = \mathbf{Z}\left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H. \qquad (147)$$

Using the two projection matrices, we can now write

$$\cos(\theta_k) = \max_{\mathbf{y}} \max_{\mathbf{z}} \frac{\mathbf{y}^H \mathbf{\Pi}_Z \mathbf{\Pi}_G \mathbf{z}}{\|\mathbf{y}\|_2 \|\mathbf{z}\|_2} = \mathbf{y}_k^H \mathbf{\Pi}_Z \mathbf{\Pi}_G \mathbf{z}_k = \sigma_k. \tag{148}$$

for $k = 1, \ldots, K$. Furthermore, we require that $\mathbf{y}^H \mathbf{y}_i = 0$ and $\mathbf{z}^H \mathbf{z}_i = 0$ for $i = 1, \ldots, k-1$. It follows that $\{\sigma_k\}$ are the singular values of $\mathbf{\Pi}_Z \mathbf{\Pi}_G$ which are related to the Frobenius norm as

$$\|\mathbf{\Pi}_Z \mathbf{\Pi}_G\|_F^2 = \mathrm{Tr}\left\{\mathbf{\Pi}_Z \mathbf{\Pi}_G \mathbf{\Pi}_G^H \mathbf{\Pi}_Z^H\right\} = \mathrm{Tr}\left\{\mathbf{\Pi}_Z \mathbf{\Pi}_G\right\} = \sum_{k=1}^{K} \sigma_k^2. \tag{149}$$

Interestingly, this can be related to the Frobenius norm of the difference between the two projection matrices, i.e.,

$$\|\mathbf{\Pi}_Z - \mathbf{\Pi}_G\|_F^2 = \mathrm{Tr}\{\mathbf{\Pi}_Z + \mathbf{\Pi}_G - 2\mathbf{\Pi}_Z \mathbf{\Pi}_G\} = M - 2\|\mathbf{\Pi}_Z \mathbf{\Pi}_G\|_F^2. \tag{150}$$

The Frobenius norm of the product $\mathbf{\Pi}_Z\mathbf{\Pi}_G$ can be rewritten as

$$\|\mathbf{\Pi}_Z\mathbf{\Pi}_G\|_F^2 = \mathrm{Tr}\left\{\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{G}\mathbf{G}^H\right\}. \tag{151}$$

This can be simplified because $\lim_{M\to\infty} M\mathbf{\Pi}_Z = \mathbf{Z}\mathbf{Z}^H$:

$$\|\mathbf{\Pi}_Z\mathbf{\Pi}_G\|_F^2 = \sum_{k=1}^{K}\sigma_k^2 \approx \frac{1}{M}\mathrm{Tr}\left\{\mathbf{Z}^H\mathbf{G}\mathbf{G}^H\mathbf{Z}\right\} = \frac{1}{M}\|\mathbf{Z}^H\mathbf{G}\|_F^2. \tag{152}$$

By averaging over all the nontrivial angles, we now arrive at (with $K = \min\{V, M - V\}$.)

$$\frac{1}{K}\sum_{k=1}^{K}\cos^2(\theta_k) = \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2 \approx \frac{1}{MK}\|\mathbf{Z}^H\mathbf{G}\|_F^2 \tag{153}$$

which can be used to measure the level of orthogonality to obtain a) parameter estimates and b) order estimates.

# Estimation using the Orthogonality Property

For the single-pitch case, the covariance matrix model is

$$\mathbf{R}_k = \mathrm{E}\left\{\mathbf{x}_k(n)\mathbf{x}_k^H(n)\right\} = \mathbf{Z}_k\mathbf{P}_k\mathbf{Z}_k^H + \sigma^2\mathbf{I} \tag{154}$$

By forming a matrix $\mathbf{Z}_k$ for different candidate frequencies and then measure the angles between the subspaces we can obtain an estimate as

$$\hat{\omega}_k = \arg\min_{\omega_k} \|\mathbf{Z}_k^H\mathbf{G}\|_F^2 = \arg\min_{\omega_k} \sum_{l=1}^{L_k} \|\mathbf{z}^H(\omega_k l)\mathbf{G}\|_2^2, \tag{155}$$

i.e., we find estimates by maximizing the angles between the subspaces $\mathcal{R}(\mathbf{Z}_k)$ and $\mathcal{R}(\mathbf{G})$. For an unknown model order we arrive at

$$\hat{\omega}_k = \arg\min_{\omega_k}\min_{L_k} \frac{1}{MK}\|\mathbf{Z}_k^H\mathbf{G}\|_F^2 \quad \text{with} \quad K = \min\{L_k, M - L_k\}. \tag{156}$$

To recapitulate, the covariance matrix model for the multi-pitch case is

$$\mathbf{R} = \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{P}_k \mathbf{Z}_k^H + \sigma^2 \mathbf{I} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \sigma^2 \mathbf{I}. \tag{157}$$

The subspace orthogonality property states that the matrix $\mathbf{Z}$ and all its sub-matrices are orthogonal to $\mathbf{G}$, i.e.,

$$\mathbf{Z}^H \mathbf{G} = \mathbf{0} \quad \text{and} \quad \mathbf{Z}_k^H \mathbf{G} = \mathbf{0} \quad \forall k. \tag{158}$$

First, assume that the model orders are known and note that $\|\mathbf{Z}^H \mathbf{G}\|_F^2 = \sum_{k=1}^{K} \|\mathbf{Z}_k^H \mathbf{G}\|_F^2$. The set of fundamental frequencies estimates are then

$$\{\hat{\omega}_k\} = \arg \min_{\{\omega_k\}} \|\mathbf{Z}^H \mathbf{G}\|_F^2 = \arg \sum_{k=1}^{K} \min_{\omega_k} \|\mathbf{Z}_k^H \mathbf{G}\|_F^2, \tag{159}$$

which allows for independent optimization over the sources.

For the case where the model orders $\{L_k\}$ (and thus the rank of $\mathbf{G}$) are unknown, the estimator becomes

$$\{\hat{\omega}_k\} = \arg \min_{\{\omega_k\}} \min_{\{L_k\}} \frac{1}{MK} \|\mathbf{Z}^H \mathbf{G}\|_F^2, \tag{160}$$

where $K = \min\{\sum_{k=1}^{K} L_k, M - \sum_{k=1}^{K} L_k\}$ since the rank of the signal and noise subspaces depend on the total number of harmonics. Fortunately, the estimator can be simplified somewhat as

$$\min_{\{\omega_k\}} \min_{\{L_k\}} \frac{1}{MK} \|\mathbf{Z}^H \mathbf{G}\|_F^2 = \min_{\{L_k\}} \sum_{k=1}^{K} \min_{\omega_k} \frac{1}{MK} \|\mathbf{Z}_k^H \mathbf{G}\|_F^2. \tag{161}$$

Simplification: Find $\mathbf{G}$ first using another method.

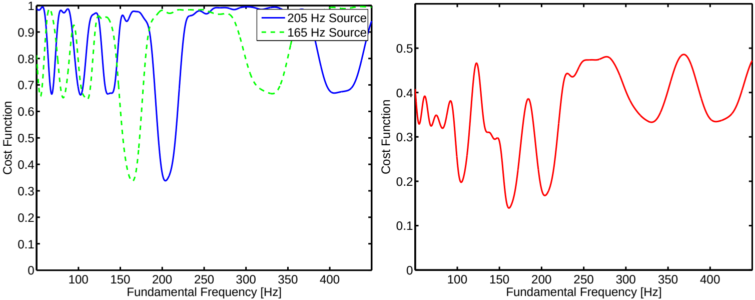


Figure: Subspace-based cost function for the two speech signals (left) and their mixture (right).

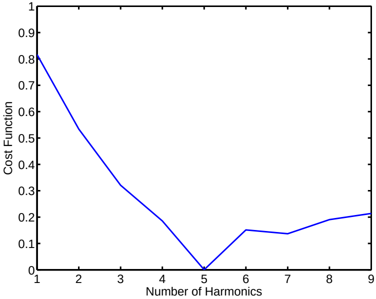


Figure: The cost function based on subspace orthogonality as a function of the model order $L_k$ for a signal where the true order is five.

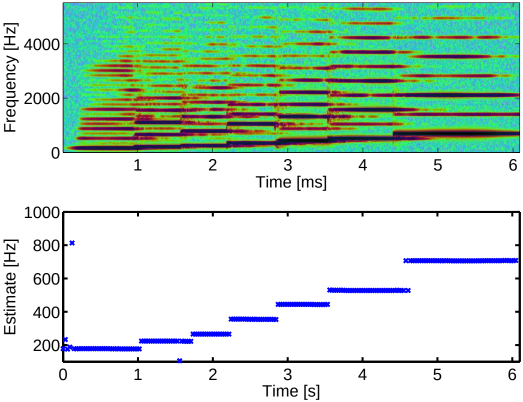


Figure: Clarinet signal spectrogram (top) and pitch estimates obtained using the joint estimator (bottom).

# Inharmonicity

Consider the unconstrained signal model and various models for $\{\psi_{k,l}\}$

$$x_k(n) = \sum_{l=1}^{L_k} A_{k,l} e^{j(\psi_{k,l} n + \phi_{k,l})} + e_k(n). \tag{162}$$

- $\psi_{k,l} l = \omega_k l$.
- $\psi_{k,l} = \omega_k l \sqrt{1 + B_k l^2}$ with $B_k \ll 1$ (pinned ends).
- $\psi_{k,l} = \omega_k l \sqrt{1 + B_k l^2} \left(1 + \frac{2}{\pi}\sqrt{B_k} + \frac{4}{\pi^2} B_k\right)$ (clamped ends).

Comments:

- We refer to the last two models as parametric inharmonicity models.
- Deviations from the harmonic model is sometimes referred to as inharmonicity (a very pronounced effect for some instruments).

It is trivial to incorporate these models in MUSIC. The $\mathbf{Z}_k$ matrix is constructed from the two nonlinear parameters as

$$\mathbf{Z}_k = \begin{bmatrix} \mathbf{z}(\omega_k\sqrt{1+B_k}) & \cdots & \mathbf{z}(\omega_k L_k\sqrt{1+B_k L_k^2}) \end{bmatrix}. \tag{163}$$

Thus, estimates can be obtained as

$$(\hat{\omega}_k, \hat{B}_k) = \arg\min_{\omega_k, B_k} \left\| \mathbf{Z}_k^H \mathbf{G} \right\|_F^2, \tag{164}$$

which has to be evaluated for a large range of combinations of the two parameters.

We will now introduce an alternative model of the inharmonicity as

$$x_k(n) = \sum_{l=1}^{L_k} a_{k,l} e^{j(\omega_k l + \Delta_{k,l})n} + e_k(n), \tag{165}$$

where $\{\Delta_{k,l}\}$ is a set of small perturbations.

(i) This model is more general than the parametric inharmonicity models (a model mismatch often leads to biased estimates).

(ii) The perturbations have to be small, otherwise the associated fundamental frequency estimate will be meaningless.

The Vandermonde matrix is now characterized by $\omega_k$ and $\{\Delta_{k,l}\}$ as

$$\mathbf{Z}_k = \left[\begin{array}{ccc} \mathbf{z}(\omega_k + \Delta_{k,1}) & \cdots & \mathbf{z}(\omega_k L_k + \Delta_{k,L_k}) \end{array}\right]. \tag{166}$$

But direct minimization of the cost function will be very computationally demanding and will not lead to any meaningful estimates.

Instead, one can use a cost function like

$$J(\omega_k, \{\Delta_{k,l}\}) = \left\| \mathbf{Z}_k^H \mathbf{G} \right\|_F^2 + P(\{\Delta_{k,l}\}), \tag{167}$$

where

- $P(\{\Delta_{k,l}\})$ is the penalty function which is a non-decreasing function of a metric with $P(\{0\}) = 0$.
- It is desirable that the penalty function is additive over the harmonics.

Therefore, a natural choice is $P(\{\Delta_{k,l}\}) = \sum_{l=1}^{L_k} \nu_l |\Delta_{k,l}|^p$ with $p \geq 1$ ($\{\nu_l\}$ is a set of positive constants).

We note that the Frobenius norm is additive over the columns of $\mathbf{Z}_k$, i.e.,

$$J(\omega_k, \{\Delta_{k,l}\}) = \sum_{l=1}^{L} \mathbf{z}^H(\omega_k l + \Delta_{k,l}) \mathbf{G}\mathbf{G}^H \mathbf{z}(\omega_k l + \Delta_{k,l}) + \sum_{l=1}^{L_k} \nu_l |\Delta_{k,l}|^p.$$

Substituting $\psi_{k,l}$ by $\omega_k l + \Delta_{k,l}$ and $\Delta_{k,l}$ by $\psi_{k,l} - \omega_k l$, and due to additive and independence, we get

$$\begin{aligned}\hat{\omega}_k &= \arg\min_{\omega_k} \min_{\{\psi_{k,l}\}} \left\{ \sum_{l=1}^{L_k} \mathbf{z}^H(\psi_{k,l})\mathbf{G}\mathbf{G}^H\mathbf{z}(\psi_{k,l}) + \nu_l|\psi_{k,l} - \omega_k l|^p \right\} \\ &= \arg\min_{\omega_k} \sum_{l=1}^{L_k} \min_{\psi_{k,l}} \left\{ \mathbf{z}^H(\psi_{k,l})\mathbf{G}\mathbf{G}^H\mathbf{z}(\psi_{k,l}) + \nu_l|\psi_{k,l} - \omega_k l|^p \right\},\end{aligned}$$

where the first term depends only on $\psi_{k,l}$ and is the reciprocal of the MUSIC pseudo-spectrum. For a given $\hat{\omega}_k$, the frequencies are simply

$$\hat{\psi}_{k,l} = \arg\min_{\psi_{k,l}} \left\{ \mathbf{z}^H(\psi_{k,l})\mathbf{G}\mathbf{G}^H\mathbf{z}(\psi_{k,l}) + \nu_l|\psi_{k,l} - \hat{\omega}_k l|^p \right\}. \tag{168}$$

About the penalty term:

- Log-prior on perturbations in the context of MAP estimation.
- For large $\nu_l$, the perturbation will be small, whereas for $\nu_l$ close to zero, the estimator will reduce to finding unconstrained frequencies.
- $\{\nu_l\}$ can be seen as Lagrange multipliers, i.e., we have a set of implicit constraints. The robust Capon beamformer is based on explicit constraints.
- $p = 1$ will result in small perturbations while allowing for a few large ones and $\nu_l \propto 1/l$ allows for large deviations for higher harmonics.
- May be worth considering an asymmetrical penalty function.

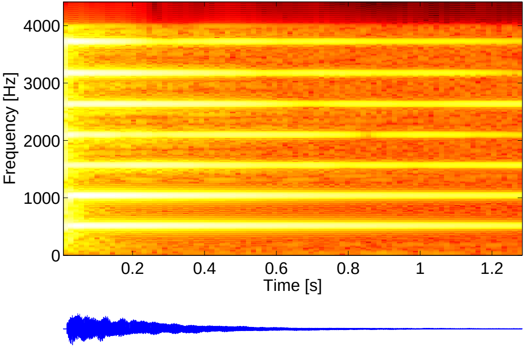


Figure: Spectrogram of signal, a piano note $C_5 \sim 523.25$ Hz.

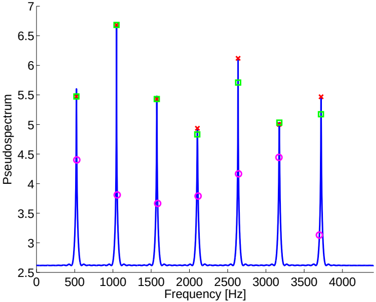


Figure: Estimates for the perfectly harmonic model (magenta), the parametric model (green), and the perturbed model (red).

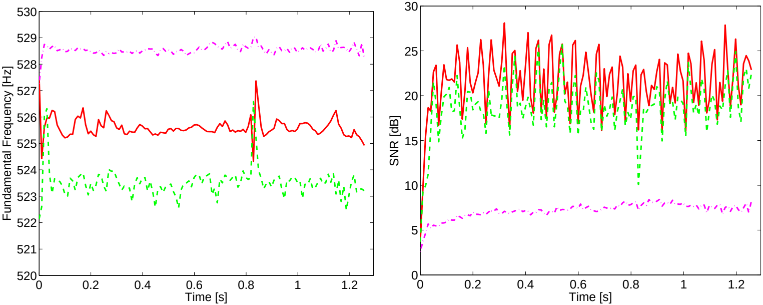


Figure: Estimates/SNR for the perfectly harmonic model (magenta), the parametric model (green), and the perturbed model (red)

# Estimation using Shift-Invariance

Sinusoidal parameters can be found by constructing the matrices $\underline{\mathbf{S}}$ and $\overline{\mathbf{S}}$ and then solving for $\boldsymbol{\Xi}$ in

$$\overline{\mathbf{S}} \approx \underline{\mathbf{S}}\boldsymbol{\Xi}, \tag{169}$$

in some sense, like

$$\widehat{\boldsymbol{\Xi}} = \arg\min_{\boldsymbol{\Xi}} \|\overline{\mathbf{S}} - \underline{\mathbf{S}}\boldsymbol{\Xi}\|_F^2 = \left(\underline{\mathbf{S}}^H \underline{\mathbf{S}}\right)^{-1} \underline{\mathbf{S}}^H \overline{\mathbf{S}}, \tag{170}$$

where the sinusoidal frequencies are found as the eigenvalues of $\widehat{\boldsymbol{\Xi}}$. Similarly, an order estimate can be obtained as (ESTER):

$$\hat{L} = \arg\min_{L} \|\overline{\mathbf{S}} - \underline{\mathbf{S}}\widehat{\boldsymbol{\Xi}}\|_2^2. \tag{171}$$

One can also use this principle to obtain a pitch estimator as follows. The sinusoidal frequencies are obtained using the empirical EVD of $\widehat{\boldsymbol{\Xi}}$, i.e.,

$$\widehat{\boldsymbol{\Xi}} = \mathbf{C}\widehat{\mathbf{D}}\mathbf{C}^{-1} \tag{172}$$

with $\mathbf{C}$ containing the empirical eigenvectors of $\widehat{\boldsymbol{\Xi}}$ and $\widehat{\mathbf{D}} = \mathrm{diag}\left(\left[\begin{array}{ccc} e^{j\hat{\psi}_1} & \cdots & e^{j\hat{\psi}_{L_k}} \end{array}\right]\right)$. Using the shift-invariance property, we can write

$$\overline{\mathbf{S}} \approx \underline{\mathbf{S}}\mathbf{C}\widehat{\mathbf{D}}\mathbf{C}^{-1}. \tag{173}$$

Defining $\widetilde{\mathbf{D}} = \mathrm{diag}\left(\left[\begin{array}{ccc} e^{j\omega_k} & \cdots & e^{j\omega_k L_k} \end{array}\right]\right)$, we introduce the cost function $J \triangleq \|\overline{\mathbf{S}} - \underline{\mathbf{S}}\mathbf{C}\widetilde{\mathbf{D}}\mathbf{C}^{-1}\|_F^2$ from which the fundamental frequency can be estimated as

$$\hat{\omega}_k = \arg\min_{\omega_k} J, \tag{174}$$

where only $\widetilde{\mathbf{D}}$ depends on $\omega_k$. Note that also the order $L_k$ can be estimated in this manner.

We could just as well have expressed the cost function as

$$J = \|\bar{\mathbf{S}}\mathbf{C} - \underline{\mathbf{S}}\mathbf{C}\widetilde{\mathbf{D}}\|_F^2 \triangleq \|\mathbf{V} - \mathbf{W}\widetilde{\mathbf{D}}\|_F^2. \tag{175}$$

The cost function in (175) can be expanded as

$$J = -2\,\mathrm{Re}\left(\mathrm{Tr}\left\{\mathbf{V}\widetilde{\mathbf{D}}^H\mathbf{W}^H\right\}\right) + \mathrm{Tr}\left\{\mathbf{V}\mathbf{V}^H\right\} + \mathrm{Tr}\left\{\mathbf{W}\widetilde{\mathbf{D}}\widetilde{\mathbf{D}}^H\mathbf{W}^H\right\}. \tag{176}$$

Introducing $\mathbf{H} = \mathbf{W}^H\mathbf{V}$ and ignoring constant terms, the cost function is redfined as

$$J \triangleq -2\,\mathrm{Re}\left(\mathrm{Tr}\left\{\mathbf{H}\widetilde{\mathbf{D}}^H\right\}\right) = -2\,\mathrm{Re}\left(\sum_{l=1}^{L_k} h_l e^{-j\omega_k l}\right) \tag{177}$$

with $h_l = [\mathbf{H}]_{ll}$. This is an extremely simple and smooth function, but it cannot easily be generalized to the multi-pitch case.

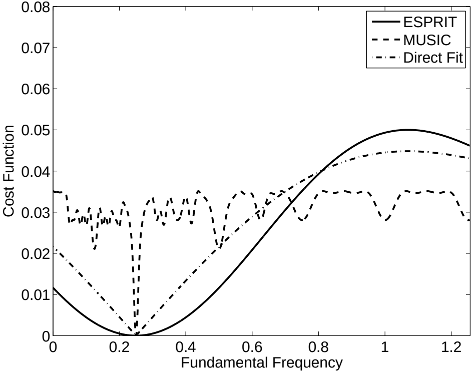


Figure: Cost function based on the shift-invariance property compared to a direct fit and subspace orthogonality.

# Experimental Results

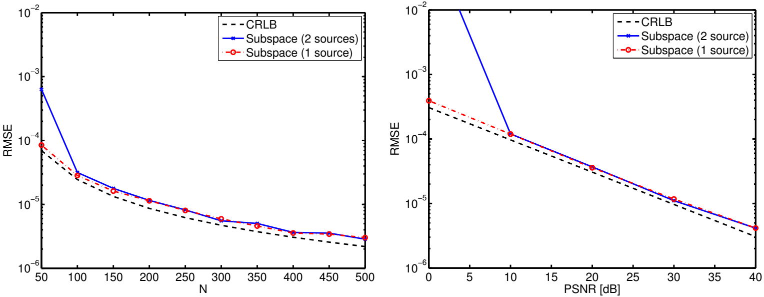


Figure: RMSE as a function of $N$ (with $PSNR = 40$ dB) and $PSNR$ (with $N = 400$).

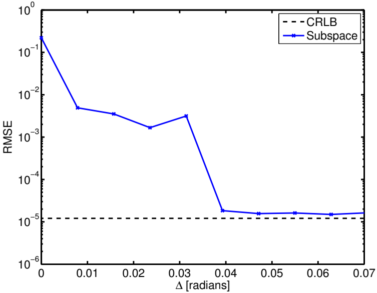


Figure: RMSE as a function of the difference between two fundamental frequencies for $N = 160$ and $PSNR = 40$ dB.

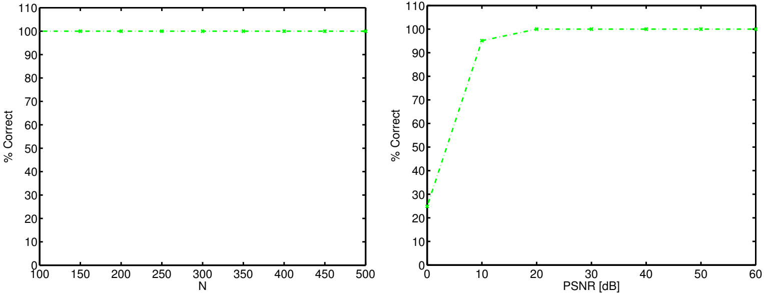


Figure: Percentage of correctly estimated model orders as a function of $N$ (with $PSNR = 40$ dB) and $PSNR$ (with $N = 500$).

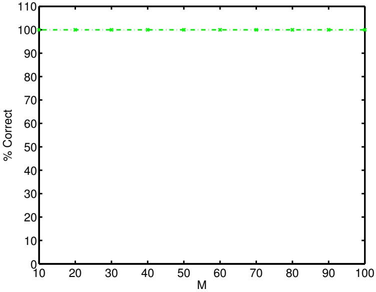


Figure: Percentage of correctly estimated model orders as a function of $M$ (with $PSNR = 40$ dB and $N = 200$).

# Discussion

- $\mathbf{Z}_k^H \mathbf{G}$ can be calculated efficiently using FFTs, i.e., subspace methods are actually quite fast once the EVD has been computed.
- $\mathbf{G}$ and $\mathbf{S}$ can be updated recursively over time using subspace trackers.
- The subspace methods are elegant solutions to the pitch estimation problem and allows for finding the order too.
- The ESPRIT-based method is sensitive in several ways while the MUSIC method is fairly robust.
- MUSIC offers partial decoupling of the multi-pitch estimation problem (except for the orders).
- They have good statistical performance but depend on a high SNR/white noise (but not the pdf).

# Amplitude Estimation

After estimating the signal's fundamental frequencies, one often wishes to estimate also the complex amplitudes of the periodic components.

With estimated amplitudes, we have a full parametrization of the signal of interest. The signal can be re-synthesized using this information.

This can be done in a number of ways. Here, we will present some different approaches, namely

- Least-squares based estimators, and
- Capon- and APES-based estimators
- Combined using WLS.

Consider the unconstrained signal model for $n = 0, \ldots, N-1$

$$x(n) = \sum_{l=1}^{L} a_l e^{j\psi_l n} + e(n), \tag{178}$$

where

(i) $L$ as well as $\{\psi_l\}_{l=1}^{L}$ are assumed *known*.
(ii) $\psi_k \neq \psi_l$ for $k \neq l$.
(iii) $e(n)$ denotes a zero mean, complex-valued, and assumed stationary (and possibly colored) additive noise.

How should one proceed to estimate $\{a_l\}_{l=1}^{L}$?

# Least-Squares Amplitude Estimation

Form

$$\begin{bmatrix} x(0) \\ \vdots \\ x(N-1) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ e^{j\psi_1} & \dots & e^{j\psi_L} \\ \vdots & \ddots & \vdots \\ e^{j\psi_1(N-1)} & \dots & e^{j\psi_L(N-1)} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_L \end{bmatrix} + \begin{bmatrix} e(0) \\ \vdots \\ e(N-1) \end{bmatrix}$$

or, using a vector-matrix notation,

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}. \tag{179}$$

Then, the LS estimator is found as

$$\hat{\mathbf{a}} = \left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}, \tag{180}$$

which is an efficient estimator for all $N \geq L$ for white Gaussian noise.

For colored Gaussian noise, LS estimators are asymptotically efficient, i.e., for sufficiently large data lengths, the variance of $\hat{\mathbf{a}}$ will reach the corresponding CRLB, given by

$$\mathrm{CRLB}(\hat{\mathbf{a}}) = \left(\mathbf{Z}^H\mathbf{Q}^{-1}\mathbf{Z}\right)^{-1}, \tag{181}$$

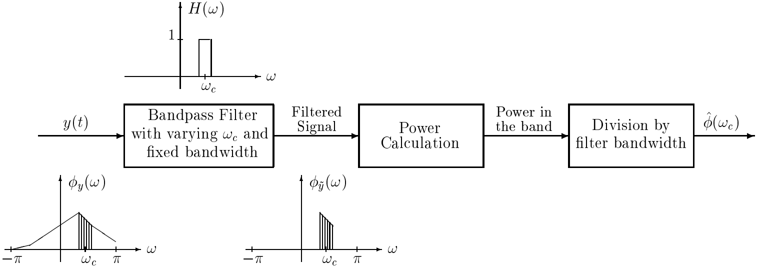
where $\mathbf{Q} = \mathrm{E}\{\mathbf{e}\mathbf{e}^H\}$, which for an additive unit variance white noise implies that $\mathbf{Q} = \mathbf{I}$.

As an alternative, an approximate LS estimate may be formed from the $L$ largest peaks of the DFT of $\{x(n)\}_{n=0}^{N-1}$, i.e.,

$$\hat{a}_l = \frac{1}{N}\sum_{n=0}^{N-1} x(n)e^{-j\psi_l n}, \quad \text{for} \quad l = 1, \ldots, L. \tag{182}$$

This estimator is also asymptotically efficient, but often (but not always) performs worse than the exact LS estimate.

# Filter-based Amplitude Estimation



An idea is to use the Capon spectral estimator for determining the amplitude of the sinusoids.

Here, we wish to do so for several components simultaneously (i.e., the harmonics). There are several ways to do so.

Form $N - M + 1$ sub-vectors of length $M$, i.e.,

$$\begin{aligned}
\mathbf{x}(n) &= \begin{bmatrix} x(n) & \dots & x(n+M-1) \end{bmatrix}^T \\
&= \begin{bmatrix} 1 & \dots & 1 \\ e^{j\psi_1} & \dots & e^{j\psi_L} \\ \vdots & \ddots & \vdots \\ e^{j\psi_1(M-1)} & \dots & e^{j\psi_L(M-1)} \end{bmatrix} \begin{bmatrix} a_1 e^{j\psi_1 n} \\ \vdots \\ a_L e^{j\psi_L n} \end{bmatrix} + \begin{bmatrix} e(n) \\ \vdots \\ e(n+M-1) \end{bmatrix} \\
&= \mathbf{Z}(n)\mathbf{a} + \mathbf{e}(n), \qquad (183)
\end{aligned}$$

where

$$\mathbf{Z}(n) = \mathbf{Z} \begin{bmatrix} e^{j\psi_1 n} & & \\ & \ddots & \\ & & e^{j\psi_L n} \end{bmatrix} = \mathbf{Z}\mathbf{D}_n. \qquad (184)$$

The magnitude squared amplitude may thus be estimated as

$$\hat{A}_l^2 = \mathrm{E}\left\{|\mathbf{h}_l^H \mathbf{x}(n)|^2\right\} = \mathbf{h}_l^H \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}(n)^H\right\} \mathbf{h}_l = \mathbf{h}_l^H \mathbf{R}\mathbf{h}_l, \tag{185}$$

where the filter of interest, $\mathbf{h}_l$, is formed as

$$\begin{aligned} \mathbf{h}_l &= \arg\min_{\mathbf{h}_l} \mathbf{h}_l^H \mathbf{R}\mathbf{h}_l \quad \text{s.t.} \quad \mathbf{h}_l^H \mathbf{z}(\psi_l) = 1 && (186) \\ &= \frac{\mathbf{R}^{-1}\mathbf{z}(\psi_l)}{\mathbf{z}^H(\psi_l)\mathbf{R}^{-1}\mathbf{z}(\psi_l)}, && (187) \end{aligned}$$

with

$$\mathbf{z}(\psi_l) = \begin{bmatrix} 1 & e^{j\psi_l} & \dots & e^{j\psi_l(M-1)} \end{bmatrix}^T \tag{188}$$

This filter constrains the current frequency of interest only, trying to minimizing the influence of the other components. This is the classical Capon amplitude (CCA) estimator

$$\hat{A}_l = \sqrt{\mathbf{h}_l^H \mathbf{R}\mathbf{h}_l} = \left(\mathbf{z}^H(\psi_l)\mathbf{R}^{-1}\mathbf{z}(\psi_l)\right)^{-1/2} \tag{189}$$

Alternatively, one may impose $L$ constraints on each filter, such that

$$\mathbf{h}_l^H \mathbf{Z} = \Big[ \underbrace{0 \quad \ldots \quad 0}_{l-1} \;\; 1 \;\; \underbrace{0 \quad \ldots \quad 0}_{L-l} \Big] = \mathbf{b}_l, \qquad (190)$$

implying that

$$\begin{aligned} \mathbf{h}_l^H \mathbf{x}(n) &= \mathbf{h}_l^H \Big[ \mathbf{Z}\mathbf{D}_n \mathbf{a} + \mathbf{e}(n) \Big] && (191) \\ &= a_l e^{j\psi_l n} + \mathbf{h}_l^H \mathbf{e}(n) && (192) \end{aligned}$$

This constraint yields the filter

$$\mathbf{h}_l = \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}^H \mathbf{R}^{-1} \mathbf{Z}\right)^{-1} \mathbf{b}_l, \qquad (193)$$

suggesting the multiple constraint Capon amplitude (MCA) estimate

$$\hat{A}_l = \sqrt{\mathbf{h}_l^H \mathbf{R} \mathbf{h}_l} = \sqrt{\mathbf{b}_l^T \left(\mathbf{Z}^H \mathbf{R}^{-1} \mathbf{Z}\right)^{-1} \mathbf{b}_l}. \qquad (194)$$

# Weighted Least-Squares Amplitude Estimation

As a third option, one may form a weighted LS estimate of the amplitude vector

$$\hat{\mathbf{a}} = \left[\sum_{n=0}^{N-M} \mathbf{Z}^H(n)\widehat{\mathbf{Q}}^{-1}\mathbf{Z}(n)\right]^{-1} \left[\sum_{n=0}^{N-M} \mathbf{Z}^H(n)\widehat{\mathbf{Q}}^{-1}\mathbf{x}(n)\right], \qquad (195)$$

where $\widehat{\mathbf{Q}}$ denotes an estimate of the noise covariance matrix. For sufficiently large $N$ and $M$, one may approximate $\widehat{\mathbf{Q}} \approx \widehat{\mathbf{R}}$, where

$$\widehat{\mathbf{R}} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{x}(n)\mathbf{x}^H(n) \qquad (196)$$

We term the resulting estimator the extended Capon amplitude (ECA) estimator.

One may improve the estimate of $\widehat{\mathbf{Q}}$ by rewriting

$$\mathbf{x}(n) = \mathbf{Z}(n)\mathbf{a} + \mathbf{e}(n) = \sum_{k=1}^{L} \underbrace{\left[a_k \mathbf{z}(\psi_k)\right]}_{\boldsymbol{\beta}_k} e^{j\psi_k n} + \mathbf{e}(n) \tag{197}$$

suggesting the *unstructured* LS estimate of $\boldsymbol{\beta}_k$

$$\hat{\boldsymbol{\beta}}_k = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{x}(n) e^{-j\psi_k n} \tag{198}$$

and the covariance matrix estimate

$$\widehat{\mathbf{Q}} = \widehat{\mathbf{R}} - \sum_{k=1}^{L} \hat{\boldsymbol{\beta}}_k \hat{\boldsymbol{\beta}}_k^H \tag{199}$$

Using this estimate yields the extended APES amplitude (EAA) estimator.

Finally, one may form a matched filterbank (MAFI) estimator using the matrix filter $\mathbf{H} = \left[ \begin{array}{ccc} \mathbf{h}_1 & \ldots & \mathbf{h}_L \end{array} \right]$, and express the design criteria as

$$\begin{aligned} \mathbf{H} &= \min_{\mathbf{H}} \operatorname{Tr}\left\{ \mathbf{H}^H \mathbf{R} \mathbf{H} \right\} \quad \text{subject to} \quad \mathbf{H}^H \mathbf{Z} = \mathbf{I} && (200) \\ &= \mathbf{R}^{-1} \mathbf{Z} \left( \mathbf{Z}^H \mathbf{R}^{-1} \mathbf{Z} \right)^{-1}. && (201) \end{aligned}$$

Then,

$$\mathbf{z}(n) = \mathbf{H}^H \mathbf{x}(n) = \mathbf{D}_n \mathbf{a} + \mathbf{H}^H \mathbf{e}(n) = \mathbf{D}_n \mathbf{a} + \mathbf{w}(n), \quad (202)$$

with the $l$th index being

$$z_l(n) = a_l e^{j\psi_l n} + w_l(n), \quad (203)$$

suggesting the MAFI amplitude estimate

$$\hat{a}_l = \frac{1}{N - M + 1} \sum_{n=0}^{N-M} z_l(n) e^{-j\psi_l n}. \quad (204)$$

# Overview and comments

We have introduced a number of amplitude estimators:

- NLS, CCA, MCA - estimating the magnitude of the amplitudes
- ECA, EAA, MAFI - estimating the complex-valued amplitudes

Most of these estimators benefit from being formed using the (per-symmetric) forward-backward averaged covariance matrix estimate in place of the forward-only estimate $\widehat{\mathbf{R}}$, i.e., using

$$\widetilde{\mathbf{R}} = \frac{1}{2}\left(\widehat{\mathbf{R}} + \mathbf{J}\widehat{\mathbf{R}}^{T}\mathbf{J}\right) \tag{205}$$

where $\mathbf{J}$ is the $M \times M$ exchange matrix. Note that EAA needs to be modified accordingly.

# Experimental Results

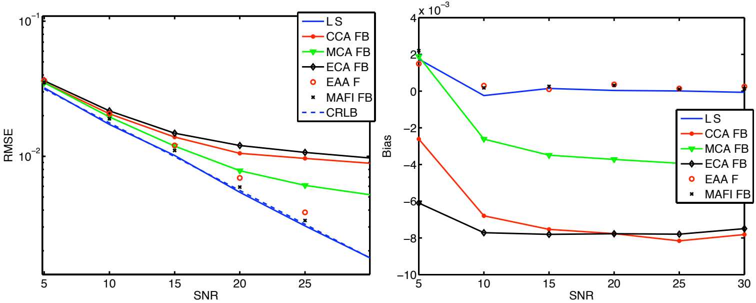


Figure: RMSE (left) and bias (right) of the discussed amplitude estimators as a function of the local SNR for $N = 160$ and $M = 40$.

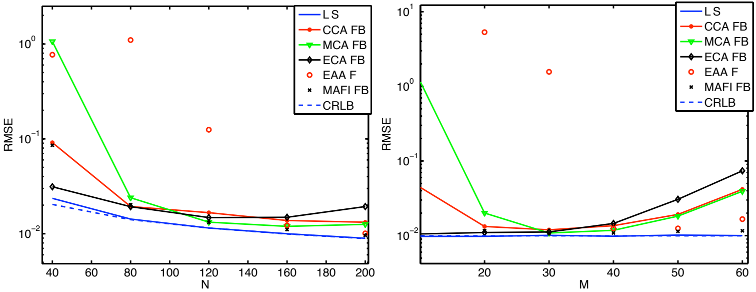


Figure: RMSE of the discussed amplitude estimators as a function of the data length, (with $M = \lfloor N/4 \rfloor$) (left) and filter length (with $N = 160$) (right).

# Comparison of Methods

In comparing the various methods, a number of parameters should be compared, namely

- Statistical efficiency, i.e., how close is the performance of the estimator to the CRLB.
- Thresholding behaviour, i.e., how does the method behave under adverse conditions like low SNR or low $N$.
- Computational complexity, i.e., how does the complexity grow as a function of $N$, $M$, $L_k$, etc.
- The complexity can be characterized in terms of a) initialization and b) per-candidate frequency complexity.

Statistical efficiency/Thresholding behaviour:

- The covariance-based methods like subspace and filtering methods are inherently suboptimal and exhibit a gap to the CRLB.
- Maximum likelihood methods like the NLS/ANLS methods are statistically efficient (for sufficiently large $N$).
- But both the Capon methods and the subspace methods work well for multi-pitch signals where the ANLS method performs poorly.
- ANLS and the Capon methods perform poorly for low fundamental frequencies, while the orthogonality-based subspace method and the NLS method perform well.
- All the considered methods are consistent!

Computational complexity (single-pitch):

- The NLS method has complexity $\mathcal{O}(L_k^2 N + L_k^3 + L_k N^2 + N^2)$ per grid point and requires no initialization (recall that $L_k \ll M \leq N$).
- The ANLS method has complexity $\mathcal{O}(L_k)$ per grid point and requires that an FFT and the power is computed as initialization.
- The WLS method has complexity $\mathcal{O}(L_k^3)$ (given unconstrained frequency estimates).
- The Capon based methods have a complexity of $\mathcal{O}(L_k^3 + ML_k^2 + M^2 L_k)$ per grid point (discounting the order-recursive implementation) and an initialization of $\mathcal{O}(M^3)$.
- The orthogonality-based subspace method has complexity $\mathcal{O}(L_k(M - L_k))$ per grid point and as initialization requires that the EVD of $\mathbf{R}_k$ is computed which is $\mathcal{O}(M^3)$ and $M$ FFTs and power computations.
- The shift-invariance-based method has complexity $\mathcal{O}(L_k)$ per grid point and an initialization of $\mathcal{O}(M^3 + L_k^3 + M^2 L_k + L_k^2 M)$.

# Open Issues and Directions for Future Research

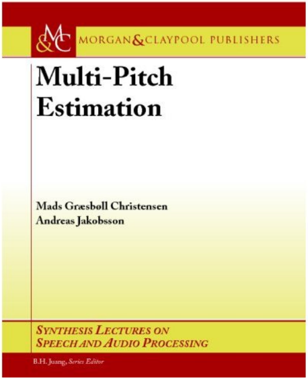
Not yet mature and tested technology. Many things need to be done:

- Extensive tests on databases
- Taking more specific knowledge into account
- Fast implementations
- Exact estimators for low frequencies
- Multiple channels
- How to take colored noise into account
- Inharmonicity (modified models)
- Spectral smoothness (e.g., filter model)
- Non-parametric vs. parametric (e.g., spectrogram modeling)
- Temporal modeling (e.g., HMMs, optimal segmentation, smoothing)
- Modified models: what do we really need?
- Order estimation and model selection (still an unsolved problem)

# Conclusions

- Fundamental frequency estimation can be solved using a number of tractable methods rooted in estimation theory.
- The parametric approach offers high-resolutions estimates and predictable behavior.
- Multi-pitch estimation is complicated and sometimes not a well-defined problem.
- Methods based on optimal filtering particularly appear especially promising for multi-pitch estimation.
- A full parametrization useful for many applications.
- In summary, the parametric methods form a quite promising alternative to the traditionally used methods.
- This is especially so in applications that require very accurate estimates (e.g., tuning, transcription/analysis of vibrato, glissando).

# Shameless Plug


MORGAN&CLAYPOOL PUBLISHERS
Multi-Pitch Estimation
Mads Græsbøll Christensen
Andreas Jakobsson
SYNTHESIS LECTURES ON SPEECH AND AUDIO PROCESSING
B.H. Juang, Series Editor

Book and MATLAB toolbox available!

Free download (if subscription) from `http://www.morganclaypool.com`

Papers and MATLAB code available at `http://imi.aau.dk/~mgc`

# Acknowledgments

A number of people who contributed to this work should be thanked:

- Andreas Jakobsson
- Søren Holdt Jensen
- Petre Stoica
- Johan Xi Zhang
- Jesper Kjær Nielsen
- Jesper Rindom Jensen
- Jesper Lisby Højvang