Statistical Parametric Speech Processing

Solving problems with the model-based approach

August 20, 2017

Mads Græsbøll Christensen, Jesper Kjær Nielsen, and Jesper Rindom Jensen

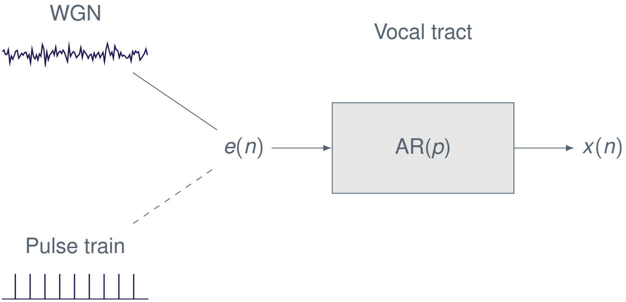
Audio Analysis Lab, AD:MT
Aalborg University
Denmark

AALBORG UNIVERSITY
DENMARK

Outline

Outline

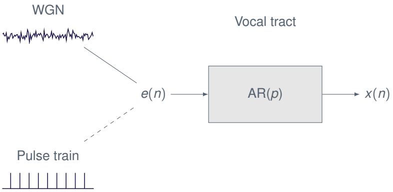
Motivation



Unvoiced speech Autoregressive process with unknown AR-parameters and excitation noise variance.

Voiced speech Periodic signal with unknown pitch, amplitudes, and phases.

Motivation



How is the model wrong?

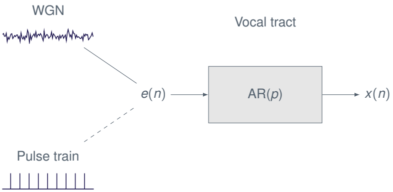
- The physics behind speech production is much more complicated
- Speech is non-stationary
- Speech might be a mixture of voiced and unvoiced sounds
- etc.

Motivation

Essentially, all models are wrong, but some are useful.

Box, 1987

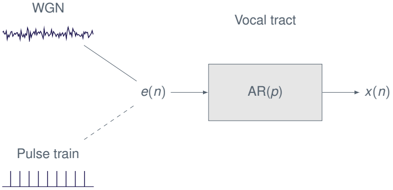
Motivation



How is a (parametric) model useful?

- A model allows us to formulate a problem in terms of the quantaties of interest (e.g., the fundamental frequency).
- A model is an explicit way of stating our assumptions.
- Models allow us to solve problems in an optimal fashion.
- Models reduce the number of unknowns from many to a few model parameters.

Motivation



Example

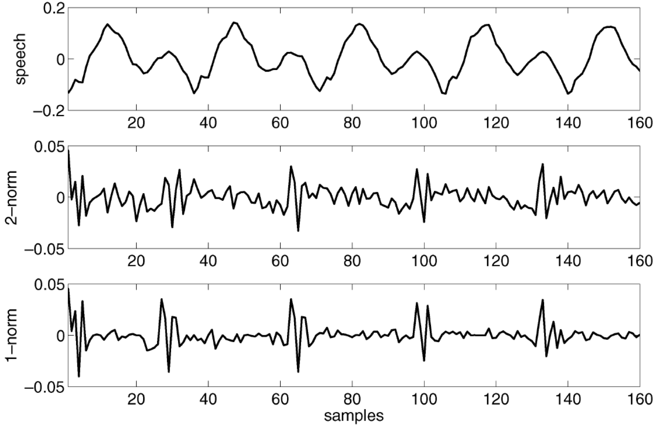
Signal model

$$\boldsymbol{x} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{e} \tag{1}$$

- Estimate $\boldsymbol{a}$ by minimising the 2-norm (Gaussian noise)
- Estimate $\boldsymbol{a}$ by minimising the 1-norm (Laplacian noise) (Giacobello 2012)

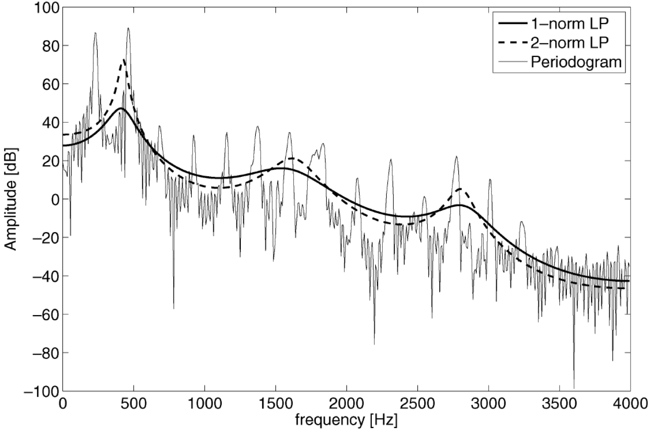
Motivation

Example



Motivation

Example



Outline

Who are we?

The Audio Analysis Lab

- Research lab founded in 2012 at AD:MT, Aalborg University, Denmark.
- Basic/applied research in signal processing theory and methods aimed at or involving analysis of audio signals.
- Our goal is to push the boundaries of current methods and increase the understanding of problems by pursuing mathematically tractable approaches.

Who are we?

The Audio Analysis Lab

People

- Four senior people
- One postdoc
- Nine Ph.D.-students
- One research assistant
- One Adjunct professor (Jacob Benesty, University of Quebec)
- One guest professors (Barry Quinn, MacQuarie University)

Current Collaborators

- GN ReSound
- Bang & Olufsen
- Bruel & Kjær
- Parkinson's Voice Initiative
- Richard Heusdens, TU Delft (former Guest Professor)
- Andreas Jakobsson, Lund University
- Jingdong Chen, Northwestern Polytechnical University

Who are we?

The Audio Analysis Lab

Current Speech-related Research Projects

- Signal processing for detecting the cocktail party problem and methods for enhancing listening in noisy situations (COCKTAIL), 2014–2018
- Signal Processing for Diagnosis of Parkinson's Disease from Noisy Speech, 2015–2019
- Spatio-Temporal Filtering Methods for Enhancement and Separation of Speech Signals, 2012–2017.

Website: http://audio.create.aau.dk

Youtube: http://tinyurl.com/yd8mo55z

Outline

Introduction

Parametric speech processing

- Parametric speech processing is processing based on parametric signal models.
- The signal models are often generative and described in terms of physically meaningful parameters.
- Parametric speech models have been around for many years (e.g., linear prediction in the 70s, sinusoidal model in the 80s).
- Skeptics argue that the models are (always) wrong and that it is not possible to estimate the model parameters well enough under adverse conditions.
- Parametric models can be used for many things and in different ways.
- An essential part of parametric speech processing is to estimate model parameters from noisy observations.

Introduction

Methodology

- Methods rooted in estimation theory.
- Based on parametric models of the signal of interest.
- Analysis of estimation and modeling problems as mathematical problems.

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?

Introduction

Some interesting questions:

- Under which conditions can a method be expected to work?
- How does performance depend on the acoustic environment?
- Is the method optimal (and what does optimal mean)?
- How do we improve the method?

Observations:

- Only possible to answer if assumptions are made explicit! Often the assumptions are sufficient conditions but not necessary.
- Non-parametric methods are hard to analyze and understand.

Outline

Outline

About Models

What's a good model?

- Captures the essence of the signal
- Physically meaningful
- As simple as possibles

Tradeoff:

- Good data fit
- As few parameters as possible (Occam's razor)

Too many parameters lead to overfitting and poorer estimates.

We will explore how we can model speech and how we can manipulate the models.

Harmonic Model

The harmonic model is given by (for $n = 0, \ldots, N-1$)

$$x(n) = s(n) + e(n) = \sum_{l=1}^{L} a_l e^{j\omega_0 l n} + e(n). \qquad (2)$$

Definitions:

- $s(n)$ is voiced speech
- $e(n)$ is the noise/stochastic parts
- ω_0 is the fundamental frequency
- $\omega_0 l$ is the frequency of the lth harmonic
- $a_l = A_l e^{j\phi_l}$ is the complex amplitude
- $\boldsymbol{\theta} = [\ \omega_0\ A_1\ \phi_1\ \cdots A_L\ \phi_L\ ]^T$

Harmonic Model

The model can be written in matrix-vector notation as

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{a} + \mathbf{e}(n) \tag{3}$$
$$= \mathbf{s}(\theta) + \mathbf{e}(n) \tag{4}$$

with the following definitions:

$$\mathbf{x}(n) = [\ x(n)\ \cdots\ x(n+M-1)\ ]^T$$
$$\mathbf{z}(\omega) = [\ 1\ e^{j\omega}\ \cdots\ e^{j\omega(M-1)}\ ]^T$$
$$\mathbf{Z} = [\ \mathbf{z}(\omega_0)\ \cdots\ \mathbf{z}(\omega_0 L)\ ]$$
$$\mathbf{a} = [\ a_1\ \cdots\ a_L\ ]^T$$

We call $\mathbf{x}(n)$ a snapshot. A collection of such snapshosts is written as $\{\mathbf{x}(n)\}$.

Harmonic Model

The model can be written in different ways:

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{Z}(n)\mathbf{a} + \mathbf{e}(n) && (5) \\ &= \mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{e}(n) && (6) \\ &= \mathbf{Z}\mathbf{a}(n) + \mathbf{e}(n), && (7) \end{aligned}$$

where $\mathbf{D}(n) = \mathbf{D}^n$ with $\mathbf{D} = \mathrm{diag}(\begin{bmatrix} e^{j\omega_0} & e^{j\omega_0 2} & \dots & e^{j\omega_0 L} \end{bmatrix})$. Notice that $\mathbf{D}(n)\mathbf{a} = \sum_{l=1}^{L} a_l e^{j\omega_0 l n}$.

This means that we can think of the time-dependency as influencing different parts. The different models are useful for different purposes!

Sometimes we also write the model as

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}, \quad (8)$$

which is a special case of the model above with $M = N$.

Harmonic Model

The covariance matrix of $\mathbf{x}(n)$ is

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\}. \tag{9}$$

Written in terms of the harmonic model, we get

$$\begin{aligned}\mathbf{R} &= \mathbf{Z}\mathrm{E}\left\{\mathbf{a}(n)\mathbf{a}^H(n)\right\}\mathbf{Z}^H + \mathrm{E}\left\{\mathbf{e}(n)\mathbf{e}^H(n)\right\} && (10)\\ &= \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \mathbf{Q}, && (11)\end{aligned}$$

which is called the covariance matrix model.

P is the covariance matrix for the amplitudes, which can be shown to be (under certain conditions)

$$\mathbf{P} \approx \mathrm{diag}\left(\left[\, A_1^2 \;\cdots\; A_L^2 \,\right]\right). \tag{12}$$

Filtering

Let the output signal $y(n)$ of a filter having coefficients $h(n)$ be defined as

$$y(n) = \sum_{m=0}^{M-1} h(m)x(n-m) = \mathbf{h}^H \mathbf{x}(n), \tag{13}$$

with $M \leq N$ and where $\mathbf{h}$ is a vector formed from $\{h(n)\}$. The output power is then

$$\mathrm{E}\left\{|y(n)|^2\right\} = \mathbf{h}^H \mathbf{R} \mathbf{h}. \tag{14}$$

Recall that the signal model was

$$\mathbf{x} = \mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{e}. \tag{15}$$

The filtered output can thus be seen to be

$$\mathbf{h}^H \mathbf{x}(n) = \mathbf{h}^H \mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{h}^H \mathbf{e}. \tag{16}$$

Filtering

The filtered observed signal **x** could be written as

$$\mathbf{h}^H\mathbf{x}(n) = \mathbf{h}^H\mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{h}^H\mathbf{e}. \tag{17}$$

This comprises two terms:

1. The speech passed throught the filter $\mathbf{h}^H\mathbf{Z}\mathbf{D}(n)\mathbf{a}$.
2. The residual noise $\mathbf{h}^H\mathbf{e}$.

Using the covariance matrix model, we can write the output power as

$$\begin{aligned} \mathrm{E}\left\{|y(n)|^2\right\} &= \mathbf{h}^H\mathbf{R}\mathbf{h} && (18)\\ &= \mathbf{h}^H\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{h} + \mathbf{h}^H\mathbf{Q}\mathbf{h}, && (19) \end{aligned}$$

where $\mathbf{h}^H\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{h}$ is the power of the filtered speech and $\mathbf{h}^H\mathbf{Q}\mathbf{h}$ is the residual noise.

Subspace Model

Recall that the model is

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{a}(n) + \mathbf{e}(n), \tag{20}$$

and that the covariance matrix then is given by

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \sigma^2\mathbf{I}, \tag{21}$$

where $\mathbf{ZPZ}^H$ has rank L and

$$\mathbf{P} = \mathrm{diag}\left(\left[\, A_1^2 \ \cdots \ A_L^2 \,\right]\right).$$

Subspace Model

Let

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H \tag{22}$$

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

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_M \end{bmatrix}, \tag{23}$$

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

$$\mathbf{S} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_V \end{bmatrix}. \tag{24}$$

The subspace that is spanned by the columns of $\mathbf{S}$ we denote $\mathcal{R}(\mathbf{S})$,

Subspace Model

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

$$\mathbf{G} = \begin{bmatrix} \mathbf{u}_{V+1} & \cdots & \mathbf{u}_M \end{bmatrix}, \tag{25}$$

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

$$\mathbf{U}\left(\mathbf{\Lambda} - \sigma^2\mathbf{I}\right)\mathbf{U}^H = \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{P}_k \mathbf{Z}_k^H. \tag{26}$$

It follows that 1) the matrices $\mathbf{Z}$ and $\mathbf{G}$ are orthogonal, i.e.,

$$\mathbf{Z}^H\mathbf{G} = \mathbf{0} \tag{27}$$

and 2) the matrices $\mathbf{S}$ and $\mathbf{Z}$ span the same space, i.e.,

$$\mathcal{R}(\mathbf{S}) = \mathcal{R}(\mathbf{Z}). \tag{28}$$

Harmonic Model

What's wrong with this model?

- It does not take non-stationarity into account
- Background noise is rarely white (and not always Gaussian)
- The model order is unknown and time-varying
- Even if stationary, speech signals are not perfectly periodic
- The model does not differentiate between background noise and unvoiced speech

Can this be dealt with? Does it matter?

On The Complex Signal Model

Often we use a complex signal model. There are a number of reasons for this:

- Simpler math
- Faster algorithms

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

- Using the Hilbert transform to calculate the analytic signal (Marple 1999).
- Do not hold for very low and high frequencies (relative to N).
- It is possible to account for real signals in estimators, but it is often not worth the trouble (Christensen 2013).

Outline

Likelihood Function

If we assume that the signal is Gaussian distributed, i.e.,

$$\mathbf{x}(n) \sim \mathcal{N}(\mathbf{s}(\theta), \mathbf{Q}) \tag{29}$$

then the likelihood function is given by

$$p(\mathbf{x}(n); \theta) = \frac{1}{\pi^M \det(\mathbf{Q})} e^{-(\mathbf{x}(n) - \mathbf{Z}\mathbf{a}(n))^H \mathbf{Q}^{-1} (\mathbf{x}(n) - \mathbf{Z}\mathbf{a}(n))}. \tag{30}$$

If the noise is i.i.d., the likelihood of $\{\mathbf{x}(n)\}_{n=0}^{G-1}$ can be written as

$$p(\{\mathbf{x}(n)\}; \theta) = \prod_{n=0}^{G-1} p(\mathbf{x}(n); \theta). \tag{31}$$

In the above, $\mathbf{Q}$ could represent the covariance of unvoiced speech, noise or both combined.

Likelihood Function

The log-likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}) = \ln p(\{\mathbf{x}(n)\}; \boldsymbol{\theta}) \tag{32}$$

$$= \sum_{n=0}^{G} \ln p(\mathbf{x}(n); \boldsymbol{\theta}). \tag{33}$$

The maximum likelihood estimator (MLE) is then given by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \tag{34}$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{n=0}^{G} \ln p(\mathbf{x}(n); \boldsymbol{\theta}). \tag{35}$$

The MLE is statistically efficient, i.e., it attains the CRLB, for sufficiently large N! Moreover, its estimates are normally distributed.

Maximum Likelihood Estimator

Let us find the MLE for pitch estimation. For white Gaussian noise ($\mathbf{Q} = \sigma^2 \mathbf{I}$) with $M = N$ the log-likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}) = -N \ln \pi - N \ln \sigma^2 - \frac{1}{\sigma^2} \|\mathbf{x} - \mathbf{Za}\|_2^2, \tag{36}$$

where $\boldsymbol{\theta} = [\omega_0 \; A_1 \; \dots A_L \; \phi_1 \; \dots \phi_L]$. The concentrated MLE is given by

(Quinn 1991)

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathcal{L}(\omega_0) = \arg\max_{\omega_0} \mathbf{x}^H \mathbf{Z} \left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{x}. \tag{37}$$

This means that we must find the ω_0 that results in the largest projection energy!

Outline

Parameter Estimation Bounds

An estimate $\hat{\theta}_i$ of θ_i (i.e., the ith element of $\boldsymbol{\theta} \in \mathbb{R}^P$) is unbiased if

$$\mathrm{E}\left\{\hat{\theta}_i\right\} = \theta_i \ \forall \theta_i, \tag{38}$$

and the difference (if any) is referred to as the bias. The Cramér-Rao lower bound (CRLB) is then given by

$$\mathrm{var}(\hat{\theta}_i) \geq \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}, \tag{39}$$

where the Fisher Information Matrix (FIM) $\mathbf{I}(\boldsymbol{\theta})$ is given by

$$[\mathbf{I}(\boldsymbol{\theta})]_{il} = -\mathrm{E}\left\{\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i \partial\theta_l}\right\}, \tag{40}$$

with $\ln p(\mathbf{x};\boldsymbol{\theta})$ being the log-likelihood function for $\mathbf{x} \in \mathbb{C}^N$.

Parameter Estimation Bounds

The CRLBs can be dervied for the harmonic model (for WGN):

$$\text{var}(\hat{\omega}_0) \geq \frac{6\sigma^2}{N(N^2-1)\sum_{l=1}^{L} A_l^2 l^2} \tag{41}$$

$$\text{var}(\hat{A}_l) \geq \frac{\sigma^2}{2N} \tag{42}$$

$$\text{var}(\hat{\phi}_l) \geq \frac{\sigma^2}{2N}\left(\frac{1}{A_l^2} + \frac{3l^2(N-1)^2}{\sum_{m=1}^{L} A_m m^2 (N^2-1)}\right). \tag{43}$$

These depend on the following quantity:

$$PSNR = 10\log_{10}\frac{\sum_{l=1}^{L} A_l^2 l^2}{\sigma^2}\ [\text{dB}]. \tag{44}$$

Parameter Estimation Bounds

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 (e.g., ω_0).
- The bounds can be used as benchmarks in simulations.
- Provide us with “rules of thumb”.

Caveat emptor: The CRLB does not accurately predict the performance of non-linear estimators under adverse conditions.

It is possible to compute *exact* CRLBs, where no asymptotic approximations are used!

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

Amplitude Estimation

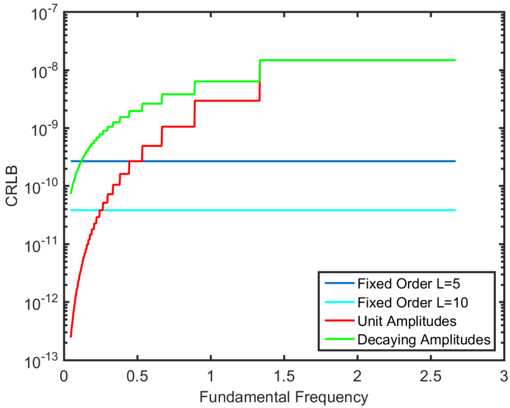


Figure: CRLB as a function of ω_0 for different cases.

Outline

General Multi-Channel Model

Define $\mathbf{x}_k(n) \in \mathbb{C}^M$ as the snapshop for the kth channel.

Each snapshot is modeled as sums of sinusoids in Gaussian noise $\mathbf{e}_k$ with covariance $\mathbf{Q}_k$ (Christensen 2012), i.e.,

$$\mathbf{x}_k(n) = \mathbf{Z}(n)\mathbf{a}_k + \mathbf{e}_k(n), \tag{45}$$

with $\mathbf{a}_k = [\ A_{k,1}e^{j\phi_{k,1}} \ \cdots \ A_{k,L}e^{j\phi_{k,L}}\ ]^T$.

Interpretation:

- Shared fundamental frequency.
- Different amplitudes and phases.
- Different noise on each channel.
- Different IR, different noise characteristics.

General Multi-Channel Model

Let θ_k be the parameter vector for the kth channel. The likelihood function is then

$$p(\mathbf{x}_k(n); \theta_k) = \frac{1}{\pi^M \det(\mathbf{Q}_k)} e^{-\mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}. \tag{46}$$

If the deterministic part is stationary and $\mathbf{e}_k(n)$ is i.i.d. over n, we get

$$p(\{\mathbf{x}_k(n)\}; \theta_k) = \frac{1}{\pi^{MG} \det(\mathbf{Q}_k)^G} e^{-\sum_{n=0}^{G-1} \mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}. \tag{47}$$

Furthermore, if it is independent over k, the combined likelihood is

$$p(\{\mathbf{x}_k(n)\}; \{\theta_k\}) = \prod_{k=1}^{K} \frac{1}{\pi^{MG} \det(\mathbf{Q}_k)^G} e^{-\sum_{n=0}^{G-1} \mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}. \tag{48}$$

General Multi-Channel Model

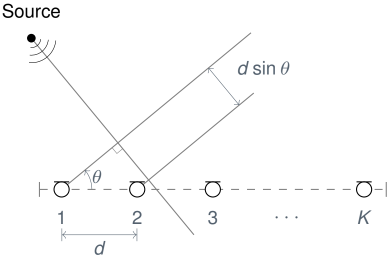
Simplifying assumptions can be made, as appropriate. For example:

- Same noise color, i.e., $\mathbf{Q}_k = \mathbf{Q} \forall k$.
- White noise, i.e., $\mathbf{Q}_k = \sigma_k^2 \mathbf{I}$.
- Only one snapshot, i.e., $G = 1$ and $M = N$.
- Same amplitudes but different phases across channels, i.e., $A_{k,l} = A_l \forall k$.

The model ignores noise correlation across channels and array geometry.

Linear Array

For a linear array and sources in the farfield:



Observations

- The delay (in samples) for adjacent microphones is $\Delta = \frac{d \sin \theta}{c} f_s$.
- What does the model look like for this case?.

Linear Array

Defining Δ_k to be the delay (in samples) between microphone 1 and k, the speech signal at microphone k is (Jensen 2014)

$$\begin{aligned} \mathbf{s}_k(n) &= \mathbf{s}(n - \Delta_k) && (49) \\ &= \mathbf{s}\left(n - \frac{d \sin\theta}{c} f_s(k-1)\right). && (50) \end{aligned}$$

Recall that $\mathbf{s}(n)$ can be written as $\mathbf{s}(n) = \mathbf{Z}\mathbf{D}(n)\mathbf{a}$ and hence

$$\mathbf{s}_k\left(n - \frac{d \sin\theta}{c} f_s(k-1)\right) = \mathbf{Z}\mathbf{D}\left(n - \frac{d \sin\theta}{c} f_s(k-1)\right)\mathbf{a}. \quad (51)$$

As we can see, it is easy to account for fractional delays in the parametric model. Other geometries can easily be incorporated too.

Linear Array

Recall that the matrix $\mathbf{D}(n)$ is given by

$$\mathbf{D}(n) = \begin{bmatrix} e^{j\omega_0 n} & 0 & \cdots & 0 \\ 0 & e^{j\omega_0 2n} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{j\omega_0 Ln} \end{bmatrix}. \tag{52}$$

and thus $\mathbf{D}(n-\Delta)$ is

$$\mathbf{D}(n-\Delta) = \begin{bmatrix} e^{j\omega_0 (n-\Delta)} & 0 & \cdots & 0 \\ 0 & e^{j\omega_0 2(n-\Delta)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{j\omega_0 L(n-\Delta)} \end{bmatrix}. \tag{53}$$

Reverberation

We can modify the model to account for reverberation.

Let $h_k(n)$ denote the impulse response from the source to the kth microphone. Then the signal at that microphone is

$$x_k(n) = s(n) * h_k(n) + e_k(n). \quad (54)$$

Assuming that the impulse response is shorter than the segment length and that the signal is stationary, then (Jensen 2016)

$$x_k(n) \approx \beta_k s(n - \Delta_k) + e_k(n). \quad (55)$$

This is due to the sinusoidal nature of $s(n)$.

Outline

Unvoiced Speech

So far, we modeled the observed signal as

$$x(n) = s(n) + e(n), \tag{56}$$

where $s(n)$ is the voiced speech and $e(n)$ is stochastic signal components (noise).

Real speech contains both voiced, unvoiced and noise components. How do we account for this?

Modified model:

$$x(n) = \underbrace{s(n)}_{voiced} + \underbrace{u(n)}_{unvoiced} + \underbrace{w(n)}_{noise}. \tag{57}$$

Unvoiced Speech

What's a good model for unvoiced speech then?

Fortunately, the good old auto-regressive (AR) model is pretty good for unvoiced speech, i.e.,

$$u(n) = \sum_{i=1}^{I} \gamma_i u(n-i) + \eta(n). \tag{58}$$

Here, $\eta(n)$ is the excitation for the unvoiced speech, which can be modeled as white Gaussian, i.e, $\eta(n) \sim \mathcal{N}\left(0, \sigma^2\right)$.

However, the AR parameters, $\{\gamma_i\}$, are now also unknown and have to be estimated along with the parameters of the harmonic model.

Colored Noise

In speech applications, the background noise is rarely white.

Even though the white noise assumption is convenient from a mathematical point of view, it is actually the worst case from an estimation theoretical point of view!

How do we deal with colored noise? Do the bounds change, etc.? These questions can be addressed in several ways.

Colored Noise

Let us examine the following signal model:

$$\mathbf{x}(n) = \mathbf{s}(n) + \mathbf{e}(n). \tag{59}$$

Suppose that the colored noise is distributed as $\mathbf{e}(n) \sim \mathcal{N}(0, \mathbf{Q})$.

We can transform the observed signal by a matrix $\mathbf{A}$ as

$$\mathbf{A}^H\mathbf{x}(n) = \mathbf{A}^H\mathbf{s}(n) + \mathbf{A}^H\mathbf{e}(n). \tag{60}$$

Then if we select $\mathbf{A}$ such that $\mathbf{v}(n) = \mathbf{A}^H\mathbf{e}(n)$ is distributed as $\mathbf{v}(n) \sim \mathcal{N}(0, \mathbf{I})$, the noise is now white.

From the above, we can deduce that $\mathbf{A}$ must be the Cholesky factor of $\mathbf{Q}^{-1}$, i.e., $\mathbf{A}\mathbf{A}^H = \mathbf{Q}^{-1}$, since $\mathbf{A}^H\mathbf{Q}\mathbf{A} = \mathbf{I}$.

Colored Noise

The voiced speech is, however, also affected by this as $\mathbf{A}^H\mathbf{s}(n)$, and the model must be modified accordingly.

Instead, consider the signal model

$$x(n) = s(n) + e(n). \tag{61}$$

Next, we apply a filter having coefficients $h(n)$, i.e.,

$$h(n) * x(n) = h(n) * s(n) + v(n), \tag{62}$$

so that $\mathbf{v} = [\ v(0)\ \cdots\ v(M-1)\ ]^H$ where $v(n) = h(n) * e(n)$ is distributed as $\mathbf{v}(n) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Colored Noise

Since

$$s(n) = \sum_{l=1}^{L} a_l e^{j\omega_0 l n} \tag{63}$$

we have that

$$h(n) * s(n) = h(n) * \sum_{l=1}^{L} a_l e^{j\omega_0 l n} \approx \sum_{l=1}^{L} \tilde{a}_l e^{j\omega_0 l n}. \tag{64}$$

This means that the model is preserved by the filter. Hence, we do not have to change it.

This principle can also be used to obtain the CRLB for the colored noise case.

Colored Noise

How to estimate the noise covariance matrix then?

- Voice activity detection
- Noise trackers (Gerkmann 2012)
- Codebook-based approach (Srinivasan 2007)
- Long-term averaged spectrum (speech, noise)
- Order-recursive estimation, APES (Nørholm 2016)

Non-Stationary Speech

Can we deal with a time-varying pitch? The harmonic chirp model aims to do just that.

For a segment of a speech signal it is given by

$$x(n) = \sum_{l=1}^{L} A_l e^{j\theta_l(n)} + e(n) \tag{65}$$

where

- $\theta_l(n)$ is the instantaneous phase of the lth harmonic.
- everything is as before.

Non-Stationary Speech

The instantaneous phase $\theta_l(\cdot)$ is given by

$$\theta_l(t) = \int_0^t l\omega_0(\tau)d\tau + \phi_l, \tag{66}$$

where $\omega_0(t)$ is the time-varying pitch and ϕ_l is the phase.
If the pitch is slowly varying, i.e., $\omega_0(t) = \alpha_0 t + \omega_0$, we get

$$\theta_l(t) = \frac{1}{2}\alpha_0 l t^2 + \omega_0 l t + \phi_l, \tag{67}$$

where α_0 is the fundamental chirp rate.

The resulting model is called the harmonic chirp model (HCM) (Christensen 2014, Nørholm 2016).

Non-Stationary Speech

We can easily put this into matrix-vector notation. Define a vector with $n_0 = -(N-1)/2$ as

$$\mathbf{x} = \begin{bmatrix} x(n_0) & x(n_0+1) & \dots & x(n_0+N-1) \end{bmatrix}. \tag{68}$$

and a matrix as

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}(\omega_0, \alpha_0) & \mathbf{z}(2\omega_0, 2\alpha_0) & \dots & \mathbf{z}(L\omega_0, L\alpha_0) \end{bmatrix}, \tag{69}$$

with columns

$$\mathbf{z}(l\omega_0, l\alpha_0) = \begin{bmatrix} e^{j(\frac{1}{2}\alpha_0 l n_0^2 + \omega_0 l n_0)} & \dots & e^{j(\frac{1}{2}\alpha_0 l (n_0+N-1)^2 + \omega_0 l (n_0+N-1))} \end{bmatrix}^T.$$

The model can now be written as before:

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e} \tag{70}$$

Note that we cannot use the trick with $\mathbf{D}(n)$ to simplify this model.

Non-Stationary Speech
Experiments

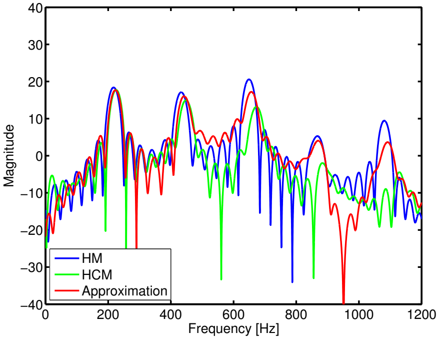


Figure: Spectrum of harmonic model, harmonic chirp model, and an approximation.

Outline

Amplitude Estimation

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

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

This can be done in a number of ways including (Stoica 2000):

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

Amplitude Estimation

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

$$x(n) = \sum_{l=1}^{L} a_l e^{j\psi_l n} + e(n), \tag{71}$$

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

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

or, using a vector-matrix notation,

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}. \tag{72}$$

Then, the LS estimator is found as

$$\hat{\mathbf{a}} = \left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}, \tag{73}$$

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

Amplitude Estimation

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

$$\mathrm{CRLB}(\hat{\mathbf{a}}) = \left(\mathbf{Z}^H \mathbf{Q}^{-1} \mathbf{Z}\right)^{-1}, \tag{74}$$

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

$$\hat{a}_l = \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-j\psi_l n}, \quad \text{for} \quad l = 1, \dots, L. \tag{75}$$

This estimator is also asymptotically efficient, but often performs worse than the exact LS estimate.

Amplitude Estimation

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

$$\begin{aligned}
\mathbf{x}(n) &= \begin{bmatrix} x(n) & \dots & x(n+M-1) \end{bmatrix}^T \\
&= \begin{bmatrix} 1 & \dots & 1 \\ e^{j\psi_1} & \dots & e^{j\psi_L} \\ \vdots & \ddots & \vdots \\ e^{j\psi_1(M-1)} & \dots & e^{j\psi_L(M-1)} \end{bmatrix} \begin{bmatrix} a_1 e^{j\psi_1 n} \\ \vdots \\ a_L e^{j\psi_L n} \end{bmatrix} + \begin{bmatrix} e(n) \\ \vdots \\ e(n+M-1) \end{bmatrix} \\
&= \mathbf{Z}(n)\mathbf{a} + \mathbf{e}(n), \qquad (76)
\end{aligned}$$

where

$$\mathbf{Z}(n) = \mathbf{Z} \begin{bmatrix} e^{j\psi_1 n} & & \\ & \ddots & \\ & & e^{j\psi_L n} \end{bmatrix} = \mathbf{ZD}(n). \qquad (77)$$

Amplitude Estimation

The squared amplitude may be estimated by applying a filter $\mathbf{h}_l$ as

$$\hat{A}_l^2 = \mathrm{E}\left\{|\mathbf{h}_l^H \mathbf{x}(n)|^2\right\} = \mathbf{h}_l^H \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}(n)^H\right\} \mathbf{h}_l = \mathbf{h}_l^H \mathbf{R} \mathbf{h}_l, \tag{78}$$

where the filter, $\mathbf{h}_l$, is given by

$$\mathbf{h}_l = \arg\min_{\mathbf{h}_l} \mathbf{h}_l^H \mathbf{R} \mathbf{h}_l \quad \text{s.t.} \quad \mathbf{h}_l^H \mathbf{z}(\psi_l) = 1 \tag{79}$$

$$= \frac{\mathbf{R}^{-1}\mathbf{z}(\psi_l)}{\mathbf{z}^H(\psi_l)\mathbf{R}^{-1}\mathbf{z}(\psi_l)}, \tag{80}$$

with

$$\mathbf{z}(\psi_l) = \begin{bmatrix} 1 & e^{j\psi_l} & \ldots & e^{j\psi_l(M-1)} \end{bmatrix}^T \tag{81}$$

This is the classical Capon amplitude (CCA) estimator

$$\hat{A}_l = \sqrt{\mathbf{h}_l^H \mathbf{R} \mathbf{h}_l} = \left(\mathbf{z}^H(\psi_l)\mathbf{R}^{-1}\mathbf{z}(\psi_l)\right)^{-1/2} \tag{82}$$

Amplitude Estimation

Alternatively, we can impose L constraints on each filter, such that

$$\mathbf{h}_l^H \mathbf{Z} = \left[\underbrace{0 \quad \dots \quad 0}_{l-1} \; 1 \; \underbrace{0 \quad \dots \quad 0}_{L-l} \right] = \mathbf{b}_l, \tag{83}$$

which means that

$$\begin{aligned} \mathbf{h}_l^H \mathbf{x}(n) &= \mathbf{h}_l^H \left[\mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{e}(n) \right] && (84) \\ &= a_l e^{j\psi_l n} + \mathbf{h}_l^H \mathbf{e}(n). && (85) \end{aligned}$$

This constraint yields the filter

$$\mathbf{h}_l = \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}^{-1}\mathbf{Z}\right)^{-1}\mathbf{b}_l, \tag{86}$$

from which we get the multiple constraints Capon amplitude (MCA) estimate

$$\hat{A}_l = \sqrt{\mathbf{h}_l^H \mathbf{R} \mathbf{h}_l} = \sqrt{\mathbf{b}_l^T \left(\mathbf{Z}^H\mathbf{R}^{-1}\mathbf{Z}\right)^{-1}\mathbf{b}_l}. \tag{87}$$

Amplitude Estimation

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

$$\hat{\mathbf{a}} = \left[\sum_{n=0}^{N-M} \mathbf{Z}^H(n) \widehat{\mathbf{Q}}^{-1} \mathbf{Z}(n) \right]^{-1} \left[\sum_{n=0}^{N-M} \mathbf{Z}^H(n) \widehat{\mathbf{Q}}^{-1} \mathbf{x}(n) \right], \tag{88}$$

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

$$\widehat{\mathbf{R}} = \frac{1}{N - M + 1} \sum_{n=0}^{N-M} \mathbf{x}(n) \mathbf{x}^H(n) \tag{89}$$

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

Amplitude Estimation

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

$$\mathbf{x}(n) = \mathbf{Z}(n)\mathbf{a} + \mathbf{e}(n) = \sum_{k=1}^{L} \underbrace{\left[a_k \mathbf{z}(\psi_k)\right]}_{\boldsymbol{\beta}_k} e^{j\psi_k n} + \mathbf{e}(n) \tag{90}$$

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

$$\hat{\boldsymbol{\beta}}_k = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{x}(n) e^{-j\psi_k n} \tag{91}$$

and the covariance matrix estimate

$$\widehat{\mathbf{Q}} = \widehat{\mathbf{R}} - \sum_{k=1}^{L} \hat{\boldsymbol{\beta}}_k \hat{\boldsymbol{\beta}}_k^H \tag{92}$$

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

Amplitude Estimation

Finally, one may form a matched filterbank (MAFI) estimator using the matrix filter $\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & \dots & \mathbf{h}_L \end{bmatrix}$, and express the design criteria as

$$\mathbf{H} = \min_{\mathbf{H}} \operatorname{Tr}\left\{\mathbf{H}^H \mathbf{R} \mathbf{H}\right\} \quad \text{subject to} \quad \mathbf{H}^H \mathbf{Z} = \mathbf{I} \tag{93}$$

$$= \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}^H \mathbf{R}^{-1} \mathbf{Z}\right)^{-1}. \tag{94}$$

Then,

$$\mathbf{z}(n) = \mathbf{H}^H \mathbf{x}(n) = \mathbf{D}(n)\mathbf{a} + \mathbf{H}^H \mathbf{e}(n) = \mathbf{D}(n)\mathbf{a} + \mathbf{w}(n), \tag{95}$$

with the lth index being

$$z_l(n) = a_l e^{j\psi_l n} + w_l(n), \tag{96}$$

from which we get the MAFI amplitude estimate as

$$\hat{a}_l = \frac{1}{N - M + 1} \sum_{n=0}^{N-M} z_l(n) e^{-j\psi_l n}. \tag{97}$$

Amplitude Estimation

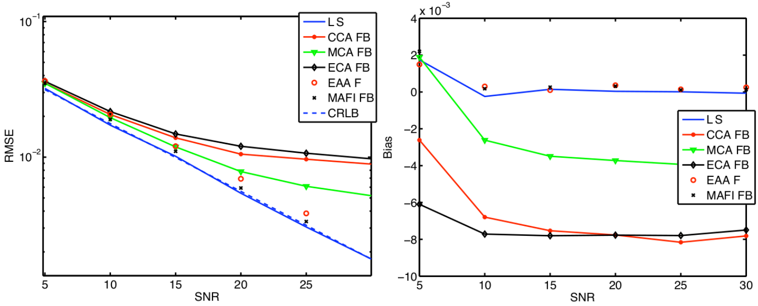


Figure: RMSE (left) and bias (right) of the discussed amplitude estimators as a function of the local SNR for $N = 160$ and $M = 40$.

Amplitude Estimation

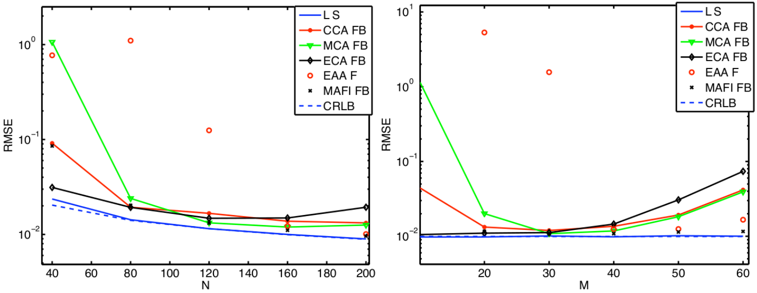


Figure: RMSE of the discussed amplitude estimators as a function of the data length, (with $M = \lfloor N/4 \rfloor$) (left) and filter length (with $N = 160$) (right).

Outline

Posterior Probabilities

Many problems require that the posterior probability be found. These include:

- Determining the model order L
- Choosing between different models
- Finding an optimal segmentation

How can this be done?

Posterior Probabilities

Let $\mathbb{Z}_q = \{0, 1, \ldots, q-1\}$ the model index and $\mathcal{M}_m, m \in \mathbb{Z}_q$ the candidate models.

The posterior probability of a model $\mathcal{M}_m$ can be written as

$$p(\mathcal{M}_m|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M}_m)p(\mathcal{M}_m)}{p(\mathbf{x})}. \tag{98}$$

The principle of MAP-based model selection is to choose the mode as (Djuric 1998)

$$\widehat{\mathcal{M}}_k = \arg \max_{\mathcal{M}_m, m\in\mathbb{Z}_q} p(\mathcal{M}_m|\mathbf{x}) = \arg \max_{\mathcal{M}_m, m\in\mathbb{Z}_q} \frac{p(\mathbf{x}|\mathcal{M}_m)p(\mathcal{M}_m)}{p(\mathbf{x})}. \tag{99}$$

The involved quantities can be computed in different ways, including sampling methods.

Posterior Probabilities

If all the models are equally probable, i.e.,

$$p(\mathcal{M}) = \frac{1}{q} \tag{100}$$

and by noting that $p(\mathbf{x})$ is constant, the MAP model selection criterion reduces to

$$\widehat{\mathcal{M}} = \arg \max_{\mathcal{M}_m, m \in \mathbb{Z}_q} p(\mathbf{x}|\mathcal{M}_m), \tag{101}$$

which is the likelihood function.

The models also depend on $\boldsymbol{\theta}$, so those have to be integrated out, i.e.,

$$p(\mathbf{x}|\mathcal{M}_m) = \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}_m) p(\boldsymbol{\theta}|\mathcal{M}_m) d\boldsymbol{\theta}. \tag{102}$$

Posterior Probabilities

Using Laplace integration, we can write (Djuric 1998)

$$\int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}_m) p(\boldsymbol{\theta}|\mathcal{M}_m) d\boldsymbol{\theta} \\ = \pi^{D/2} \det\left(\widehat{\mathbf{H}}\right)^{-1/2} p(\mathbf{x}|\hat{\boldsymbol{\theta}}, \mathcal{M}_m) p(\hat{\boldsymbol{\theta}}|\mathcal{M}_m), \tag{103}$$

where D is the number of parameters and

$$\widehat{\mathbf{H}} = -\left.\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}_m)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{104}$$

is the Hessian of the log-likelihood function evaluated at $\hat{\boldsymbol{\theta}}$ (i.e., the observed information matrix).

Posterior Probabilities

Taking the logarithm and ignoring constant terms, we get

$$\widehat{\mathcal{M}} = \arg \min_{\mathcal{M}_m, m \in \mathbb{Z}_q} -\underbrace{\ln p(\mathbf{x}|\hat{\theta}, \mathcal{M}_m)}_{\text{log-likelihood}} + \underbrace{\frac{1}{2} \ln \det \left(\widehat{\mathbf{H}} \right)}_{\text{penalty}}, \tag{105}$$

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

Note that $\widehat{\mathbf{H}}$ is the Fischer information matrix evaluated in $\hat{\theta}$.

Using a normalization matrix, $\mathbf{K}$, such that $\mathbf{K}\widehat{\mathbf{H}}\mathbf{K} = \mathcal{O}(1)$, we can write

$$\ln \det \left(\widehat{\mathbf{H}} \right) = \ln \det \left(\mathbf{K}^{-2} \right) + \ln \det \left(\mathbf{K}\widehat{\mathbf{H}}\mathbf{K} \right). \tag{106}$$

Posterior Probabilities

For the harmonic model, we introduce

$$\mathbf{K} = \begin{bmatrix} N^{-3/2} & \mathbf{0} \\ \mathbf{O} & N^{-1/2}\mathbf{I} \end{bmatrix} \tag{107}$$

where $\mathbf{I}$ is an $2L_k \times 2L_k$ identity matrix. From this we obtain

$$\ln\det\left(\widehat{\mathbf{H}}\right) = \ln\det\left(\mathbf{K}^{-2}\right) + \ln\det\left(\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}\right) \tag{108}$$

$$= 3\ln N + 2L\ln N + \mathcal{O}(1). \tag{109}$$

Using this principle, model selection rules can be applied. Different normalization matrices must be found for different models.

Detection

The generalized likelihood ratio test (GLRT) principle (Kay 1993) can easily be adopted for voice activity detection!

Model:

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e} \tag{110}$$

Hypotheses:

$$\mathcal{H}_0 : \mathbf{a} = \mathbf{0} \tag{111}$$

$$\mathcal{H}_1 : \mathbf{a} \neq \mathbf{0} \tag{112}$$

Test statistic:

$$T(\mathbf{x}) = \frac{N-L}{L} \frac{\mathbf{x}^H \mathbf{Z} \left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{x}}{\mathbf{x}^H \left(\mathbf{I} - \mathbf{Z} \left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H\right) \mathbf{x}} \tag{113}$$

Detection

The detection rule is then to choose $\mathcal{H}_1$ when

$$T(\mathbf{x}) > \gamma' \tag{114}$$

and $\mathcal{H}_0$ otherwise. The threshold, γ', is then chosen according to a desired false alarm (FA) rate as

$$P_{\mathrm{FA}} = Q_{F_L,N-L}(\gamma') \tag{115}$$

where $Q_{F_L,N-L}(\cdot)$ is the F distribution with L numerator and N-L denominator degress of freedom.

This is an optimal detector for the harmonic model in white Gaussian noise.

Optimal Segmentation

The problem of joint model selection and optimal segmentation can be be solved using dynamic programming (Prandoni 1997)!

It requires that optimality can be specified in terms of an (additive) cost that can be optimized for independent for each segment.

In a statistical sense, an optimal segmentation should be optimal in terms of the posterior probability. We have just seen how to compute posterior probabilities for different models!

Be aware that since the segment length, N, is now variable, terms including N should now be included!

Optimal Segmentation

Let J_{xy} be the cost (i.e., the posterior probability) of a segment starting at block x and ending at block y.

Costs (white Gaussian noise):

$$\begin{aligned} \text{Chirp} \quad & N \ln \pi + N \ln \hat{\sigma}^2 + N + \tfrac{3}{2} \ln N + \tfrac{5}{2} \ln N + L \ln N \\ \text{Harmonic} \quad & N \ln \pi + N \ln \hat{\sigma}^2 + N + \tfrac{3}{2} \ln N + L \ln N \\ \text{Noise} \quad & N \ln \pi + N \ln \hat{\sigma}^2 + N \end{aligned}$$

where $\hat{\sigma}^2$ is the noise variance estimate for the particular model. The

voiced/unvoiced detection can also be done by use of the generalised likelihood ratio test (GLRT)

Optimal Segmentation

$m = 1$:

$J_1 = J_{11}$

$m = 2$:

$J_1 = J_{12}$

$J_2 = J_{11} + J_{22}$

$m = 3$:

$J_1 = J_{13}$

$J_2 = J_{11} + J_{23}$

$J_3 = J_{12} + J_{33}$

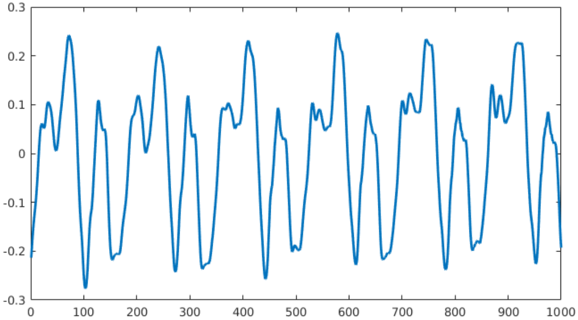
$J_4 = J_{11} + J_{22} + J_{33}$

Conclusion

- As we have seen, it is quite easy to modify the basic model to take more complicated phenomena into account or generalize it.
- We have seen that it can easily be extended to multiple channels for different array geometries.
- It is also fairly easy to incorporate an unvoiced model.
- Colored noise can be accounted for either by modifying the model or via pre-whitening.
- The model can also account for changes in the pitch which results in polynomial instantaneous phase.
- Posterior probabilities can be computed to compare or choose between models/orders and to find the optimal segmentation.

Outline

Periodic Signals



Periodic signals

A periodic signal repeats itself after some period τ or, equivalently with some frequency ω_0.

$$x(n) = x(n - \tau) = x(n - 2\pi/\omega_0) \tag{116}$$

Periodic Signals

Some examples of periodic signals and applications:

- Voiced speech and singing
  - Are people singing on-key?
  - Diagnosis of the Parkinson's disease
- Many musical instruments (e.g., guitar, violin, flute, trumpet, piano)
  - Tuning of instruments
  - Music transcription
- Electrocardiographic (ECG) signals
  - Measure your heart rate
  - Heart defect diagnosis
- Rotating machines
  - Vibration analysis
  - Rotation speed

Outline

The Autocorrelation Method

For a periodic signal $x(n)$ with a period τ, we have that

$$e(n) = x(n) - x(n-\tau) = 0 \tag{117}$$

- Unfortunately, τ is unknown so we have to try out different τ's to find one that satisfies the above equation.
- Real-world signals are not perfectly periodic so we might never find one.
- Instead, the estimate of τ is the value which minimises some objective.

The Autocorrelation Method

Consider the objective

$$G(a,\tau) = \mathbb{E}\left[|e(n)|^2\right] = \mathbb{E}\left[|x(n) - ax(n-\tau)|^2\right] \tag{118}$$

where a allows the amplitude to change. We can rewrite the objective as

$$G(a,\tau) = \sigma_x^2 + a^2\sigma_x^2 - 2ar_x(\tau) \tag{119}$$

where $r_x(\tau) = \mathbb{E}\left[x(n)x(n-\tau)\right]$ is the autocorrelation function. Since the first two terms do not depend on τ, we have that

$$\hat{\tau} = \underset{\tau\in[\tau_{\mathrm{MIN}},\tau_{\mathrm{MAX}}]}{\operatorname{argmax}} \; r_x(\tau) \tag{120}$$

The Autocorrelation Method

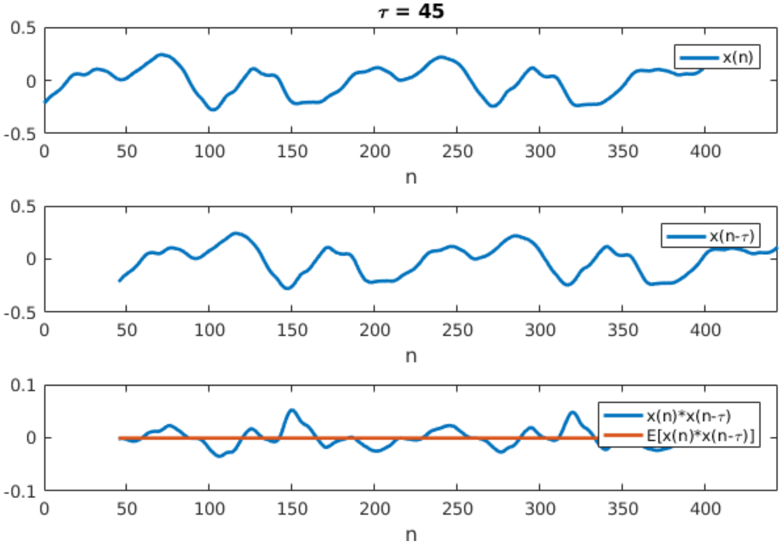
For a segment of data $\{x(n)\}_{n=0}^{N-1}$, we estimate the mean $\mathbb{E}[\cdot]$ as

$$r_x(\tau) = \mathbb{E}\left[x(n)x(n-\tau)\right] \approx \frac{1}{N-\tau}\sum_{n=\tau}^{N-1} x(n)x(n-\tau) \tag{121}$$

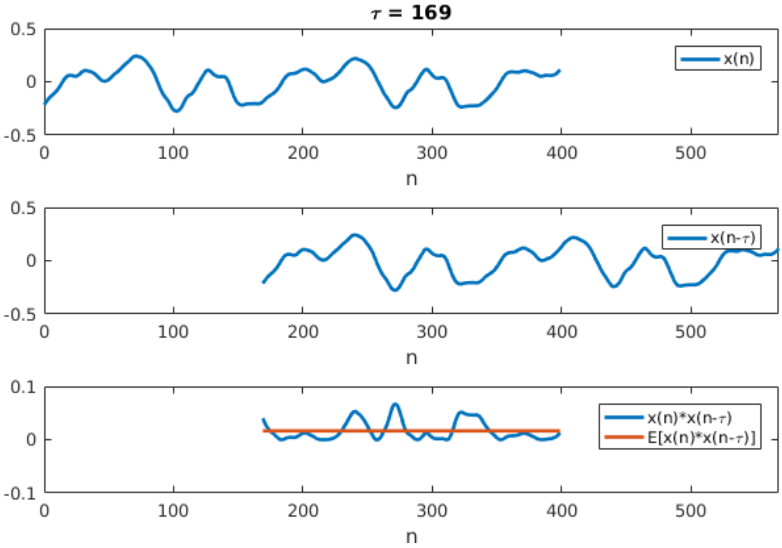
For every $\tau \in [\tau_{\text{MIN}}, \tau_{\text{MAX}}]$, we now do the following:

1. Shift the data by τ samples
2. Trim the ends of $x(n)$ and $x(n-\tau)$ so that all samples overlap.
3. Multiply the trimmed versions $x(n)$ and $x(n-\tau)$ and compute the mean of these products.

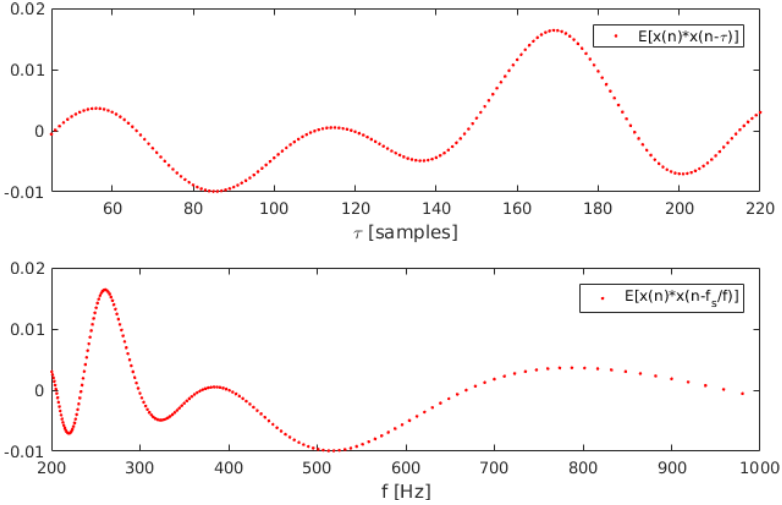
The Autocorrelation Method



The Autocorrelation Method



The Autocorrelation Method



Outline

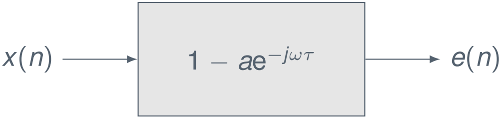
The Comb Filtering Method

In the comb filtering method, we consider the objective

$$J(a,\tau) = \frac{1}{N-\tau}\sum_{n=\tau}^{N-1} |e(n)|^2 \tag{122}$$

for a segment of data $\{x(n)\}_{n=0}^{N-1}$ where

$$e(n) = x(n) - ax(n-\tau) \tag{123}$$



The Comb Filtering Method

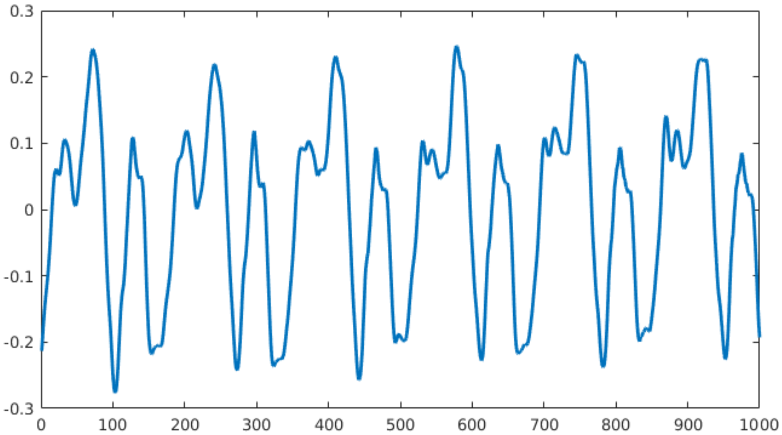
Given τ, the optimal value for a is

$$\hat{a} = \frac{\sum_{n=\tau}^{N-1} x(n)x(n-\tau)}{\sum_{n=\tau}^{N-1} x^2(n-\tau)} \tag{124}$$

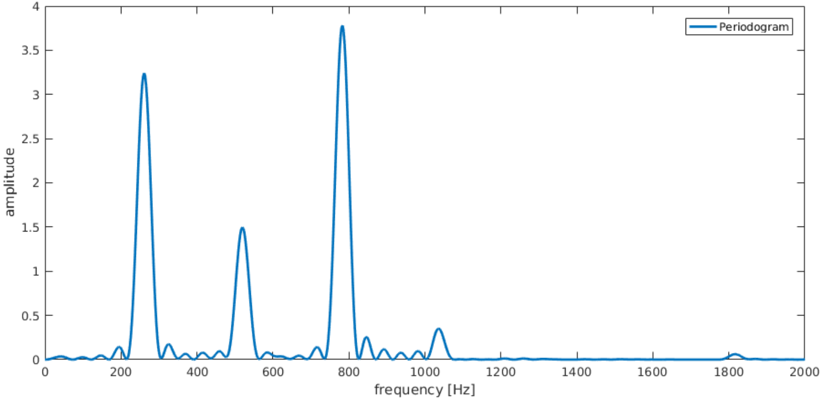
Inserting this into the objective $J(a,\tau)$ yields the estimator

$$\hat{\tau} = \underset{\tau \in [\tau_{\text{MIN}},\tau_{\text{MAX}}]}{\operatorname{argmin}} \frac{1}{N-\tau} \left[\sum_{n=\tau}^{N-1} x^2(n) - \frac{\left[\sum_{n=\tau}^{N-1} x(n)x(n-\tau)\right]^2}{\sum_{n=\tau}^{N-1} x^2(n-\tau)} \right] \tag{125}$$

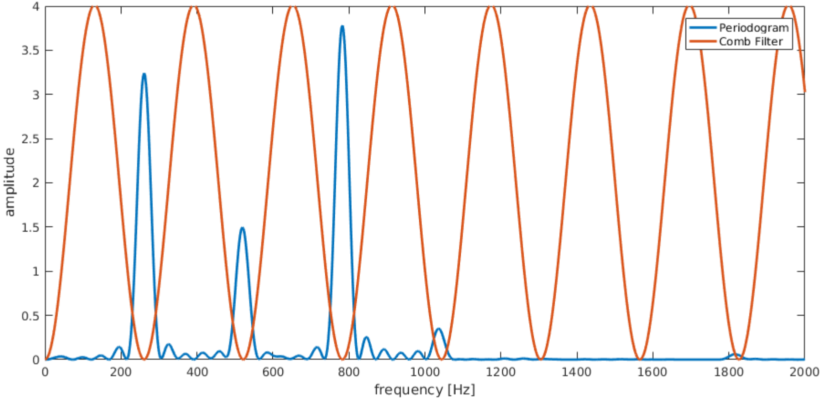
The Comb Filtering Method



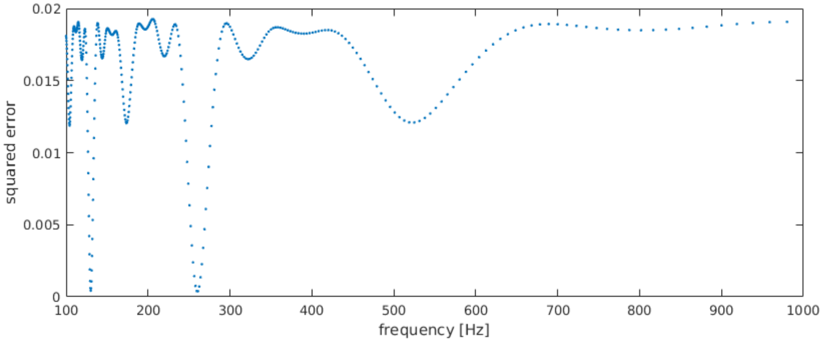
The Comb Filtering Method



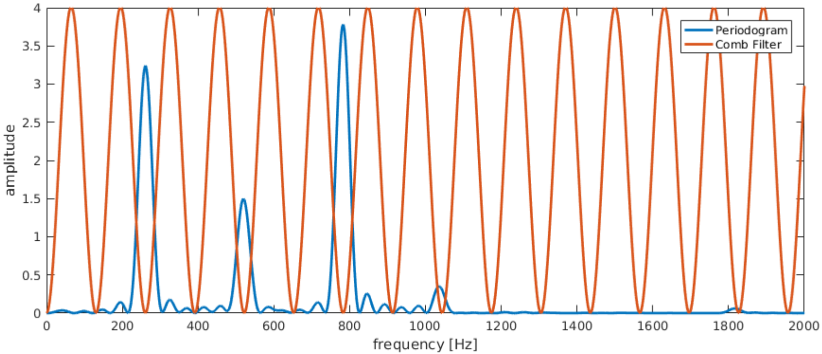
The Comb Filtering Method



The Comb Filtering Method

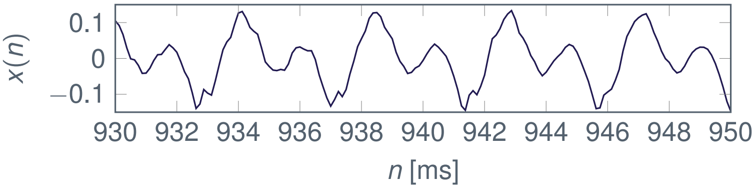


The Comb Filtering Method

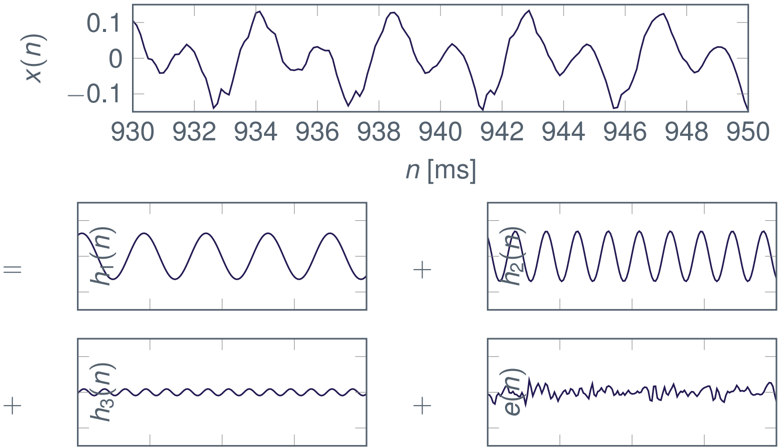


Outline

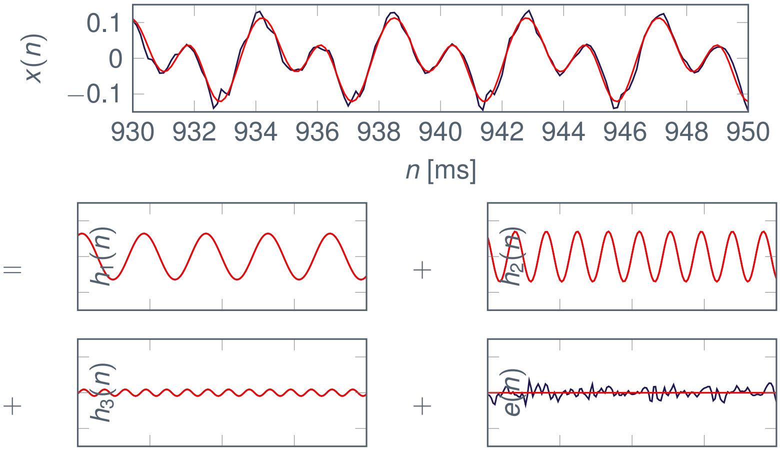
Harmonic Model



Harmonic Model



Harmonic Model



Harmonic Model

Mathematical Model

The signal model for any periodic signal is

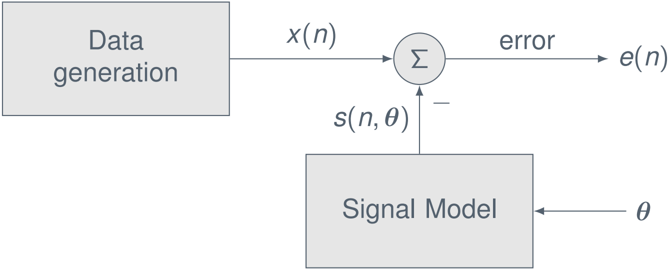
$$s(n) = \sum_{l=1}^{L} h_l(n) = \sum_{l=1}^{L} A_l \cos(\omega_0 l n + \phi_l) \tag{126}$$

where

- A_l real amplitude of the lth harmonic
- ϕ_l phase of the lth harmonic
- ω_0 fundamental frequency in radians/sample
- L the number of harmonics/model order

Method of Least Squares

The method of least-squares



- The vector θ contains the model parameters
- The signal $s(n,\theta)$ is produced by the signal model
- The signal $x(n)$ is the observed data
- The error consists of noise and model inaccuracies

Method of Least Squares

From the figure (on the previous slide), we have that

$$e(n) = x(n) - s(n, \boldsymbol{\theta}) \,, \qquad n = 0, 1, \ldots, N-1 \tag{127}$$

where $s(n, \boldsymbol{\theta})$ is a harmonic model given by

$$s(n, \boldsymbol{\theta}) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l) \tag{128}$$

$$\boldsymbol{\theta} = \begin{bmatrix} A_1 & \cdots & A_L & \phi_1 & \cdots & \phi_L & \omega_0 \end{bmatrix}^T \tag{129}$$

Method of Least Squares

The method of least squares (LS) is that of solving

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, J(\theta) \tag{130}$$

where $J(\theta)$ measures the squared error

$$J(\theta) = \sum_{n=0}^{N-1} |e(n)|^2 \tag{131}$$

Solving this problem is very computationally demanding since the fundamental frequency is a nonlinear parameter.

Outline

The Harmonic Summation Method

From Parseval's theorem, we have that

$$\lim_{N\to\infty} \sum_{n=0}^{N-1} |e(n)|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(\omega)|^2 d\omega \tag{132}$$

where

$$E(\omega) = X(\omega) - 2\pi \sum_{l=1}^{L} [\alpha_l \delta(\omega - \omega_0 l) + \alpha_l^* \delta(\omega + \omega_0 l)] \tag{133}$$

$$\alpha_l = A_l \exp(j\phi_l)/2 \tag{134}$$

The Harmonic Summation Method

Given ω_0, the optimal value for α_l is

$$\hat{\alpha}_l = \frac{1}{2\pi} X(\omega_0 l) \tag{135}$$

Inserting this into the error $E(\omega)$ yields the objective

$$H(\omega_0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(\omega)|^2 d\omega - \frac{2}{2\pi} \sum_{l=1}^{L} |X(\omega_0 l)|^2 \tag{136}$$

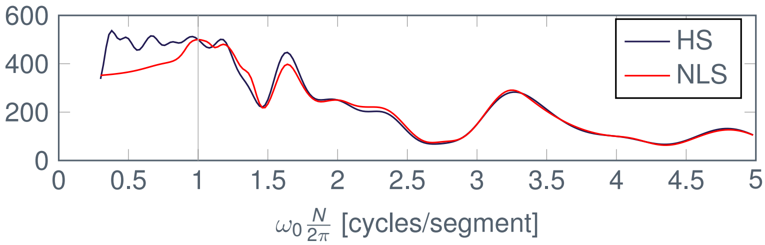
The harmonic summation (HS) estimator is

$$\hat{\omega}_0 = \underset{\omega_0 \in [\omega_{\mathrm{MIN}}, \omega_{\mathrm{MAX}}]}{\operatorname{argmax}} \sum_{l=1}^{L} |X(\omega_0 l)|^2 \tag{137}$$

The Harmonic Summation Method

Some remarks:

- The HS method can be implemented very efficiently using a single FFT.
- The HS method is an approximate NLS method and is, therefore, often referred to as aNLS.
- The HS method works very well, unless the fundamental frequency is low.



Outline

NLS Estimator

The harmonic model

$$x(n) = \sum_{l=1}^{L} \left[a_l \cos(l\omega_0 n) - b_l \sin(l\omega_0 n) \right] + e(n) \tag{138}$$

for $n = n_0, n_0 + 1, \ldots, n_0 + N - 1$ can be written as

$$\boldsymbol{x} = \boldsymbol{Z}_L(\omega_0)\alpha_L + \boldsymbol{e}\,. \tag{139}$$

where

$$\boldsymbol{Z}_L(\omega) = \begin{bmatrix} \boldsymbol{c}(\omega) & \boldsymbol{c}(2\omega) & \cdots & \boldsymbol{c}(L\omega) & \boldsymbol{s}(\omega) & \boldsymbol{s}(2\omega) & \cdots & \boldsymbol{s}(L\omega) \end{bmatrix}$$

$$\boldsymbol{c}(\omega) = \begin{bmatrix} \cos(\omega n_0) & \cdots & \cos(\omega(n_0 + N - 1)) \end{bmatrix}^T$$

$$\boldsymbol{s}(\omega) = \begin{bmatrix} \sin(\omega n_0) & \cdots & \sin(\omega(n_0 + N - 1)) \end{bmatrix}^T$$

$$\alpha_l = \begin{bmatrix} \boldsymbol{a}_L^T & -\boldsymbol{b}_L^T \end{bmatrix}^T, \ \boldsymbol{a}_L = \begin{bmatrix} a_1 & \cdots & a_L \end{bmatrix}^T, \ \boldsymbol{b}_L = \begin{bmatrix} b_1 & \cdots & b_L \end{bmatrix}^T$$

NLS Estimator

The least squares error is

$$\sum_{n=0}^{N-1} e^2(n) = \mathbf{e}^T \mathbf{e} = [\mathbf{x} - \mathbf{Z}_L(\omega_0)\boldsymbol{\alpha}_L]^T [\mathbf{x} - \mathbf{Z}_L(\omega_0)\boldsymbol{\alpha}_L] \tag{140}$$

Given ω_0, the estimate of $\boldsymbol{\alpha}_L$ is

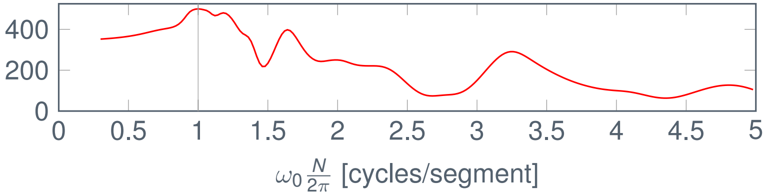
$$\hat{\boldsymbol{\alpha}}_L = \left[\mathbf{Z}_L^T(\omega_0)\mathbf{Z}_L(\omega_0)\right]^{-1} \mathbf{Z}_L^T(\omega_0)\mathbf{x} \tag{141}$$

Inserting this back into the objective yields the NLS estimator

$$\hat{\omega}_{0,L} = \underset{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]}{\operatorname{argmax}} \; \mathbf{x}^T \mathbf{Z}_L(\omega_0) \left[\mathbf{Z}_L^T(\omega_0)\mathbf{Z}_L(\omega_0)\right]^{-1} \mathbf{Z}_L^T(\omega_0)\mathbf{x} \tag{142}$$

The NLS estimator has been known since Quinn and Thomson (1991), but is costly to compute.

NLS Estimator

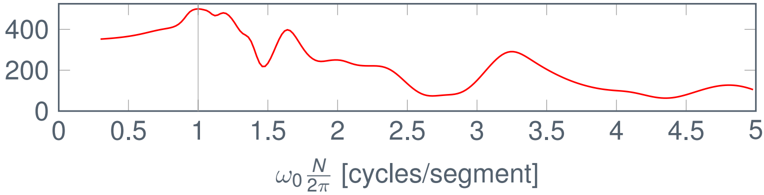


1. Compute NLS cost function

$$\hat{\omega}_{0,L} = \underset{\omega_0 \in [\omega_{\mathrm{MIN}}, \omega_{\mathrm{MAX}}]}{\operatorname{argmax}} \boldsymbol{x}^T \boldsymbol{Z}_L(\omega_0) \left[\boldsymbol{Z}_L^T(\omega_0)\boldsymbol{Z}_L(\omega_0)\right]^{-1} \boldsymbol{Z}_L^T(\omega_0)\boldsymbol{x} \quad (143)$$

on an F/L-point uniform grid for all model orders $L \in \{1, \dots, L_{\mathrm{MAX}}\}$.

NLS Estimator

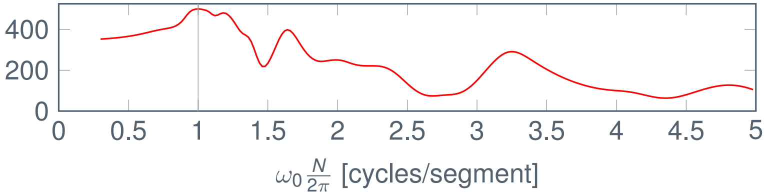


1. Compute NLS cost function

$$\hat{\omega}_{0,L} = \underset{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]}{\text{argmax}} \; \boldsymbol{x}^T \boldsymbol{Z}_L(\omega_0) \left[\boldsymbol{Z}_L^T(\omega_0) \boldsymbol{Z}_L(\omega_0) \right]^{-1} \boldsymbol{Z}_L^T(\omega_0) \boldsymbol{x} \quad (143)$$

   on an F/L-point uniform grid for all model orders $L \in \{1, \dots, L_{\text{MAX}}\}$.
2. Optionally refine the L_{MAX} grid estimates.

NLS Estimator



1. Compute NLS cost function

$$\hat{\omega}_{0,L} = \underset{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]}{\text{argmax}} \; \boldsymbol{x}^T \boldsymbol{Z}_L(\omega_0) \left[\boldsymbol{Z}_L^T(\omega_0) \boldsymbol{Z}_L(\omega_0) \right]^{-1} \boldsymbol{Z}_L^T(\omega_0) \boldsymbol{x} \quad (143)$$

   on an F/L-point uniform grid for all model orders $L \in \{1, \dots, L_{\text{MAX}}\}$.
2. Optionally refine the L_{MAX} grid estimates.
3. Do model comparison.

NLS Estimator

Nonlinear least squares (NLS) estimator:

$$\hat{\omega}_{0,L} = \underset{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]}{\text{argmax}} \; \boldsymbol{x}^T \boldsymbol{Z}_L(\omega_0) \left[\boldsymbol{Z}_L^T(\omega_0) \boldsymbol{Z}_L(\omega_0) \right]^{-1} \boldsymbol{Z}_L^T(\omega_0) \boldsymbol{x} \tag{144}$$

Harmonic summation (HS) estimator:

$$\hat{\omega}_{0,L} = \underset{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]}{\text{argmax}} \; \boldsymbol{x}^T \boldsymbol{Z}_L(\omega_0) \boldsymbol{Z}_L^T(\omega_0) \boldsymbol{x} \tag{145}$$

Complexities

Order of complexity of step 1. (on previous slide)

NLS $\mathcal{O}(F \log F) + \mathcal{O}(F L_{\text{MAX}}^3)$

HS $\mathcal{O}(F \log F) + \mathcal{O}(F L_{\text{MAX}})$

We have recently decreased the complexity of NLS to that of HS.

Fast Nonlinear Least Squares Estimator

Fast NLS Algorithm

A MATLAB implementation

```
% create an estimator object (the data independent step is computed)
f0Estimator = fastF0Nls(nData, maxNoHarmonics, f0Bounds);
% analyse a segment of data
[f0Estimate, estimatedNoHarmonics, estimatedLinParam] = ...
    f0Estimator.estimate(data);
```

Fast Nonlinear Least Squares Estimator
Fast NLS Algorithm

A MATLAB implementation

```
% create an estimator object (the data independent step is computed)
f0Estimator = fastF0Nls(nData, maxNoHarmonics, f0Bounds);
% analyse a segment of data
[f0Estimate, estimatedNoHarmonics, estimatedLinParam] = ...
    f0Estimator.estimate(data);
```

- The algorithm includes model comparison in a Bayesian framework using numerical integration (see our paper *Default Bayesian Estimation of the Fundamental Frequency, T-ASLP, 2013* for more details)
- The algorithm also includes refinement of the grid-estimates. Can be controlled using an optional user-parameter.
- The algorithm can also be set-up to work for a model with a non-zero DC-value.
- Can be downloaded from `http://tinyurl.com/fastF0Nls`.

Fast Nonlinear Least Squares Estimator

Fast NLS Algorithm

A MATLAB implementation (example)

```
% load the mono speech signal
[speechSignal, samplingFreq] = audioread('roy.wav');
nData = length(speechSignal);

% set up
segmentTime = 0.025; % seconds
segmentLength = round(segmentTime*samplingFreq); % samples
nSegments = floor(nData/segmentLength);
f0Bounds = [80, 400]/samplingFreq; % cycles/sample
maxNoHarmonics = 15;
f0Estimator = fastF0Nls(segmentLength, maxNoHarmonics, f0Bounds);

% do the analysis
idx = 1:segmentLength;
f0Estimates = nan(1,nSegments); % cycles/sample
for ii = 1:nSegments
    speechSegment = speechSignal(idx);
    f0Estimates(ii) = f0Estimator.estimate(speechSegment);
    idx = idx + segmentLength;
end
```

Outline

Comparison of Methods

What could be evaluated?

1. Estimation accuracy
2. Computational complexity
3. Robustness to noise
4. Time-frequency resolution

How to evaluate a pitch estimator?

- Pitch detection, tracking, or estimation?
- Synthetic signals vs. real speech data
- Component vs. system evaluation
- White vs. coloured noise.

Comparison of Methods

How can we hope to solve the complex problems if we cannot solve the simple ones?

1. Analyse each component individually - not only the entire system
2. Quantify the performance on synthetic signals - not only on speech signals

Outline

Comparison of Methods
Estimation Accuracy

Cramér-Rao Lower Bound (CRLB)

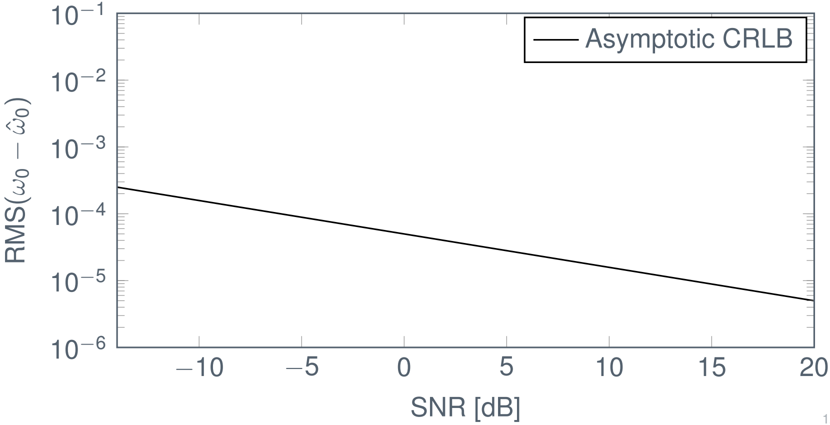
- Lower bound on the variance of any unbiased estimator.
- Does not depend on the data - only the model structure (the likelihood function) and the model parameters.
- If the variance of an estimator attains the bound, it is statistically optimal.
- The bound tells us how the performance can be expected to depend on various quantities.
- The bound can be used as a benchmark in simulations.

Asymptotic CRLB for the pitch in WGN

$$\mathrm{var}(\hat{\omega}_0) \geq \frac{24\sigma^2}{N(N^2-1)\sum_{l=1}^{L} A_l^2 l^2} \tag{146}$$

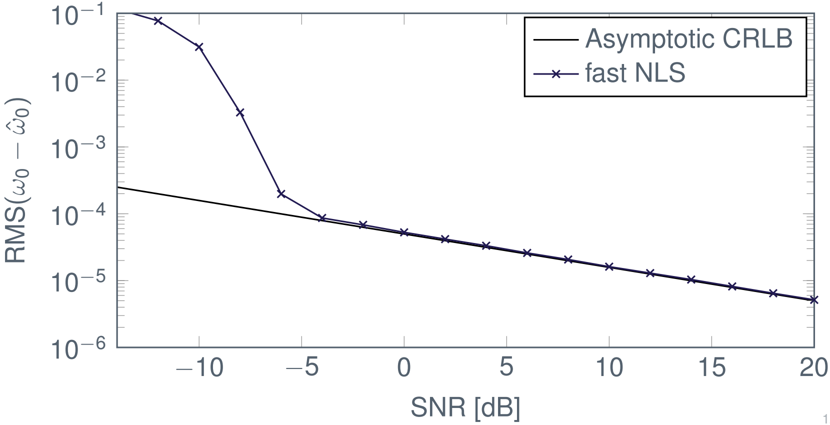
Comparison of Methods
Estimation Accuracy

Setup: $N = 500$, $L = 10$, 1000 repetitions, random, but not low, fundamental frequency, random phases, and constant amplitudes.



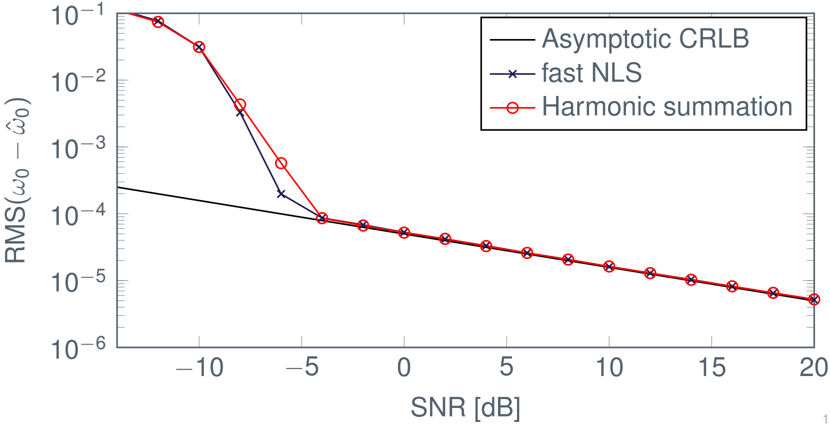
Comparison of Methods
Estimation Accuracy

Setup: $N = 500$, $L = 10$, 1000 repetitions, random, but not low, fundamental frequency, random phases, and constant amplitudes.



Comparison of Methods
Estimation Accuracy

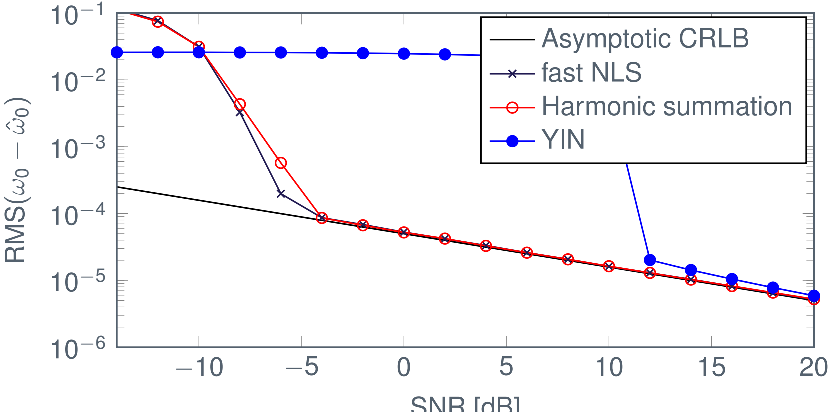
Setup: $N = 500$, $L = 10$, 1000 repetitions, random, but not low, fundamental frequency, random phases, and constant amplitudes.



Comparison of Methods

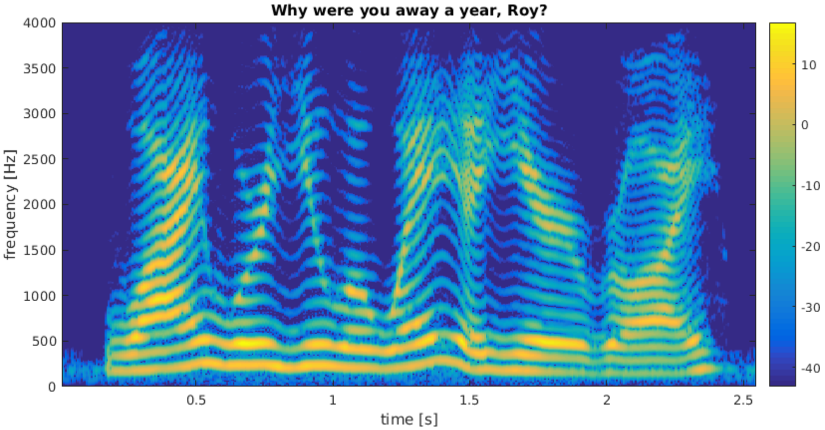
Estimation Accuracy

Setup: $N = 500$, $L = 10$, 1000 repetitions, random, but not low, fundamental frequency, random phases, and constant amplitudes.



Outline

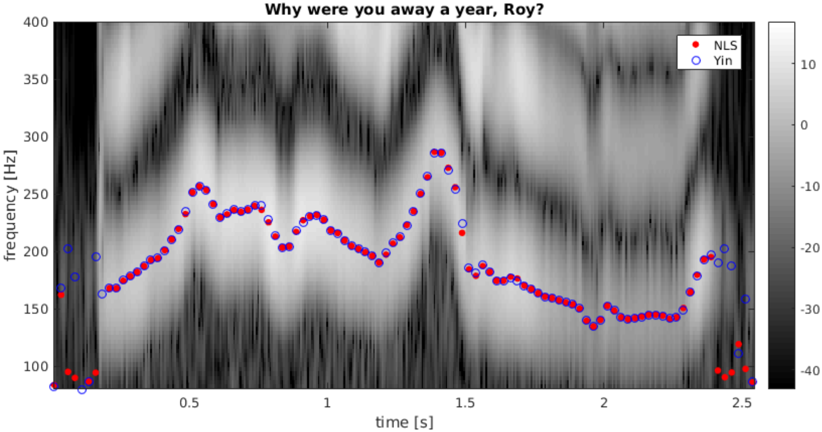
Comparison of Methods
Robustness to noise



Comparison of Methods

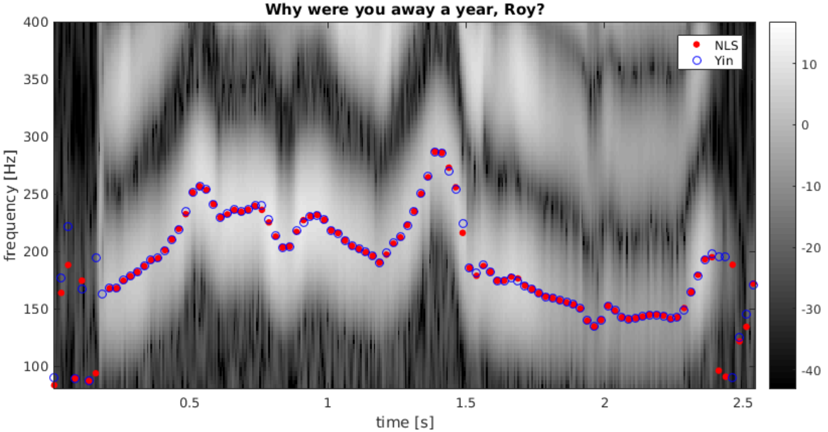
Robustness to noise

No noise and window size of 25 ms.



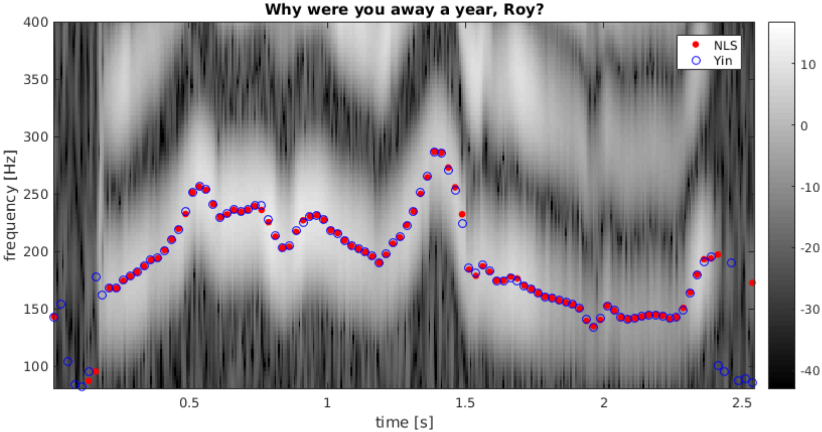
Comparison of Methods
Robustness to noise

30 dB SNR and window size of 25 ms.



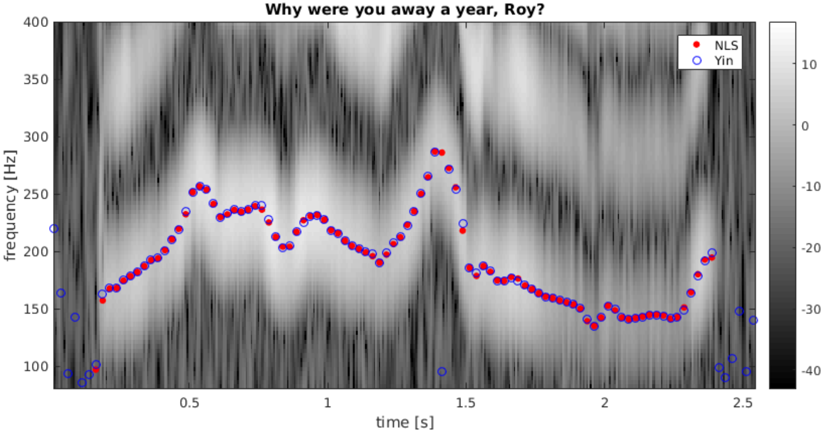
Comparison of Methods
Robustness to noise

20 dB SNR and window size of 25 ms.



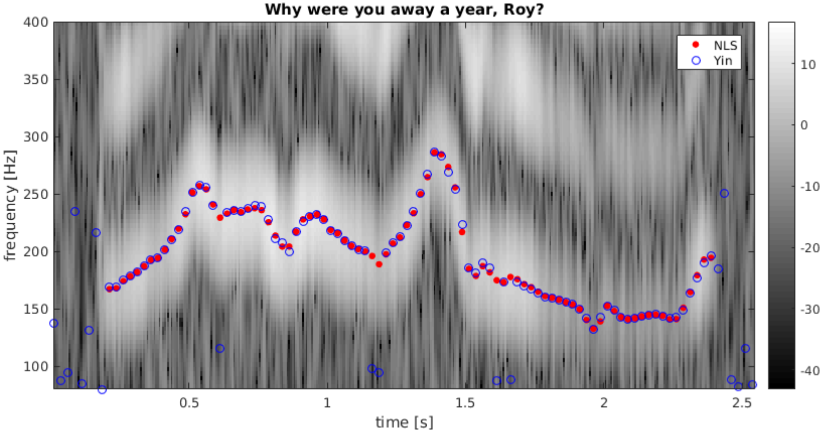
Comparison of Methods
Robustness to noise

15 dB SNR and window size of 25 ms.



Comparison of Methods
Robustness to noise

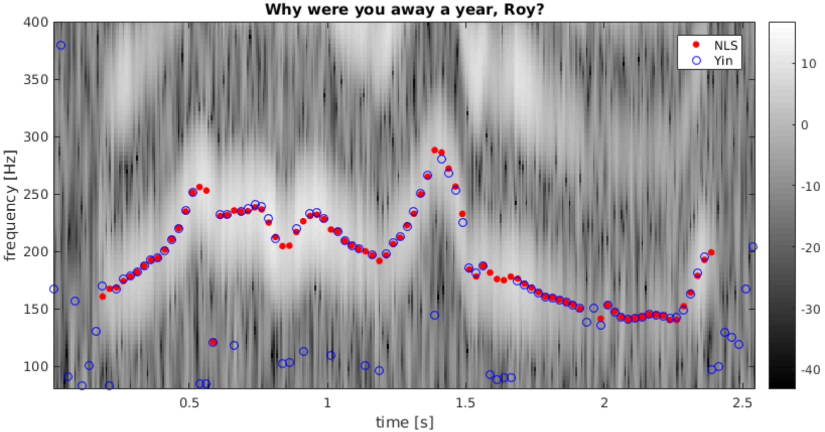
10 dB SNR and window size of 25 ms.



Comparison of Methods

Robustness to noise

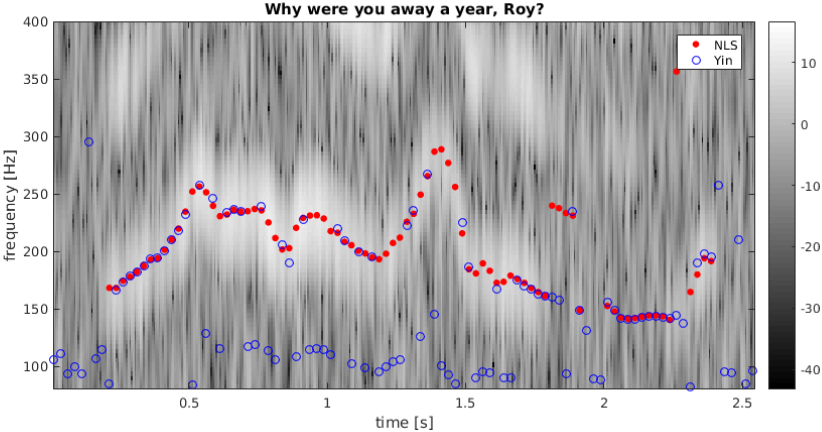
5 dB SNR and window size of 25 ms.



Comparison of Methods

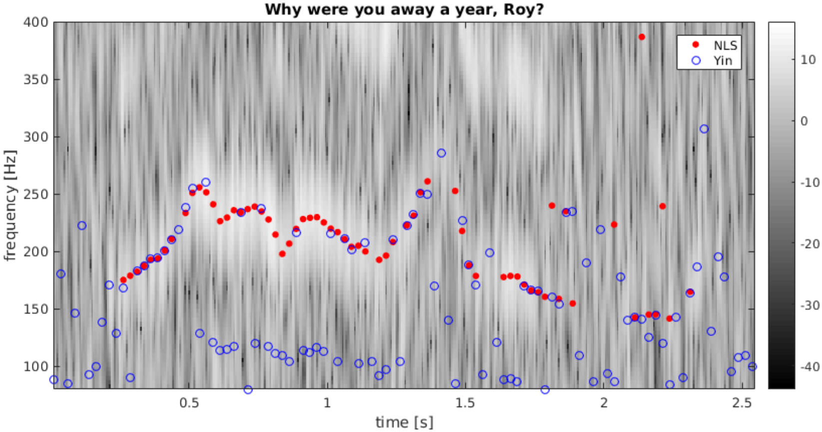
Robustness to noise

0 dB SNR and window size of 25 ms.



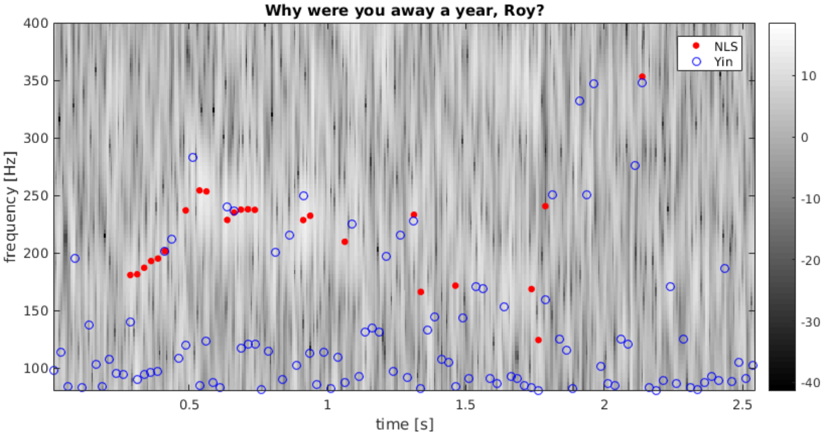
Comparison of Methods
Robustness to noise

-5 dB SNR and window size of 25 ms.



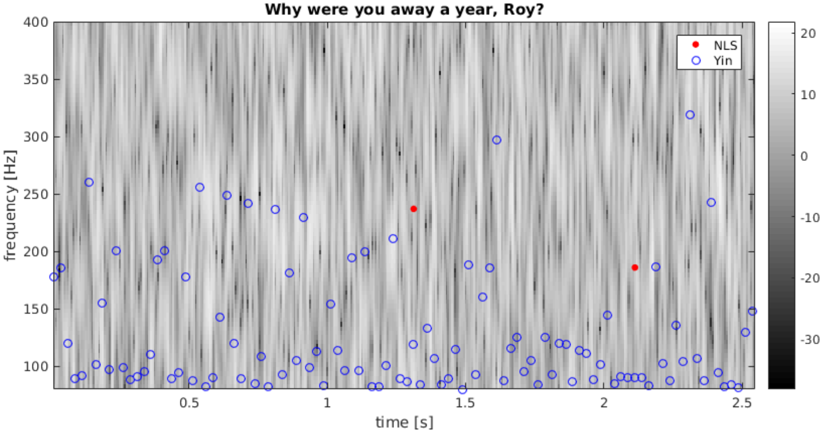
Comparison of Methods
Robustness to noise

-10 dB SNR and window size of 25 ms.



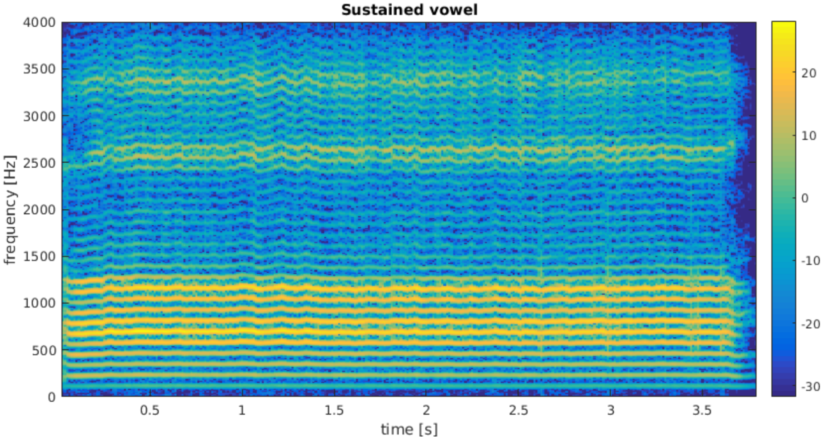
Comparison of Methods
Robustness to noise

-15 dB SNR and window size of 25 ms.



Outline

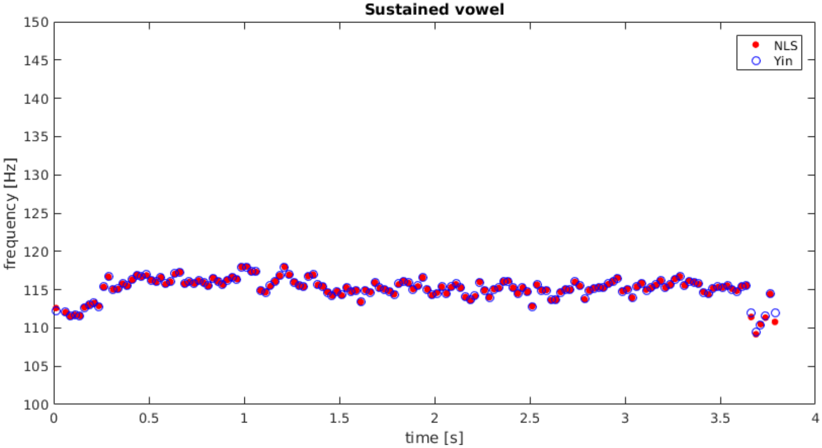
Comparison of Methods
Time-frequency resolution



Comparison of Methods

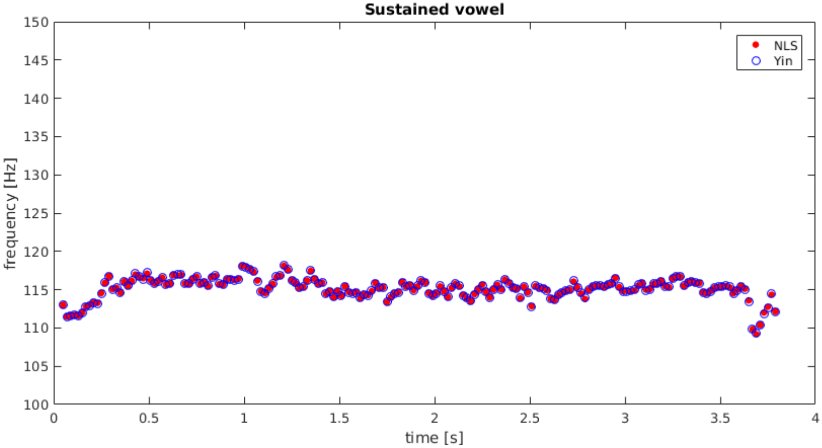
Time-frequency resolution

Window size of 25 ms and no noise.



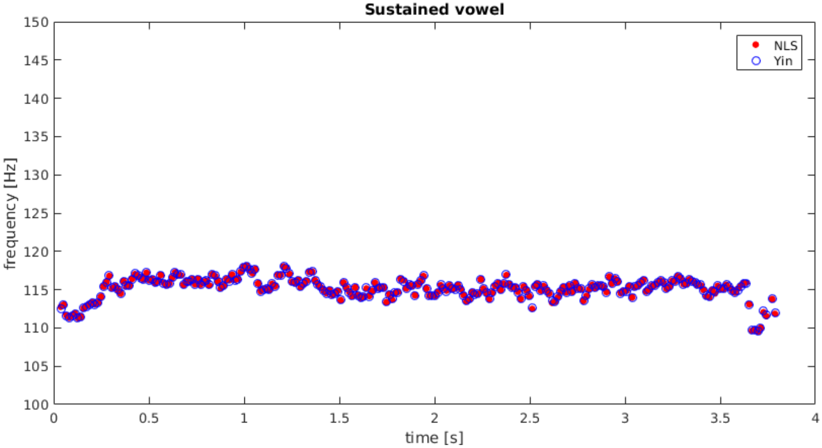
Comparison of Methods
Time-frequency resolution

Window size of 20 ms and no noise.



Comparison of Methods
Time-frequency resolution

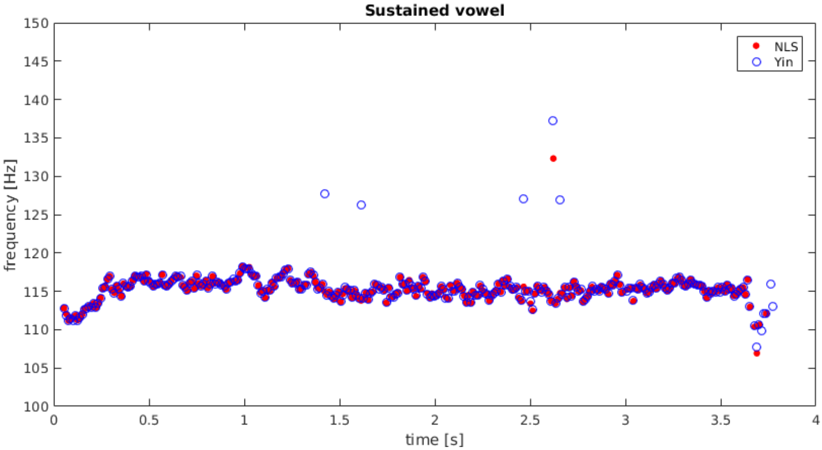
Window size of 15 ms and no noise.



Comparison of Methods

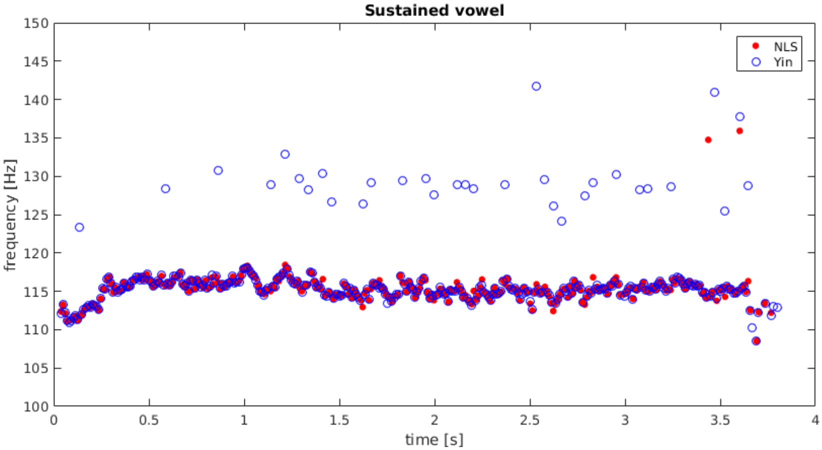
Time-frequency resolution

Window size of 12 ms and no noise.



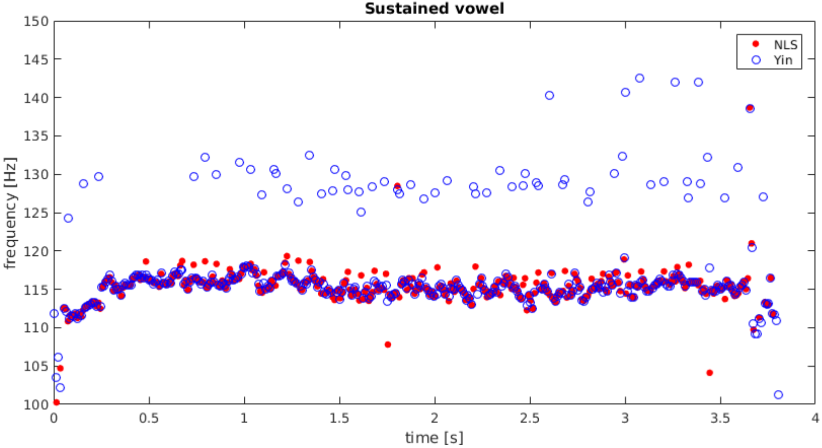
Comparison of Methods
Time-frequency resolution

Window size of 11 ms and no noise.



Comparison of Methods
Time-frequency resolution

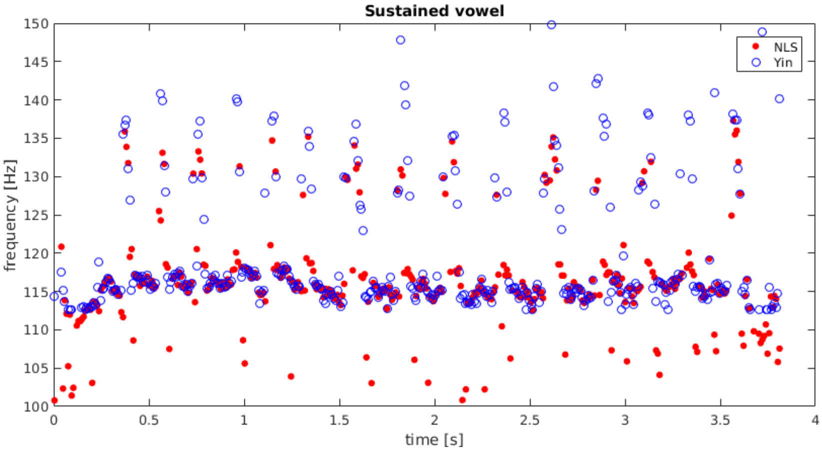
Window size of 10 ms and no noise.



Comparison of Methods

Time-frequency resolution

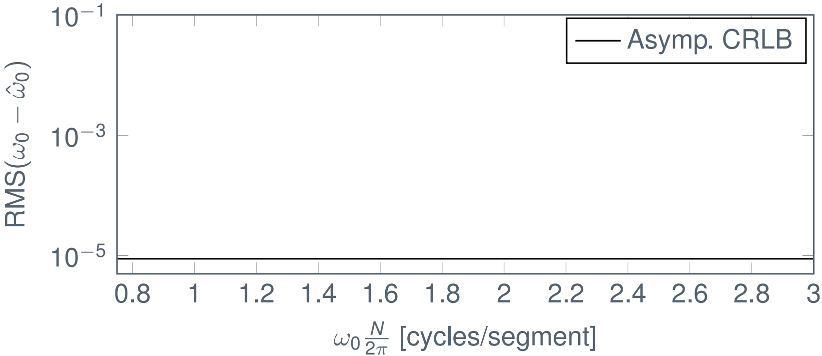
Window size of 9 ms and no noise.



Comparison of Methods

Time-frequency resolution

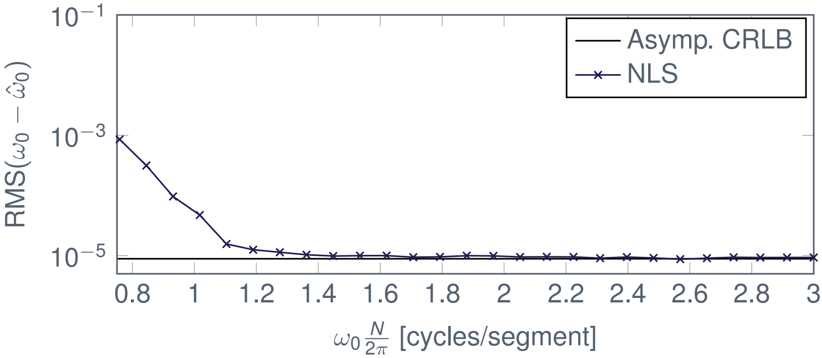
Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 15 dB



Comparison of Methods

Time-frequency resolution

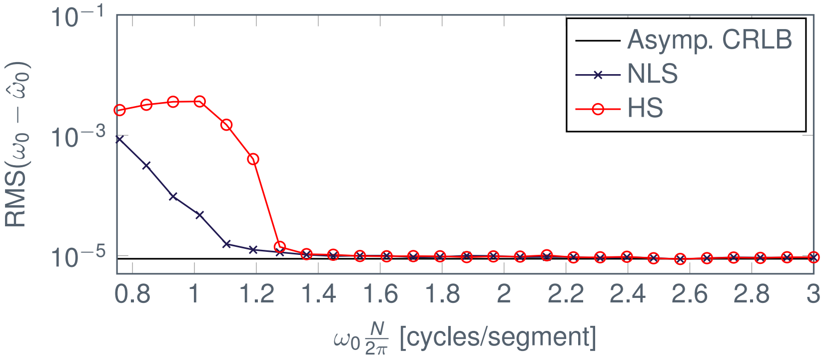
Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 15 dB



Comparison of Methods

Time-frequency resolution

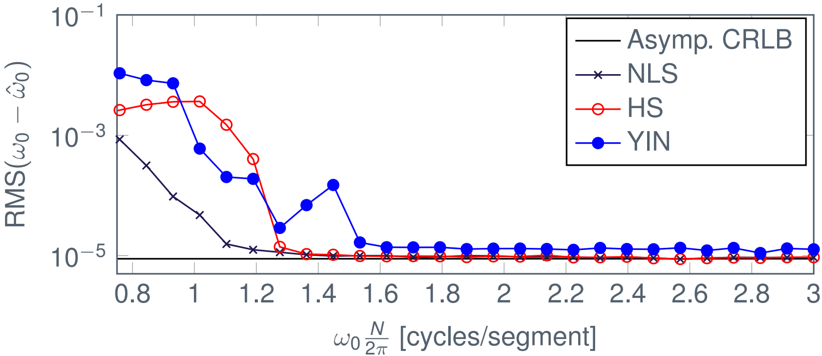
Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 15 dB



Comparison of Methods

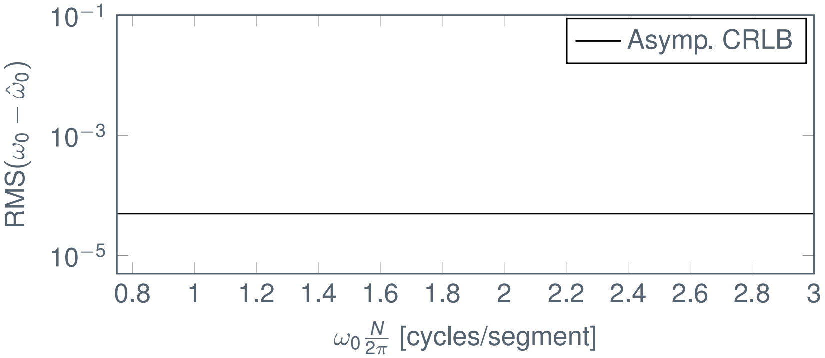
Time-frequency resolution

Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 15 dB



Comparison of Methods
Time-frequency resolution

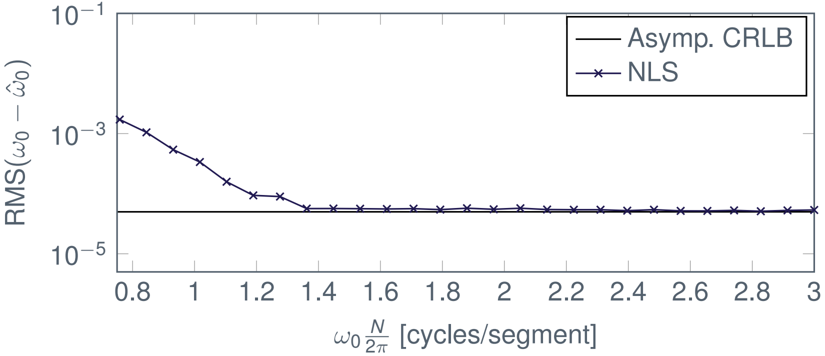
Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 0 dB



Comparison of Methods

Time-frequency resolution

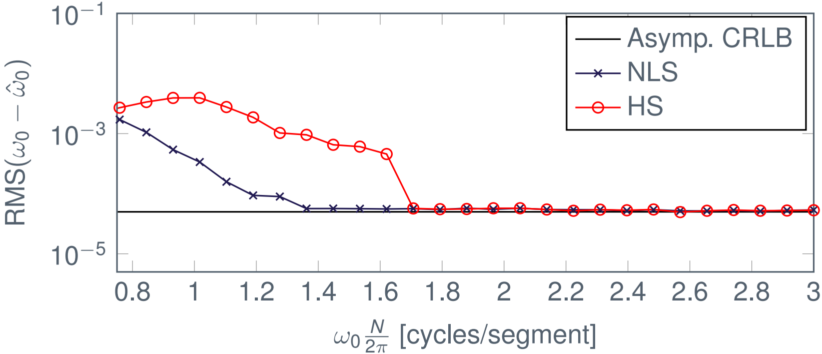
Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 0 dB



Comparison of Methods

Time-frequency resolution

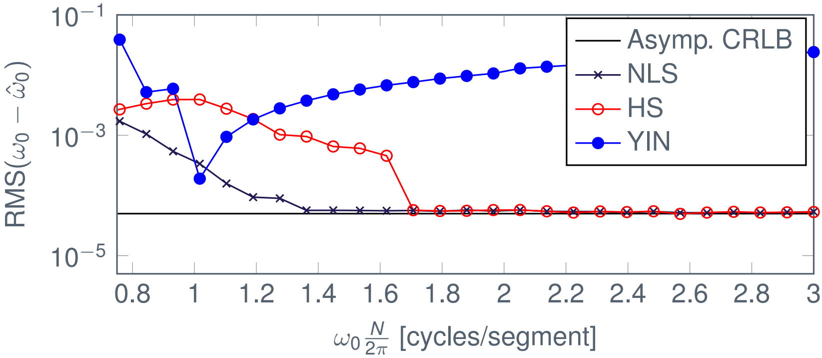
Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 0 dB



Comparison of Methods

Time-frequency resolution

Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 0 dB



Outline

Comparison of Methods
Summary

Correlation-based Methods
A periodic signal satisfies that

$$x(n) = x(n - \tau) \tag{147}$$

where $\tau = 2\pi/\omega_0$ is the period.

+ Intuitive and simple
+ Low computational complexity

+/- No need to estimate the model order

- Poor time-frequency resolution
- Are very sensitive to noise
- Interpolation needed for fractional delay estimation

Comparison of Methods
Summary

Parametric Methods
Estimate the parameters in

$$x(n) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l) + e(n) \tag{148}$$

- \+ High estimation accuracy
- \+ Work very well in even noisy conditions
- \+ Good time-frequency resolution
- +/- The model order has to be estimated
- \- Might produce over-optimistic results
- \- High computational complexity

Outline

Non-stationary Pitch Estimation

- Speech is non-stationary since the fundamental frequency is continuously changing.
- The harmonic model assumes that the the fundamental frequency is constant in a segment of data
- We can extend the model of the phase of the lth harmonic component to
$$\theta_l(n) \approx \phi_l + l\omega_0 n + l\beta_0 n^2/2 \tag{149}$$
where β_0 is the fundamental chirp rate.
- We refer to this model as the harmonic chirp model

$$s(n) = \sum_{l=1}^{L} A_l \cos(l\beta_0 n^2/2 + l\omega_0 n + \phi_l) \tag{150}$$

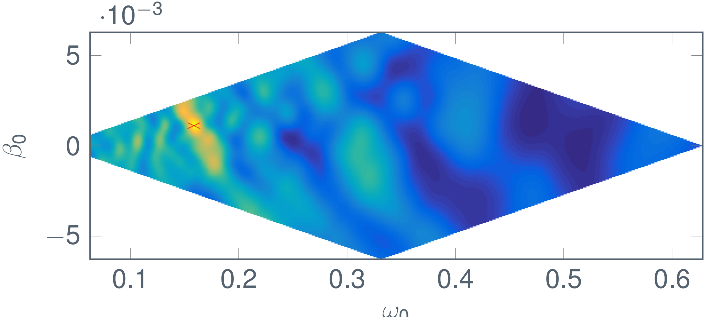
Non-stationary Pitch Estimation

Nonlinear least squares (NLS) objective

$$J_L(\omega_0, \beta_0) = \boldsymbol{x}^T \boldsymbol{Z}_L(\omega_0, \beta_0) \left[\boldsymbol{Z}_L^T(\omega_0, \beta_0) \boldsymbol{Z}_L(\omega_0, \beta_0) \right]^{-1} \boldsymbol{Z}_L^T(\omega_0, \beta_0) \boldsymbol{x} \quad (151)$$

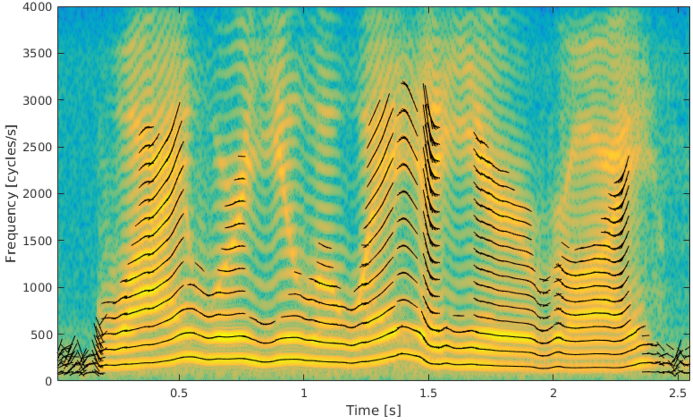
Harmonic chirp summation objective:

$$J_L(\omega_0, \beta_0) = \boldsymbol{x}^T \boldsymbol{Z}_L(\omega_0, \beta_0) \boldsymbol{Z}_L^T(\omega_0, \beta_0) \boldsymbol{x} \quad (152)$$



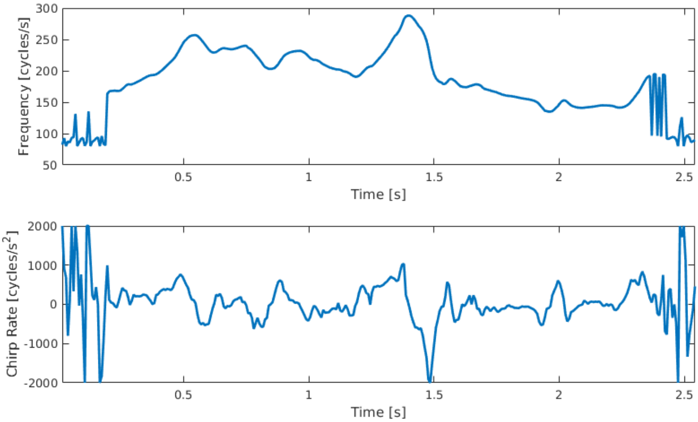
Non-stationary Pitch Estimation

Window size of 30 ms, 75 % overlap, and no noise



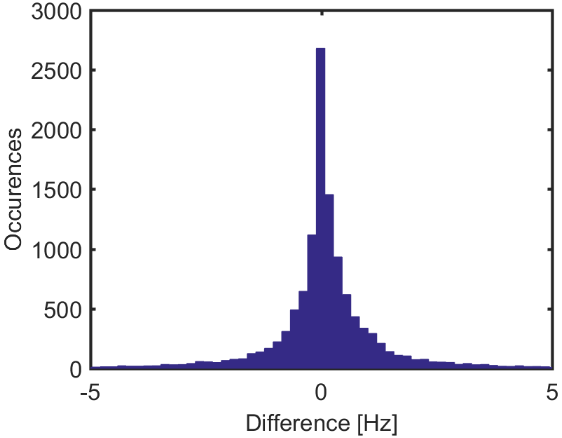
Non-stationary Pitch Estimation

Window size of 30 ms, 75 % overlap, and no noise



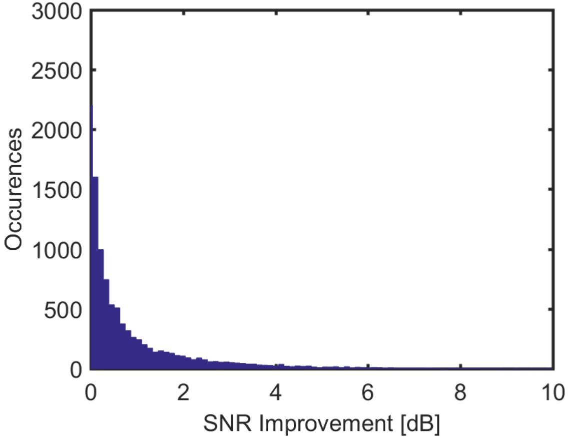
Non-stationary Pitch Estimation

Window size of 30 ms and no noise

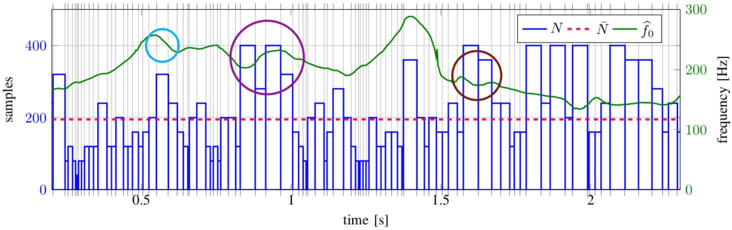


Non-stationary Pitch Estimation

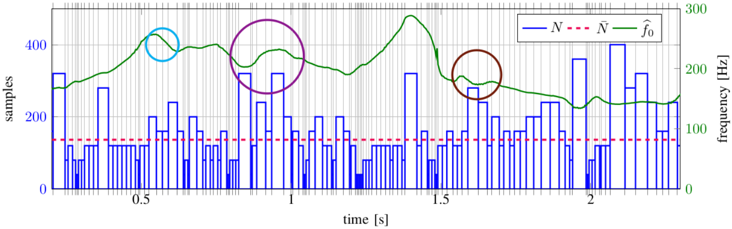
Window size of 30 ms and no noise



Non-stationary Pitch Estimation



(a) Harmonic chirp model



(b) Traditional harmonic model

Outline

Multi-channel pitch estimation

- In, e.g., a hearing aid, we might have K channels.
- For every channel, we use the harmonic model and obtain

$$x_k(n) = \sum_{l=1}^{L} A_{l,k} \cos(l\omega_0 n + \phi_{l,k}) + e_k(n) \tag{153}$$

- If we assume the same noise variance in every channel, we obtain the NLS objective

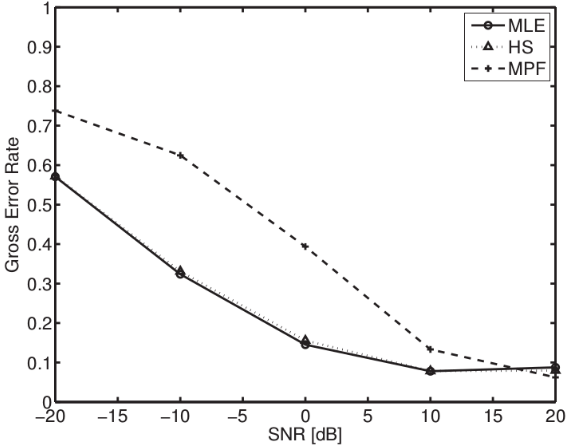
$$J_L(\omega_0) = \sum_{k=1}^{K} \boldsymbol{x}_k^T \boldsymbol{Z}_L(\omega_0) \left[\boldsymbol{Z}_L^T(\omega_0)\boldsymbol{Z}_L(\omega_0)\right]^{-1} \boldsymbol{Z}_L^T(\omega_0)\boldsymbol{x}_k$$

- If we assume independent noise variances in every channel, we obtain the NLS objective

$$J_L(\omega_0) = \sum_{k=1}^{K} \ln\left\{\boldsymbol{x}_k^T\boldsymbol{x}_k - \boldsymbol{x}_k^T \boldsymbol{Z}_L(\omega_0) \left[\boldsymbol{Z}_L^T(\omega_0)\boldsymbol{Z}_L(\omega_0)\right]^{-1} \boldsymbol{Z}_L^T(\omega_0)\boldsymbol{x}_k\right\}$$

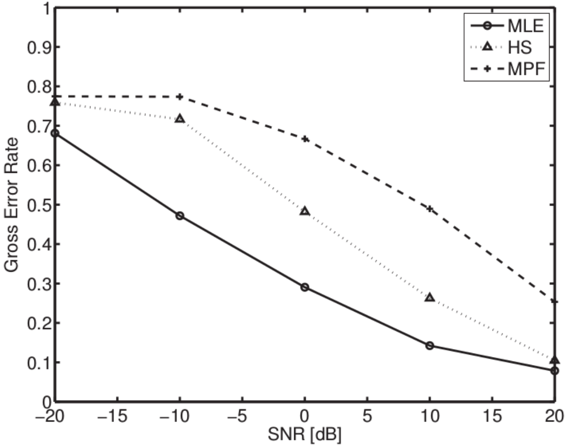
Multi-channel pitch estimation

Same noise variance on every channel.



Multi-channel pitch estimation

Different noise variances on every channel.



Outline

Summary

- Parametric pitch estimation methods typically outperform non-parametric methods in terms of estimation accuracy, noise robustness, and time-frequency resolution.
- However, parametric method are still more computationally costly.
- The modelling assumptions are explicit in parametric methods.
- Consequently, we can easily extend the model to take more complex phenomena into account.

Outline

Outline

Traditional TDOA Estimation

- Time-difference of arrival (TDOA) estimation important in many microphone array applications:
  - array calibration
  - room geometry estimation
  - noise reduction
  - source localization, etc.
- Traditionally, solved using cross-correlation method.
- Can be shown to be special case of more general method.
- This method can yield better estimation accuracy.

Fundamental Models

- Problems with traditional methods become clear from model.
- Time domain model of 2 microphone recordings:

$$\begin{aligned} x_1(n) &= s(n) + e_1(n), \\ x_2(n) &= \beta s(n-\eta) + e_2(n), \quad n = 0, \ldots, N-1. \end{aligned} \tag{154}$$

- For periodic signal ($\omega_0 = m2\pi/N$), (154) equals frequency domain model:

$$\begin{aligned} X_1(k) &= S(k) + E_1(k), \\ X_2(k) &= \beta S(k) e^{-j2\pi k\eta/N} + E_2(k). \end{aligned} \tag{155}$$

Problems with Model

- Frequency domain often too restrictive:
  - Source signal often periodic on short time-scale, but ω_0 assumption not satisfied.
  - Leads to edge effects.
  - Can be reduced through zero-padding, but colors noise spectrum.
- Frequency domain model aren't suited for fractional TDOA estimation:
  - Fractional delay corresponds to a complex valued sensor signal.

Improvements

- Can be improved using a periodic model without $\omega_0 = m2\pi/N$ assumption.
- Leads to more general model, having traditional one as special case.
- Reveals conditions where cross-correlation method is statistically efficient.
- A maximum likelihood estimator for joint fundamental frequency and TDOA estimation is formed based on the model.
- Yields fractional delay estimates without interpolation.

Periodic model

Assuming periodic signal:

$$s(n) = \sum_{k=1}^{L} A_k \cos(\omega_0 kn + \phi_k) = \sum_{k=-L}^{L} \alpha_k e^{j\omega_0 kn}, \tag{156}$$

with

A_k/α_k: real/complex amplitude ($A_k > 0$, $\alpha_k = \frac{A_k}{2} e^{j\phi_k}$),
ϕ_k: phase ($\phi_k \in [-\pi, \pi[$),
ω_0: fundamental frequency.

Signal delay by η gives

$$s(n-\eta) = \sum_{k=-L}^{L} \alpha_k e^{j\omega_0 kn} e^{-j\omega_0 \eta k} = \sum_{k=-L}^{L} \alpha_k e^{j\omega_0 kn} e^{-j\xi k}. \tag{157}$$

Matrix-vector model

Model in matrix-vector notation:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{Z}(\omega_0) \\ \beta\mathbf{Z}(\omega_0)\mathbf{D}(\xi) \end{bmatrix} \boldsymbol{\alpha} + \mathbf{e} = \mathbf{H}(\beta, \xi, \omega_0)\boldsymbol{\alpha} + \mathbf{e} \tag{158}$$

with

$$\begin{aligned}
\mathbf{z}(\omega) &= \begin{bmatrix} 1 & e^{j\omega} & \cdots & e^{j\omega(N-1)} \end{bmatrix}^T, \\
\mathbf{Z}(\omega_0) &= [\mathbf{z}(-L\omega_0) \ \cdots \ \mathbf{z}(-\omega_0) \ \mathbf{z}(\omega_0) \ \mathbf{z}(L\omega_0)], \\
\mathbf{D}(\xi) &= \mathrm{diag}\left(e^{jL\xi}, \ldots, e^{j\xi}, e^{-j\xi}, \ldots, e^{-jL\xi}\right), \\
\boldsymbol{\alpha} &= [\alpha_{-L} \ \cdots \ \alpha_{-1} \ \alpha_1 \ \cdots \ \alpha_L]^T, \\
\mathbf{H}(\beta, \xi, \omega_0) &= \begin{bmatrix} \mathbf{Z}(\omega_0) \\ \beta\mathbf{Z}(\omega_0)\mathbf{D}(\xi) \end{bmatrix},
\end{aligned}$$

e: white Gaussian with pdf $\mathcal{N}(\mathbf{H}(\beta, \xi, \omega_0)\boldsymbol{\alpha}, \sigma^2\mathbf{I}_{2N})$.

Maximum Likelihood Estimator

ML estimates obtained using non-linear least squares. Solving for linear parameters first yields:

$$(\widehat{\beta}, \widehat{\xi}, \widehat{\omega}_0) = \arg\max_{\beta,\xi,\omega_0} J(\beta, \xi, \omega_0), \tag{159}$$

with

$$J(\beta, \eta, \omega_0) = \mathbf{x}^H \mathbf{H} \left(\mathbf{H}^H \mathbf{H}\right)^{-1} \mathbf{H}^H \mathbf{x}.$$

Computationally complex due to non-convexity $\rightarrow$ 3D search required.

Complexity can be reduced through approximations.

Approximate ML Method

For large N we have

$$\mathbf{Z}^H(\omega_0)\mathbf{Z}(\omega_0) \approx N\mathbf{I}_{2L}. \tag{160}$$

Approximation exact for $N \to \infty$. Then,

$$\mathbf{H}^H(\beta, \xi, \omega_0)\mathbf{H}(\beta, \xi, \omega_0) \approx (1 + \beta^2)N\mathbf{I}_{2L}, \tag{161}$$

and as a result:

$$J(\beta, \xi, \omega_0) = \frac{1}{N(1+\beta^2)}\Big[\mathbf{x}_1^H\mathbf{Z}(\omega_0)\mathbf{Z}^H(\omega_0)\mathbf{x}_1 + \beta^2\mathbf{x}_2\mathbf{Z}(\omega_0)\mathbf{Z}^H(\omega_0)\mathbf{x}_2 \\ +2\beta\mathbf{x}_1^H\mathbf{Z}(\omega_0)\mathbf{D}^*(\xi)\mathbf{Z}^H(\omega_0)\mathbf{x}_2\Big]. \tag{162}$$

Important Special Case

For $\omega_0 = 2\pi/N$ and $L = \lceil N/2 \rceil - 1$ the large sample approx. is exact.

Cost function for $\eta = \xi/\omega_0$ thus doesn't depend on β, so:

$$J(\eta) = \mathbf{x}_1^H \mathbf{Z}(2\pi/N)\mathbf{D}^*(2\pi\eta/N)\mathbf{Z}^H(2\pi/N)\mathbf{x}_2 \tag{163}$$

$$= \sum_{k=-\lceil N/2 \rceil+1}^{\lceil N/2 \rceil-1} X_1^*(k)X_2(k)e^{j2\pi k\eta/N}. \tag{164}$$

For N being even:

$$J(\eta) = \sum_{k=0}^{N-1} X_1^*(k)X_2(k)e^{j2\pi k\eta/N}. \tag{165}$$

This resembles the cross-correlation (CC) TDOA estimator!

Cross-Correlation TDOA Estimator

Thus, the CC TDOA estimator is statistically efficient when:

1. Source signal periodic with zero-mean.
2. Fundamental frequency of source signal is $2\pi/N$.
3. Number of harmonics of the source signal is $\lceil N/2 \rceil - 1$.
4. Delay is integer valued.

Fractional TDOA Estimation

Assume no noise and η_0 being true delay, then:

$$X_2(k) = X_1(k)e^{-j2\pi k\eta_0/N}, \quad k = 0, \dots, N-1. \tag{166}$$

Inserting in cross-correlation cost function gives complex value for fractional delays.

Traditionally, solved using interpolation, fractional delay filters, or fraction Fourier transform.

Problem avoided by using:

$$J(\eta) = \sum_{k=-\lceil N/2\rceil+1}^{\lceil N/2\rceil-1} X_1^*(k)X_2(k)e^{j2\pi k\eta/N}. \tag{167}$$

Experiments
Setup

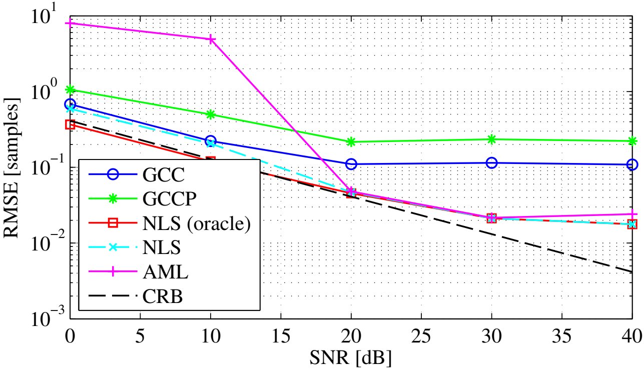
Synthetic data experiments:

- signal 1: harmonic signal with $\omega_0 \sim \mathcal{U}(0.1, 0.15)$, $L = 5$, unit amp. harmonics with random phase,
- signal 2: white Gaussian noise (N-periodic), i.e., $\omega_0 = 2\pi/N$, $L = N/2 - 1$,
- $N = 100$, $f_s = 8$ kHz.

Speech data experiments

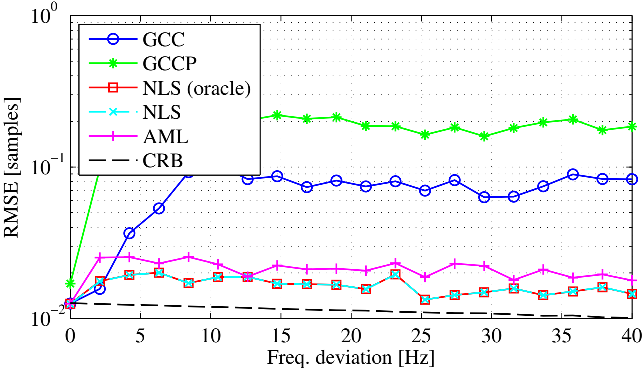
- $\sim$2.2 s of female speech (mainly voiced),
- stereo recording made using RIR generator,
- $\eta = 0$ samples, no reverb.,
- $N = 100$, $f_s = 8$ kHz.

Experiments
Results on synthetic data

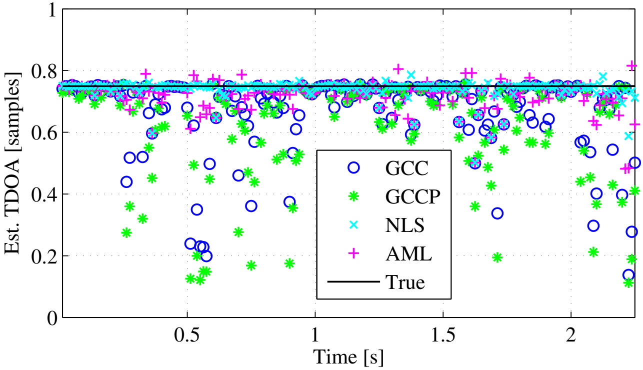


Experiments

Results on synthetic data



Experiments
Results on speech data



Outline

DOA and Pitch Estimation

- DOA estimation possible with $K \geq 2$ microphones.
- Has applications in beamforming, autonomous steering, surveillance, etc.
- In speech applications, traditional DOA estimators are based on generic broadband model.
- Examples of such methods are: steered response power, TDOA-based, and subspace-based.
- More accurate estimates obtainable by exploiting a more accurate signal model.
- Periodic signal model can be used for, e.g., short voiced speech segments.

Why Model-based DOA Estimation?

- For periodic signal, joint pitch and DOA estimation can have significant advantages.
- In multi-source scenarios, sources are better resolvable, especially, with overlapping parameters.
- Another strategy is to: 1) estimate DOA, 2) extract signal from DOA, and 3) estimate pitch from extracted signal.
- Corresponds to transformation, which likely increases Cramér-Rao bound (CRB).
- Taking pitch structure into account, decreases CRB.
- Using multiple microphones, pitch estimation CRB is decreased.

Signal Model

Recording N samples of sound source in noise using K microphones:

$$y_k(n) = \beta_k s(n - f_s \tau_k) + e_k(n) \tag{168}$$
$$= x_k(n) + e_k(n), \tag{169}$$

with

- β_k: attenuation of source to mic k,
- $s(n)$: periodic signal to be localized,
- f_s: sampling frequency,
- τ_k: delay of source at mic k,
- $e_k(n)$: additive noise (background noise, sensor noise, etc.).

Signal Model

We choose reference point for which:

$$s_{\mathrm{ref}}(n) = \beta_{\mathrm{ref}} s(n - f_{\mathrm{s}}\tau_{\mathrm{ref}}). \tag{170}$$

With this, and inv. sq. for sound propagation:

$$x_k(n) = \frac{r_{\mathrm{ref}}}{r_k} s_{\mathrm{ref}}(n - f_{\mathrm{s}}\tau_{\mathrm{ref},k}) + e_k(n), \tag{171}$$

$$= \frac{r_{\mathrm{ref}}}{r_k} s_{\mathrm{ref}}\left(n - f_{\mathrm{s}}\frac{r_k - r_{\mathrm{ref}}}{c}\right) + e_k(n), \tag{172}$$

where

$\tau_{\mathrm{ref,k}}$: TDOA of source between mic 0 and k.

Complex Periodic Signal Model

Clean desired signal modeled as

$$s_{\mathrm{ref}}(n) = \sum_{l=1}^{L} \gamma_l e^{jl\omega_0 n}, \tag{173}$$

with

$$\begin{aligned} \gamma_l &= \beta_{\mathrm{ref}}\alpha_l, \\ \alpha_l&\text{: complex amplitude of } l\text{'th harmonic,} \\ \omega_0&\text{: fundamental frequency [rad/sample],} \\ L&\text{: model order, i.e., number of harmonics.} \end{aligned}$$

Uniform Linear Array Model

With array center as reference points, law of cosines dictates that the range from source to mic k is:

$$r_k(r_c, \theta) = \sqrt{g_k^2 d^2 + r_c^2 - 2 g_k d r_c \sin\theta} \tag{174}$$

with

$$g_k = \frac{K-1}{2} - k + 1,$$

r_c: source-to-array-center distance (SAD).

Then,

$$x_k(n) = \frac{r_c}{r_k} \sum_{l=1}^{L} \gamma_l e^{jl\omega_0 n} e^{-jf_s l\omega_0 \frac{r_k - r_c}{c}} + e_k(n). \tag{175}$$

Matrix-Vector Model

Consider vector of N samples from mic k:

$$\begin{aligned} \mathbf{x}_k &= \begin{bmatrix} x_k(0) & x_k(1) & \cdots & x_k(N-1) \end{bmatrix}^T, \\ &= \mathbf{Z}(\omega_0)\mathbf{D}_k(r_k)\gamma + \mathbf{e}_k, \end{aligned} \tag{176}$$

where

$$\begin{aligned} \mathbf{Z}(\omega_0) &= [\mathbf{z}(\omega_0)\ \mathbf{z}(2\omega_0)\ \cdots\ \mathbf{z}(L\omega_0)], \\ \mathbf{z}(\omega) &= [1\ e^{j\omega}\ \cdots\ e^{j(N-1)\omega}]^T, \\ [\mathbf{D}_k(r_k)]_{pq} &= \begin{cases} \frac{r_c}{r_k} e^{-jf_s p\omega_0 \frac{r_k - r_c}{c}}, & p = q, \\ 0, & \text{otherwise}, \end{cases} \\ \gamma &= [\gamma_1\ \gamma_2\ \cdots\ \gamma_L]^T. \end{aligned}$$

Likelihood Function

Likelihood function useful for finding optimal estimators and CRB's.

Assuming WGN which is not correlated across mics:

$$\mathcal{L} = \ln p(\{\mathbf{x}_k\}; \boldsymbol{\nu}) = -N\left(K \ln \pi + \sum_{k=0}^{K-1} \ln \sigma_k^2\right) - \sum_{k=0}^{K-1} \frac{\|\mathbf{e}_k\|^2}{\sigma_k^2}, \tag{177}$$

with

$\boldsymbol{\nu}$: vector containing unknown parameters of interest,
σ_k^2: variance of noise at mic k.

Asymptotic CRBs

In far-field, the following asymptotic bounds ($N \to \infty$) can be found:

$$\mathrm{CRB}(\omega_0) \approx \frac{6}{N^3 K}\mathrm{PSNR}^{-1}, \tag{178}$$

$$\mathrm{CRB}(\theta) \approx \left[\left(\frac{c}{\omega_0 f_\mathrm{s} d \cos\theta}\right)^2 \frac{6}{NK^3} + \left(\frac{\tan\theta}{\omega_0}\right)^2 \frac{6}{N^3 K}\right]\mathrm{PSNR}^{-1},$$

$$\mathrm{PSNR} = \frac{\sum_{l=1}^{L} l^2 A_l^2}{\sigma^2}. \tag{179}$$

Oberservations

- ω_0 CRB decreases with both N and K but independent on θ,
- θ CRB decreases with increasing ω_0, N and K.
- both ω_0 and θ CRBs decreases by exploiting harmonic structure.

Maximum Likelihood Estimators

First, closed-form solutions for γ and σ_k^2 's minimizing $\mathcal{L}$ can be found:

$$\widehat{\gamma} = \left(\sum_{k=0}^{K-1} \frac{D_k^H \mathbf{Z}^H \mathbf{Z} \mathbf{D}_k}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{\mathbf{D}_k^H \mathbf{Z}^H \mathbf{x}_k}{\sigma_k^2}, \tag{180}$$

$$\widehat{\sigma}_k^2 = \frac{\|\mathbf{x}_k - \mathbf{Z}\mathbf{D}_k \gamma\|^2}{N}. \tag{181}$$

Estimates depend on each other $\rightarrow$ estimated iteratively!

Resulting estimator after inserting closed-form solutions:

$$\{\widehat{\omega_0}, \widehat{r}_c, \widehat{\theta}\} = \arg\min \sum_{k=0}^{K-1} \ln \|\mathbf{x}_k - \mathbf{Z}\mathbf{D}_k \widehat{\gamma}\|^2. \tag{182}$$

Approximate ML Estimator

For large sample sizes, it holds that

$$\lim_{N \to \infty} \frac{1}{N} \mathbf{Z}^H \mathbf{Z} = \mathbf{I}. \tag{183}$$

With this approximation:

$$\hat{\gamma} = \left(\sum_{k=0}^{K-1} \frac{r_c^2}{r_k^2} \frac{N}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{\mathbf{D}_k \mathbf{Z}^H \mathbf{x}_k}{\sigma_k^2}. \tag{184}$$

Main computational complexity is $\mathbf{Z}^H \mathbf{x}_k$, but replaceable with FFT.

ML Estimator

Special case: equal noise levels

With the same noise level at each mic:

$$\widehat{\gamma} = \left(\sum_{k=0}^{K-1} \mathbf{D}_k^H \mathbf{Z}^H \mathbf{Z} \mathbf{D}_k \right)^{-1} \sum_{k=0}^{K-1} \mathbf{D}_k^H \mathbf{Z}^H \mathbf{x}_k. \tag{185}$$

With the large sample approximation, it reduces to

$$\widehat{\gamma} = \left(\sum_{k=0}^{K-1} \frac{r_c^2}{r_k^2} N \right)^{-1} \sum_{k=0}^{K-1} \mathbf{D}_k \mathbf{Z}^H \mathbf{x}_k. \tag{186}$$

ML Estimator

Special case: far-field scenarios

In far-field, following approximations hold

$$\frac{r_{\mathrm{c}}}{r_k} \approx 1, \quad \text{and} \quad \tau_{\mathrm{c},k} \approx g_k \frac{d \sin\theta}{c}. \tag{187}$$

Amplitude and noise estimates are then:

$$\widehat{\boldsymbol{\gamma}} = \left(\sum_{k=0}^{K-1} \frac{\widetilde{\mathbf{D}}_k^H \mathbf{Z}^H \mathbf{Z} \widetilde{\mathbf{D}}_k}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{\widetilde{\mathbf{D}}_k^H \mathbf{Z}^H \mathbf{x}_k}{\sigma_k^2}, \tag{188}$$

$$\widehat{\sigma}_k^2 = \frac{\|\mathbf{x}_k - \mathbf{Z}\widetilde{\mathbf{D}}_k \boldsymbol{\gamma}\|}{N}, \tag{189}$$

with

$$[\widetilde{\mathbf{D}}_k]_{pq} = \begin{cases} e^{-jf_{\mathrm{s}} p \omega_0 \tau_{\mathrm{c},k}}, & \text{for } p = q, \\ 0, & \text{otherwise.} \end{cases}$$

ML Estimator

Special case: far-field scenarios

Far-field assumption can be combined with equal noise variance assumption:

$$\widehat{\gamma} = \left(\sum_{k=0}^{K-1} \widetilde{\mathbf{D}}_k^H \mathbf{Z}^H \mathbf{Z} \widetilde{\mathbf{D}}_k \right)^{-1} \sum_{k=0}^{K-1} \widetilde{\mathbf{D}}_k^H \mathbf{Z}^H \mathbf{x}_k. \tag{190}$$

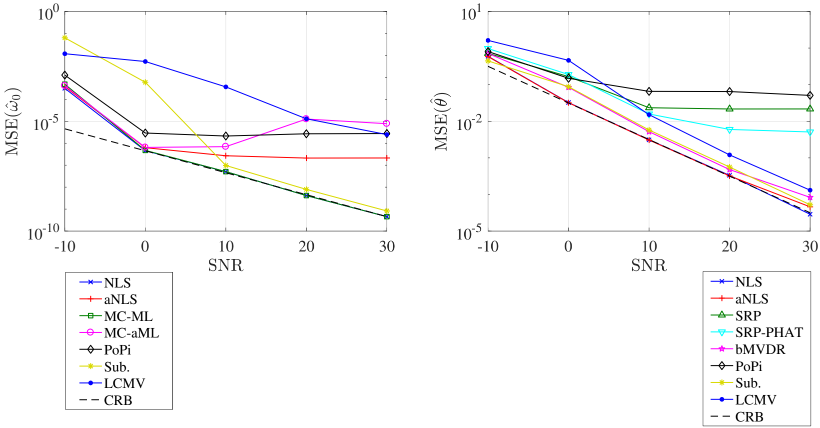
large sample approximation:

$$\widehat{\gamma} = \left(\sum_{k=0}^{K-1} \frac{N}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{\widetilde{\mathbf{D}}_k \mathbf{Z}^H \mathbf{x}_k}{\sigma_k^2}. \tag{191}$$

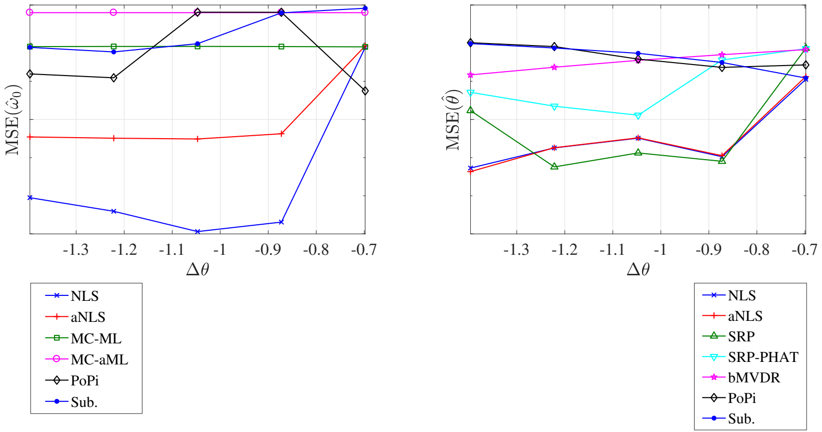
or both:

$$\widehat{\gamma} = \frac{1}{NK} \sum_{k=0}^{K-1} \mathbf{D}_k \mathbf{Z}^H \mathbf{x}_k. \tag{192}$$

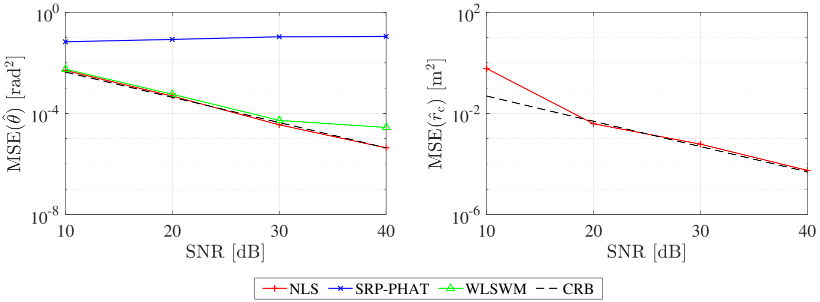
Experimental Results
Synthetic source in far-field



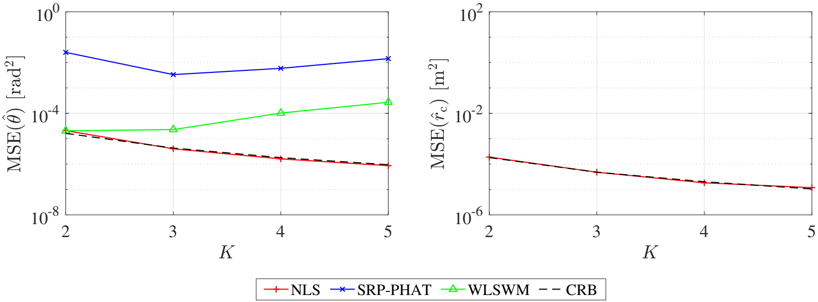
Experimental Results
Synthetic source in far-field



Experimental Results
Synthetic source in near-field

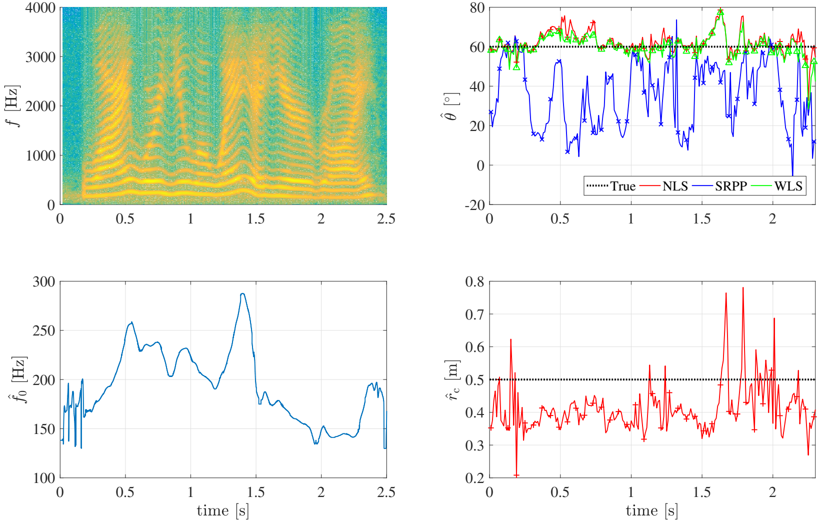


Experimental Results
Synthetic source in near-field



Experimental Results

Real speech in near-field



Outline

Introduction

- DOA of audio/speech useful for, e.g., surveillance and beamforming.
- Reverberation have a detrimental impact on estimation.
- Most existing DOA estimator do not (explicitly) account for reverberation.
- Performance with reverberation is therefore limited.
- Some methods (e.g., SRP-PHAT) are relatively robust against reverb without accounting for it directly.
- Robust DOA estimators based on simple reverb model was proposed.
- **Model:** direct-path + early reflections + noise.

Signal Model

An acoustic source is sampled using a microphone array:

$$y_k(n) = (s' * g_k)(n) + v'_k(n) = s_k(n) + v'_k(n), \tag{193}$$

where

$s'(n)$: clean source signal

$g_k(n)$: room impulse response from source to mic k

$v'_k(n)$: additive noise (interferers, sensor noise, etc.)

Remarks:

- Focus on reverb robust DOA estimation.
- Noise, $v'_k(n)$ assumed white Gaussian.

Vector Model

With K microphones recording N samples each, we get

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T & \cdots & \mathbf{y}_K^T \end{bmatrix}^T = \mathbf{s} + \mathbf{v}'. \tag{194}$$

where

$$\mathbf{y}_k = \begin{bmatrix} y_k(0) & \cdots & y_k(N-1) \end{bmatrix}^T$$

$\mathbf{s}$ & $\mathbf{v}'$: desired signal and noise vectors (defined as $\mathbf{y}$)

Further model specifications:

- desired signal assumed quasi-periodic,
- a ULA structure is assumed.

Periodic Signal Model

Clean desired signal modeled as

$$s'(n) = \sum_{l=-L}^{L} \alpha_l e^{jl\omega_0 n}, \tag{195}$$

with

α_l: complex amplitude of l'th harmonic,

ω_0: fundamental frequency [rad/sample],

L: model order, i.e., number of harmonics.

Important observation:

Widely used broadband model is a special case of (195), i.e., for

$$\omega_0 = 2\pi/N \qquad \wedge \qquad L = \lfloor N/2 \rfloor. \tag{196}$$

Array Model

We assume the source of interest to be in the far-field.

For a ULA, TDOA of source r between mic 1 and k is then

$$\tau_{r,k} = k\frac{d\sin\theta_r}{c} = k\eta_r, \tag{197}$$

with

d: spacing between two adjacent mics,

θ_r: DOA of source r,

c: sound propagation speed.

Complete Signal Model

Observation modeled as multiple early reflections in noise:

$$\mathbf{y} = \sum_{r=1}^{R} \mathbf{H}(\eta_r)\boldsymbol{\alpha}_r + \mathbf{v}, \tag{198}$$

where

$$\begin{aligned} R&: \text{ number of early reflections} \\ \mathbf{H}(\eta_r) &= [\mathbf{Z}^T \ (\mathbf{Z}\mathbf{D}_2(\eta_r))^T \ \cdots \ (\mathbf{Z}\mathbf{D}_K(\eta_r))^T]^T \\ \mathbf{Z} &= \begin{bmatrix} \mathbf{z}_1 & \cdots & \mathbf{z}_L & \mathbf{z}_1^* & \cdots & \mathbf{z}_L^* \end{bmatrix} \\ \mathbf{z}_l &= \begin{bmatrix} 1 & e^{jl\omega_0} & \cdots & e^{j(N-1)l\omega_0} \end{bmatrix}^T \\ \mathbf{D}_k(\eta_r) &= \operatorname{diag}\left(\begin{bmatrix} \mathbf{d}_k^T(\eta_r) & \mathbf{d}_k^H(\eta_r) \end{bmatrix}\right) \\ \mathbf{d}_k(\eta_r) &= \begin{bmatrix} e^{-j\omega_0 k\eta_r} & \cdots & e^{-jL\omega_0 k\eta_r} \end{bmatrix}^T \end{aligned}$$

Estimation problem: find η_1 from observations!

Reverb Robust DOA Estimation

- Two methods for DOA estimation with reverb were proposed.
- Idea is to estimate DOAs of both direct-path and early reflections.
- Bias of direct-path estimate reduced in this way.
- Both methods are based on nonlinear least squares:
  1. a method where amplitudes of direct-path and reflections are assumed independent.
  2. a method where the relation between the amplitudes is modeled.
- Estimation of multiple DOAs facilitated by an iterative approach.

Nonlinear Least Squares
Unstructured amplitudes

With unstructured amplitudes, the NLS estimator is

$$\{\widehat{\boldsymbol{\eta}}, \widehat{\overline{\boldsymbol{\alpha}}}\} = \arg\min_{\{\boldsymbol{\eta},\overline{\boldsymbol{\alpha}}\}} \|\mathbf{y} - \overline{\mathbf{H}}(\boldsymbol{\eta})\overline{\boldsymbol{\alpha}}\|_2^2, \tag{199}$$

with

$$\begin{aligned} \boldsymbol{\eta} &= [\eta_1 \ \cdots \ \eta_R]^T \\ \overline{\mathbf{H}}(\boldsymbol{\eta}) &= \begin{bmatrix} \mathbf{H}(\eta_1) & \cdots & \mathbf{H}(\eta_R) \end{bmatrix} \\ \overline{\boldsymbol{\alpha}} &= \begin{bmatrix} \boldsymbol{\alpha}_1^T & \cdots & \boldsymbol{\alpha}_R^T \end{bmatrix}^T \end{aligned}$$

Solving for $\overline{\boldsymbol{\alpha}}$ gives

$$\widehat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} \Big\| \big(\underbrace{\mathbf{I} - \overline{\mathbf{H}}(\boldsymbol{\eta})(\overline{\mathbf{H}}(\boldsymbol{\eta})^H\overline{\mathbf{H}}(\boldsymbol{\eta}))^{-1}\overline{\mathbf{H}}(\boldsymbol{\eta})^H}_{\mathbf{P}^{\perp}_{\overline{\mathbf{H}}(\boldsymbol{\eta})}} \big)\mathbf{y} \Big\|_2^2. \tag{200}$$

Iterative Procedure
Unstructured amplitudes

Consider a modified observed signal model:

$$\mathbf{y}_r = \mathbf{y} - \sum_{q=1, q\neq r}^{R} \mathbf{H}(\widehat{\eta}_q)\widehat{\alpha}_q, \tag{201}$$

This suggests:

$$\widehat{\alpha}_r = (\mathbf{H}^H(\eta_r)\mathbf{H}(\eta_r))^{-1}\mathbf{H}(\eta_r)^H\mathbf{y}_r, \tag{202}$$

$$\widehat{\eta}_r = \arg\min_{\eta_r} \|\mathbf{P}^{\perp}_{\mathbf{H}(\eta_r)}\mathbf{y}_r\|_2^2. \tag{203}$$

This enables iterative DOA estimation [Li&Stoica,1996], termed RNLS.

Algorithm

Unstructured amplitudes

Step (1): Assume $R = 1$. Estimate η_1 and $\boldsymbol{\alpha}_1$ from $\mathbf{y}_1 = \mathbf{y}$ as described before.

Step (2): Assume $R = 2$. Estimate η_2 and $\boldsymbol{\alpha}_2$ from $\mathbf{y}_2$ using parameter estimates from Step (1). Re-estimate η_1 and $\boldsymbol{\alpha}_1$ from $\mathbf{y}_1$. Iterate until "practical convergence".

Step (3): Assume $R = 3$. Estimate η_3 and $\boldsymbol{\alpha}_3$ from $\mathbf{y}_3$ using parameters from Step (2). Re-estimate η_1 and $\boldsymbol{\alpha}_1$ from $\mathbf{y}_1$. Re-estimate η_2 and $\boldsymbol{\alpha}_2$ from $\mathbf{y}_2$. Iterate until "practical convergence".

Remaining steps: Continue similarly to the previous steps until R is equal to the number of early reflections.

Nonlinear Least Squares
Structured amplitudes

An alternative model with amplitude relations can be formulated

$$\mathbf{y} = \sum_{r=1}^{R} \gamma_r \mathbf{H}(\eta_r)\mathbf{T}_r \boldsymbol{\alpha} + \mathbf{v}, \tag{204}$$

where

γ_r: attenuation of reflection r ($\gamma_1 = 1$)

η_r: delay of reflection r ($\eta_1 = 0$)

$\boldsymbol{\alpha}$: direct-path harmonic amplitudes

$$\mathbf{T}_r = \operatorname{diag}\left(\begin{bmatrix}\mathbf{t}_r^T & \mathbf{t}_r^H\end{bmatrix}\right)$$

$$\mathbf{t}_r = \begin{bmatrix} e^{j\omega_0\xi_r} & \cdots & e^{jL\omega_0\xi_r}\end{bmatrix}^T$$

Iterative Procedure

Structured amplitudes

Again, consider a modified observed signal model:

$$\mathbf{y}_r = \mathbf{y} - \sum_{q=1, q\neq r}^{R} \widehat{\gamma}_q \mathbf{H}(\widehat{\eta}_q)\widehat{\boldsymbol{\alpha}} \tag{205}$$

With this, LS amplitudes and attenuations estimates are

$$\widehat{\boldsymbol{\alpha}} = [\mathbf{H}^H(\eta_1)\mathbf{H}(\eta_1)]^{-1}\mathbf{H}^H(\eta_1)\mathbf{y}_1 \quad (r = 1) \tag{206}$$

$$\widehat{\gamma}_r = \frac{\mathrm{Re}\{\widehat{\boldsymbol{\alpha}}^H \mathbf{T}_r^H \mathbf{H}^H(\eta_r)\mathbf{y}_r\}}{\widehat{\boldsymbol{\alpha}}^H \mathbf{T}_r^H \mathbf{H}^H(\eta_r)\mathbf{H}(\eta_r)\mathbf{T}_r\widehat{\boldsymbol{\alpha}}} \quad (r = 2, \ldots, R). \tag{207}$$

Iterative Procedure
Structured amplitudes

DOA of direct-path is then estimated by ($r = 1$)

$$\widehat{\eta}_1 = \arg\min_{\eta_1} \|\mathbf{P}^{\perp}_{\mathbf{H}(\eta_1)} \mathbf{y}_1\|_2^2. \tag{208}$$

Early reflection DOAs and delays estimated jointly ($r = 2, \dots, R$)

$$\{\widehat{\eta}_r, \widehat{\xi}_r\} = \arg\min_{\eta_r, \xi_r} \|\mathbf{y}_r - \widehat{\gamma}_r \mathbf{H}(\eta_r) \mathbf{T}_r \widehat{\alpha}\|_2^2. \tag{209}$$

This method is termed RNLS-S.

Remarks

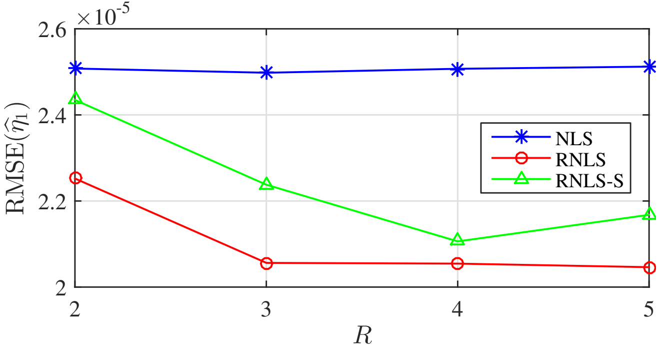
- Implemented using iterative procedure as for RNLS.
- More complex (2d estimation for reflections), but more realistic.

Experimental Results
Synthetic data

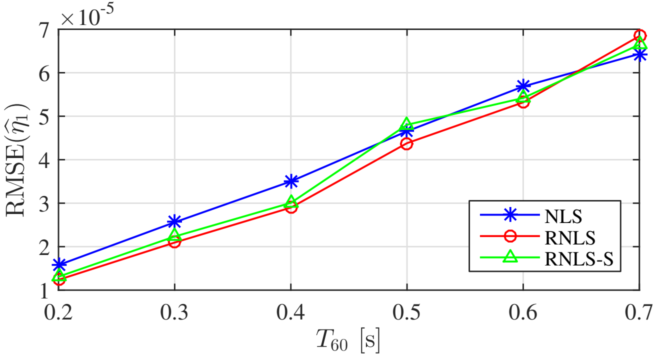
- Evaluated the method on synthetic data.
- Setup:
  - $f_0 = 255.2$ Hz, $f_s = 8$ kHz
  - $L = 6$ (unit amplitude + random phase)
  - f_0 assumed known
  - signal synthesized spatially using RIR generator
  - $d = 0.05$ cm, SNR$= 40$ dB, $N = 200$
  - source DOA varied $(-80^\circ, -75^\circ, \ldots, 80^\circ)$
  - source-array distance: 2.5 m.
- Average results depicted to the right.

Experimental Results

Synthetic data



Experimental Results
Synthetic data



Experimental Results
Real data

- Also evaluated on a real and moving speech source.
- Four seconds of female speech used (synthesized spatially using RIR generator).
- Pitch and model order estimated using an NLS estimator [Christensen,2009].
- Setup: $R = 4$, $T_{60} = 0.3$ s, $K = 4$.

NLS	RNLS	RNLS-S	SRP-PHAT
$3.8 \cdot 10^{-5}$	$3.6 \cdot 10^{-5}$	$3.6 \cdot 10^{-5}$	$5.4 \cdot 10^{-5}$

Outline

Model-based Enhancement

- Different model-based approaches to speech enhancement have been proposed.
- Model-based approach ease the computation of needed second-order statistics, which are otherwise difficult to obtain.
- A recent one is based on speech production (voiced+unvoiced) and noise models driving a Kalman filter.
- Has been applied for hearing-aids application with speech in babble noise.
- Shows improvement in intelligibility as opposed to many existing methods!

Signal Model

Speech assumed to be binaurally recorded with additive and statistically uncorrelated noise:

$$z_{l/r}(n) = s_{l/r}(n) + w_{l/r}(n) \qquad \forall n = 0, 1, 2 \ldots . \tag{210}$$

with

l/r: indicates left or right channel,
$s_{l,r}, w_{l/r}$: speech and noise signal.

Speech can be modeled as an autoregressive (AR) process:

$$s(n) = \left(\sum_{i=1}^{P} a_i(n) s(n-i) \right) + u(n), \tag{211}$$

with P being model order, and $u(n)$ the excitation:

$$u(n) = b(n, p_n) u(n - p_n) + d(n). \tag{212}$$

Noise Model

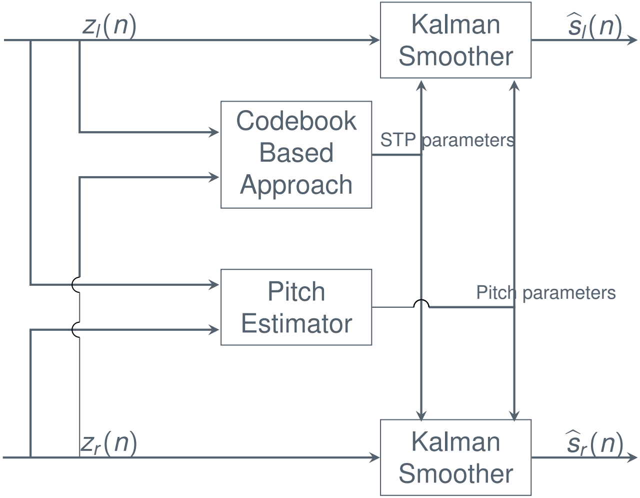
Noise signal modelled as an autoregressive process

$$\begin{aligned} w(n) &= \sum_{i=1}^{Q} c_i(n) w(n-i) + v(n) \\ &= \mathbf{c}(n)^T \mathbf{w}(n-1) + v(n), \end{aligned} \tag{213}$$

where $v(n)$ is WGN with zero mean and excitation variance $\sigma_v^2(n)$.

AR parameters and excitation variance termed short term predictor (STP) parameters. Assumed constant in 25 ms frames.

Block Diagram of Method



Binaural Estimation of STP Parameters

- Usage of a fixed lag Kalman smoother for speech enhancement requires the speech and noise STP parameters to be estimated.
- This approach uses a priori information about the spectral envelopes of speech and noise stored in codebooks.
- Single-channel codebook based estimation of STP parameters can be used [Srinivasan,2007].

Binaural Estimation of STP Parameters

The random variables of parameters to be estimated are $\theta = [\mathbf{a}; \mathbf{c}; \sigma_u^2; \sigma_v^2]$.

The MMSE estimate of the parameter is written as

$$\widehat{\theta} = \mathrm{E}(\theta|\mathbf{z}_l, \mathbf{z}_r), \tag{214}$$

where $\mathbf{z}_l, \mathbf{z}_r$ denotes a frame of noisy samples at left, right ears.

Using the Bayes Theorem:

$$\widehat{\theta} = \int_\Theta \theta p(\theta|\mathbf{z}_l, \mathbf{z}_r) d\theta = \int_\Theta \theta \frac{p(\mathbf{z}_l, \mathbf{z}_r|\theta)p(\theta)}{p(\mathbf{z}_l, \mathbf{z}_r)} d\theta. \tag{215}$$

Binaural Estimation of STP Parameters

With $\theta_{ij} = [\mathbf{a}_i; \sigma^{2,ML}_{u,ij}; \mathbf{c}_j; \sigma^{2,ML}_{v,ij}]$, discrete counterpart, (215) is:

$$\hat{\theta} = \sum_{i=1}^{N_s} \sum_{j=1}^{N_w} \theta_{ij} \frac{p(\mathbf{z}_l, \mathbf{z}_r | \theta_{ij}) p(\theta_{ij})}{p(\mathbf{z}_l, \mathbf{z}_r)}, \tag{216}$$

where the MMSE estimate is expressed as weighted linear combination of θ_{ij} with weights $p(\mathbf{z}_l, \mathbf{z}_r | \theta_{ij})$.

Assuming conditional independence for the left and right noisy signal:

$$p(\mathbf{z}_l, \mathbf{z}_r | \theta_{ij}) = p(\mathbf{z}_l | \theta_{ij}) p(\mathbf{z}_r | \theta_{ij}). \tag{217}$$

Binaural Estimation of STP Parameters

Log-likelihood can be expressed using negative of Itakura Saito distortion between noisy and modelled noisy spectral envelopes:

$$\begin{aligned} p(\mathbf{z}_l, \mathbf{z}_r|\theta_{ij}) &= p(\mathbf{z}_l|\theta_{ij})p(\mathbf{z}_r|\theta_{ij}) \\ &= e^{-\left(d_{\text{IS}}(P_{z_l}(\omega), \widehat{P}_z^{ij}(\omega)) + d_{\text{IS}}(P_{z_r}(\omega), \widehat{P}_z^{ij}(\omega))\right)} \end{aligned} \tag{218}$$

where

$P_{z_l/z_r}(\omega)$: noisy spectral envelope at the left, right ear,

$$\widehat{P}_z^{ij}(\omega) = \frac{\sigma_{u,ij}^{2,ML}}{|A_s^i(\omega)|^2} + \frac{\sigma_{v,ij}^{2,ML}}{|A_w^j(\omega)|^2}.$$

Binaural Estimation of STP Parameters

Dual channel estimation of noise PSD proposed by [Dorbecker,1996].

Assumes a homogeneous noise field, i.e.

$$\Phi_{ww}(\omega) = \Phi_{w_l w_l}(\omega) = \Phi_{w_r w_r}(\omega), \Phi_{w_l w_r}(\omega) = 0. \tag{219}$$

Shown that noise PSD estimate at frame number k is obtained as

$$\hat{\Phi}_{ww}(\omega, k) = \sqrt{\Phi_{z_l z_l}(\omega, k)\Phi_{z_r z_r}(\omega, k)} - |\Phi_{z_l z_r}(\omega, k)| \tag{220}$$

Dual channel noise PSD estimate then used to find LPC coefficients and variance of spectral envelope, which are appended to the noise codebook.

Binaural Pitch Estimator

Noisy signals modeled as

$$\mathbf{y} = \begin{bmatrix} \mathbf{z}_l \\ \mathbf{z}_r \end{bmatrix} = \begin{bmatrix} \mathbf{Z}\mathbf{D}_l \\ \mathbf{Z}\mathbf{D}_r \end{bmatrix} \alpha + \begin{bmatrix} \mathbf{w}_l \\ \mathbf{w}_r \end{bmatrix} = \mathbf{H}\alpha + \mathbf{w}. \tag{221}$$

with

- $\mathbf{Z}$: matrix of Fourier vectors for harmonics,
- $\mathbf{D}_{l/r}$: diag. directivity matrices for left/right mic (phase shift and gain scaling),
- α: complex harmonic amplitudes.

ML estimate of amplitude vector α:

$$\widehat{\alpha} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} \tag{222}$$

Binaural Pitch Estimator

ML noise variance estimates for the left and right channels:

$$\widehat{\sigma}^2_{l/r} = \frac{1}{N}\|\mathbf{z}_{l/r} - \mathbf{Z}\mathbf{D}_{l/r}\widehat{\boldsymbol{\alpha}}\|^2 \tag{223}$$

Using amplitude and noise estimates, fundamental frequency estimated as:

$$\left\{\widehat{\omega}_0, \widehat{L}\right\} = \underset{\{L,\omega_0\}}{\operatorname{argmin}}\, N \ln \widehat{\sigma}^2_l \widehat{\sigma}^2_r + L \ln 2N. \tag{224}$$

Fixed Lag Kalman Smoother (FLKS)

Speech model can be written as concatenated state space equation:

$$\begin{bmatrix} \mathbf{s}_{l/r}(n) \\ \mathbf{u}(n+1) \\ \mathbf{w}_{l/r}(n) \end{bmatrix} = \begin{bmatrix} \mathbf{A}(n) & \mathbf{\Gamma}_1\mathbf{\Gamma}_2^T & \mathbf{0} \\ \mathbf{0} & \mathbf{B}(n+1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}(n) \end{bmatrix} \begin{bmatrix} \mathbf{s}_{l/r}(n-1) \\ \mathbf{u}(n) \\ \mathbf{w}_{l/r}(n-1) \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{\Gamma}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_3 \end{bmatrix} \begin{bmatrix} d(n+1) \\ v(n) \end{bmatrix} \Leftrightarrow, \quad (225)$$

$$\mathbf{x}_{l/r}(n+1) = \mathbf{F}(n)\mathbf{x}_{l/r}(n) + \mathbf{\Gamma}_4 g(n+1) \quad (226)$$

where $\mathbf{x}_{l/r}(n)$ is the concatenated state space vector, $\mathbf{F}(n)$ is the concatenated state evolution matrix.

Measurement equation given by

$$z_{l/r}(n) = \mathbf{\Gamma}^T \mathbf{x}_{l/r}(n). \quad (227)$$

Eventually, clean speech is then predicted using Kalman filter.

Experimental Setup

- Objective measures such as STOI and PESQ and used to evaluate the algorithm.
- Test set of clean signals taken from CHiME and Eurom databases.
- Binaural noisy signals generated by convolving with anechoic binaural HRIR obtained from [Kayser,2009] and adding with binaural noise signals.
- Noise codebook generated using a training sample of 2 minutes of babble.
- Speaker codebook generated using 2-5 minutes of speech from the speaker of interest.

Experimental Setup

Parameters

fs	Frame Size	N_{spk}	N_w	P	Q
8 Khz	200 (25ms)	64	8	14	14

Compared Methods

UV: binaural enhancement using the conventional AR model,

V-UV: binaural enhancement using the voiced-unvoiced model (includes pitch information from noisy signal).

Experimental Results

STOI scores

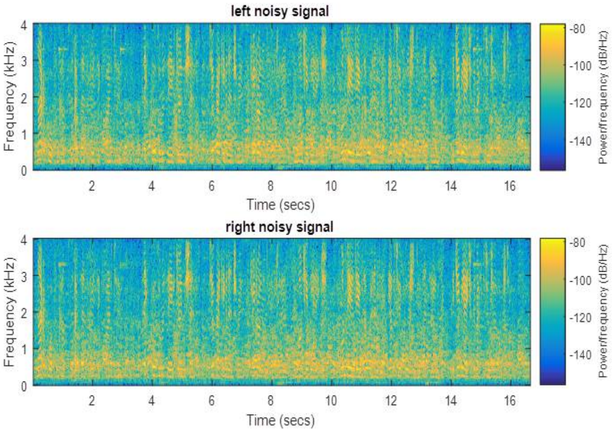
		SNR (dB)			
		-5	-2	1	4
Noisy	male	0.6486	0.7178	0.7827	0.8387
	female	0.6305	0.7003	0.7668	0.8259
UV	male	0.6882	0.7655	0.8334	0.8841
	female	0.6568	0.7332	0.8035	0.8603
V-UV	male	**0.7036**	**0.7803**	**0.8436**	**0.8893**
	female	**0.6857**	**0.7635**	**0.8277**	**0.8762**

Experimental Results

PESQ scores

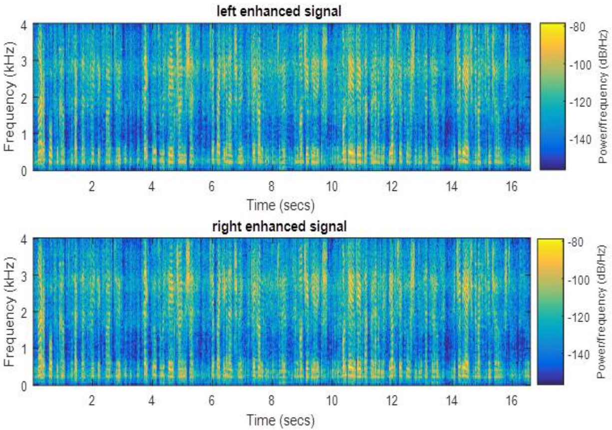
		SNR (dB)			
		-5	-2	1	4
Noisy	male	1.6313	1.7754	1.8434	1.9577
	female	1.2924	1.4868	1.6331	1.7941
UV	male	1.8666	2.0467	2.2978	2.5326
	female	1.5066	1.7273	1.9487	2.1744
V-UV	male	1.8720	2.1006	2.3396	2.5489
	female	1.6088	1.8429	2.0625	2.2626

Experimental Results
Examples



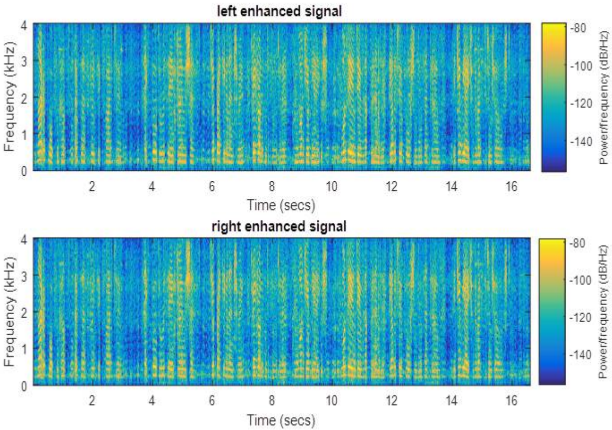
Noisy signal

Experimental Results
Examples



Enhanced signal (UV)

Experimental Results
Examples



Enhanced signal (V-UV)

Outline

Voiced Speech Enhancement

Model-based single-channel methods

- Voiced speech exhibits quasi-periodic structure.
- Has been exploited to derive model- and filtering-based enhancement methods.
- Compared to traditional speech enhancement method, these can be guaranteed distortionless!
- Also, they do not require noise statistics estimates.

Signal Model

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

$$x_k(n) = \sum_{l=1}^{L_k} \alpha_{k,l} e^{j\omega_k l n} + e_k(n), \tag{228}$$

where

- ω_k is the fundamental frequency,
- $\alpha_{k,l} = A_{k,l} e^{j\phi_{k,l}}$ is the complex amplitude,
- $e_k(n)$ is the observation noise.

Matrix-Vector Model

Define vectors of M consecutive observations (with $M \leq N$):

$$\mathbf{x}(n) = [x(n)\ x(n-1)\ \cdots\ x(n-M+1)]^T, \tag{229}$$

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

$$\mathbf{x}(n) = \sum_{k=1}^{K} \mathbf{Z}_k \begin{bmatrix} e^{-j\omega_k 1 n} & & 0 \\ & \ddots & \\ 0 & & e^{-j\omega_k L_k n} \end{bmatrix} \boldsymbol{\alpha}_k + \mathbf{e}(n) \tag{230}$$

$$\tag{231}$$

where $\mathbf{Z}_k = [\ \mathbf{z}(\omega_k)\ \cdots\ \mathbf{z}(\omega_k L_k)\ ]$, $\mathbf{z}(\omega) = [\ 1\ e^{-j\omega}\ \cdots\ e^{-j\omega(M-1)}\ ]^T$, and $\boldsymbol{\alpha}_k = [\ \alpha_{k,1}\ \cdots\ \alpha_{k,L_k}\ ]^T$.

MSE-based Filters

Enhancement filters for voiced speech derived from MSE between filter output, $y_k(n)$ and desired output, $\widehat{y}_k(n)$,

$$P = \frac{1}{G} \sum_{n=M-1}^{N-1} \left| y_k(n) - \hat{y}_k(n) \right|^2, \tag{232}$$

$$P = \frac{1}{G} \sum_{n=M-1}^{N-1} |\mathbf{h}_k^H \mathbf{x}(n) - \alpha_k^H \mathbf{w}_k(n)|^2,$$

with

$$\begin{aligned} \mathbf{h}_k &= \left[h_k(0) \ \cdots \ h_k(M-1) \right]^T, \\ \boldsymbol{\alpha}_k &= \left[\alpha_{k,1} \ \cdots \ \alpha_{k,L} \right]^T, \\ \mathbf{w}_k(n) &= \left[e^{j\omega_k n} \ \ldots \ e^{j\omega_k L n} \right]^T. \end{aligned}$$

MSE-based Filters

MSE can be written out as:

$$P = \mathbf{h}_k^H \widehat{\mathbf{R}} \mathbf{h}_k - \boldsymbol{\alpha}_k^H \mathbf{G}_k \mathbf{h}_k - \mathbf{h}_k^H \mathbf{G}_k^H \boldsymbol{\alpha}_k + \boldsymbol{\alpha}_k^H \mathbf{W}_k \boldsymbol{\alpha}_k, \tag{233}$$

where

$$\begin{aligned} \mathbf{G}_k &= \frac{1}{G} \sum_{n=M-1}^{N-1} \mathbf{w}_k(n) \mathbf{x}^H(n), \\ \mathbf{W}_k &= \frac{1}{G} \sum_{n=M-1}^{N-1} \mathbf{w}_k(n) \mathbf{w}_k^H(n). \end{aligned}$$

Can be solved for the unknown amplitudes:

$$\widehat{\boldsymbol{\alpha}}_k = \mathbf{W}_k^{-1} \mathbf{G}_k \mathbf{h}_k, \quad (G \geq L_k, G \geq M). \tag{234}$$

MSE-based Filters

Inserting amplitude estimates, we get:

$$P = \mathbf{h}_k^H \widehat{\mathbf{R}}_k \mathbf{h}_k - \mathbf{h}_k^H \mathbf{G}_k^H \mathbf{W}_k^{-1} \mathbf{G}_k \mathbf{h}_k, \tag{235}$$

$$P = \mathbf{h}_k^H \left(\widehat{\mathbf{R}}_k - \mathbf{G}_k^H \mathbf{W}_k^{-1} \mathbf{G}_k \right) \mathbf{h}_k, \tag{236}$$

$$P \triangleq \mathbf{h}_k^H \widehat{\mathbf{Q}}_k \mathbf{h}_k. \tag{237}$$

where

$$\widehat{\mathbf{Q}}_k = \widehat{\mathbf{R}}_k - \mathbf{G}_k^H \mathbf{W}_k^{-1} \mathbf{G}_k \tag{238}$$

can be thought of as a *modified* or *noise* covariance matrix estimate.

MSE-based Filters

Optimal filters then derived by minimizing residual noise with no signal distortion as:

$$\min_{\mathbf{h}_k} \mathbf{h}_k^H \widehat{\mathbf{Q}}_k \mathbf{h}_k \quad \text{s.t.} \quad \mathbf{h}_k^H \mathbf{Z}_k = \mathbf{1}. \tag{239}$$

Solution to optimization problem:

$$\widehat{\mathbf{h}}_k = \widehat{\mathbf{Q}}_k^{-1} \mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{Q}}_k^{-1} \mathbf{Z}_k \right)^{-1} \mathbf{1}. \tag{240}$$

These filteres are termed SF-APES filters.

One can make filterbank equivalents (filter for each harmonic). These are termed FB-*.

Simplifications
#1

Replacing $\mathbf{W}_k$ by $\mathbf{I}$ in (237), the usual noise covariance matrix estimate is obtained. As before:

$$\widehat{\mathbf{h}}_k = \widehat{\mathbf{Q}}_k^{-1}\mathbf{Z}_k\left(\mathbf{Z}_k^H\widehat{\mathbf{Q}}_k^{-1}\mathbf{Z}_k\right)^{-1}\mathbf{1}, \tag{241}$$

but the modified covariance matrix estimate is now:

$$\widehat{\mathbf{Q}}_k = \widehat{\mathbf{R}} - \mathbf{G}_k^H\mathbf{G}_k. \tag{242}$$

Computationally simpler as it does not require the inversion of $\mathbf{W}_k$ for each candidate frequency.

We refer to this design as SF-APES (appx).

Simplifications

#2 and #3

Capon-like filters can be obtained as a special case ($\widehat{\mathbf{Q}}_k = \widehat{\mathbf{R}}$):

$$\widehat{\mathbf{h}}_k = \widehat{\mathbf{R}}^{-1}\mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{R}}^{-1}\mathbf{Z}_k\right)^{-1} \mathbf{1}, \tag{243}$$

which we refer to as SF-Capon.

A simpler set of filters obtained by assuming WGN (i.e., $\widehat{\mathbf{R}} = \sigma^2\mathbf{I}$):

$$\widehat{\mathbf{h}}_k = \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{1}, \tag{244}$$

which is fully specified by $\mathbf{Z}_k$. Referred to as SF-WNC.

Simplification
#4

Further simplification possible through large sample approximation:

$$\lim_{M\to\infty} M\mathbf{Z}_k \left(\mathbf{Z}_k^H\mathbf{Z}_k\right)^{-1} = \mathbf{Z}_k \lim_{M\to\infty} \left(\frac{1}{M}\mathbf{Z}_k^H\mathbf{Z}_k\right)^{-1} \tag{245}$$

$$= \mathbf{Z}_k. \tag{246}$$

Leads to trivial filter design:

$$\hat{\mathbf{h}}_k = \frac{1}{M}\mathbf{Z}_k\mathbf{1}, \tag{247}$$

i.e., the normalized sum over a set of filters defined by Fourier vectors. Referred to as SF-WNC (appx).

Some Results

Experimental details:

- The first part of the experiments is based on synthetic signals with a periodic signal buried in noise and with another periodic signal interfering.
- We then vary the signal-to-noise ratio (SNR) and the signal-to-interference ratio (SIR) and measure the signal-to-distortion ratio.
- We then also demonstrate how the optimal filters can be used for processing real non-stationary speech signals.

Experiments on Synthetic Data

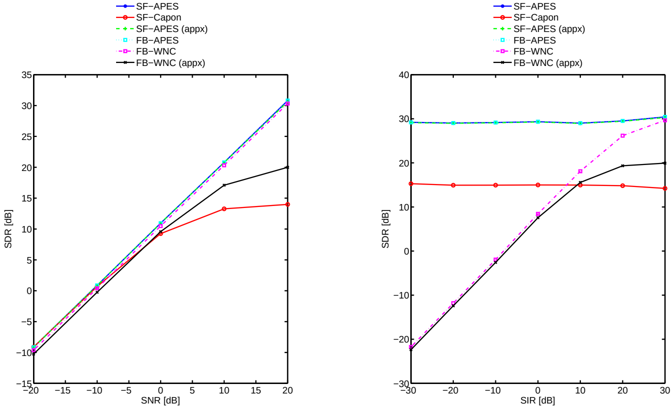


Figure: SDR versus (left) SNR and (right) SIR with an interfering source present (SNR of 10 dB).

Experiments on Synthetic Data

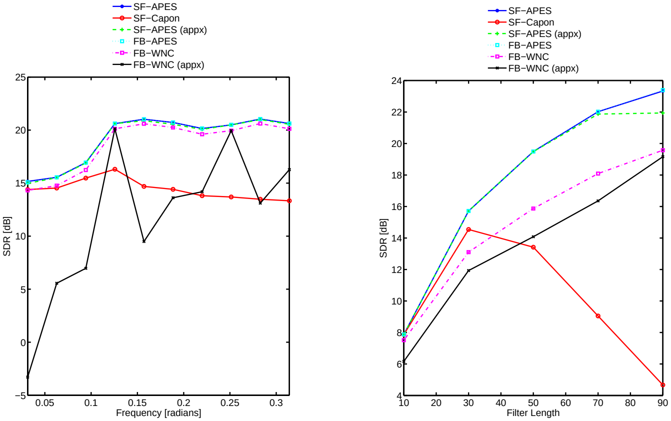


Figure: SDR versus (left) fundamental frequency, and (right) filter length with an interfering source present.

Experiments on Speech Data

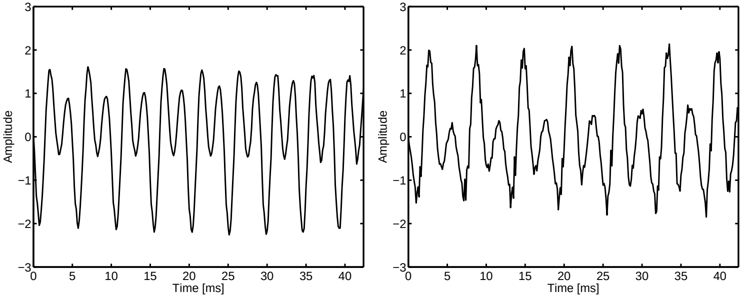
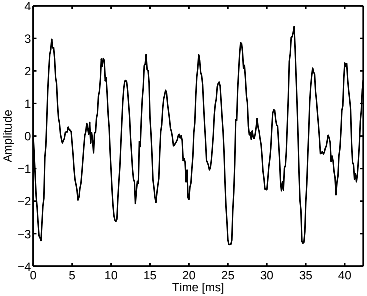


Figure: Plots of: voiced speech signal of sources (left) 1 and (right) 2.

Experiments on Speech Data



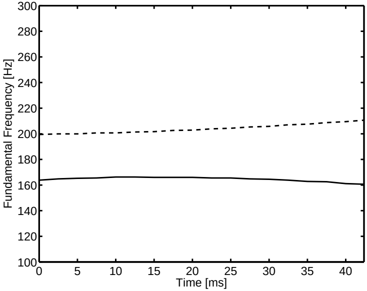


Figure: Plots of: (left) mixture of the two signals and (right) estimated pitch tracks for source 1 (dashed) and 2 (solid).

Experiments on Speech Data

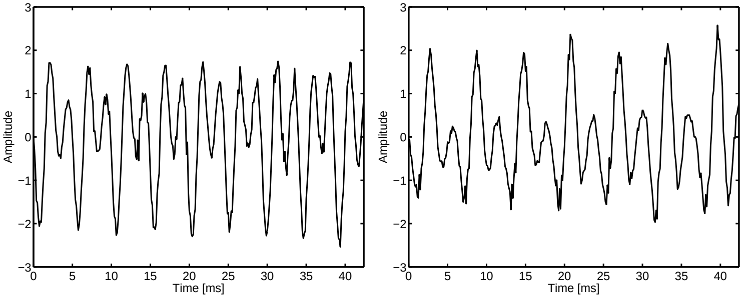


Figure: Plots of: estimate of sources (left) 1 and (right) 2 obtained from mixture.

Extension to Multichannel

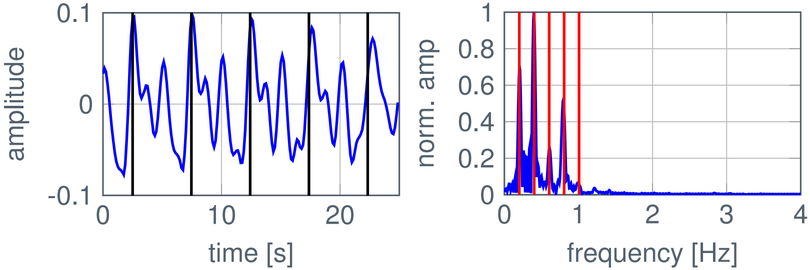
- The filtering method was extended to the multichannel case in [Jensen2017].
- Idea is to use APES principle on each channel and do weighted average based on MSEs.
- Two approaches were considered:
  - a method which is dependent on geometry,
  - a method being independent on geometry.
- Results, in terms of PESQ scores, showed that geometry-based approach is best for larger arrays, but worse when significant DOA errors are present.

Outline

Enhancement of Non-Stationary Speech

- Many traditional enhancement methods rely on strict assumptions on stationarity.
- E.g., stationarity is assumed when estimating second-order statistics for Wiener filtering.
- Stationarity assumptions on speech never hold, even in short time frames.
- For voiced speech, harmonic chirp models can account for this.
- Recently, enhancement filters based on this model were proposed.
- Second-order statistics of noise found as by-product, and can be used in traditional enhancement methods.

Harmonic Modeling of Speech

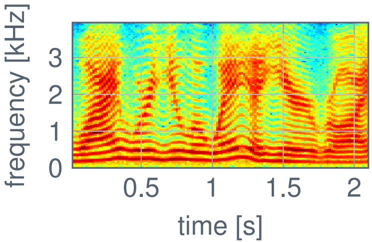


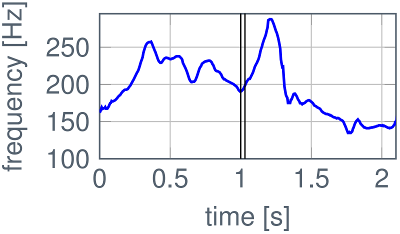
Harmonic model for speech:

$$x(n) = s(n) + v(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n} + v(n), \quad \alpha_l = A_l e^{j\phi_l}. \qquad (248)$$

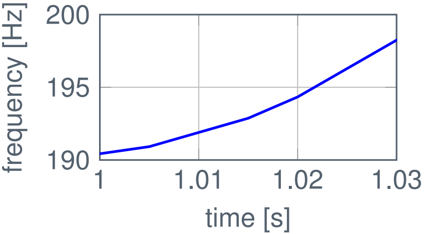
$\omega_0 = 2\pi f_0 / f_s$: fundamental frequency, L: number of harmonics.

Introduction
Motivation





In the 30 ms segment
$\Delta f_0 \approx 8\,\text{Hz} \Rightarrow$
Harmonic chirp model



Harmonic Chirp Model

Speech modelled using chirp model with linearly time-varying instantaneous frequencies of harmonics:

$$\omega_l(n) = l(\omega_0 + kn). \tag{249}$$

where k is the chirp parameter.

Instantaneous phase is integral of instantaneous frequency:

$$\theta_l(n) = l\left(\omega_0 n + \frac{1}{2}kn^2\right) + \phi_l. \tag{250}$$

Leads to harmonic chirp model:

$$x(n) = \sum_{l=1}^{L} \alpha_l e^{jl(\omega_0 n + k/2n^2)} + v(n). \tag{251}$$

LCMV Filtering

LCMV filter, $\mathbf{h}$, minimises its output power without signal distortion. Mathematically equivalent to:

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}_x \mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{1}, \tag{252}$$

where

$$\begin{aligned}
\mathbf{h} &= [h(0)\, h(1)\, \dots\, h(M-1)]^H, \\
\mathbf{R}_x&\text{: covariance matrix of } \mathbf{x}(n), \\
\mathbf{x}(n) &= [x(n)\, x(n+1)\, \cdots\, x(n+M-1)]^T, \\
\mathbf{Z} &= [\mathbf{z}(\omega_0, k)\, \mathbf{z}(2\omega_0, 2k)\, \cdots\, \mathbf{z}(L\omega_0, Lk)], \\
\mathbf{z}(l\omega_0, lk) &= [1\; e^{jl(\omega_0 + k/2)}\; \dots\; e^{jl(\omega_0(M-1) + k/2(M-1)^2)}]^T, \\
M&\text{: filter length}, \\
\mathbf{1} &= [1 \cdots 1].
\end{aligned}$$

LCMV Filtering

The solution is

$$\mathbf{h} = \mathbf{R}_x^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{R}_x^{-1}\mathbf{Z})^{-1}\mathbf{1}. \tag{253}$$

Often, $\mathbf{R}_x$ unknown but can be estimated:

$$\widehat{\mathbf{R}}_x = \frac{1}{N-M+1}\sum_{n=0}^{N-M}\mathbf{x}(n)\mathbf{x}^H(n), \tag{254}$$

where N is the segment length, and $M \leq N/2$ ensures invertibility.

Assumes a stationary signal within the length N segment!

APES Filtering

Amplitude and Phase EStimation (APES) based filter uses signal model to obtain noise covariance matrix estimate.

APES filter minimises the mean square error (MSE)

$$\text{MSE} = \frac{1}{N-M+1} \sum_{n=0}^{N-M+1} |\mathbf{h}^H \mathbf{x}(n) - \boldsymbol{\alpha}^H \mathbf{w}(n)|^2, \qquad (255)$$

with

$$\begin{aligned} \boldsymbol{\alpha} &= [\alpha_1 \; \alpha_2 \; \dots \; \alpha_L]^H, \\ \mathbf{w}(n) &= [e^{j(\omega_0 n + k/2n^2)} \; e^{j2(\omega_0 n + k/2n^2)} \; \dots \; e^{jL(\omega_0 n + k/2n^2)}]^T. \end{aligned}$$

APES Filtering

The solution is

$$\mathbf{h} = \mathbf{Q}^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{Q}^{-1}\mathbf{Z})^{-1}\mathbf{1}, \tag{256}$$

with

$$\begin{aligned}
\mathbf{Q} &= \mathbf{R}_x - \mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}, \\
\mathbf{G} &= \frac{1}{N-M+1}\sum_{n=0}^{N-M+1}\mathbf{w}(n)\mathbf{x}^H(n), \\
\mathbf{W} &= \frac{1}{N-M+1}\sum_{n=0}^{N-M+1}\mathbf{w}(n)\mathbf{w}^H(n).
\end{aligned}$$

Interestingly, only difference between LCMV and APES is the covariance matrix used.

Simulations

Filters compared:

- **$\mathbf{LCMV_{opt}}$**: chirp LCMV filter with $\widehat{\mathbf{R}}_x$ replaced by $\widehat{\mathbf{R}}_v$ estimated directly from the noise signal.
- **$\mathbf{LCMV_{h}}$**: harmonic LCMV filter, $k = 0$.
- **$\mathbf{LCMV_{c}}$**: chirp LCMV filter.
- **$\mathbf{APES_{h}}$**: harmonic APES filter, $k = 0$.
- **$\mathbf{APES_{c}}$**: chirp APES filter.
- **$\mathbf{APES_{hc}}$**: APES filter with $\mathbf{Z}$ based on chirp model and $\mathbf{Q}$ on the harmonic model with $k = 0$.

Simulations
Performance measures

Filters evaluated as function of input SNR:

$$\mathrm{iSNR} = \frac{\sigma_s^2}{\sigma_v^2}, \tag{257}$$

Performance measured using output SNR and signal reduction factor:

$$\mathrm{oSNR}(\mathbf{h}) = \frac{\sigma_{s,\mathrm{nr}}^2}{\sigma_{v,\mathrm{nr}}^2} = \frac{\mathbf{h}^H \mathbf{R}_s \mathbf{h}}{\mathbf{h}^H \mathbf{R}_v \mathbf{h}}, \tag{258}$$

$$\xi_{\mathrm{sr}}(\mathbf{h}) = \frac{\sigma_s^2}{\sigma_{s,\mathrm{nr}}^2} = \frac{\sigma_s^2}{\mathbf{h}^H \mathbf{R}_s \mathbf{h}}, \tag{259}$$

where

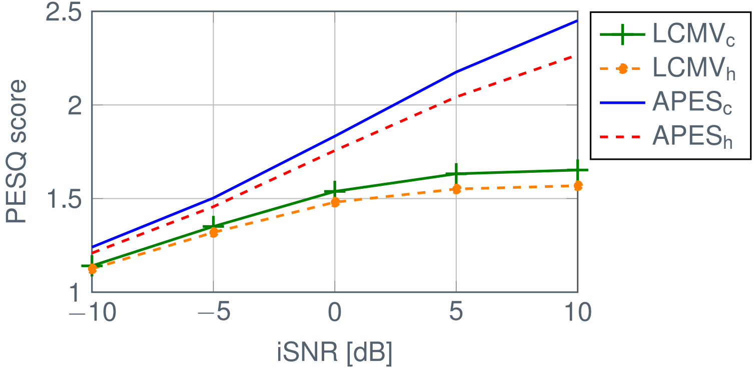
$\sigma_s^2, \sigma_v^2, \sigma_{s,\mathrm{nr}}^2, \sigma_{v,\mathrm{nr}}^2$: variances of the desired signal and noise before and after noise reduction

$\mathbf{R}_s, \mathbf{R}_v$: covariance matrices of desired signal and noise.

Simulations
Speech signal

- Female speaker uttering: "Why were you away a year, Roy?".
- ω_0, k and L estimated with NLS estimator.
- Filter length, $M = 50$.
- Segment length $N = 200$.
- The iSNR is -10 dB to 10 dB in steps of 2.5 dB.
- The noise is white Gaussian.
- 50 Monte Carlo simulations.

Simulations
Results



Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement

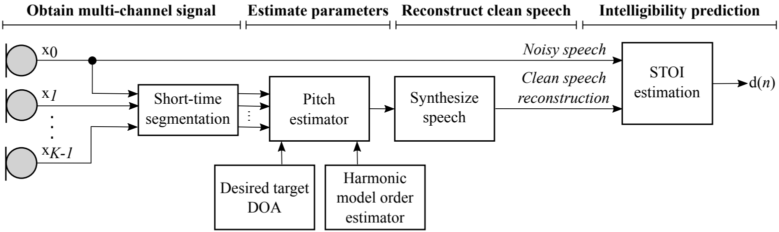
- Parametric vs. Non-Parametric TDOA Estimation
- Joint DOA and Pitch Estimation
- Reverb Robust Speech Localization
- Model-based Speech Enhancement
- Enhancement of Non-Stationary Speech
- Non-Intrusive Speech Intelligibility Prediction

Summary and Conclusion

Speech Intelligibility Prediction

- Most objective speech intelligibility metrics requires clean reference signal.
- Pitch-based method for short-time objective intelligibility (STOI) was proposed that does not require this.
- Reference signal replaced by a reconstructed clean speech estimate.
- The reconstruction based pitch features of desired source.
- Simulations show high correlation between the non-intrusive and intrusive STOI estimators.

Non-Intrusive STOI Estimator



Signal Model

Assume K microphones records N samples each:

$$\mathbf{x}_k) = \beta_k \mathbf{Z}\mathbf{D}(k)\boldsymbol{\alpha} + \mathbf{e}_k \tag{260}$$

$$= \begin{bmatrix} x_k(0) & x_k(1) & \cdots & x_k(N-1) \end{bmatrix}^T, \tag{261}$$

with

$$\begin{aligned}
\mathbf{Z} &= [\mathbf{z}(\omega_0) \ \cdots \ \mathbf{z}(L\omega_0)], \\
\mathbf{z}(l\omega_0) &= [1 \ e^{jl\omega_0} \ \cdots \ e^{jl\omega_0(N-1)}], \\
\mathbf{D}(k) &= \mathrm{diag}\{e^{-j\omega_0 f_s \tau_k}, \dots, e^{-jL\omega_0 f_s \tau_k}\}, \\
\boldsymbol{\alpha} &= [\alpha_1 \ \cdots \ \alpha_L]^T \text{ (complex harmonic amplitudes)},
\end{aligned}$$

ω_0, f_s: fundamental and sampling frequencies,

τ_k, β_k: delay and attenuation of speech between mics. 0 and k,

L: number of harmonics.

Maximum Likelihood Pitch estimation

With zero-mean WGN at each mic with variance σ_k^2, log-likelihood is:

$$\ln p(\mathbf{x}_k;\psi) = -NK\ln\pi - N\sum_{k=0}^{K-1}\ln\sigma_k^2 - \sum_{k=0}^{K-1}\frac{\|\mathbf{e}_k\|^2}{\sigma_k^2}, \tag{262}$$

where ψ contains signal parameters.

Unknown, linear, parameters found iteratively using:

$$\widehat{\boldsymbol{\alpha}}_k = \left[\sum_{k=0}^{K-1}\frac{\beta_k^2}{\sigma_k^2}\mathbf{D}^H(k)\mathbf{Z}^H\mathbf{Z}\mathbf{D}(k)\right]^{-1}\sum_{k=0}^{K-1}\frac{\beta_k}{\sigma_k^2}\mathbf{D}^H(k)\mathbf{Z}^H\mathbf{x}_k, \tag{263}$$

$$\widehat{\beta}_k = \frac{\mathrm{Re}\{\boldsymbol{\alpha}^H\mathbf{D}^H(k)\mathbf{Z}\mathbf{x}_k\}}{\boldsymbol{\alpha}\mathbf{D}^H(k)\mathbf{Z}^H\mathbf{Z}\mathbf{D}(k)\boldsymbol{\alpha}}, \tag{264}$$

$$\widehat{\sigma}_k^2 = \frac{\|\mathbf{x}_k - \beta_k\mathbf{Z}\mathbf{D}(k)\boldsymbol{\alpha}\|^2}{N}. \tag{265}$$

Maximum Likelihood Pitch estimation

For known θ (i.e., τ_k's), the ML pitch estimator is then

$$\widehat{\omega}_0 = \arg\min_{\omega_0} \sum_{k=0}^{K-1} \ln \|\mathbf{x}_k - \widehat{\beta}_k \mathbf{Z}\mathbf{D}(k)\widehat{\boldsymbol{\alpha}}\|^2. \tag{266}$$

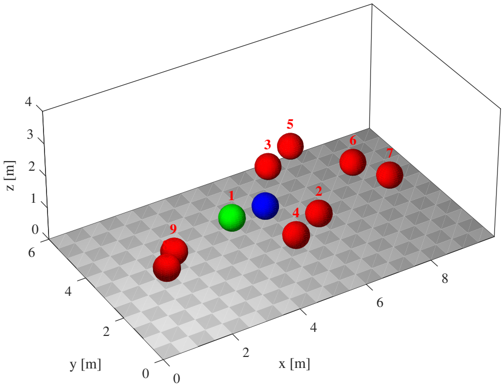
Amplitude estimate for channel k can be used to reconstruct signal:

$$\widehat{s}_k(n) = \mathbf{Z}\widehat{\boldsymbol{\alpha}}_k. \tag{267}$$

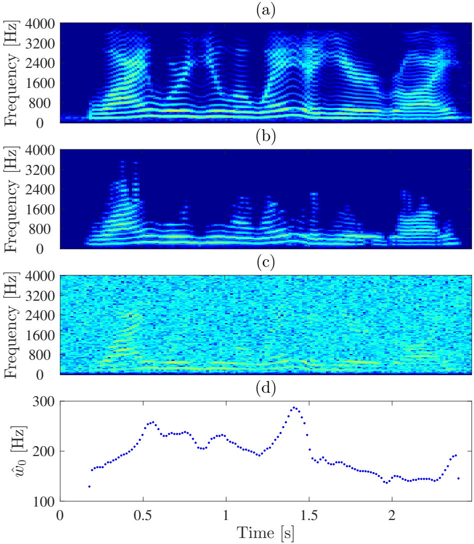
A final estimate can be obtained through averaging:

$$\widehat{\mathbf{s}}(n) = \frac{1}{K} \sum_{k=0}^{K-1} \widehat{\mathbf{s}}_k(n). \tag{268}$$

Experimental Setup



Experimental Results
Pitch Estimates and Reconstruction



Experimental Results

Intrusive vs non-intrusive STOI estimates

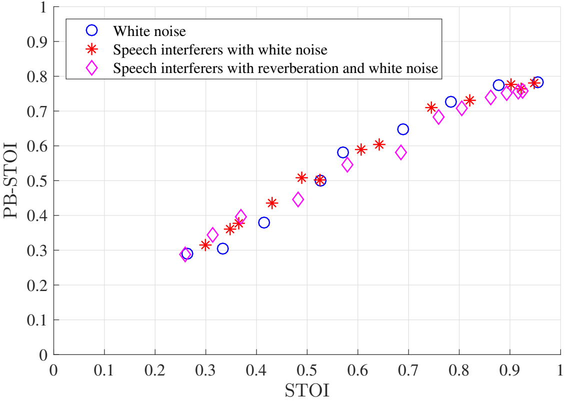


Figure: Results from PB-STOI using a ULA setup.

Experimental Results
Intrusive vs non-intrusive STOI estimates

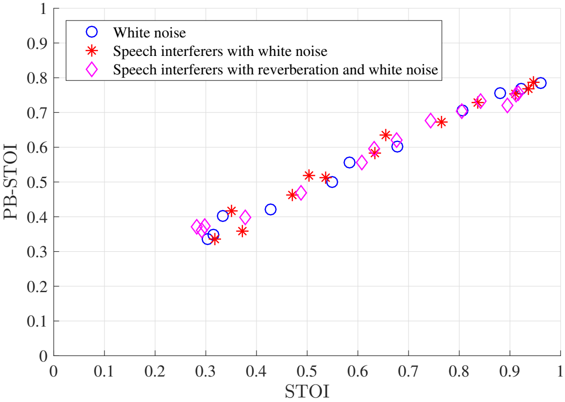


Figure: Results from PB-STOI using a BTE HA setup.

Outline

Applications

The ideas presented here are/can be used in many applications, including:

- Hearing aids
- Voice over IP
- Telecommunication
- Reproduction systems
- Voice analysis
- Biomedical engineering
- Surveillance
- Music equipment/software
- Sound and vibration

Some Other Results

- Parametric models can be used for speech/audio compression (van Schijndel 2008).
- Model-based interpolation/extrapolation can be used for packet losses/corrupt data (Rødbro 2003, Nielsen 2011).
- Feedback cancellation can be improved using a model of the near-end signal .(Ngo 2011)
- It is possible to take common panning techniques in stereo into account (Hansen 2017).

Conclusion

- Parametric models have shown promise for several problems, but they are not (yet) widespread.
- An argument against the usage of such models is that they do not take various phenomena into account.
- However, we can only have this discussion because the assumptions are explicit.
- And it is often fairly easy to improve the model and methods, if needed.
- There are many more speech processing problems that could probably benefit from this approach!
- These include applications with multiple channels, adverse conditions or where the fine details matter.

Acknowledgments

Thanks to our collaborators:

- Søren Holdt Jensen
- Andreas Jakobsson
- Petra Stoica
- Tobias Lindstrøm Jensen
- Mathew Shaji Kavalekalam
- Charlotte Sørensen
- Jesper Boldt
- Sidsel Marie Nørholm
- Sam Karimian-Azari
- Liming Shi

I. Single- and Multi-Pitch estimation

[1] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen. "Estimating Multiple Pitches Using Block Sparsity". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2013, pp. 6220–6224.

[2] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen. "Multi-Pitch Estimation Exploiting Block Sparsity". In: *Signal Processing* 109 (Apr. 2015), pp. 236–247.

[3] L. Armani and M. Omologo. "Weighted autocorrelation-based f0 estimation for distant-talking interaction with a distributed microphone network". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Vol. 1. 2004, pp. 113–116.

[4] R. Badeau, V. Emiya, and B. David. "Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2009, pp. 3073–3076.

[5] D. Chazan, Y. Stettiner, and D. Malah. "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Vol. 2. 1993, pp. 728–731. DOI: 10.1109/ICASSP.1993.319415.

[6] A. de Cheveignï£¡ï£¡ and H. Kawahara. "YIN, a fundamental frequency estimator for speech and music". In: *J. Acoust. Soc. Am.* 111(4) (Apr. 2002), pp. 1917–1930.

[7] M. G. Christensen. "A Method for Low-Delay Pitch Tracking and Smoothing". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2012, pp. 345–348.

[8] M. G. Christensen. "Accurate Estimation of Low Fundamental Frequencies". In: *IEEE Trans. Audio, Speech, Language Process.* 21(10) (2013), pp. 2042–2056.

[9] M. G. Christensen. "An Exact Subspace Method For Fundamental Frequency Estimation". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2013, pp. 6802–6806.

[10] M. G. Christensen. "Multi-Channel Maximum Likelihood Pitch Estimation". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2012, pp. 409–412.

[11] E. Conte, A. Filippi, and S. Tomasin. "ML Period Estimation With Application to Vital Sign Monitoring". In: *IEEE Signal Process. Lett.* 17.11 (2010), pp. 905–908. DOI: 10.1109/LSP.2010.2071382.

[12] M. Davy, S. Godsill, and J. Idier. "Bayesian Analysis of Western Tonal Music". In: *J. Acoust. Soc. Am.* 119(4) (Apr. 2006), pp. 2498–2517.

[13] Manuel Davy. "Multiple Fundamental Frequency Estimation Based on Generative Models". In: *Signal Processing Methods for Music Transcription*. 2006, pp. 203–227.

[14] D. J. Hermes. "Measurement of pitch by subharmonic summation". In: *J. Acoust. Soc. Am.* 83(1) (1988), pp. 257–264.

[15] W. Hess. "Pitch and Voicing Determination". In: *Advances in Speech Signal Processing*. Ed. by S. Furui and M. M. Sohndi. Marcel Dekker, New York, 1992, pp. 3–48.

[16] W. Hess. *Pitch Determination of Speech Signals*. Springer-Verlag, Berlin, 1983.

[17] J. K. Nielsen, M. G. Christensen, and S. H. Jensen. "An Approximate Bayesian Fundamental Frequency Estimator". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2012, pp. 4617–4620.

[18] J. K. Nielsen, M. G. Christensen, and S. H. Jensen. "Default Bayesian Estimation of the Fundamental Frequency". In: *IEEE Trans. Audio, Speech, Language Process.* 21(3) (2013), pp. 598–610.

[19] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen. "A Fast Algorithm for Maximum Likelihood-based Fundamental Frequency Estimation". In: *Proc. European Signal Processing Conf.* 2015.

[20] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen. "Fast and Statistically Efficient Fundamental Frequency Estimation". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2016, pp. 86–90.

[21] J. R. Jensen, J. K. Nielsen, M. G. Christensen, S. H., Jensen and T. Larsen. "On Fast Implementation of Harmonic MUSIC for Known and Unknown Model Orders". In: *Proc. European Signal Processing Conf.* 2008.

[22] J. R. Jensen, M. G. Christensen and S. H. Jensen. "A Single Snapshot Optimal Filtering Method for Fundamental Frequency Estimation". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2011, pp. 4272–4275.

[23] J. R. Jensen, M. G. Christensen, and S. H. Jensen. "Fundamental Frequency Estimation using Polynomial Rooting of a Subspace-Based Method". In: *Proc. European Signal Processing Conf.* 2010.

[24] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen. "A Robust and Computationally Efficient Subspace-based Fundamental Frequency Estimator". In: *IEEE Trans. Audio, Speech, Language Process.* 18(3) (2010), pp. 487–497.

[25] J. R. Jensen et al. "Fast LCMV-based Methods for Fundamental Frequency Estimation". In: *IEEE Trans. Signal Process.* 61(12) (2013), pp. 3159–3172.

[26] A. Klapuri. "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness". In: *IEEE Trans. Speech Audio Process.* 11(6) (2003), pp. 804–816.

[27] D. Kundu and S. Nandi. "A note on estimating the fundamental frequency of a periodic function". In: *Signal Processing* 84 (2004), pp. 653–661.

[28] H. Li, P. Stoica, and J. Li. "Computationally efficient parameter estimation for harmonic sinusoidal signals". In: *Signal Processing* 80 (2000), pp. 1937–1944.

[29] M. G. Christensen, A. Jakobsson and S. H. Jensen. "Fundamental Frequency Estimation using the Shift-Invariance Property". In: *Rec. Asilomar Conf. Signals, Systems, and Computers*. 2007, pp. 631–635.

[30] M. G. Christensen, A. Jakobsson and S. H. Jensen. “Joint High-Resolution Fundamental Frequency and Order Estimation”. In: *IEEE Trans. Audio, Speech, Language Process.* 15(5) (2007), pp. 1635–1644.

[31] M. G. Christensen, A. Jakobsson and S. H. Jensen. “Multi-Pitch Estimation using Harmonic MUSIC”. In: *Rec. Asilomar Conf. Signals, Systems, and Computers*. 2006, pp. 521–525.

[32] M. G. Christensen and A. Jakobsson. “Improved Subspace-based Frequency Estimation for Real-Valued Data using Angles between Subspaces”. In: *Proc. European Signal Processing Conf.* 2010.

[33] M. G. Christensen and A. Jakobsson. *Multi-Pitch Estimation*. Vol. 5. Synthesis Lectures on Speech & Audio Processing. 160 pages. Morgan & Claypool Publishers, 2009, p. 160. ISBN: 9871598298383.

[34] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen. “Joint Fundamental Frequency and Order Estimation using Optimal Filtering”. In: *EURASIP J. on Advances in Signal Process.* 2011(1) (2011), pp. 1–13. ISSN: 1687-6180.

[35] M. G. Christensen, P. Stoica, A. Jakobsson and S. H. Jensen. “Multi-Pitch Estimation”. In: *Signal Processing* 88(4) (Apr. 2008), pp. 972–983.

[36] M. G. Christensen, P. Stoica, A. Jakobsson and S. H. Jensen. “The Multi-Pitch Estimation Problem: Some New Solutions”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Vol. 3. 2007, pp. 1221–1224.

[37] M. G. Christensen, P. Vera-Candeas, S. D. Somasundaram and A. Jakobsson. “Robust Subspace-based Fundamental Frequency Estimation”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2008, pp. 101–104.

[38] M. G. Christensen, S. H. Jensen, S. V. Andersen and A. Jakobsson. “Subspace-based Fundamental Frequency Estimation”. In: *Proc. European Signal Processing Conf.* 2004, pp. 637–640.

[39] M. W. Hansen, J. R. Jensen, and M. G. Christensen. “Estimation of Multiple Pitches in Stereophonic Mixtures using a Codebook-based Approach”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2017.

[40] M. W. Hansen, J. R. Jensen, and M. G. Christensen. “Multi-Pitch Estimation of Audio Recordings Using a Codebook-Based Approach”. In: *Proc. European Signal Processing Conf.* 2016.

[41] M. W. Hansen, J. R. Jensen, and M. G. Christensen. “Pitch Estimation of Stereophonic Mixtures of Delay and Amplitude Panned Signals”. In: *Proc. European Signal Processing Conf.* 2015.

[42] Y. Medan, E. Yair, and D. Chazan. “Super resolution pitch determination of speech signals”. In: *IEEE Trans. Signal Process.* 39.1 (1991), pp. 40–48. DOI: 10.1109/78.80763.

[43] A. Nehorai and B. Porat. “Adaptive Comb Filtering for Harmonic Signal Enhancement”. In: *IEEE Trans. Acoust., Speech, Signal Process.* 34(5) (Oct. 1986), pp. 1124–1138.

[44] J. K. Nielsen et al. “Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient”. In: *Signal Processing* 135 (2017), pp. 188–197.

[45] A. M. Noll. “Cepstrum pitch determination”. In: *J. Acoust. Soc. Am.* 41(2) (1967), pp. 293–309.

[46] M. Noll. “Pitch Determination of Human Speech by Harmonic Product Spectrum, the harmonic sum, and a maximum likelihood estimate”. In: *Proc. Symposium on Computer Processing Communications*. 1969, pp. 779–797.

[47] G. Ogden et al. “Frequency domain trackin of passive vessel harmonics”. In: *J. Acoust. Soc. Am.* 126 (2009), p. 2249.

[48] L. Rabiner. “On the use of autocorrelation analysis for pitch detection”. In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 25.1 (1977), pp. 24–33.

[49] M. Ross et al. “Average magnitude difference function pitch extractor”. In: *IEEE Trans. Acoust., Speech, Signal Process.* 22.5 (Oct. 1974), pp. 353–362.

[50] S. Karimian-Azari, A. Jakobsson, J. R. Jensen, and M. G. Christensen. “Multi-Pitch Estimation and Tracking Using Bayesian Inference in Block Sparsity”. In: *Proc. European Signal Processing Conf.* 2015.

[51] S. Karimian-Azari, J. R. Jensen and M. G. Christensen. “Robust Pitch Estimation Using an Optimal Filter on Frequency Estimates”. In: *Proc. European Signal Processing Conf.* 2014, pp. 1557–1561.

[52] S. Karimian-Azari, N. Mohammadiha, J. R. Jensen and M. G. Christensen. “Pitch Estimation and Tracking with Harmonic Emphasis On The Acoustic Spectrum”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2015, pp. 4330–4334.

[53] H.C. So et al. “Linear prediction approach for efficient frequency estimation of multiple real sinusoids: algorithms and analyses”. In: *IEEE Trans. Signal Process.* 53.7 (2005), pp. 2290–2305. DOI: 10.1109/TSP.2005.849154.

[54] J. Tabrikian, S. Dubnov, and Y. Dickalov. “Maximum a posteriori probability pitch tracking in noisy environments using harmonic model”. In: *IEEE Trans. Audio, Speech, Language Process.* 12(1) (2004), pp. 76–87.

[55] D. Talkin. “A robust algorithm for pitch tracking (RAPT)”. In: *Speech Coding and Synthesis*. Ed. by W. B. Kleijn and K. K. Paliwal. Elsevier Science B.V., 1995. Chap. 5, pp. 495–518.

II. Harmonic Chirp Estimation

[1] Y. Doweck, A. Amar, and I. Cohen. “Joint Model Order Selection and Parameter Estimation of Chirps With Harmonic Components”. In: *IEEE Trans. Signal Process.* 63.7 (2015), pp. 1765–1778. ISSN: 1053-587X. DOI: 10.1109/TSP.2015.2391075.

[2] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen. “Fast Harmonic Chirp Summation”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2017.

[3] T. L. Jensen et al. “A Fast Algorithm for Maximum Likelihood Estimation of Harmonic Chirp Parameters”. In: *IEEE Trans. Signal Process.* 65.19 (2017).

[4] M. G. Christensen and J. R. Jensen. “Pitch Estimation for Non-Stationary Speech”. In: *Rec. Asilomar Conf. Signals, Systems, and Computers*. 2014, pp. 1400–1404.

[5] F. Myburg, A. C. den Brinker, and S. van Eijndhoven. “Multi-Component Chirp Analysis in Parametric Audio Coding”. In: *Fourth IEEE Benelux Signal Processing Symposium*. 2004.

[6] Y. Pantazis, O. Rosec, and Y. Stylianou. “Chirp rate estimation of speech based on a time-varying quasi-harmonic model”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2009, pp. 3985–3988. DOI: 10.1109/ICASSP.2009.4960501.

[7] S. M. Nørholm, J. R. Jensen, and M. G. Christensen. “Instantaneous Pitch Estimation with Optimal Segmentation for Non-Stationary Voiced Speech”. In: *IEEE Trans. Audio, Speech, Language Process.* 24(12) (2016), pp. 2354–2367.

III. Speech Enhancement

[1] J. Chen et al. “New insights into the noise reduction Wiener filter”. In: *IEEE Trans. Audio, Speech, Language Process.* 14.4 (2006), pp. 1218–1234. ISSN: 1558-7916. DOI: 10.1109/TSA.2005.860851.

[2] M. G. Christensen et al. “Spatio-Temporal Filtering Methods for Enhancement and Separation of Speech Signals”. In: *Proc. IEEE China Summit & Int. Conf. on Signal and Information Process.* 2013, pp. 303–307.

[3] B. Cornelis et al. “Theoretical Analysis of Binaural Multimicrophone Noise Reduction Techniques”. In: *IEEE Trans. Audio, Speech, Language Process.* 18.2 (2010), pp. 342–355. ISSN: 1558-7916. DOI: 10.1109/TASL.2009.2028374.

[4] S. Doclo and M. Moonen. “GSVD-based optimal filtering for single and multimicrophone speech enhancement”. In: *IEEE Trans. Signal Process.* 50(9) (2002), pp. 2230–2244. DOI: 10.1109/TSP.2002.801937.

[5] J. R. Jensen, J. Benesty, and M. G. Christensen. “Noise Reduction with Optimal Variable Span Linear Filter”. In: *IEEE Trans. Audio, Speech, Language Process.* 24(4) (2016), pp. 631–644.

[6] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen. “Enhancement of Single-Channel Periodic Signals in the Time-Domain”. In: *IEEE Trans. Audio, Speech, Language Process.* 20(7) (2012), pp. 1948–1963.

[7] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen. “Non-Causal Time-Domain Filters for Single-Channel Noise Reduction”. In: *IEEE Trans. Audio, Speech, Language Process.* 20(5) (2012), pp. 1526–1541.

[8] J. R. Jensen, M. G. Christensen, and A. Jakobsson. “Harmonic Minimum Mean Squared Error Filters for Multichannel Speech Enhancement”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2017.

[9] J. R. Jensen, M. G. Christensen, and S. H. Jensen. “An Optimal Spatio-Temporal Filter for Extraction and Enhancement of Multi-Channel Periodic Signals”. In: *Rec. Asilomar Conf. Signals, Systems, and Computers*. 2010, pp. 1846–1850.

[10] J. R. Jensen, M. G. Christensen, J. Benesty, and S. H. Jensen. “Joint Filtering Scheme for Nonstationary Noise Reduction”. In: *Proc. European Signal Processing Conf.* 2012, pp. 2323–2327.

[11] J. Jensen and J. H. L. Hansen. “Speech enhancement using a constrained iterative sinusoidal model”. In: *IEEE Trans. Speech Audio Process.* 9 (Oct. 2001), pp. 731–740.

[12] M. Krawczyk and T. Gerkmann. “STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement”. In: *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 22.12 (2014), pp. 1931–1940. ISSN: 2329-9290. DOI: 10.1109/TASLP.2014.2354236.

[13] Junfeng Li et al. “Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication”. In: *Speech Communication* 53.5 (2011). Perceptual and Statistical Audition, pp. 677–689. ISSN: 0167-6393. DOI: http://dx.doi.org/10.1016/j.specom.2010.04.009. URL: http://www.sciencedirect.com/science/article/pii/S0167639310000981.

[14] M. G. Christensen and A. Jakobsson. “Optimal Filter Designs for Separating and Enhancing Periodic Signals”. In: *IEEE Trans. Signal Process.* 58(12) (2010), pp. 5969–5983.

[15] M. G. Christensen and A. Jakobsson. “Optimal Filters for Extraction and Separation of Periodic Sources”. In: *Rec. Asilomar Conf. Signals, Systems, and Computers*. 2009, pp. 376–379.

[16] M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt. “Binaural Speech Enhancement using a Codebook based Approach”. In: *Proc. Int. Workshop on Acoustic Signal Enhancement*. 2016.

[17] M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt. “Model based Binaural Enhancement of Voiced and Unvoiced Speech”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2017.

[18] S. M. Nørholm, J. R. Jensen, and M. G. Christensen. “On the Influence of Inharmonicities in Model-Based Speech Enhancement”. In: *Proc. European Signal Processing Conf.* 2013, pp. 1–5.

[19] S. M. Nørholm, J. R. Jensen, and M. G. Christensen. “Spatio-Temporal Audio Enhancement Based on IAA Noise Covariance Matrix Estimates”. In: *Proc. European Signal Processing Conf.* 2014, pp. 934–938.

[20] P. Mowlaee, M. G. Christensen, and S. H. Jensen. “New Results on Single-Channel Speech Separation Using Sinusoidal Modeling”. In: *IEEE Trans. Audio, Speech, Language Process.* 19(5) (2011), pp. 1265–1277.

[21] K. Paliwal and A. Basu. “A speech enhancement method based on Kalman filtering”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Vol. 12. 1987, pp. 177 –180. DOI: 10.1109/ICASSP.1987.1169756.

[22] S. Karimian-Azari, J. Benesty, J. R. Jensen and M. G. Christensen. "A Broadband Beamformer Using Controllable Constraints and Minimum Variance". In: *Proc. European Signal Processing Conf.* 2014, pp. 666–670.

[23] S. M. Nørholm, J. R. Jensen, and M. G. Christensen. "Enhancement and Noise Statistics Estimation for Non-Stationary Voiced Speech". In: *IEEE Trans. Audio, Speech, Language Process.* 24(4) (2016), pp. 645–658.

[24] S. Srinivasan, J. Samuelsson, and W. B. Kleijn. "Codebook-Based Bayesian Speech Enhancement for Nonstationary Environments". In: *IEEE Trans. Audio, Speech, Language Process.* 15.2 (2007), pp. 441–452. ISSN: 1558-7916. DOI: 10.1109/TASL.2006.881696.

[25] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, and J. Benesty. "A Framework for Speech Enhancement with Ad Hoc Microphone Arrays". In: *IEEE Trans. Audio, Speech, Language Process.* 24(6) (2016), pp. 1038–1051.

IV. TDOA, DOA, and Source Localisation

[1] P. Bestagini et al. "TDOA-based acoustic source localization in the space–range reference frame". In: *Multidimensional Systems and Signal Process.* 25.2 (2014), pp. 337–359. ISSN: 1573-0824. DOI: 10.1007/s11045-013-0233-8. URL: https://doi.org/10.1007/s11045-013-0233-8.

[2] S. T. Birchfield and R. Gangishetty. "Acoustic localization by interaural level difference". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Vol. 4. 2005, pp. 1109–1112. DOI: 10.1109/ICASSP.2005.1416207.

[3] M. S. Brandstein. "On the use of explicit speech modeling in microphone array applications". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Vol. 6. 1998, pp. 3613–3616.

[4] M. S. Brandstein and H. F. Silverman. "A robust method for speech signal time-delay estimation in reverberant rooms". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Vol. 1. 1997, pp. 375–378. DOI: 10.1109/ICASSP.1997.599651.

[5] K. C. Ho and M. Sun. "Passive Source Localization Using Time Differences of Arrival and Gain Ratios of Arrival". In: *IEEE Trans. Signal Process.* 56.2 (2008), pp. 464–477. ISSN: 1053-587X. DOI: 10.1109/TSP.2007.906728.

[6] J. R. Jensen, J. K. Nielsen, M. G. Christensen and S. H. Jensen. "On Frequency Domain Models for TDOA Estimation". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2015, pp. 11–15.

[7] J. R. Jensen, J. K. Nielsen, R. Heusdens, and M. G. Christensen. "DOA Estimation of Audio Sources in Reverberant Environments". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2016, pp. 176–80.

[8] J. R. Jensen, M. G. Christensen, and S. H. Jensen. "Joint DOA and Fundamental Frequency Estimation Methods based on 2-D Filtering". In: *Proc. European Signal Processing Conf.* 2010.

[9] J. R. Jensen, M. G. Christensen, and S. H. Jensen. "Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation". In: *IEEE Trans. Audio, Speech, Language Process.* 21(5) (2013), pp. 923–933.

[10] J. R. Jensen, M. G. Christensen, J. Benesty and S. H. Jensen. "Joint Spatio-Temporal Filtering Methods for DOA and Fundamental Frequency Estimation". In: *IEEE Trans. Audio, Speech, Language Process.* 23(1) (2015), pp. 174–185.

[11] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen. "Joint DOA and Multi-Pitch Estimation based on Subspace Techniques". In: *EURASIP J. on Advances in Signal Process.* 2012(1) (2012), pp. 1–11.

[12] J. R. Jensen and M. G. Christensen. "DOA and Pitch Estimation of Audio Sources using IAA-based Filtering". In: *Proc. European Signal Processing Conf.* 2014, pp. 900–904.

[13] J. R. Jensen and M. G. Christensen. "Near-field Localization of Audio: A Maximum Likelihood Approach". In: *Proc. European Signal Processing Conf.* 2014, pp. 895–899.

[14] J. R. Jensen, M. G. Christensen, and S. H. Jensen. "Statistically Efficient Methods for Pitch and DOA Estimation". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2013, pp. 3900–3904.

[15] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen. "Fast Joint DOA and Pitch Estimation Using a Broadband MVDR Beamformer". In: *Proc. European Signal Processing Conf.* 2013, pp. 1–5.

[16] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen. "Fundamental Frequency and Model Order Estimation Using Spatial Filtering". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2014, pp. 5964–5968.

[17] M. Képesi, L. Ottowitz, and T. Habib. "Joint Position-Pitch Estimation for Multiple Speaker Scenarios". In: *Proc. Hands-Free Speech Communication and Microphone Arrays*. 2008, pp. 85–88. DOI: 10.1109/HSCMA.2008.4538694.

[18] C. Knapp and G. Carter. "The generalized correlation method for estimation of time delay". In: *IEEE Trans. Acoust., Speech, Signal Process.* 24.4 (1976), pp. 320–327. ISSN: 0096-3518.

[19] M. W. Hansen, J. R. Jensen and M. G. Christensen. "Localizing Near and Far Field Acoustic Sources with Distributed Microphone Arrays". In: *Rec. Asilomar Conf. Signals, Systems, and Computers*. 2014, pp. 491–495.

[20] M. W. Hansen, J. R. Jensen and M. G. Christensen. "Pitch and TDOA-based Localization of Acoustics Sources with Distributed Arrays". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2015, pp. 2664–2668.

[21] L. Y. Ngan et al. "Joint time delay and pitch estimation for speaker localization". In: *Proc. IEEE Int. Symp. Circuits and Systems*. Vol. 3. 2003, pp. 722–725. DOI: 10.1109/ISCAS.2003.1205121.

[22] X. Qian and R. Kumaresan. "Joint Estimation of Time Delay and Pitch of Voiced Speech Signals". In:

Rec. Asilomar Conf. Signals, Systems, and Computers (1996).

[23] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen. "Computationally Efficient and Noise Robust DOA and Pitch Estimation". In: *IEEE Trans. Audio, Speech, Language Process.* 24(9) (2016), pp. 1613–1625.

[24] S. Karimian-Azari, J. R. Jensen and M. G. Christensen. "Robust DOA Estimation of Harmonic Signals Using Constrained Filters on Phase Estimates". In: *Proc. European Signal Processing Conf.* 2014, pp. 1930–1934.

[25] X. Sheng and Y.-H. Hu. "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks". In: *IEEE Trans. Signal Process.* 53.1 (2005), pp. 44–53. ISSN: 1053-587X. DOI: 10.1109/TSP.2004.838930.

[26] M. Wohlmayr and M. Képesi. "Joint Position-Pitch Extraction from Multichannel Audio". In: *Proc. Interspeech*. 2007, pp. 1629–1632.

[27] Y. Wu et al. "Joint Pitch and DOA Estimation Using the ESPRIT Method". In: *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 23.1 (2015), pp. 32–45. ISSN: 2329-9290. DOI: 10.1109/TASLP.2014.2367817.

[28] Z. Zhou, M. G. Christensen, and H. C. So. "Two Stage DOA and Fundamental Frequency Estimation Based on Subspace Techniques". In: *Proc. IEEE Int. Conf. Signal Processing*. 2012, pp. 210–213.

[29] Z. Zhou et al. "Joint DOA and fundamental frequency estimation based on relaxed iterative adaptive approach and optimal filtering". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2013, pp. 6812–6816. DOI: 10.1109/ICASSP.2013.6638981.

V. Speech and Audio Coding

[1] L. Almeida and J. Tribolet. "Harmonic coding: A low bit-rate, good-quality speech coding technique". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Vol. 7. 1982, pp. 1664–1667.

[2] L. B. Almeida and F. M. Silva. "Variable-Frequency Synthesis: An Improved Harmonic Coding Scheme". In: *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing*. Dec. 1984, 27.5.1.

[3] B. S. Atal and J. R. Remde. "A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 1982.

[4] M. G. Christensen. "Metrics for Vector Quantization-based Parametric Speech Enhancement and Separation". In: *J. Acoust. Soc. Am.* 133(5) (2013), pp. 3062–3071.

[5] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen. "Sparse Linear Predictors for Speech Processing". In: *Proc. Interspeech*. 2008.

[6] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen. "Sparse Linear Prediction and Its Applications to Speech Processing". In: *IEEE Trans. Audio, Speech, Language Process.* 20(5) (2012), pp. 1644–1657.

[7] D. Giacobello, T. van Waterschoot, M. G. Christensen, S. H. Jensen, and M. Moonen. "High-Order Sparse Linear Predictors for Audio Processing". In: *Proc. European Signal Processing Conf.* 2010.

[8] D. Giacobello et al. "Stable 1-norm error minimization based linear predictors for speech modeling". In: *IEEE Trans. Audio, Speech, Language Process.* 22(5) (2014), pp. 912–922.

[9] P. Hedelin. "A Sinusoidal LPC Vocoder". In: *IEEE Workshop on Speech Coding*. 2000, 2–4.

[10] P. Hedelin. "A tone oriented voice excited vocoder". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 1981, pp. 205–208.

[11] R. Heusdens et al. "Bit-Rate Scalable Intraframe Sinusoidal Audio Coding Based on Rate-Distortion Optimization". In: *J. Audio Eng. Soc.* (Mar. 2006), pp. 167–188.

[12] P. Korten, J. Jensen, and R. Heusdens. "High-Resolution Spherical Quantization of Sinusoidal Parameters". In: *IEEE Trans. Audio, Speech, Language Process.* 15(3) (2007), pp. 966–981.

[13] J. Lindblom. "A Sinusoidal Voice Over Packet Coder Tailored for the Frame-Erasure Channel". In: (2004).

[14] J. Lindblom. "Coding Speech for Packet Networks". PhD thesis. Chalmers University of Technology, 2003.

[15] M. G. Christensen. "Estimation and Modeling Problems in Parametric Audio Coding". Award: Spar Nord Foundation's Research Prize. PhD thesis. Aalborg University, July 2005. ISBN: 87-90834-80-1.

[16] M. G. Christensen. "On Perceptual Distortion Measures and Parametric Modeling". In: *In Proc. Acoustics'08 Paris*. 2008.

[17] M. H. Larsen, M. G. Christensen and S. H. Jensen. "Variable Dimension Trellis-Coded Quantization of Sinusoidal Parameters". In: *IEEE Signal Process. Lett.* 15 (2008), pp. 17–20.

[18] R. J. McAulay and T. F. Quatieri. "Sinusoidal Coding". In: *Speech Coding and Synthesis*. Ed. by W. B. Kleijn and K. K. Paliwal. Elsevier Science B.V., 1995. Chap. 4.

[19] K. K. Paliwal and B. S. Atal. "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame". In: *IEEE Trans. Speech Audio Process.* 1 (1993), pp. 3–14.

[20] H. Purnhagen and N. Meine. "HILN - The MPEG-4 Parametric Audio Coding Tools". In: *IEEE International Symposium on Circuits and Systems*. 2000.

[21] T. L. Jensen, D. Giacobello, T. van Waterschoot, and M. G. Christensen. "Fast Algorithms for High-Order Sparse Linear Prediction with Applications to Speech Processing". In: *Speech Communication* 76 (Feb. 2016), pp. 143–156.

[22] K. Vos et al. "High-Quality Consistent Analysis-Synthesis in Sinusoidal Coding". In: *Proc. Audio Eng. Soc. 17th Conf: High Quality Audio Coding*. 1999, pp. 244–250.

VI. Model Comparison

[1] P. M. Djuric. “Asymptotic MAP Criteria for Model Selection”. In: *IEEE Trans. Signal Process.* 46 (Oct. 1998), pp. 2726–2735.

[2] E. J. Hannan. “Developments in Time Series Analysis”. In: Chapman and Hall, 1993. Chap. Determining the number of jumps in a spectrum, pp. 127–138.

[3] L. Kavalieris and E. J. Hannan. “Determining the number of terms in a trigonometric regression”. In: *J. on Time Series Analysis* 15(6) (1994), pp. 613–625.

[4] M. G. Christensen, A. Jakobsson, and S. H. Jensen. “Sinusoidal Order Estimation using Angles between Subspaces”. In: *EURASIP J. on Advances in Signal Process.* (2009). Article ID 948756, pp. 1–11.

[5] M. G. Christensen, A. Jakobsson and S. H. Jensen. “Sinusoidal Order Estimation using the Subspace Orthogonality and Shift-Invariance Properties”. In: *Rec. Asilomar Conf. Signals, Systems, and Computers*. 2007, pp. 651–655.

[6] M. G. Christensen and S. H. Jensen. “Variable Order Harmonic Sinusoidal Parameter Estimation for Speech and Audio Signals”. In: *Rec. Asilomar Conf. Signals, Systems, and Computers*. 2006, pp. 1126–1130.

[7] J. K. Nielsen, M. G. Christensen, and S. H. Jensen. “Bayesian Model Comparison and the BIC for Regression Models”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2013, pp. 6362–6366.

[8] J. K. Nielsen, M. G. Christensen, and S. H. Jensen. “Model Selection and Comparison for Independents Sinusoids”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2014, pp. 1891–1895.

[9] J. K. Nielsen et al. “Bayesian Model Comparison with the g-Prior”. In: *IEEE Trans. Signal Process.* 62(1) (2014), pp. 225–238.

[10] B. G. Quinn. “Estimating the number of terms in a sinusoidal regression”. In: *J. on Time Series Analysis* 10(1) (1989), pp. 71–75.

[11] P. Stoica and Y. Selen. “Model-order selection: a review of information criterion rules”. In: *IEEE Signal Process. Mag.* 21(4) (July 2004), pp. 36–47.

VII. Speech and Audio Modeling

[1] L. B. Almeida and J. M. Tribolet. “A Model for Short-Time Phase Prediction of Speech”. In: *Proc. Int. Conf. Acoust., Speech, Signal Processing*. 1981, pp. 213–216.

[2] R. Boyer and K. Abed-Meraim. “Audio Modeling Based on Delayed Sinusoids”. In: *IEEE Trans. Speech Audio Process.* 12(2) (Mar. 2004), pp. 110 –120.

[3] D. Clark et al. “Multi-Object Tracking of Sinusoidal Components in Audio with the Gaussian Mixture Probability Hypothesis Density Filter”. In: *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2007, pp. 339–342. DOI: 10.1109/ASPAA.2007.4393009.

[4] B. David and R. Badeau. “Fast Sequential LS Estimation for Sinusoidal Modeling and Decomposition of Audio Signals”. In: *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.* 2007, pp. 211–214.

[5] A. El-Jaroudi and J. Makhoul. “Discrete All-Pole Modeling”. In: *IEEE Trans. Signal Process.* 39 (1991), pp. 411–423.

[6] E. B. George and M. J. T. Smith. “Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones”. In: *J. Audio Eng. Soc.* 40(6) (1992), pp. 497–516.

[7] E. B. George and M. J. T. Smith. “Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model”. In: *IEEE Trans. Speech Audio Process.* 5(5) (Sept. 1997), pp. 389–406.

[8] S. Godsill and M. Davy. “Bayesian Computational Models for Inharmonicity in Musical Instruments”. In: *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.* 2005, pp. 283–286.

[9] S. Godsill and M. Davy. “Bayesian harmonic models for musical pitch estimation and analysis”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2002.

[10] M. M. Goodwin. “Adaptive Signal Models: Theory, Algorithms, and Audio Applications”. PhD thesis. University of California, Berkeley, 1997.

[11] M. M. Goodwin. “Residual Modeling in Music Analysis-Synthesis”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Vol. 2. 1996, pp. 1005–1008.

[12] R. Gribonval and E. Bacry. “Harmonic Decomposition of Audio Signals with Matching Pursuit”. In: *IEEE Trans. Signal Process.* Vol. 51(1). Jan. 2003.

[13] R. Heusdens, R. Vafin, and W. B. Kleijn. “Sinusoidal Modeling using Psychoacoustic-adaptive matching pursuits”. In: *IEEE Signal Process. Lett.* 9(8) (Aug. 2002), pp. 262–265.

[14] G. Li, L. Qiu, and L. K. Ng. “Signal Representation Based on Instantaneous Amplitude Models with Application to Speech Synthesis”. In: *IEEE Trans. Speech Audio Process.* 8(3) (2000), pp. 353–357.

[15] P. Maragos, J. F. Kaiser, and T. F. Quatieri. “Energy Separation in Signal Modulations with Application to Speech Analysis”. In: *IEEE Trans. Signal Process.* 41(10) (Oct. 1993), pp. 3024–3051.

[16] R. J. McAulay and T. F. Quatieri. “Speech Analysis/Synthesis Based on a Sinusoidal Representation”. In: *IEEE Trans. Acoust., Speech, Signal Process.* 34(4) (Aug. 1986), pp. 744–754.

[17] R. J. McAulay and T. F. Quatieri. “Speech Transformation Based on a Sinusoidal Representation”. In: *IEEE Trans. Acoust., Speech, Signal Process.* 34 (Dec. 1986), pp. 1449–1464.

[18] J. Nieuwenhuijse, R. Heusdens, and E. F. Deprettere. “Robust Exponential Modeling of Audio Signals”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 1998, pp. 3581–3584.

[19] M. R. Portnoff. “Time-frequency representation of digital signals and systems based on short-time Fourier analysis”. In: *IEEE Trans. Acoust., Speech, Signal Process.* 28 (Feb. 1980), pp. 55–69.

[20] M. R. Portnoff. “Time-Scale Modification of Speech Based on Short-Time Fourier Analysis”. In: *IEEE*

Trans. Acoust., Speech, Signal Process. 29 (June 1981), pp. 374–390.

[21] P. Prandoni. “Optimal Segmentation Techniques for Piecewise Stationary Signals”. PhD thesis. Ecole Polytechnique Federale de Lausanne, 1999.

[22] P. Prandoni, M. M. Goodwin, and M. Vetterli. “Optimal time segmentation for signal modeling and compression”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 1997, pp. 2029–2032.

[23] P. Prandoni and M. Vetterli. “R/D optimal linear prediction”. In: *IEEE Trans. Speech Audio Process.* (2000), pp. 646–655.

[24] C. A. Rødbro and S. H. Jensen. “Time-scaling of Sinusoids for Intelligent Jitter Buffer in Packet Based Telephony”. In: *Proc. IEEE Workshop on Speech Coding for Telecommunications*. 2002, pp. 71–73.

[25] B. Santhanam and P. Maragos. “Demodulation of Discrete Multicomponent AM-FM Signals using Periodic Algebraic Separation and Energy Demodulation”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 1997.

VIII. Estimation Theory

[1] G. Bienvenu. “Influence of the spatial coherence of the background noise on high resolution passive methods”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 1979, pp. 306–309.

[2] J. Capon. “High-resolution frequency-wavenumber spectrum analysis”. In: *Proc. IEEE* 57(8) (1969), pp. 1408–1418.

[3] M. Feder and E. Weinstein. “Parameter Estimation of Superimposed Signals using the EM Algorithm”. In: *IEEE Trans. Acoust., Speech, Signal Process.* 36(4) (Apr. 1988), pp. 477–489.

[4] O. L. Frost III. “An algorithm for linearly constrained adaptive array processing”. In: *Proc. IEEE* 60(8) (1972), pp. 926–935. ISSN: 0018-9219.

[5] R. M. Gray. “Toeplitz and Circulant Matrices: A review”. In: *Foundations and Trends in Communications and Information Theory* 2(3) (2006), pp. 155–239.

[6] E. J. Hannan and B. Wahlberg. “Convergence Rates for Inverse Toeplitz Matrix Forms”. In: *J. Multivariate Analysis* 31 (1989), pp. 127–135.

[7] A. Jakobsson. “Model-Based and Matched-Filterbank Signal Analysis”. PhD thesis. Uppsala University, Feb. 2000.

[8] K. Kim and G. Shevlyakov. “Why Gaussianity?” In: *IEEE Signal Process. Mag.* 25.2 (2008), pp. 102–113. DOI: 10.1109/MSP.2007.913700.

[9] J. Li and P. Stoica. “Efficient Mixed-Spectrum Estimation with Application to Target Feature EXtraction”. In: *IEEE Trans. Signal Process.* 44(2) (Feb. 1996), pp. 281–295.

[10] J. Makhoul. “Linear Prediction: A Tutorial Review”. In: *Proc. IEEE* 63(4) (Apr. 1975), pp. 561–580.

[11] S. L. Marple. “Computing the discrete-time "analytic" signal via FFT”. In: *IEEE Trans. Signal Process.* 47 (Sept. 1999), pp. 2600–2603.

[12] S. O. Rice. “Mathematical Analysis of Random Noise”. In: *The Bell Systems Technical Journal* 3 (1944), 282–332.

[13] R. Roy and T. Kailath. “ESPRIT - Estimation of Signal Parameters via Rotational Invariance Techniques”. In: *IEEE Trans. Acoust., Speech, Signal Process.* 22 (1989), pp. 353–362.

[14] T. Shu and X. Liu. “Robust and Computationally Efficient Signal-Dependent Method for Joint DOA and Frequency Estimation”. In: *EURASIP J. on Advances in Signal Processing* 2008 (2008), 16 pages.

[15] P. Stoica, A. Jakobsson, and J. Li. “Cisoid Parameter Estimation in the Colored Noise Case: Asymptotic Cramér-Rao Bound, Maximum Likelihood and Nonlinear Least-Squares”. In: *IEEE Trans. Signal Process.* 45 (1997), pp. 2048–2059.

[16] P. Stoica, H. Li, and J. Li. “Amplitude Estimation of Sinusoidal Signals: Survey, New Results and an Application”. In: *IEEE Trans. Signal Process.* 48(2) (Feb. 2000), pp. 338–352.

[17] P. Stoica and A. Nehorai. “MUSIC, Maximum Likelihood, and Cramer-Rao Bound”. In: *IEEE Trans. Acoust., Speech, Signal Process.* 37(5) (May 1989), pp. 720–741.

[18] P. Stoica and A. Nehorai. “MUSIC, Maximum Likelihood, and Cramer-Rao Bound; further results and comparisons”. In: *IEEE Trans. Acoust., Speech, Signal Process.* 38(12) (Dec. 1990), pp. 2140–2150.

[19] H. L. Van Trees. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, Inc., 2002.

[20] A. J. van der Veen, M. Vanderveen, and A. Paulraj. “Joint angle and delay estimation using shift invariance techniques”. In: *IEEE Trans. Signal Process.* 46(2) (1998), pp. 405–418.

[21] H. Wang and M. Kaveh. “Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources”. In: *IEEE Trans. Acoust., Speech, Signal Process.* 33.4 (1985), pp. 823–831. ISSN: 0096-3518.

[22] M.-Y. Zou, C. Zhenming, and R. Unbehauen. “Separation of Periodic Signals by using an algebraic method”. In: *Proc. IEEE Int. Symp. Circuits and Systems*. Vol. 5. 1991, pp. 2427–2430.

IX. Noise Tracking and Estimation

[1] I. Cohen. “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging”. In: *IEEE Trans. Audio, Speech, Language Process.* 11(5) (2003), pp. 466–475.

[2] D. Ealey, H. Kelleher, and D. Pearce. “Harmonic tunnelling: tracking non-stationary noises during speech”. In: *EuroSpeech*. 2001, pp. 437–440.

[3] T. Gerkmann and R. C. Hendriks. “Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay”. In: *IEEE Trans. Audio, Speech, Language Process.* 20(4) (2012), pp. 1383–1393.

X. Spectral Estimation and Frequency Estimation

[1] D. Huang. “On low and high frequency estimation”. In: *J. of Time Series Analysis* 17(4) (1996), pp. 351–365.

[2] J. H. Jensen, M. G. Christensen, and S. H. Jensen. “An Amplitude and Covariance Matrix Estimator for Signals in Colored Gaussian Noise”. In: *Proc. European Signal Processing Conf.* 2009, pp. 2485–2488.

[3] J. K. Nielsen, P. Smaragdis, M. G. Christensen, and S. H. Jensen. “An Amplitude Spectral Capon Estimator with a Variable Filter Length”. In: *Proc. European Signal Processing Conf.* 2012, pp. 430–434.

[4] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen. “Grid Size Selection for Nonlinear Least-Squares Optimization in Spectral Estimation and Array Processing”. In: *Proc. European Signal Processing Conf.* 2016.

[5] J. X. Zhang, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen. “Frequency-Domain Parameter Estimations for Binary Masked Signals”. In: *Proc. Interspeech*. 2008.

[6] L. Shi, J. R. Jensen, and M. G. Christensen. “Least 1-Norm Pole-Zero Modeling with Sparse Deconvolution for Speech Analysis.” In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2017.

[7] B. G. Quinn and J. M. Fernandes. “A Fast Efficient Technique for the Estimation of Frequency”. In: *Biometrika* 78(3) (Sept. 1991), pp. 489–497.

[8] B. G. Quinn and E. J. Hannan. *The Estimation and Tracking of Frequency*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2001.

[9] B. G. Quinn and P. J. Thomson. “Estimating the frequency of a periodic function”. In: *Biometrika* 78(1) (1991), pp. 65–74.

[10] S. Tretter. “Estimating the frequency of a noisy sinusoid by linear regression (Corresp.)” In: *IEEE Trans. Inf. Theory* 31.6 (1985), pp. 832–835. ISSN: 0018-9448. DOI: 10.1109/TIT.1985.1057115.

[11] Z. Zhou, H. C. So, and M. G. Christensen. “Parametric Modeling for Damped Sinusoids from Multiple Channels”. In: *IEEE Trans. Signal Process.* 61(15) (2013), pp. 3895–3907.

XI. Adaptive Feedback Cancellation

[1] K. Ngo, T. van Waterschoot, M. G. Christensen, M. Moonen, S. H. Jensen, and J. Wouters. “Adaptive Feedback Cancellation in Hearing Aids using a Sinusoidal Near-End Model”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2010, pp. 181–184.

[2] K. Ngo, T. van Waterschoot, M. G. Christensen, M. Moonen, S. H. Jensen, J. Wouters. “Prediction-Error-Method-based Adaptive Feedback Cancellation in Hearing Aids using Pitch Estimation”. In: *Proc. European Signal Processing Conf.* 2010.

[3] K. Ngo et al. “Improved Prediction Error Filters for Adaptive Feedback Cancellation in Hearing Aids”. In: *Signal Processing* 91(11) (2013), pp. 3062–3075.

XII. Packet-loss Concealment

[1] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, S. J. Godsill, and S. H. Jensen. “Bayesian Interpolation and Parameter Estimation in a Dynamic Sinusoidal Model”. In: *IEEE Trans. Audio, Speech, Language Process.* 19(7) (2011), pp. 1986–1998.

[2] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, S. J. Godsill, and S. H. Jensen. “Bayesian Interpolation in a Dynamic Sinusoidal Model with Application to Packet-loss Concealment”. In: *Proc. European Signal Processing Conf.* 2010.

XIII. Speech Intelligibility

[1] C. Sørensen, A. Xenaki, J. B. Boldt, and M. G. Christensen. “Pitch-Based Non-Instrusive Intelligibility Prediction”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2017.

[2] F. Chen, O. Hazrati, and P. C. Loizou. “Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure”. In: *Biomedical Signal Processing and Control* 8.3 (2013), pp. 311–314. ISSN: 1746-8094.

[3] T. H. Falk, C. Zheng, and W. Y. Chan. “A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech”. In: *IEEE Trans. Audio, Speech, Language Process.* 18.7 (2010), pp. 1766–1774. ISSN: 1558-7916.

[4] M. Karbasi, A. H. Abdelaziz, and D. Kolossa. “Twin-HMM-based non-intrusive speech intelligibility prediction”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 2016, pp. 624–628. DOI: 10.1109/ICASSP.2016.7471750.

[5] C. H. Taal et al. “An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech”. In: *IEEE Trans. Audio, Speech, Language Process.* 19.7 (2011), pp. 2125–2136. ISSN: 1558-7916. DOI: 10.1109/TASL.2011.2114881.

XIV. Databases

[1] J. K. Nielsen, J. R. Jensen, S. H. Jensen and M. G. Christensen. “The Single- and Multichannel Audio Recordings Database (SMARD)”. In: *Proc. Int. Workshop on Acoustic Signal Enhancement*. 2014, pp. 40–44.