Model-based Speech and Audio Processing

ICASSP 2018 Tutorial

April 16 2018

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

Audio Analysis Lab, CREATE
Aalborg University, Denmark
Website: http://audio.create.aau.dk
Youtube: http://tinyurl.com/yd8mo55z

AALBORG UNIVERSITY
DENMARK

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Outline

Introduction
Who are we?
Motivation
Model-based speech and audio processing
Statistical Speech and Audio Models
Model-based Pitch Estimation
Model-based Single-Channel Enhancement
Model-based Array Processing and Enhancement
Summary and Conclusion

Audio Analysis Lab

- Research lab founded in 2012 at CREATE, Aalborg University, Denmark.
- 4 faculty, 13 junior researchers.
- Research in audio and acoustic signal processing.
- Our goal is to push the boundaries of current methods and increase the understanding of problems by pursuing mathematically tractable approaches.
- Major research projects in hearing aids, voice analysis, and microphone arrays.
- Collaborations with, e.g., GN Hearing, Bang & Olufsen, Brüel & Kjær, and Parkinson's Voice Initiative

Outline

Introduction
- Who are we?
- Motivation
- Model-based speech and audio processing

Statistical Speech and Audio Models

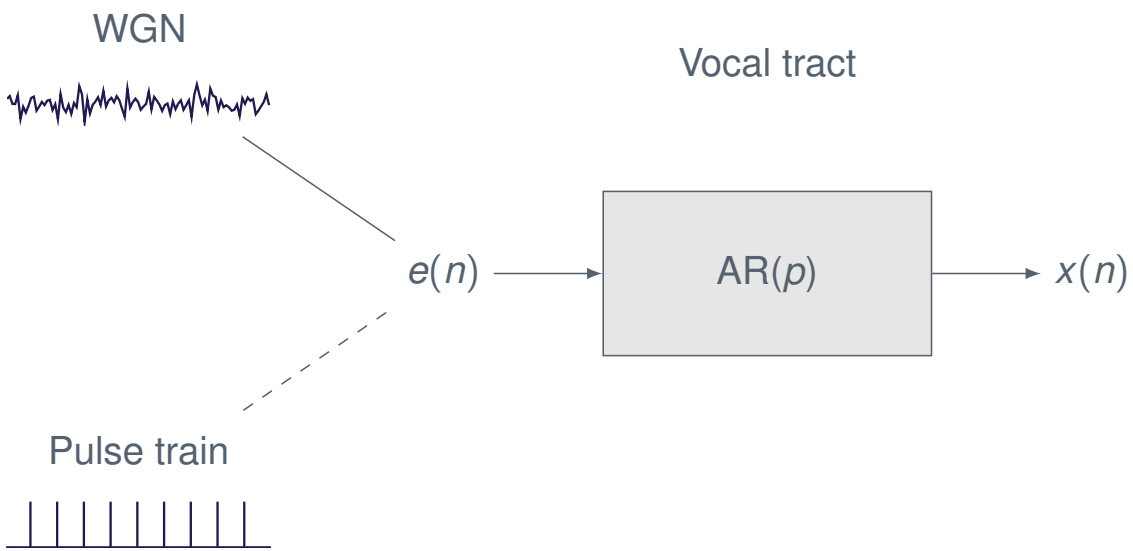
Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Motivation



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

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

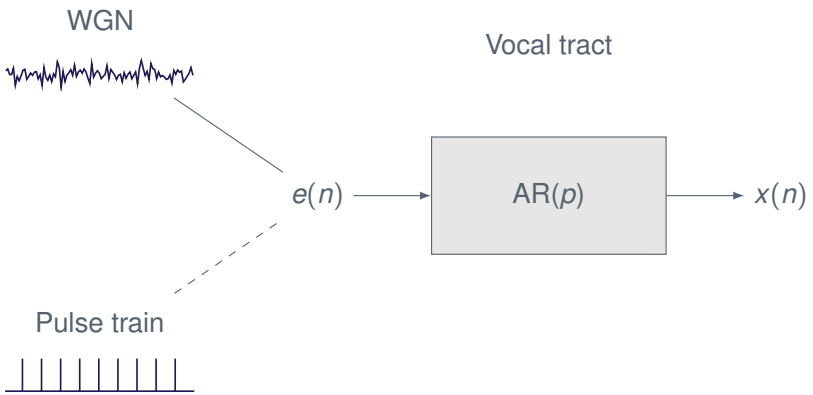
How is the model wrong?

Motivation

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

Box, 1987

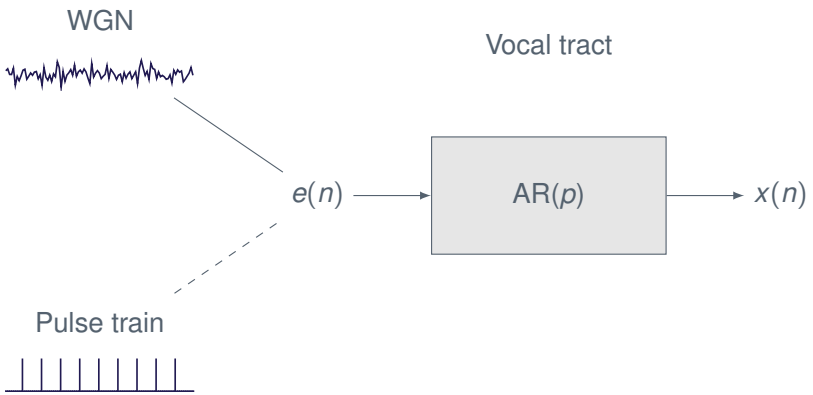
Motivation



How is a model useful?

- A model allows us to state problems in terms of the quantaties of interest (e.g., fundamental frequency, AR parameters).
- 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



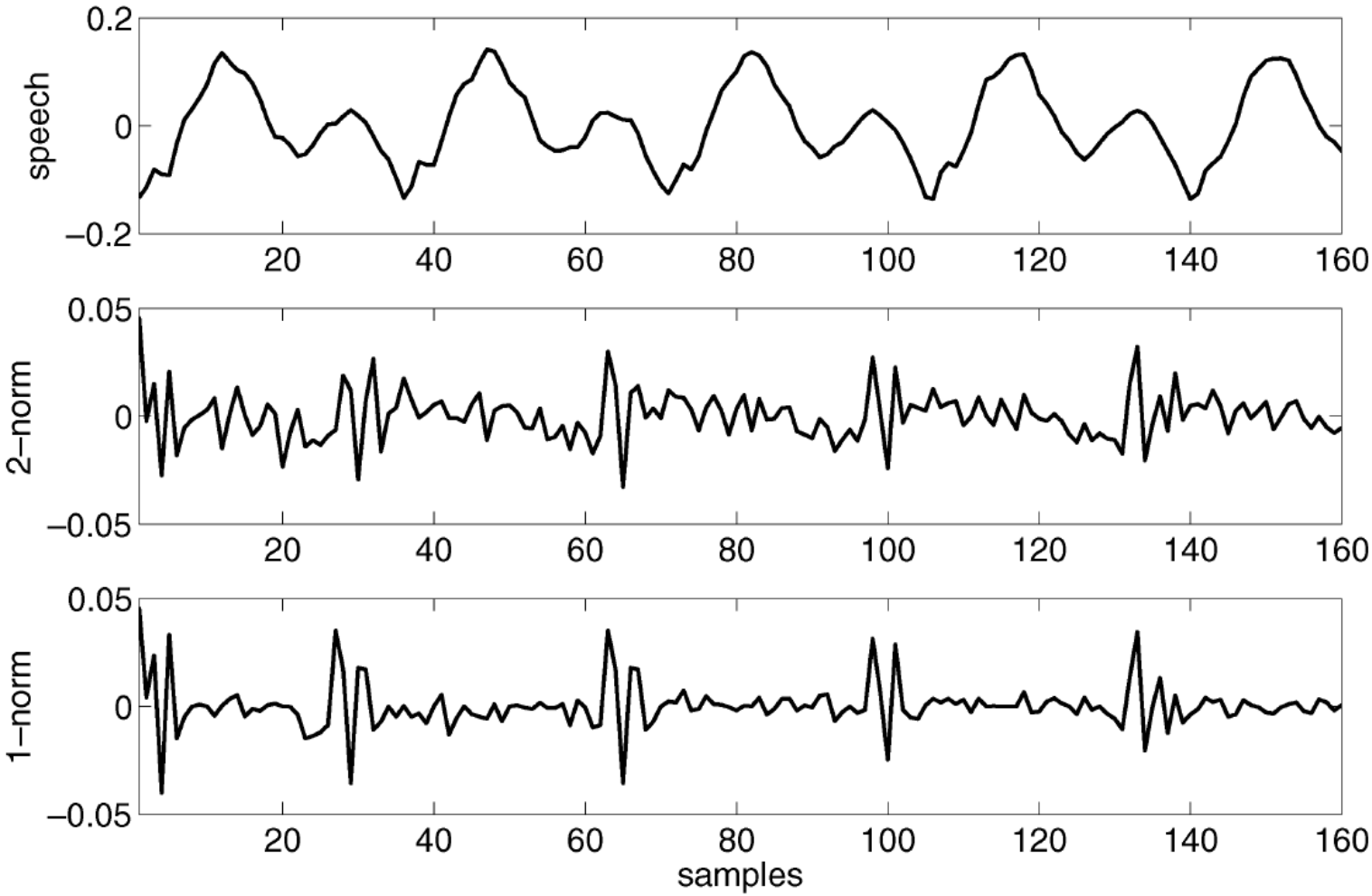
Example

Signal model

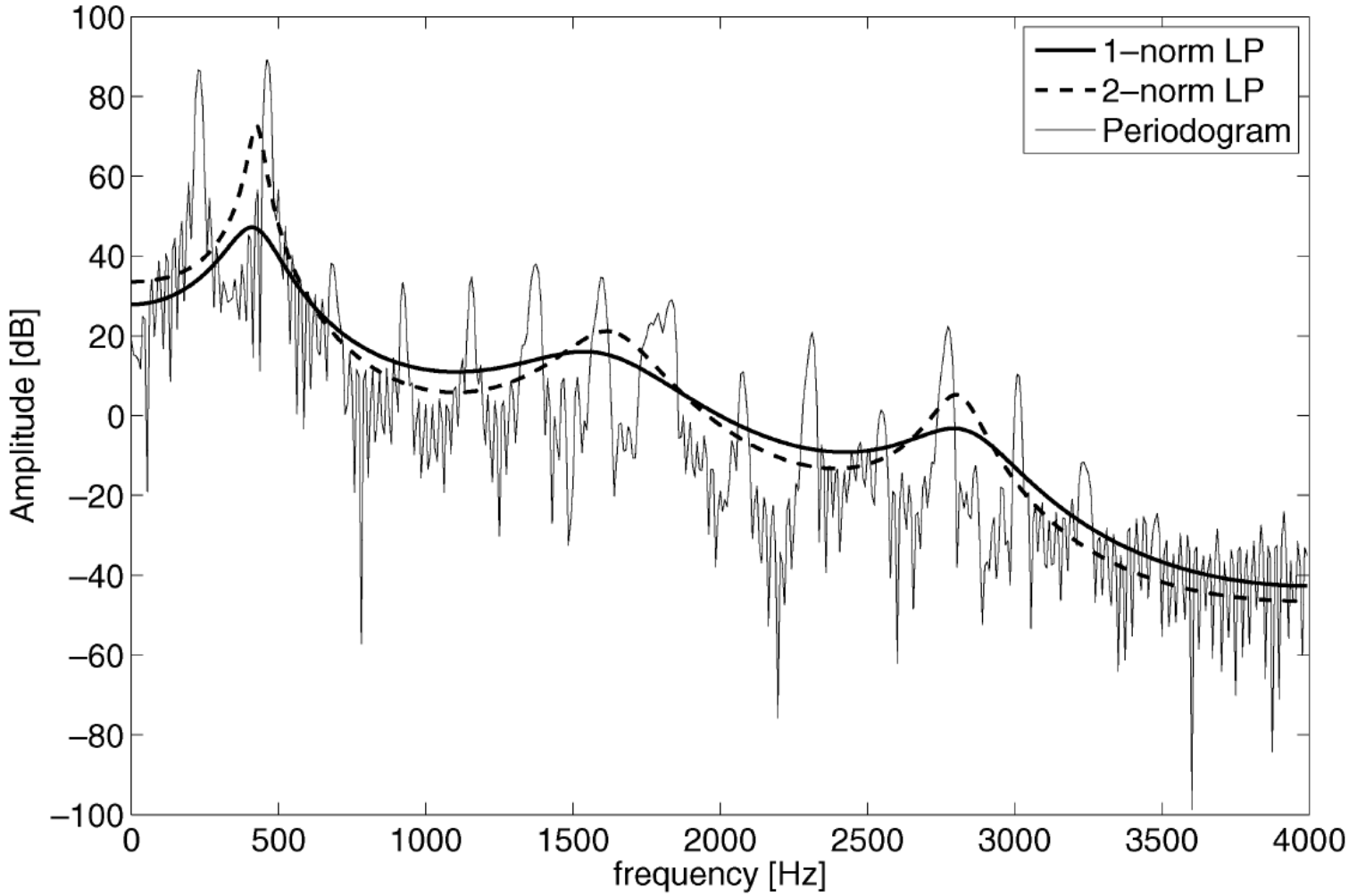
$$\boldsymbol{x} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{e} \tag{1}$$

- 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

Introduction
- Who are we?
- Motivation
- Model-based speech and audio processing

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Introduction

Model-based speech and audio processing

- Model-based processing is based on signal models.
- The signal models are often generative and described in terms of physically meaningful parameters.
- Speech and audio 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.
- Models can be used for many things and in different ways.
- An essential part of model-based 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 model-based methods?

- They lead to robust, tractable methods that can be improved and whose properties can be analyzed and understood.
- A full parametrization of the signal of interest is obtained.
- Fast implementations due to structure often exist.
- How can we hope to solve complicated problems if we cannot solve the simple ones?

Introduction

Some questions:

- Under which conditions can a method be expected to work?
- How does performance depend on the acoustic environment?
- Is the method optimal and if so in what sense?
- How do we improve the method if it does not work?

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.

Introduction

How about deep learning?

- The models we here talk about are *physically* meaningful.
- They can be interpreted, modeled and manipulated because of it.
- The models are *low-dimensional*, neural networks are high-dimensional.
- Signal models are beneficial when little data is available, the (unstructured) problem is high-dimensional, or the result should be *interpretable*!
- Signal models can of course be combined with machine learning.
- Neural networks can be useful in situations where the model is not obvious.

Outline

Introduction

Statistical Speech and Audio Models
- Basic Model
- Likelihood Function
- Estimating Parameters
- Multi-Channel Models
- Modified Models
- Amplitude Estimation
- Model Selection and Detection

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Outline

Introduction

Statistical Speech and Audio Models
- Basic Model
  - Likelihood Function
  - Estimating Parameters
  - Multi-Channel Models
  - Modified Models
  - Amplitude Estimation
  - Model Selection and Detection

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

About Models

What's a good model?

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

Tradeoff:

- Good data fit
- As few parameters as possible (Occam's razor)
- Too many parameters lead to overfitting and poorer estimates.

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

Harmonic Model

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

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

Definitions:

- $s(n)$ is the deterministic component
- $e(n)$ is the stochastic/noise component
- 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

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

with the following definitions:

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

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

Harmonic Model

The model can be written in different ways:

$$\mathbf{x}(n) = \mathbf{Z}(n)\mathbf{a} + \mathbf{e}(n) \tag{5}$$
$$= \mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{e}(n) \tag{6}$$
$$= \mathbf{Z}\mathbf{a}(n) + \mathbf{e}(n), \tag{7}$$

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

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

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}, \tag{8}$$

which is a special case of the model above with $M = N$ and $n = 0$.

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.

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

$$\mathbf{P} \approx \mathrm{diag}\left(\left[\begin{array}{ccc} A_1^2 & \cdots & A_L^2 \end{array}\right]\right). \tag{12}$$

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 $\mathbf{x}$ 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 audio 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

$$\mathrm{E}\left\{|y(n)|^2\right\} = \mathbf{h}^H\mathbf{R}\mathbf{h} \tag{18}$$

$$= \mathbf{h}^H\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{h} + \mathbf{h}^H\mathbf{Q}\mathbf{h}, \tag{19}$$

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

Subspace Model

Recall that $\mathbf{x}(n) = \mathbf{Z}\mathbf{a}(n) + \mathbf{e}(n)$ and

$$\mathbf{R} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \sigma^2\mathbf{I} \quad \text{where} \quad \mathbf{P} = \operatorname{diag}\left(\left[\, A_1^2 \;\; \cdots \;\; A_L^2 \,\right]\right).$$

where $\mathbf{Z}\mathbf{P}\mathbf{Z}^H$ has rank L. Let the EVD of $\mathbf{R}$ be

$$\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H. \tag{20}$$

$\mathbf{U}$ contains the M orthonormal eigenvectors of $\mathbf{R}$, i.e.,

$$\mathbf{U} = \left[\, \mathbf{u}_1 \quad \cdots \quad \mathbf{u}_M \,\right], \tag{21}$$

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

$$\mathbf{S} = \left[\, \mathbf{u}_1 \quad \cdots \quad \mathbf{u}_L \,\right]. \tag{22}$$

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

$$\mathbf{G} = \begin{bmatrix} \mathbf{u}_{L+1} & \cdots & \mathbf{u}_M \end{bmatrix}, \tag{23}$$

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

$$\mathbf{U}\left(\mathbf{\Lambda} - \sigma^2\mathbf{I}\right)\mathbf{U}^H = \mathbf{Z}\mathbf{P}\mathbf{Z}^H. \tag{24}$$

It follows that

$$\mathbf{Z}^H\mathbf{G} = \mathbf{0} \quad \text{and} \quad \mathcal{R}(\mathbf{S}) = \mathcal{R}(\mathbf{Z}). \tag{25}$$

These properties can be exploited for various purposes (as in, e.g., MUSIC, ESPRIT)

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 and audio signals are not perfectly periodic
- The model does not differentiate between background noise and stochastic components
- Multiple components can be present at the same time

Can this be dealt with? Does it matter?

Outline

Introduction

Statistical Speech and Audio Models
- Basic Model
- Likelihood Function
- Estimating Parameters
- Multi-Channel Models
- Modified Models
- Amplitude Estimation
- Model Selection and Detection

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Likelihood Function

If we assume that the signal is Gaussian distributed, i.e., $\mathbf{x}(n) \sim \mathcal{N}(\mathbf{s}(\theta), \mathbf{Q})$ then the likelihood function is given by

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

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

$$p(\{\mathbf{x}(n)\};\theta) = \prod_{n=0}^{G-1} p(\mathbf{x}(n);\theta). \tag{27}$$

In the above, $\mathbf{Q}$ could represent the covariance of stochastic components, background noise or both combined.

Likelihood Function

The log-likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}) = \ln p(\{\mathbf{x}(n)\}; \boldsymbol{\theta}) = \sum_{n=0}^{G} \ln p(\mathbf{x}(n); \boldsymbol{\theta}). \tag{28}$$

The maximum likelihood estimator (MLE) is then given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=0}^{G} \ln p(\mathbf{x}(n); \boldsymbol{\theta}). \tag{29}$$

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

$$\mathcal{L}(\boldsymbol{\theta}) = -N \ln \pi - N \ln \sigma^2 - \frac{1}{\sigma^2} \|\mathbf{x} - \mathbf{Z}\mathbf{a}\|_2^2, \tag{30}$$

where $\boldsymbol{\theta} = [\omega_0 \; A_1 \; \ldots A_L \; \phi_1 \; \ldots \phi_L]$.

The concentrated MLE is given by (Quinn 1991)

$$\hat{\omega}_0 = \underset{\omega_0}{\operatorname{argmax}} \, \mathcal{L}(\omega_0) = \underset{\omega_0}{\operatorname{argmax}} \, \mathbf{x}^H \mathbf{Z} \left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{x}. \tag{31}$$

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

Outline

Introduction

Statistical Speech and Audio Models

- Basic Model
- Likelihood Function
- Estimating Parameters
- Multi-Channel Models
- Modified Models
- Amplitude Estimation
- Model Selection and Detection

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

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

$$\mathrm{E}\left\{\hat{\theta}_i\right\} = \theta_i \; \forall \theta_i, \tag{32}$$

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

$$\mathrm{var}(\hat{\theta}_i) \geq \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}, \tag{33}$$

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

$$[\mathbf{I}(\boldsymbol{\theta})]_{il} = -\mathrm{E}\left\{\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i \partial\theta_l}\right\}, \tag{34}$$

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):

$$\mathrm{var}(\hat{\omega}_0) \geq \frac{6\sigma^2}{N(N^2-1)\sum_{l=1}^{L} A_l^2 l^2}, \tag{35}$$

$$\mathrm{var}(\hat{A}_l) \geq \frac{\sigma^2}{2N}, \tag{36}$$

$$\mathrm{var}(\hat{\phi}_l) \geq \frac{\sigma^2}{2N}\left(\frac{1}{A_l^2} + \frac{3l^2(N-1)^2}{\sum_{m=1}^{L} A_m^2 m^2 (N^2-1)}\right). \tag{37}$$

The CRLB of the fundamental frequency and phase both depend on the following quantity:

$$\mathrm{PSNR} = 10\log_{10}\frac{\sum_{l=1}^{L} A_l^2 l^2}{\sigma^2}\ [\mathrm{dB}]. \tag{38}$$

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*.

Fundamental Frequency Estimation

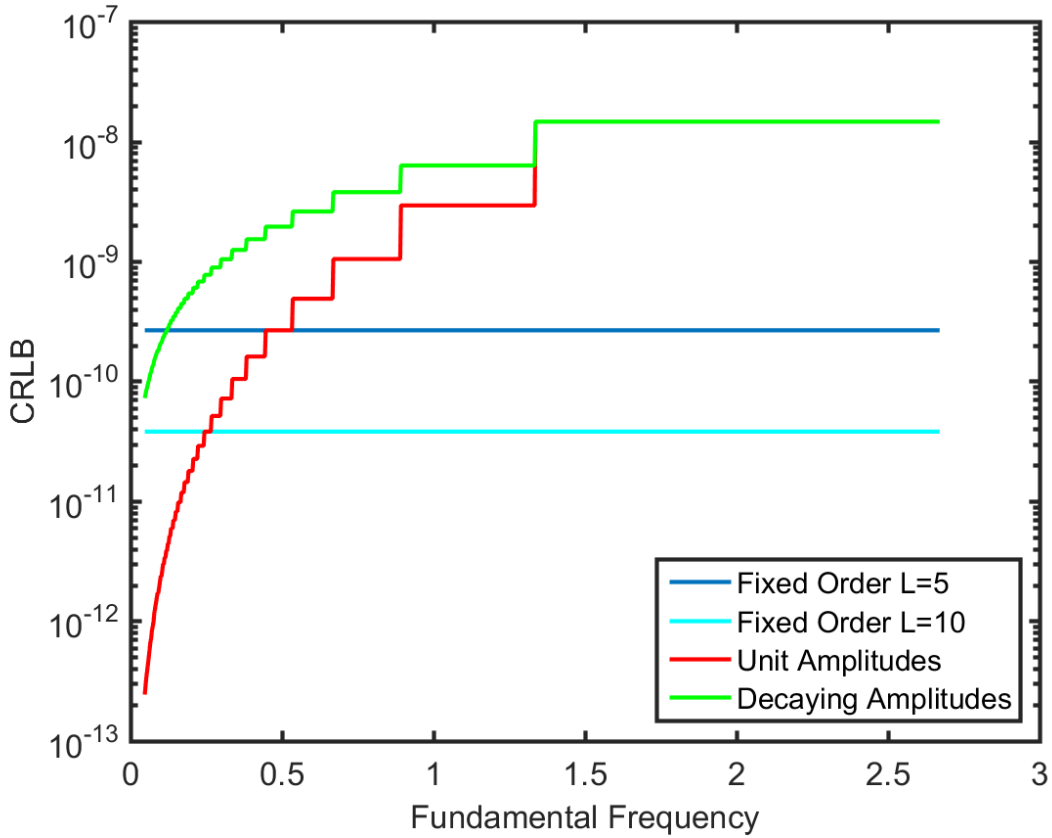


Figure: CRLB as a function of ω_0 for different cases.

Outline

Introduction

Statistical Speech and Audio Models
- Basic Model
- Likelihood Function
- Estimating Parameters
- Multi-Channel Models
- Modified Models
- Amplitude Estimation
- Model Selection and Detection

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

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.,

$$\mathbf{x}_k(n) = \mathbf{Z}(n)\mathbf{a}_k + \mathbf{e}_k(n), \tag{39}$$

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 parameters for the kth channel. The likelihood function is then

$$p(\mathbf{x}_k(n);\theta_k) = \frac{1}{\pi^M \det(\mathbf{Q}_k)} e^{-\mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}. \tag{40}$$

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

$$p(\{\mathbf{x}_k(n)\};\{\theta_k\}) = \prod_{k=1}^{K} \frac{1}{\pi^{MG} \det(\mathbf{Q}_k)^G} e^{-\sum_{n=0}^{G-1} \mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}. \tag{41}$$

General Multi-Channel Model

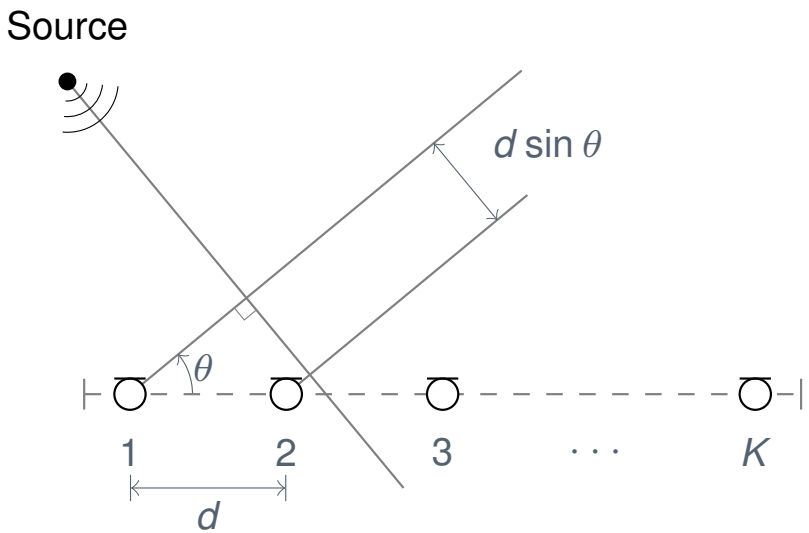
Simplifying assumptions can be made, as appropriate. For example:

- 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.

Uniform Linear Array

For a uniform 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?

Uniform 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)

$$\mathbf{s}_k(n) = \mathbf{s}(n - \Delta_k) \tag{42}$$

$$= \mathbf{s}\left(n - \frac{d\sin\theta}{c} f_s(k-1)\right). \tag{43}$$

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

$$\mathbf{s}_k\left(n - \frac{d\sin\theta}{c} f_s(k-1)\right) = \mathbf{Z}\mathbf{D}\left(n - \frac{d\sin\theta}{c} f_s(k-1)\right)\mathbf{a}. \tag{44}$$

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

$$\mathbf{D}(n) = \begin{bmatrix} e^{j\omega_0 n} & 0 & \cdots & 0 \\ 0 & e^{j\omega_0 2n} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{j\omega_0 Ln} \end{bmatrix}. \tag{45}$$

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

$$\mathbf{D}(n-\Delta) = \begin{bmatrix} e^{j\omega_0 (n-\Delta)} & 0 & \cdots & 0 \\ 0 & e^{j\omega_0 2(n-\Delta)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{j\omega_0 L(n-\Delta)} \end{bmatrix}. \tag{46}$$

Reverberation

We can modify the multi-channel 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

$$x_k(n) = s(n) * h_k(n) + e_k(n). \tag{47}$$

Assuming that the impulse response is shorter than the segment length and that the signal is stationary, then

$$h_k(n) * s(n) = h_k(n) * \sum_{l=1}^{L} a_l e^{j\omega_0 ln} \approx \sum_{l=1}^{L} \tilde{a}_{k,l} e^{j\omega_0 ln}, \tag{48}$$

due to the sinusoidal nature of $s(n)$. For anechoic environments, a simpler model is (Jensen 2016)

$$x_k(n) \approx \beta_k s(n - \Delta_k) + e_k(n). \tag{49}$$

Outline

Introduction

Statistical Speech and Audio Models
- Basic Model
- Likelihood Function
- Estimating Parameters
- Multi-Channel Models
- Modified Models
- Amplitude Estimation
- Model Selection and Detection

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Multiple Sources

How do we model multiple sources? This can be done by introducing a source index k. Then

$$x_k(n) = s_k(n) + e_k(n) = \sum_{l=1}^{L_k} a_{k,l} e^{j\omega_k l n} + e_k(n). \tag{50}$$

and then we have that

- The observed signal is then $x(n) = \sum_{k=1}^{K} x_k(n)$.
- The subvector is now $\mathbf{x}_k(n) = \mathbf{Z}_k(n)\mathbf{a}_k + \mathbf{e}_k(n)$.
- The covariance matrix is $\mathbf{R} = \sum_{k=1}^{K} \mathbf{R}_k$.

Stochastic Signals

So far, we modeled the observed signal as

$$x(n) = s(n) + e(n), \tag{51}$$

where $s(n)$ is the deterministic and $e(n)$ is all stochastic signal components.

Real speech and audio contains both harmonic and stochastic components as well as background noise. How do we account for this? Modified model:

$$x(n) = \underbrace{s(n)}_{\textit{deterministic}} + \underbrace{u(n)}_{\textit{stochastic}} + \underbrace{w(n)}_{\textit{background noise}}. \tag{52}$$

What's a good model for stochastic signals then?

Stochastic Signals

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

$$u(n) = \sum_{i=1}^{I} \gamma_i u(n-i) + \eta(n). \tag{53}$$

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 and audio applications, the background noise is rarely white.

Even though the white noise assumption is mathematically convenient, 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. Let us examine the following signal model:

$$\mathbf{x}(n) = \mathbf{s}(n) + \mathbf{e}(n). \tag{54}$$

Colored Noise

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

$$\mathbf{A}^H \mathbf{x}(n) = \mathbf{A}^H \mathbf{s}(n) + \mathbf{A}^H \mathbf{e}(n). \tag{55}$$

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 $\mathbf{A}^H \mathbf{Q} \mathbf{A} = \mathbf{I}$.

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

Colored Noise

Instead, consider the signal model $x(n) = s(n) + e(n)$ to which we apply a filter having coefficients $h(n)$, i.e.,

$$h(n) * x(n) = h(n) * s(n) + v(n), \tag{56}$$

so that $\mathbf{v} = [\ v(0)\ \cdots\ v(M-1)\ ]^H$ where $v(n) = h(n) * e(n)$ is distributed as $\mathbf{v}(n) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Since $s(n) = \sum_{l=1}^{L} a_l e^{j\omega_0 l n}$ we have that

$$h(n) * s(n) = h(n) * \sum_{l=1}^{L} a_l e^{j\omega_0 l n} \approx \sum_{l=1}^{L} \tilde{a}_l e^{j\omega_0 l n}. \tag{57}$$

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)
- Nonnegative matrix factorisation (NMF).

Sum of Autoregressive Processes

We model the signal $\mathbf{x} = [x(0) \ldots x(N-1)]^T$ as a sum of $U = U_s + U_w$ AR processes $\mathbf{c}_u$, i.e.,

$$\mathbf{x} = \sum_{u=1}^{U} \mathbf{c}_u = \underbrace{\sum_{u=1}^{U_s} \mathbf{c}_u}_{\text{signal of interest}} + \underbrace{\sum_{u=U_s+1}^{U} \mathbf{c}_u}_{\text{background noise}}, \tag{58}$$

where the first U_s AR processes are the signal of interest and the remaining U_w AR processes are background noise.

Each of the AR processes is expressed as a multivariate Gaussian (Srinivasan 2006)

$$\mathbf{c}_u \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{Q}_u). \tag{59}$$

Sum of Autoregressive Processes

$\mathbf{Q}_u$ can be asymptotically approximated as a circulant matrix which can be diagonalised as (Gray 2006)

$$\mathbf{Q}_u = \frac{1}{N}\mathbf{F}\mathbf{D}_u\mathbf{F}^H \quad \text{and} \quad \mathbf{Q}_u^{-1} = \frac{1}{N}\mathbf{F}\mathbf{D}_u^{-1}\mathbf{F}^H \tag{60}$$

where $\mathbf{F}$ is the DFT matrix defined as $[\mathbf{F}]_{k,n} = \exp(j2\pi nk/N)$, for $n, k = 0 \ldots N-1$ and

$$\mathbf{D}_u = \text{diag}\left([d_u(0), \ldots, d_u(K-1)]\right) = (\mathbf{\Lambda}_u^H\mathbf{\Lambda}_u)^{-1} \tag{61}$$

where

$$\mathbf{\Lambda}_u = \text{diag}\left(\mathbf{F}^H \begin{bmatrix}\mathbf{a}_u \\ \mathbf{0}\end{bmatrix}\right) \tag{62}$$

and $\mathbf{a}_u = [1\ a_u(1) \ldots a_u(P)]^T$ contains the AR coefficients of the u^{th} P-order process. The likelihood is

$$p(\mathbf{x}|\{\sigma_u\}, \{\mathbf{a}_u\}) \sim \mathcal{N}(\mathbf{0}, \sum_{u=1}^{U} \sigma_u^2\mathbf{Q}_u). \tag{63}$$

Sum of Autoregressive Processes

We can estimate the non-negative variances as

$$\{\hat{\sigma}_u\} = \underset{\sigma_u \geq 0 \forall u}{\operatorname{argmax}} \ln p(\mathbf{x}|\{\sigma_u\},\{\mathbf{a}_u\}) \triangleq \underset{\sigma_u \geq 0 \forall u}{\operatorname{argmax}} \mathcal{L}(\{\sigma_u\},\{\mathbf{a}_u\}), \tag{64}$$

where the log-likelihood function can be written as

$$\begin{aligned}\mathcal{L}(\{\sigma_u\},\{\mathbf{a}_u\}) = &-\frac{K}{2}\ln 2\pi + \ln \prod_{k=1}^{K} \left(\sum_{u=1}^{U} \sigma_u^2 d_u(k)\right)^{-\frac{1}{2}} \\ &- \frac{1}{2N}\mathbf{x}^T\mathbf{F}\left[\sum_{u=1}^{U}\sigma_u^2\mathbf{D}_u\right]^{-1}\mathbf{F}^H\mathbf{x}.\end{aligned} \tag{65}$$

Note that $\mathbf{F}^H\mathbf{x}$ is the Fourier transform of $\mathbf{x}$.

Sum of Autoregressive Processes

Using $\Phi(k) = \frac{1}{N}|\sum_{n=0}^{N-1} x(n)\exp(-j2\pi\frac{nk}{N})|^2$ and $\hat{\Phi}_u(k) = \sigma_u^2 d_u(k)$, we get

$$\mathcal{L}(\{\sigma_u\},\{\mathbf{a}_u\}) = -\frac{K}{2}\ln 2\pi + \ln\prod_{k=1}^{K}\left(\sum_{u=1}^{U}\hat{\Phi}_u(k)\right)^{-\frac{1}{2}} - \frac{1}{2}\sum_{k=1}^{K}\frac{\Phi(k)}{\sum_{u=1}^{U}\hat{\Phi}_u(k)}.$$

Finally the log-likelihood can be written as

$$\mathcal{L}(\{\sigma_u\},\{\mathbf{a}_u\}) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=1}^{K}\left(\frac{\Phi(k)}{\sum_{u=1}^{U}\hat{\Phi}_u(k)} + \ln\sum_{u=1}^{U}\hat{\Phi}_u(k)\right) \quad (66)$$

where $\sum_{u=1}^{U}\hat{\Phi}_u(k) = \sum_{u=1}^{U}\sigma_u^2 d_u(k)$. This looks very familiar!

Sum of Autoregressive Processes

Comments:

- Maximising the likelihood is equivalent to minimising the Itakura-Saito (IS) divergence!
- This model and its estimation problems are equivalent to those of supervised IS-NMF.
- Only, the spectral basis are parametrized by AR coefficients. We call this parametric NMF.
- Traditional IS-NMF (Fevotte, 2009) is a special case of parametric NMF with $P = N - 1$.
- It is better than NMF for low-delay applications and when few training data is available.
- Note that the NMF model is correct for sums of autoregressive processes (unlike for deterministic models).

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

$$x(n) = \sum_{l=1}^{L} A_l e^{j\theta_l(n)} + e(n) \tag{67}$$

where $\theta_l(n)$ is the instantaneous phase of the lth harmonic which is a continuous function, and everything else is as before. $\theta_l(\cdot)$ is given by

$$\theta_l(t) = \int_0^t l\omega_0(\tau)d\tau + \phi_l, \tag{68}$$

where $\omega_0(t)$ is the time-varying pitch and ϕ_l is the phase.

Non-Stationary Speech

If the pitch is slowly varying, i.e., $\omega_0(t) = \alpha_0 t + \omega_0$, we get

$$\theta_l(t) = \frac{1}{2}\alpha_0 l t^2 + \omega_0 l t + \phi_l, \tag{69}$$

where α_0 is the fundamental chirp rate. This is the harmonic chirp model (HCM) (Nørholm 2016). The model can be written as before:

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

With the definitions:

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

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

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

Non-Stationary Speech
Experiments

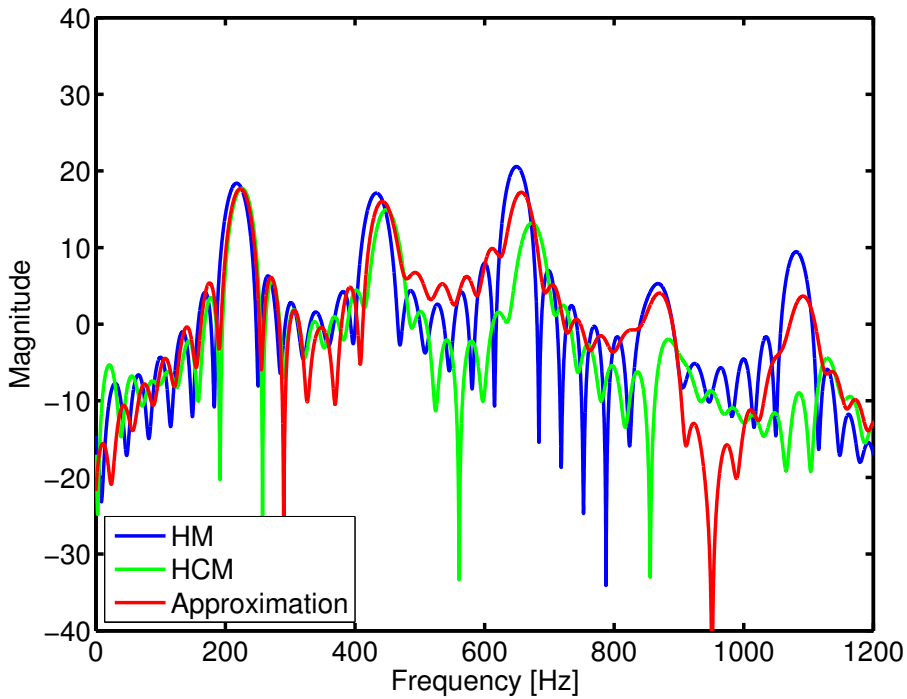


Figure: Spectrum of harmonic model, harmonic chirp model, and an approximation.

Outline

Introduction

Statistical Speech and Audio Models

- Basic Model
- Likelihood Function
- Estimating Parameters
- Multi-Channel Models
- Modified Models
- Amplitude Estimation
- Model Selection and Detection

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Amplitude Estimation

After estimating the signal's fundamental frequencies, one often wishes to estimate also the amplitudes of the harmonics having frequencies $\{\psi_l\}_{l=1}^{L}$.

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

- Least-squares (LS)
- Optimal filtering (APES, Capon)
- Combinations (WLS)

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

Amplitude Estimation

Consider the unconstrained signal model $x(n) = \sum_{l=1}^{L} a_l e^{j\psi_l n} + e(n)$, which can be written as

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}. \tag{71}$$

Then, the LS estimator is $\hat{\mathbf{a}} = \left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}$, which is an efficient estimator for all $N \geq L$ for WGN and asymptotically efficient for colored noise. For sub-vectors the model can be written as

$$\mathbf{x}(n) = \mathbf{Z}(n)\mathbf{a} + \mathbf{e}(n) = \mathbf{Z}\mathbf{D}(n) + \mathbf{e}(n). \tag{72}$$

We can also estimate of the amplitude vector using WLS as

$$\hat{\mathbf{a}} = \left[\sum_{n=0}^{N-M} \mathbf{Z}^H(n)\widehat{\mathbf{Q}}^{-1}\mathbf{Z}(n)\right]^{-1}\left[\sum_{n=0}^{N-M} \mathbf{Z}^H(n)\widehat{\mathbf{Q}}^{-1}\mathbf{x}(n)\right], \tag{73}$$

where $\widehat{\mathbf{Q}}$ denotes an estimate of the noise covariance matrix.

Amplitude Estimation

For sufficiently large N and M, we can use $\widehat{\mathbf{Q}} \approx \widehat{\mathbf{R}}$, but the estimate of $\widehat{\mathbf{Q}}$ may be improved by rewriting

$$\mathbf{x}(n) = \mathbf{Z}(n)\mathbf{a} + \mathbf{e}(n) = \sum_{k=1}^{L} \underbrace{\left[a_k \mathbf{z}(\psi_k)\right]}_{\boldsymbol{\beta}_k} e^{j\psi_k n} + \mathbf{e}(n) \tag{74}$$

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

$$\hat{\boldsymbol{\beta}}_k = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{x}(n) e^{-j\psi_k n} \tag{75}$$

and the covariance matrix estimate

$$\widehat{\mathbf{Q}} = \widehat{\mathbf{R}} - \sum_{k=1}^{L} \hat{\boldsymbol{\beta}}_k \hat{\boldsymbol{\beta}}_k^H \tag{76}$$

Using this estimate yields an APES-like estimator.

Amplitude Estimation

A matched filterbank (MAFI) estimator can be designed using the filterbank matrix $\mathbf{H}$ and the design criteria

$$\mathbf{H} = \min_{\mathbf{H}} \mathrm{Tr}\left\{\mathbf{H}^H\mathbf{R}\mathbf{H}\right\} \quad \text{subject to} \quad \mathbf{H}^H\mathbf{Z} = \mathbf{I} \tag{77}$$

This has the solution $\mathbf{H} = \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}^{-1}\mathbf{Z}\right)^{-1}$ Then,

$$\mathbf{z}(n) = \mathbf{H}^H\mathbf{x}(n) = \mathbf{D}(n)\mathbf{a} + \mathbf{H}^H\mathbf{e}(n) = \mathbf{D}(n)\mathbf{a} + \mathbf{w}(n), \tag{78}$$

with the lth index being $z_l(n) = a_l e^{j\psi_l n} + w_l(n)$ from which we get the MAFI amplitude estimate as

$$\hat{a}_l = \frac{1}{N-M+1}\sum_{n=0}^{N-M} z_l(n)e^{-j\psi_l n}. \tag{79}$$

Outline

Introduction

Statistical Speech and Audio Models

- Basic Model
- Likelihood Function
- Estimating Parameters
- Multi-Channel Models
- Modified Models
- Amplitude Estimation
- Model Selection and Detection

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

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

$$p(\mathcal{M}_m|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M}_m)p(\mathcal{M}_m)}{p(\mathbf{x})}. \tag{80}$$

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

$$\widehat{\mathcal{M}}_k = \underset{\mathcal{M}_m, m\in\mathbb{Z}_q}{\text{argmax}}\, p(\mathcal{M}_m|\mathbf{x}) = \underset{\mathcal{M}_m, m\in\mathbb{Z}_q}{\text{argmax}}\, \frac{p(\mathbf{x}|\mathcal{M}_m)p(\mathcal{M}_m)}{p(\mathbf{x})}. \tag{81}$$

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

Posterior Probabilities

If all the models are equally probable, i.e., $p(\mathcal{M}) = \frac{1}{q}$ and by noting that $p(\mathbf{x})$ is constant, the MAP model selection criterion reduces to

$$\widehat{\mathcal{M}} = \underset{\mathcal{M}_m, m \in \mathbb{Z}_q}{\text{argmax}}\ p(\mathbf{x}|\mathcal{M}_m), \tag{82}$$

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

$$p(\mathbf{x}|\mathcal{M}_m) = \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}_m) p(\boldsymbol{\theta}|\mathcal{M}_m) d\boldsymbol{\theta}. \tag{83}$$

This can be done in several ways, including numerically.

Posterior Probabilities

Using Laplace integration, we can write (Djuric 1998)

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

where $\hat{\theta}$ is the MLE and D is the number of real parameters and

$$\widehat{\mathbf{H}} = - \left. \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}_m)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{84}$$

is the Hessian of the log-likelihood function evaluated at $\hat{\theta}$. Taking the logarithm and ignoring constant terms, we get

$$\widehat{\mathcal{M}} = \arg \min_{\mathcal{M}_m, m \in \mathbb{Z}_q} - \underbrace{\ln p(\mathbf{x}|\hat{\boldsymbol{\theta}}, \mathcal{M}_m)}_{\text{log-likelihood}} + \underbrace{\frac{1}{2} \ln \det\left(\widehat{\mathbf{H}}\right)}_{\text{penalty}}, \tag{85}$$

which can be used directly for selecting between various models.

Posterior Probabilities

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

$$\ln\det\left(\widehat{\mathbf{H}}\right) = \ln\det\left(\mathbf{K}^{-2}\right) + \ln\det\left(\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}\right). \tag{86}$$

For the harmonic model, we introduce

$$\mathbf{K} = \begin{bmatrix} N^{-3/2} & \mathbf{0} \\ \mathbf{O} & N^{-1/2}\mathbf{I} \end{bmatrix}, \tag{87}$$

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

$$\ln\det\left(\widehat{\mathbf{H}}\right) = 3\ln N + 2L\ln N + \mathcal{O}(1). \tag{88}$$

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:

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e} \tag{89}$$

Hypotheses:

$$\mathcal{H}_0 : \mathbf{a} = \mathbf{0} \tag{90}$$
$$\mathcal{H}_1 : \mathbf{a} \neq \mathbf{0} \tag{91}$$

Test statistic:

$$T(\mathbf{x}) = \frac{N-L}{L} \frac{\mathbf{x}^H \mathbf{Z} \left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{x}}{\mathbf{x}^H \left(\mathbf{I} - \mathbf{Z} \left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H\right) \mathbf{x}} \tag{92}$$

Detection

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

$$T(\mathbf{x}) > \gamma' \tag{93}$$

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

$$P_{\text{FA}} = Q_{F_{L,N-L}}(\gamma') \tag{94}$$

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

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

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 a model of stochastic components.
- 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

Introduction

Statistical Speech and Audio Models

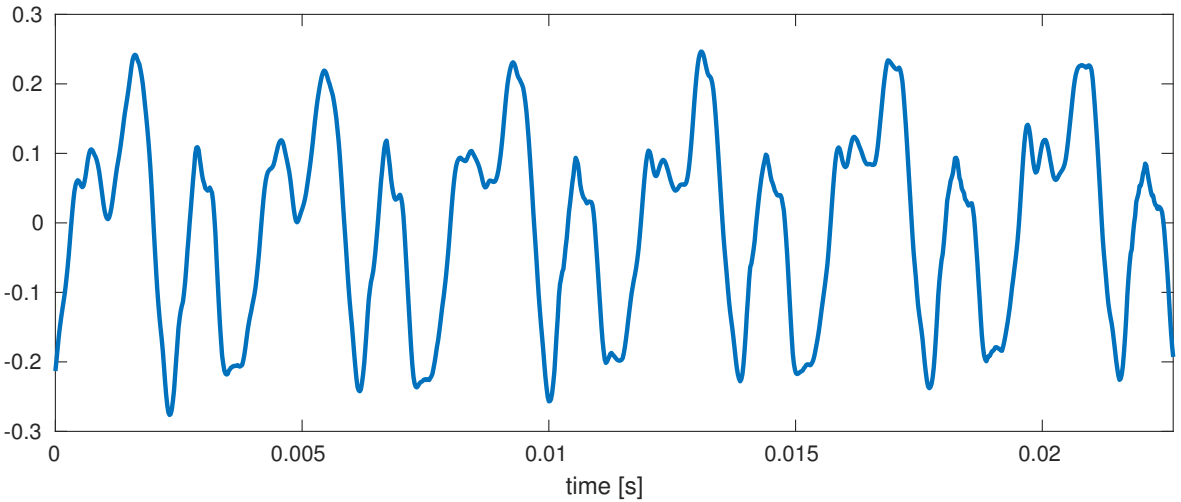
Model-based Pitch Estimation
- Correlation-based Methods
- Nonlinear Least Squares Methods
- Comparison of Methods
- Non-stationary Pitch Estimation
- Multi-channel Pitch Estimation
- Summary

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Periodic Signals



Periodic signals

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

$$x(n) = x(n-\tau) = x(n-2\pi/\omega_0) \tag{95}$$

where ω_0 is the fundamental frequency or pitch in radians/sample.

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 or heart rate variability
  - Heart defect diagnosis
- Rotating machines
  - Vibration analysis
  - Rotation speed

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation
- Correlation-based Methods
- Nonlinear Least Squares Methods
- Comparison of Methods
- Non-stationary Pitch Estimation
- Multi-channel Pitch Estimation
- Summary

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Correlation-based Methods

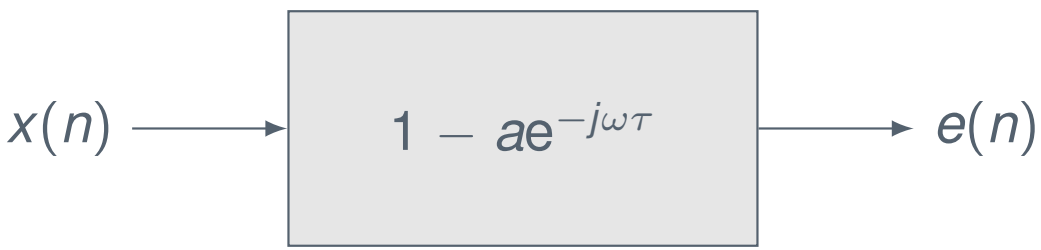
Consider the objective

$$J(a,\tau) = \sum_{n=\tau_{\text{MAX}}}^{N-1} |e(n)|^2 \tag{96}$$

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

$$e(n) = x(n) - ax(n-\tau)\,, \quad a > 0 \ \wedge \ \tau \in [\tau_{\text{MIN}}, \tau_{\text{MAX}}] \tag{97}$$

Often referred to as comb-filtering.



Correlation-based Methods

Conditioned on τ, the optimal value for a is

$$\hat{a}(\tau) = \max\left(\frac{\sum_{n=\tau_{\text{MAX}}}^{N-1} x(n)x(n-\tau)}{\sum_{n=\tau_{\text{MAX}}}^{N-1} x^2(n-\tau)}, 0\right) \tag{98}$$

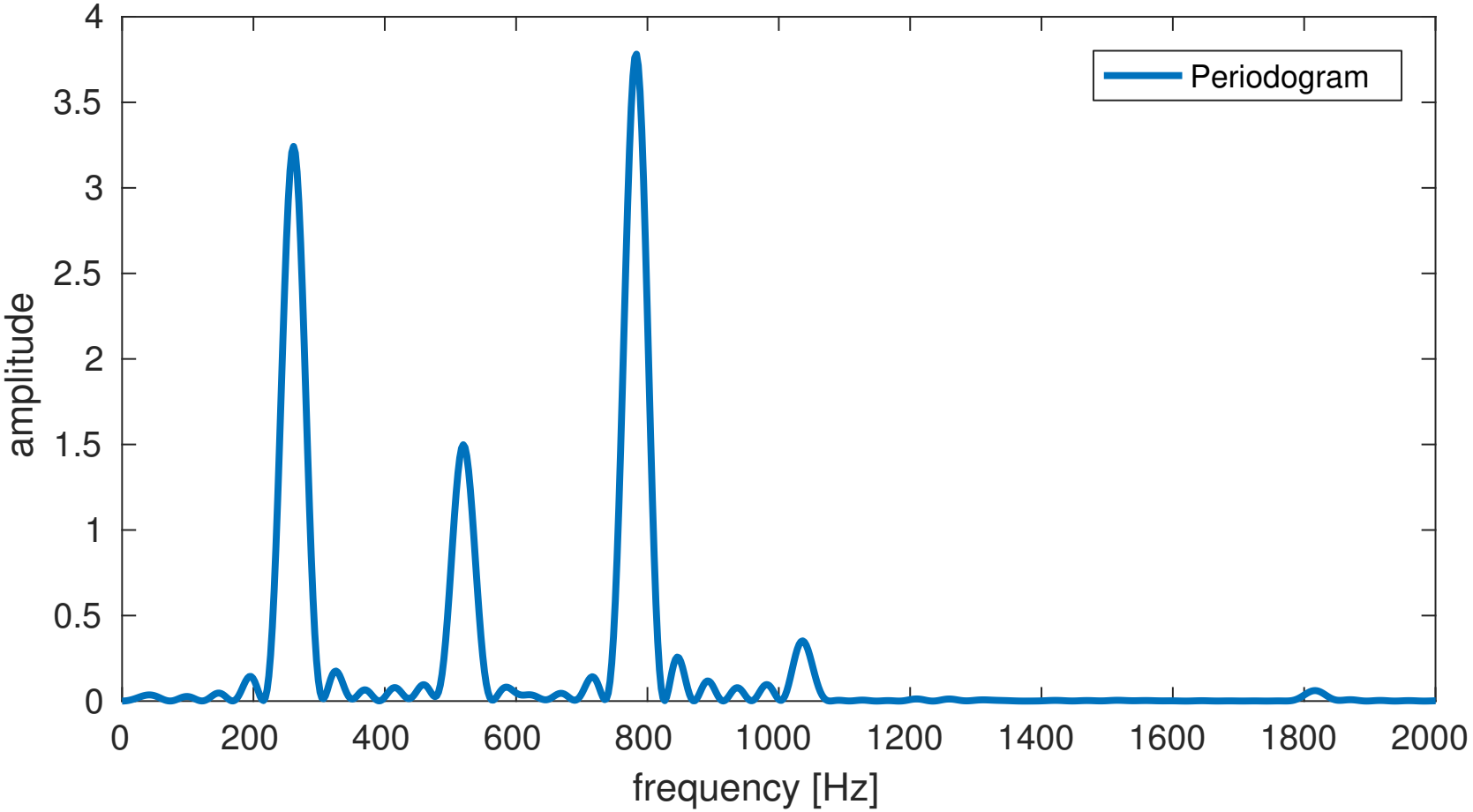
Inserting this into the objective $J(a,\tau)$ yields the estimator

$$\hat{\tau} = \underset{\tau\in[\tau_{\text{MIN}},\tau_{\text{MAX}}]}{\text{argmax}} \ \max\left(\phi(\tau), 0\right) \tag{99}$$

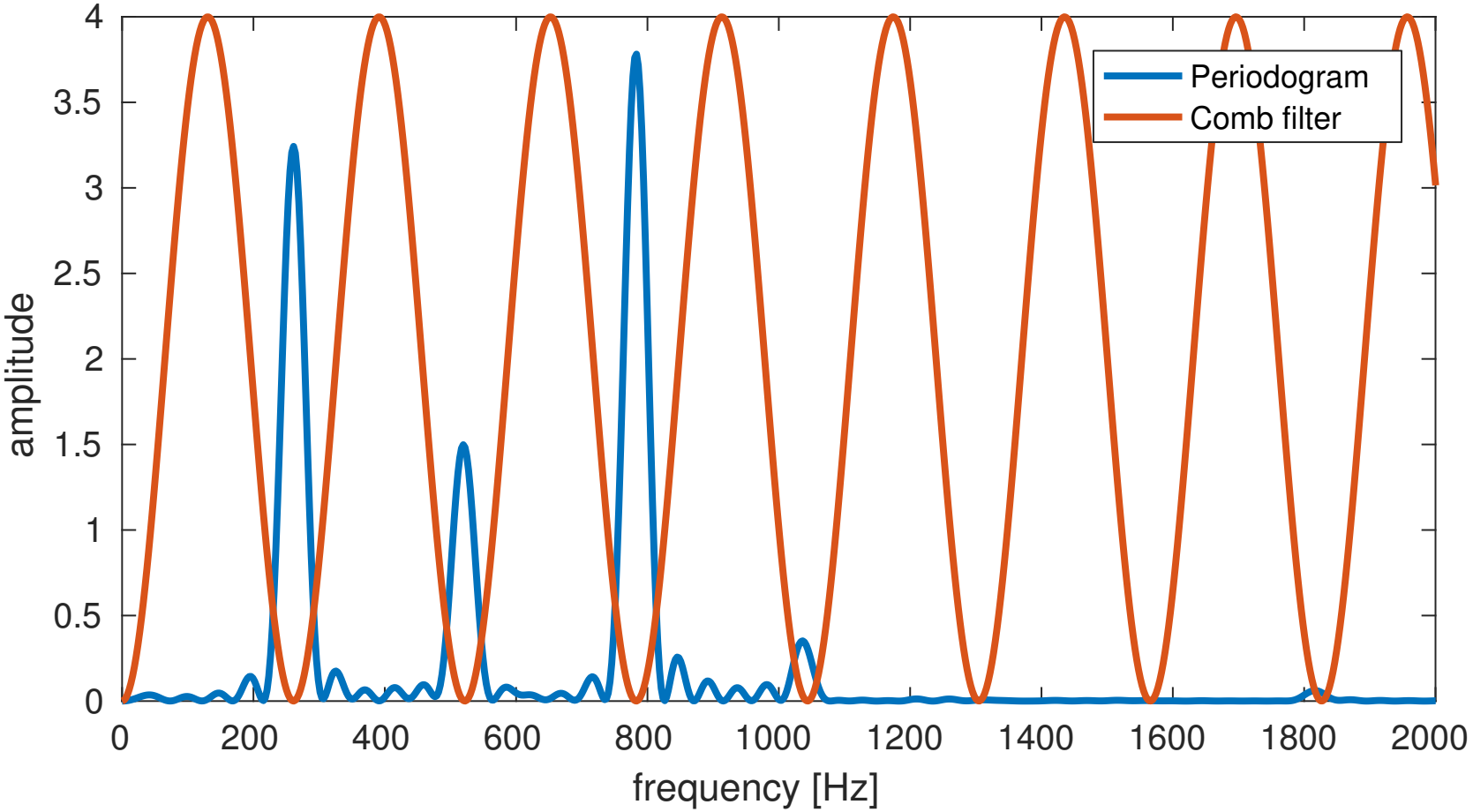
where $\phi(\tau)$ is the normalised cross correlation function given by

$$\phi(\tau) = \frac{\sum_{n=\tau_{\text{MAX}}}^{N-1} x(n)x(n-\tau)}{\sqrt{\sum_{n=\tau_{\text{MAX}}}^{N-1} x^2(n) \sum_{n=\tau_{\text{MAX}}}^{N-1} x^2(n-\tau)}} \tag{100}$$

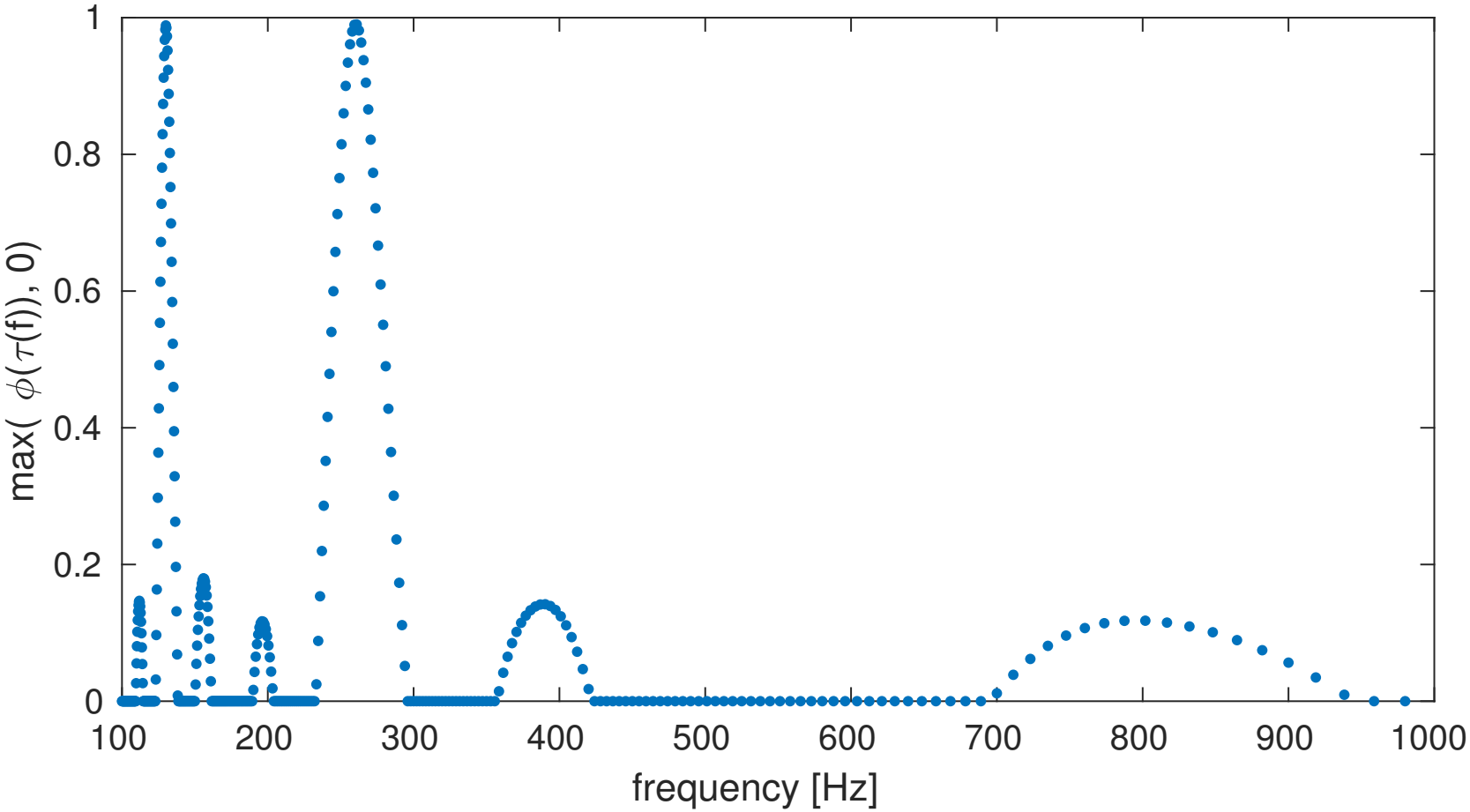
Correlation-based Methods



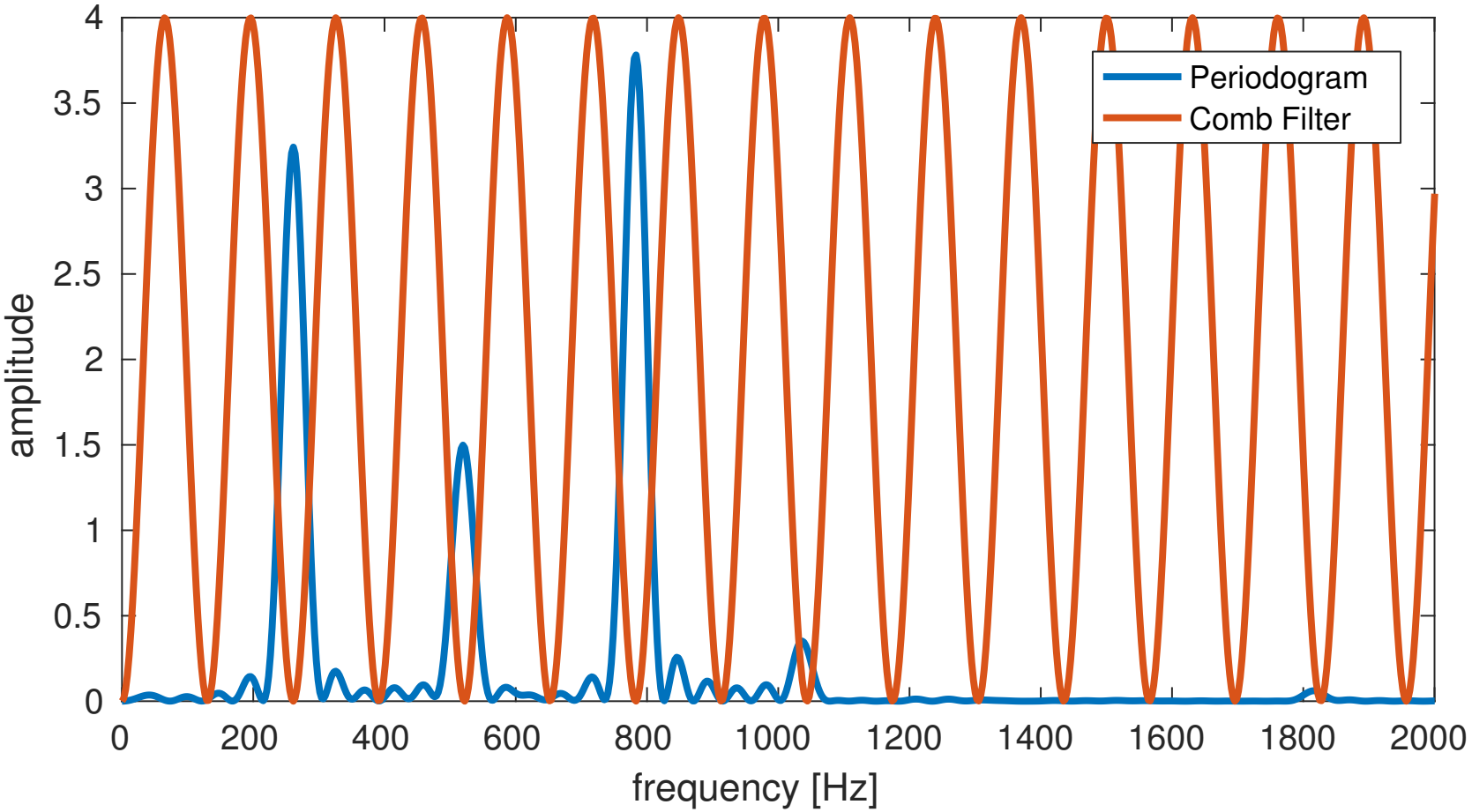
Correlation-based Methods



Correlation-based Methods



Correlation-based Methods



Correlation-based Methods

.... but is anyone actually using the comb filtering method?

PRAAT: (Boersma, 1993), well over 1000 citations (Google Scholar)
Maximises a windowed normalised cross-correlation function

RAPT: (Talkin, 1995), nearly 1000 citations (Google Scholar)
Maximises a normalised cross-correlation function

YIN: (Cheveigné, 2002), nearly 2000 citations (Google Scholar)
Minimises the comb filtering error for $a = 1$

Kaldi: (Ghahremani et al., 2014), nearly 150 citations (Google Scholar)
Maximises a normalised cross-correlation function

Correlation-based Methods

Typical components of a correlation-based pitch estimator

1. Compute the (normalised) cross-correlation function
2. Interpolate the (normalised) cross-correlation function to a desired frequency resolution
3. Do something about subharmonic errors
4. Perform interframe smooting (i.e., pitch tracking)

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

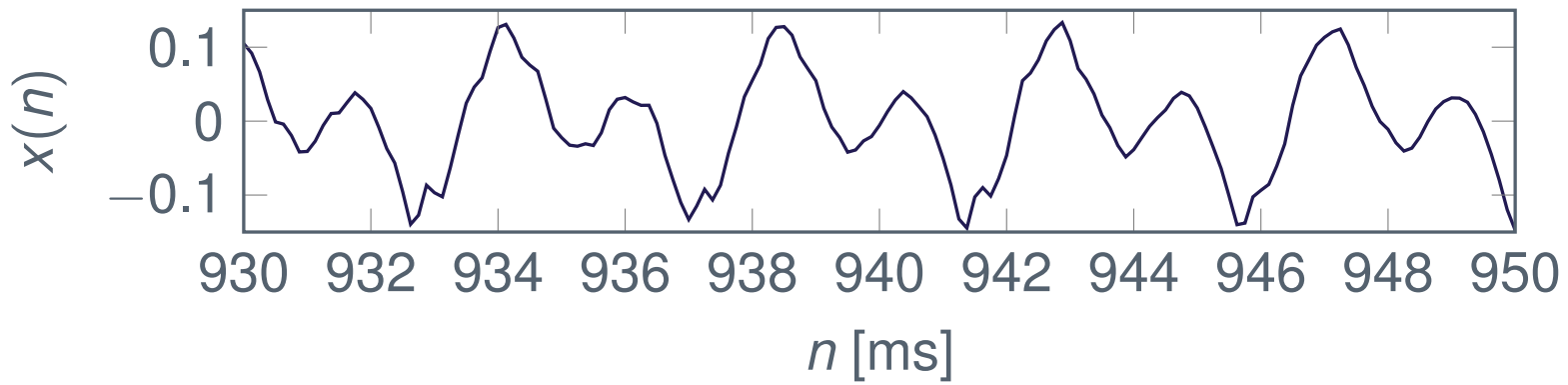
- Correlation-based Methods
- Nonlinear Least Squares Methods
- Comparison of Methods
- Non-stationary Pitch Estimation
- Multi-channel Pitch Estimation
- Summary

Model-based Single-Channel Enhancement

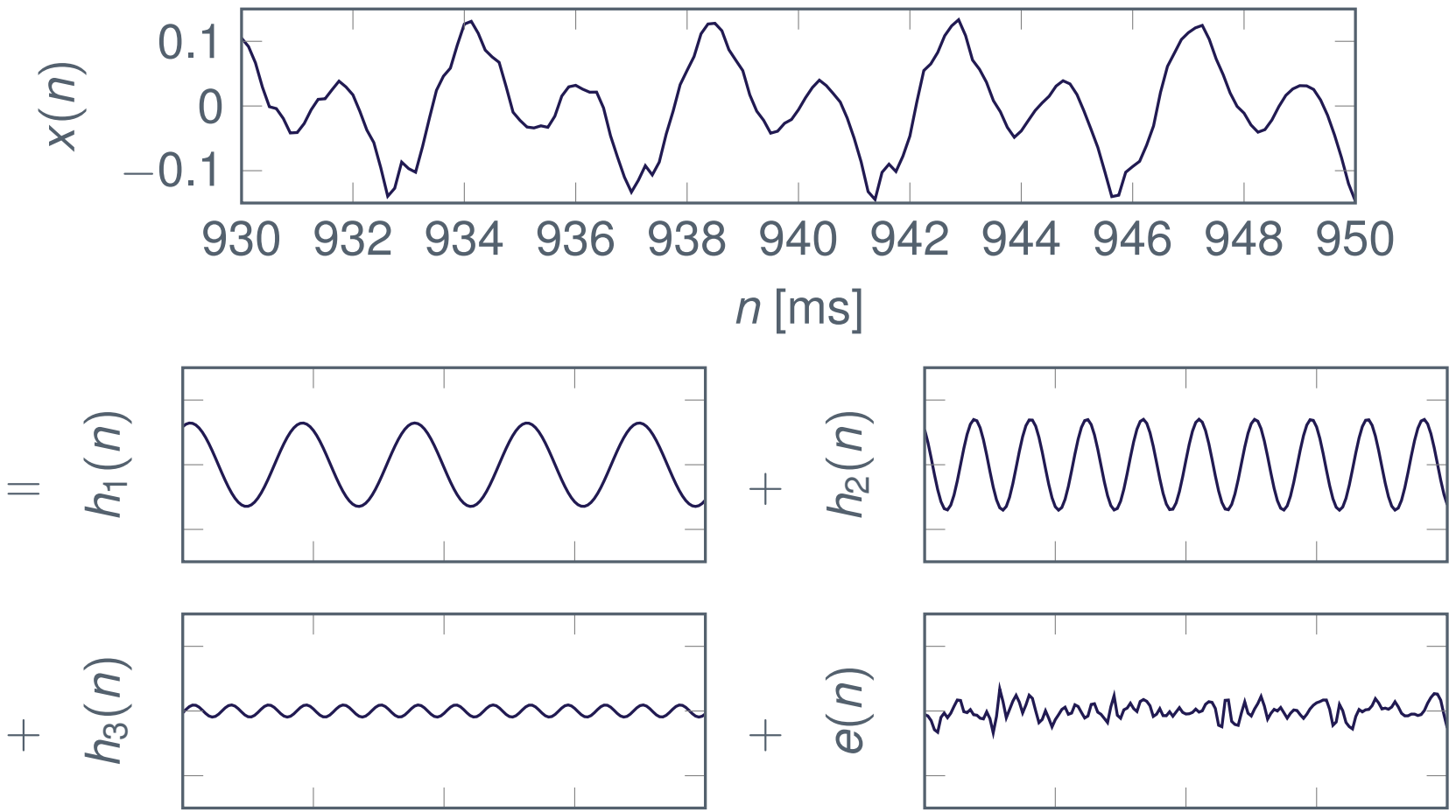
Model-based Array Processing and Enhancement

Summary and Conclusion

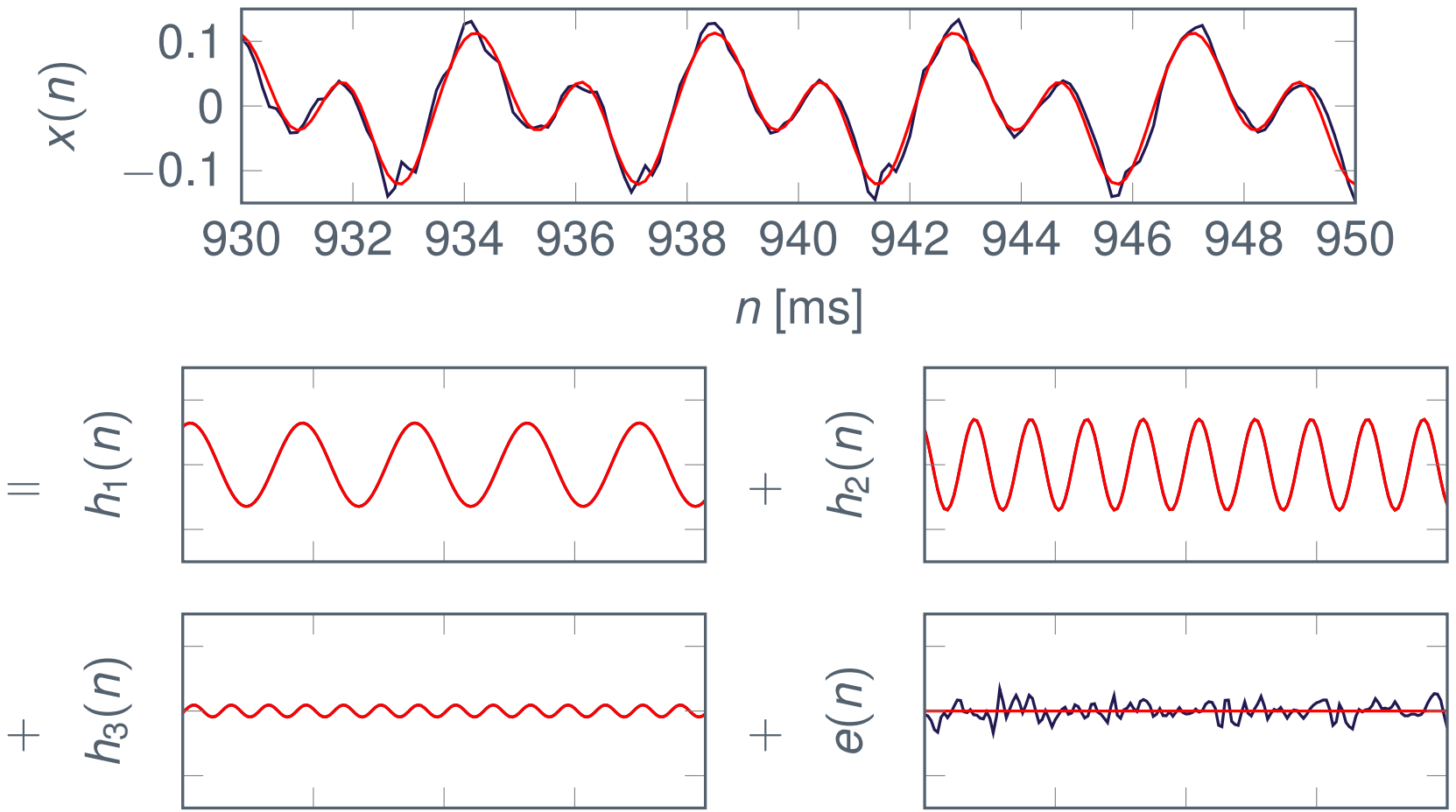
Harmonic Model



Harmonic Model



Harmonic Model



Harmonic Model

Mathematical Model

The signal model for any periodic signal is

$$s(n) = \sum_{l=1}^{L} h_l(n) = \sum_{l=1}^{L} A_l \cos(\omega_0 l n + \phi_l) \tag{101}$$

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

Instead of considering the comb-filtering error

$$e(n) = x(n) - ax(n-\tau)\,, \tag{102}$$

we consider the least-squares error

$$e(n) = x(n) - s(n,\boldsymbol{\theta})\,, \qquad n = 0, 1, \ldots, N-1 \tag{103}$$

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

$$s(n,\boldsymbol{\theta}) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l) \tag{104}$$

$$\boldsymbol{\theta} = \begin{bmatrix} A_1 & \cdots & A_L & \phi_1 & \cdots & \phi_L & \omega_0 \end{bmatrix}^T \tag{105}$$

Method of Least Squares

The nonlinear least squares (NLS) method is that of solving

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, J(\theta) \tag{106}$$

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

$$J(\theta) = \sum_{n=0}^{N-1} |e(n)|^2 = \sum_{n=0}^{N-1} |x(n) - s(n,\theta)|^2 \tag{107}$$

- Solving this problem naïvely is very computationally demanding since the fundamental frequency is a nonlinear parameter.
- Asymptotically, however, an efficient solution exists which for historical reasons is called harmonic summation (Noll, 1969).

NLS Estimator

The harmonic model

$$x(n) = \sum_{l=1}^{L} \left[a_l \cos(l\omega_0 n) - b_l \sin(l\omega_0 n) \right] + e(n) \tag{108}$$

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

$$\boldsymbol{x} = \boldsymbol{Z}_L(\omega_0)\boldsymbol{\alpha}_L + \boldsymbol{e} \tag{109}$$

where

$$\begin{aligned}
\boldsymbol{Z}_L(\omega) &= \begin{bmatrix} \boldsymbol{c}(\omega) & \boldsymbol{c}(2\omega) & \cdots & \boldsymbol{c}(L\omega) & \boldsymbol{s}(\omega) & \boldsymbol{s}(2\omega) & \cdots & \boldsymbol{s}(L\omega) \end{bmatrix} \\
\boldsymbol{c}(\omega) &= \begin{bmatrix} \cos(\omega n_0) & \cdots & \cos(\omega(n_0 + N - 1)) \end{bmatrix}^T \\
\boldsymbol{s}(\omega) &= \begin{bmatrix} \sin(\omega n_0) & \cdots & \sin(\omega(n_0 + N - 1)) \end{bmatrix}^T \\
\boldsymbol{\alpha}_l &= \begin{bmatrix} \boldsymbol{a}_L^T & -\boldsymbol{b}_L^T \end{bmatrix}^T, \; \boldsymbol{a}_L = \begin{bmatrix} a_1 & \cdots & a_L \end{bmatrix}^T, \; \boldsymbol{b}_L = \begin{bmatrix} b_1 & \cdots & b_L \end{bmatrix}^T
\end{aligned}$$

NLS Estimator

The least squares error is

$$\sum_{n=0}^{N-1} e^2(n) = \boldsymbol{e}^T\boldsymbol{e} = \left[\boldsymbol{x} - \boldsymbol{Z}_L(\omega_0)\boldsymbol{\alpha}_L\right]^T\left[\boldsymbol{x} - \boldsymbol{Z}_L(\omega_0)\boldsymbol{\alpha}_L\right] \tag{110}$$

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

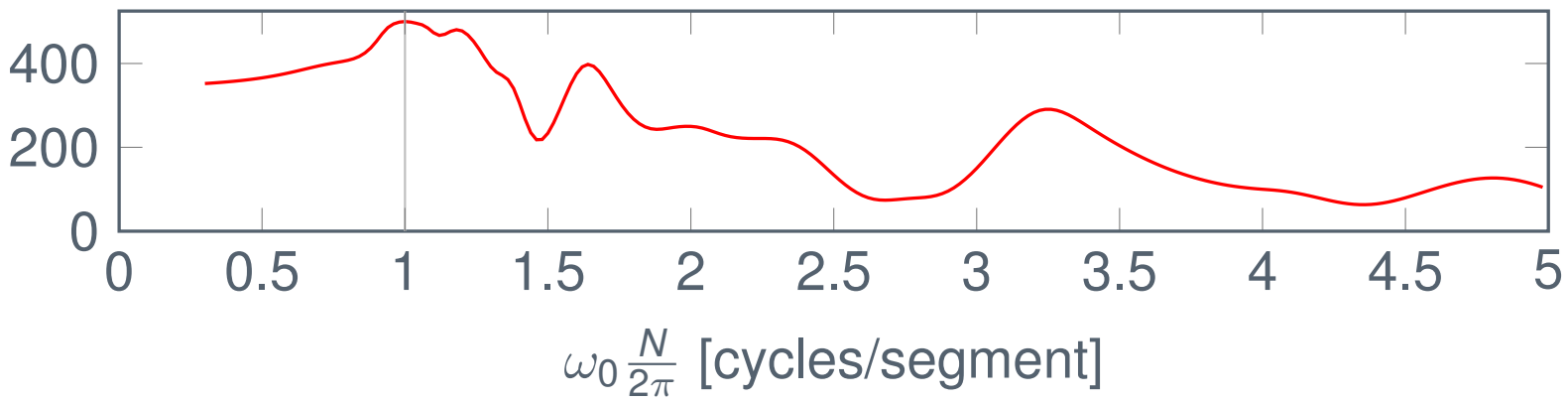
$$\hat{\boldsymbol{\alpha}}_L(\omega_0) = \left[\boldsymbol{Z}_L^T(\omega_0)\boldsymbol{Z}_L(\omega_0)\right]^{-1}\boldsymbol{Z}_L^T(\omega_0)\boldsymbol{x} \tag{111}$$

Inserting this back into the objective yields the NLS estimator

$$\hat{\omega}_{0,L} = \underset{\omega_0\in[\omega_{\text{MIN}},\omega_{\text{MAX}}]}{\text{argmax}}\ \boldsymbol{x}^T\boldsymbol{Z}_L(\omega_0)\left[\boldsymbol{Z}_L^T(\omega_0)\boldsymbol{Z}_L(\omega_0)\right]^{-1}\boldsymbol{Z}_L^T(\omega_0)\boldsymbol{x} \tag{112}$$

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

NLS Estimator



1. Compute NLS cost function

$$\hat{\omega}_{0,L} = \underset{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]}{\text{argmax}} \; \boldsymbol{x}^T \boldsymbol{Z}_L(\omega_0) \left[\boldsymbol{Z}_L^T(\omega_0) \boldsymbol{Z}_L(\omega_0) \right]^{-1} \boldsymbol{Z}_L^T(\omega_0) \boldsymbol{x} \quad (113)$$

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

Approximate NLS estimator

Harmonic summation (HS) estimator

Asymptotically,

$$\lim_{N\to\infty} \frac{2}{N}\boldsymbol{Z}_L^T(\omega_0)\boldsymbol{Z}_L(\omega_0) = \boldsymbol{I}_L \ . \tag{114}$$

Using this limit as an approximation gives the harmonic summation estimator (Noll, 1969)

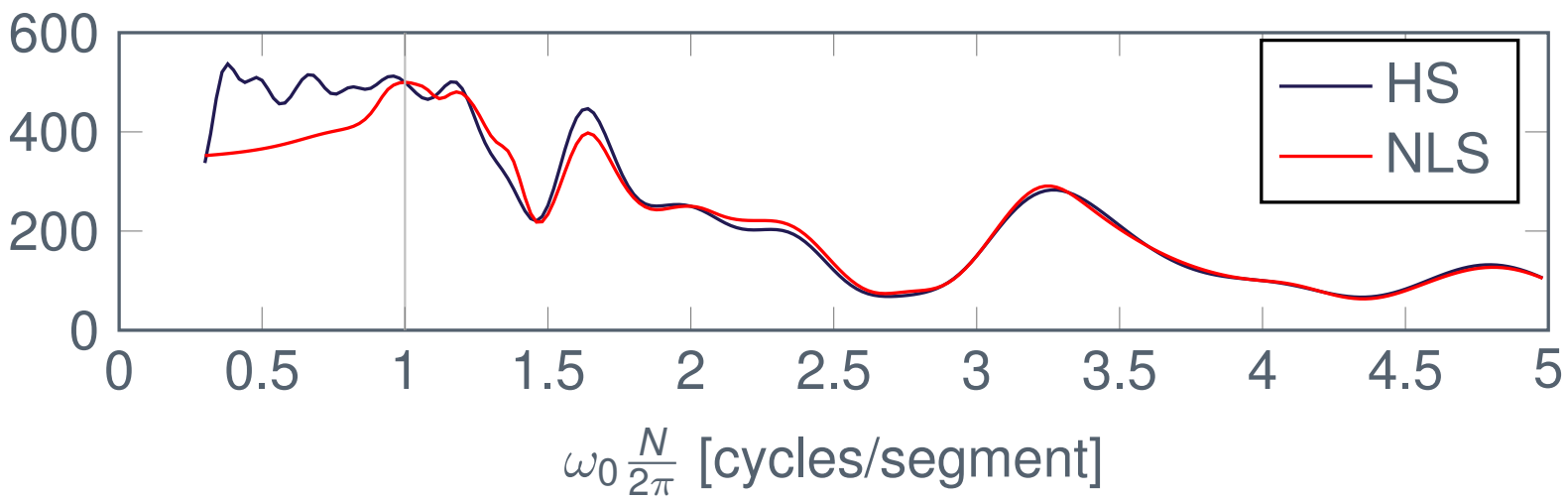
$$\hat{\omega}_{0,L} = \underset{\omega_0\in[\omega_{\text{MIN}},\omega_{\text{MAX}}]}{\text{argmax}} \ \boldsymbol{x}^T\boldsymbol{Z}_L(\omega_0)\boldsymbol{Z}_L^T(\omega_0)\boldsymbol{x} = \underset{\omega_0\in[\omega_{\text{MIN}},\omega_{\text{MAX}}]}{\text{argmax}} \sum_{l=1}^{L} |X(\omega_0 l)|^2$$

The HS estimator is also referred to as approximate NLS (aNLS).

Exact NLS vs HS

Some remarks:

- The HS method works very well, unless the fundamental frequency is low or the maximum harmonic component is close to the Nyquist frequency.
- The HS method can be implemented very efficiently using a single FFT.
- The order of complexity for NLS has recently been decreased to that of HS (Nielsen et al., 2017).



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 also includes model comparison.
- The algorithm can also be set-up to work for a model with a non-zero DC-value.
- A C++-implementation is also available.
- Can be downloaded from `https://github.com/jkjaer/fastF0Nls`.

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation
- Correlation-based Methods
- Nonlinear Least Squares Methods
- Comparison of Methods
  - Robustness to noise
  - Time-frequency resolution
  - Summary
- Non-stationary Pitch Estimation
- Multi-channel Pitch Estimation
- Summary

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Comparison of Methods

What could be evaluated?

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

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

- Correlation-based Methods
- Nonlinear Least Squares Methods
- Comparison of Methods
  - Robustness to noise
  - Time-frequency resolution
  - Summary
- Non-stationary Pitch Estimation
- Multi-channel Pitch Estimation
- Summary

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

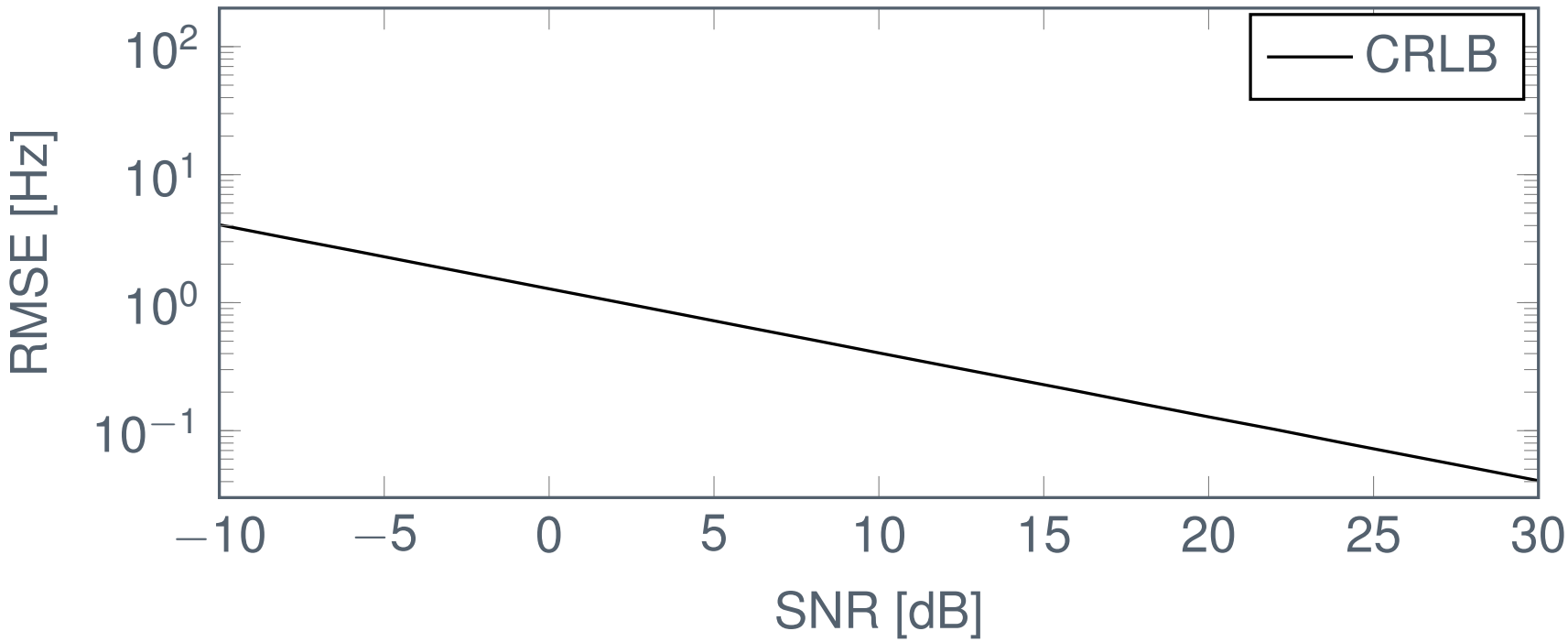
Comparison of Methods
Robustness to noise

Simulation setup

- Segment size of 25 ms at a sampling frequency of 8000 Hz.
- Estimate the pitch from 1000 Monte Carlo runs for every SNR.
- In each run, the true pitch is randomly selected from $[90, 380]$ Hz and the true phases are also generated at random.
- The true amplitudes are exponentially decreasing.
- The true model order is 7.
- Each method searches for a pitch in the range $[80, 400]$ Hz.
- The maximum model order in NLS is set to 15.
- The noise is white and Gaussian.
- No pitch tracking used in any method.

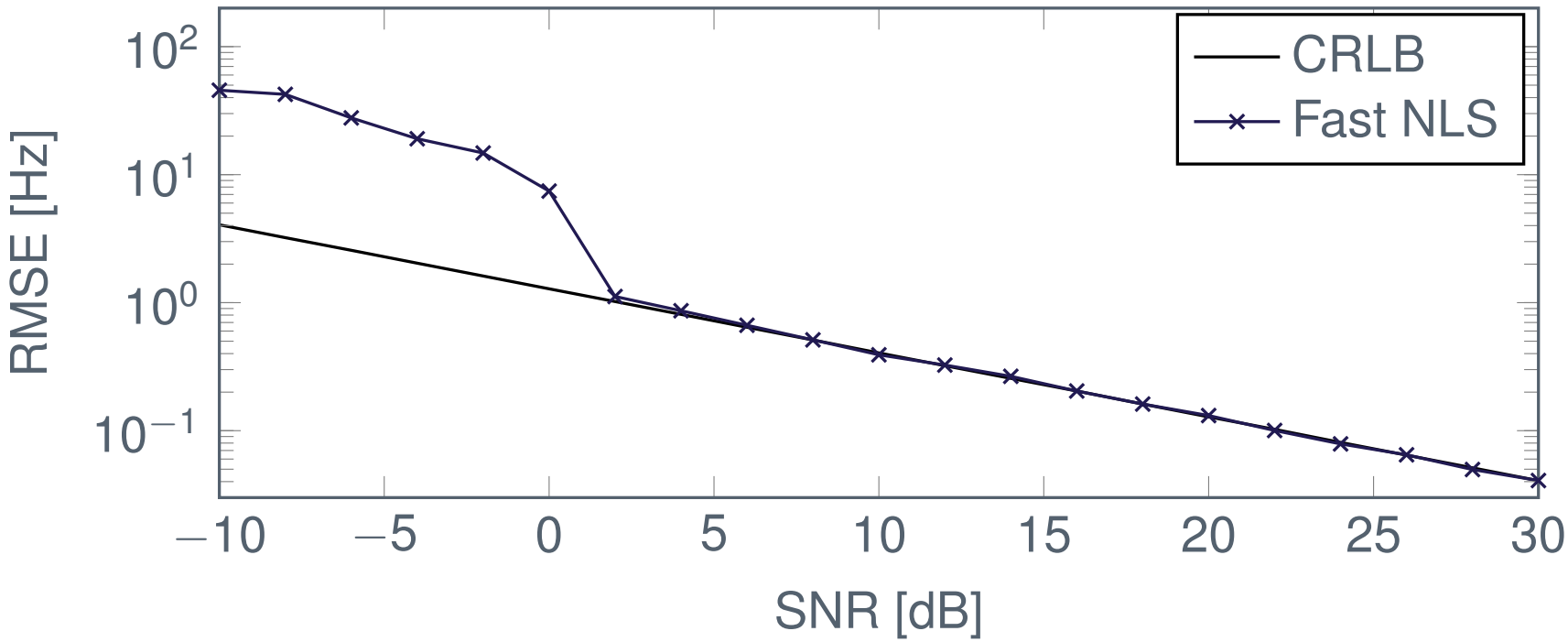
Comparison of Methods

Robustness to noise



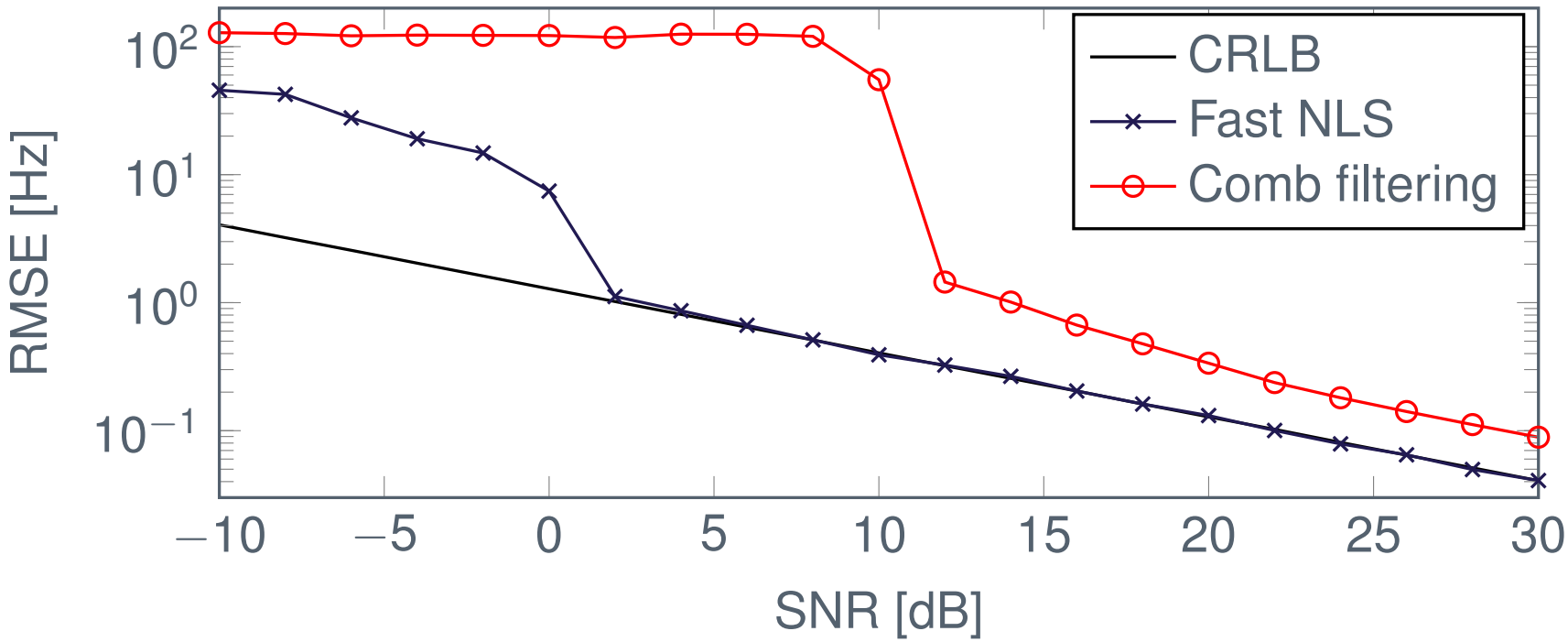
Comparison of Methods

Robustness to noise



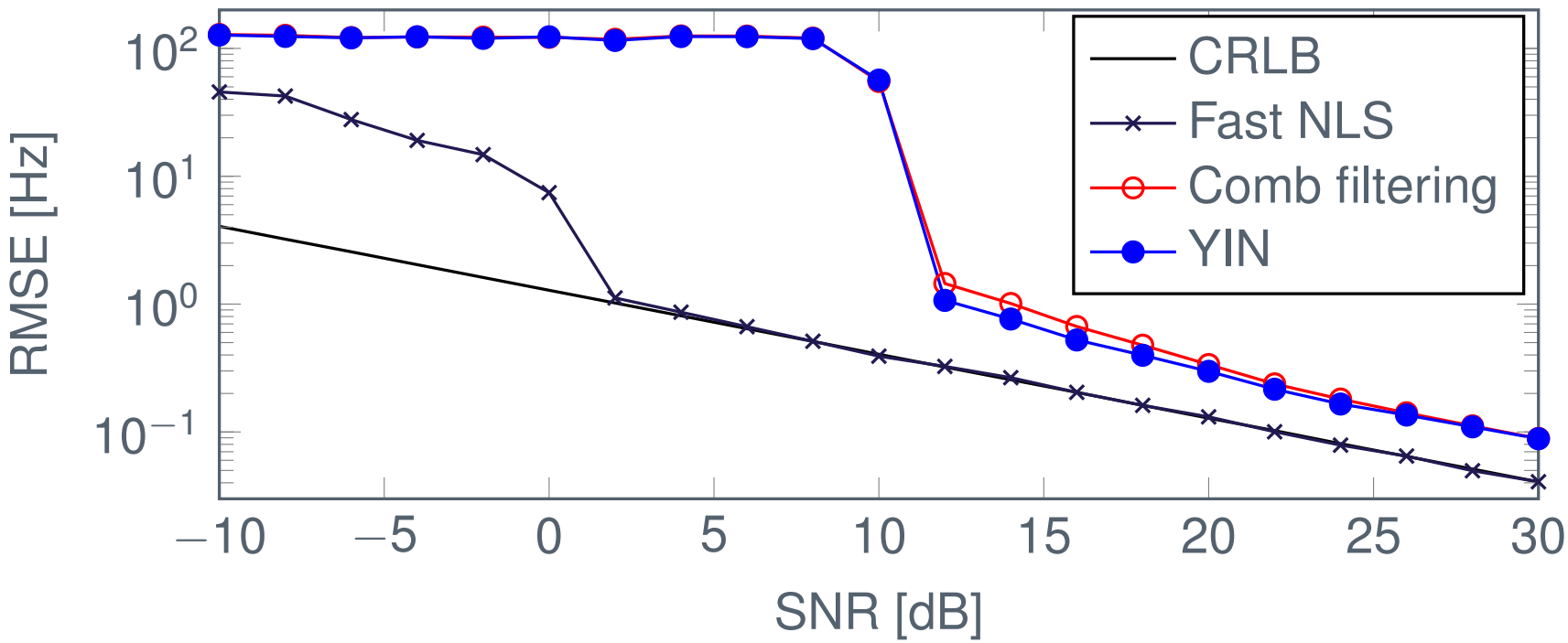
Comparison of Methods

Robustness to noise



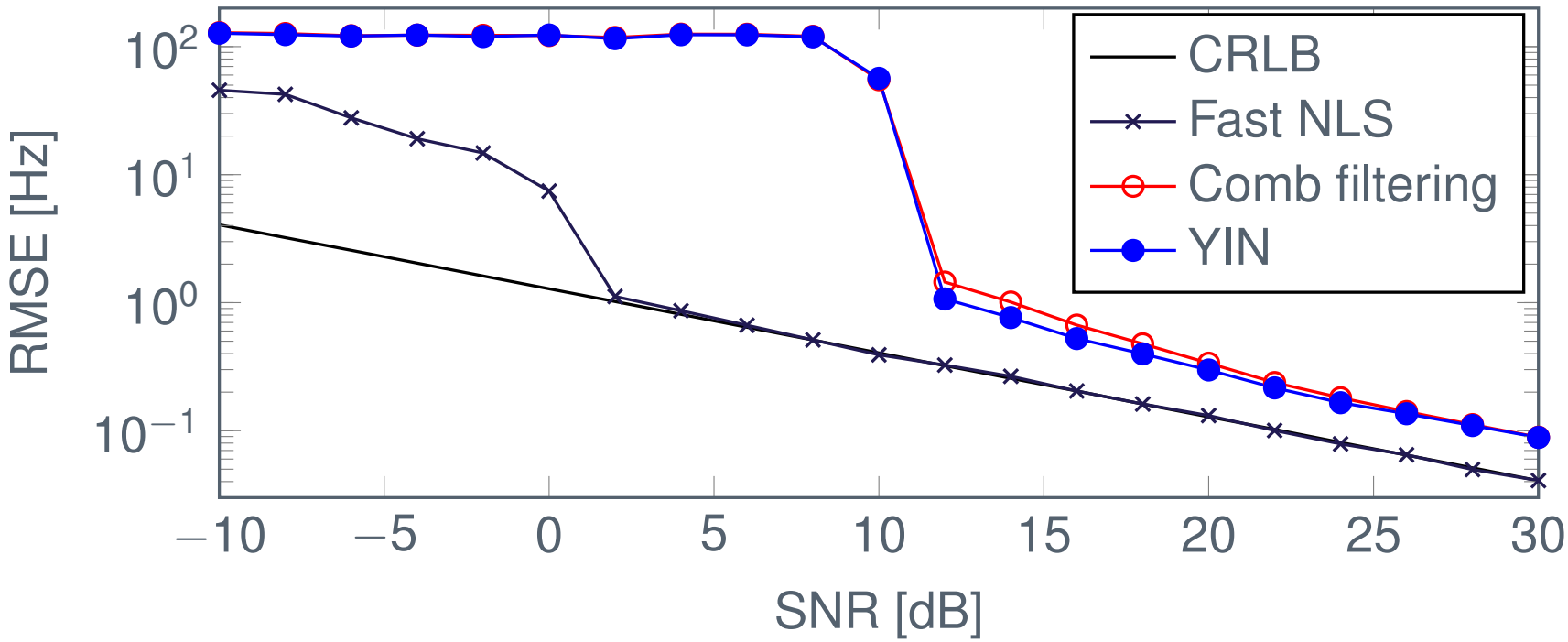
Comparison of Methods

Robustness to noise



Comparison of Methods

Robustness to noise

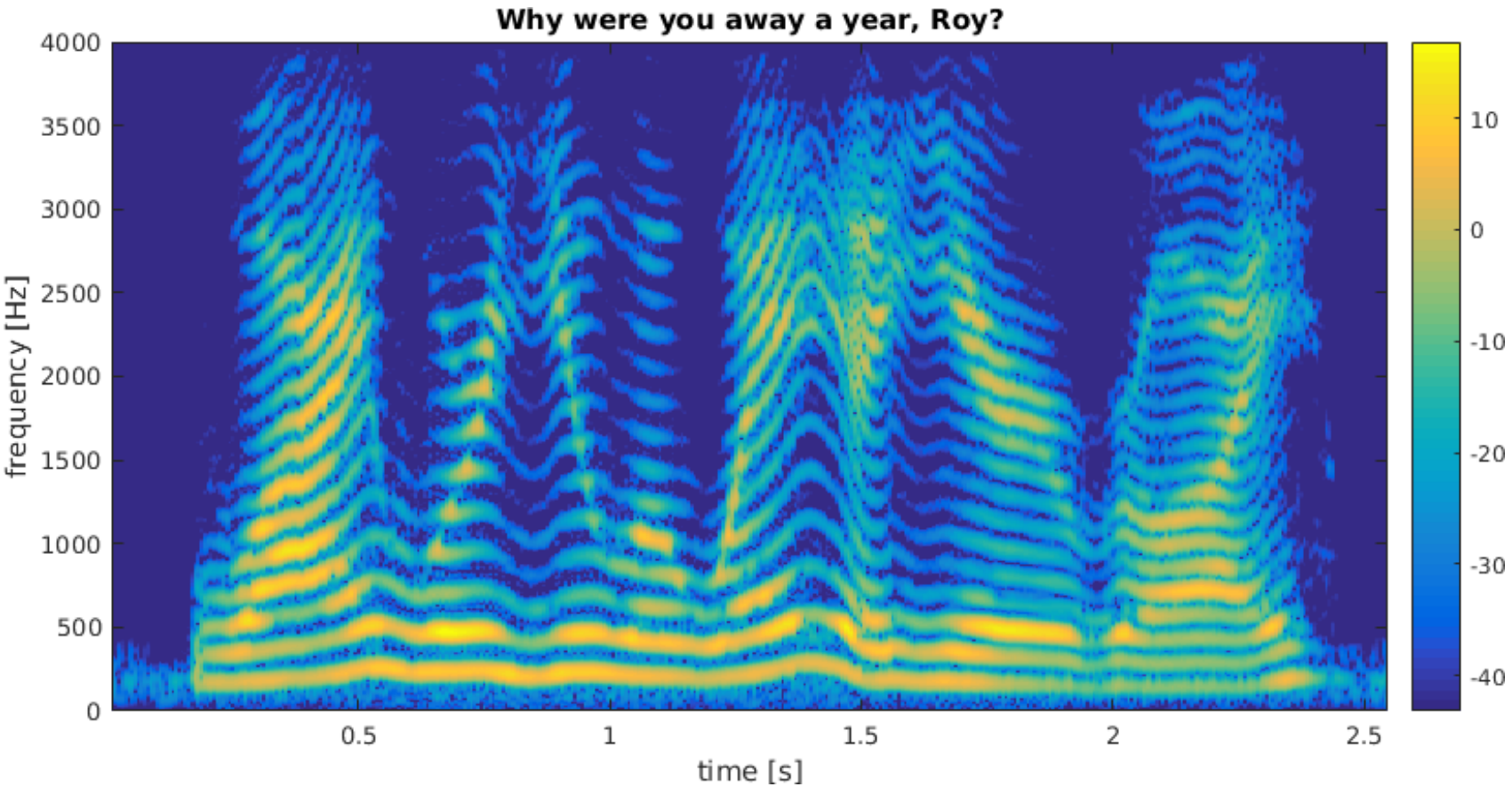


Average computation times in MATLAB

Fast NLS: 7.6 ms, Comb filter: 2.4 ms, YIN: 0.7 ms

Comparison of Methods

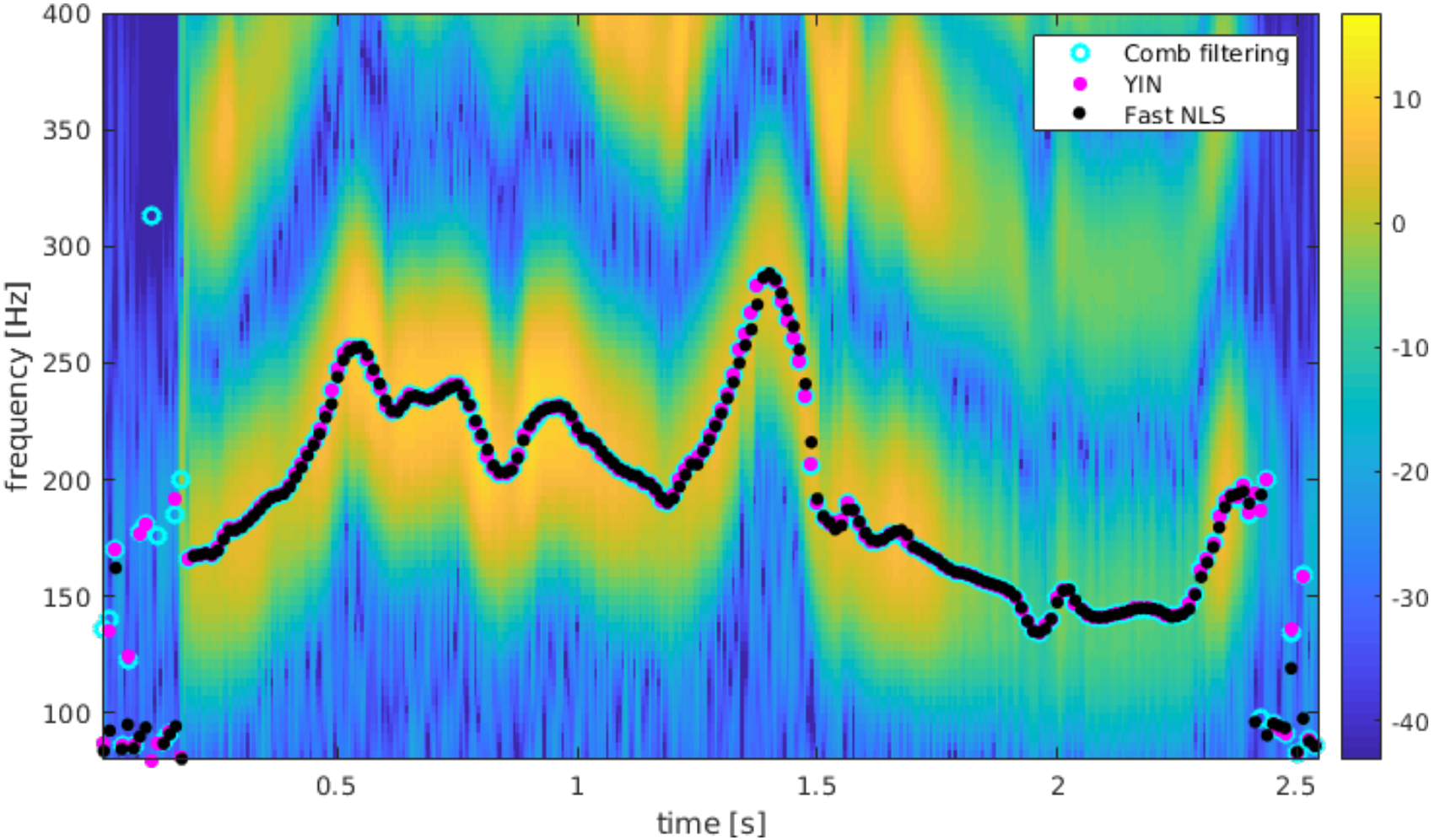
Robustness to noise



Comparison of Methods

Robustness to noise

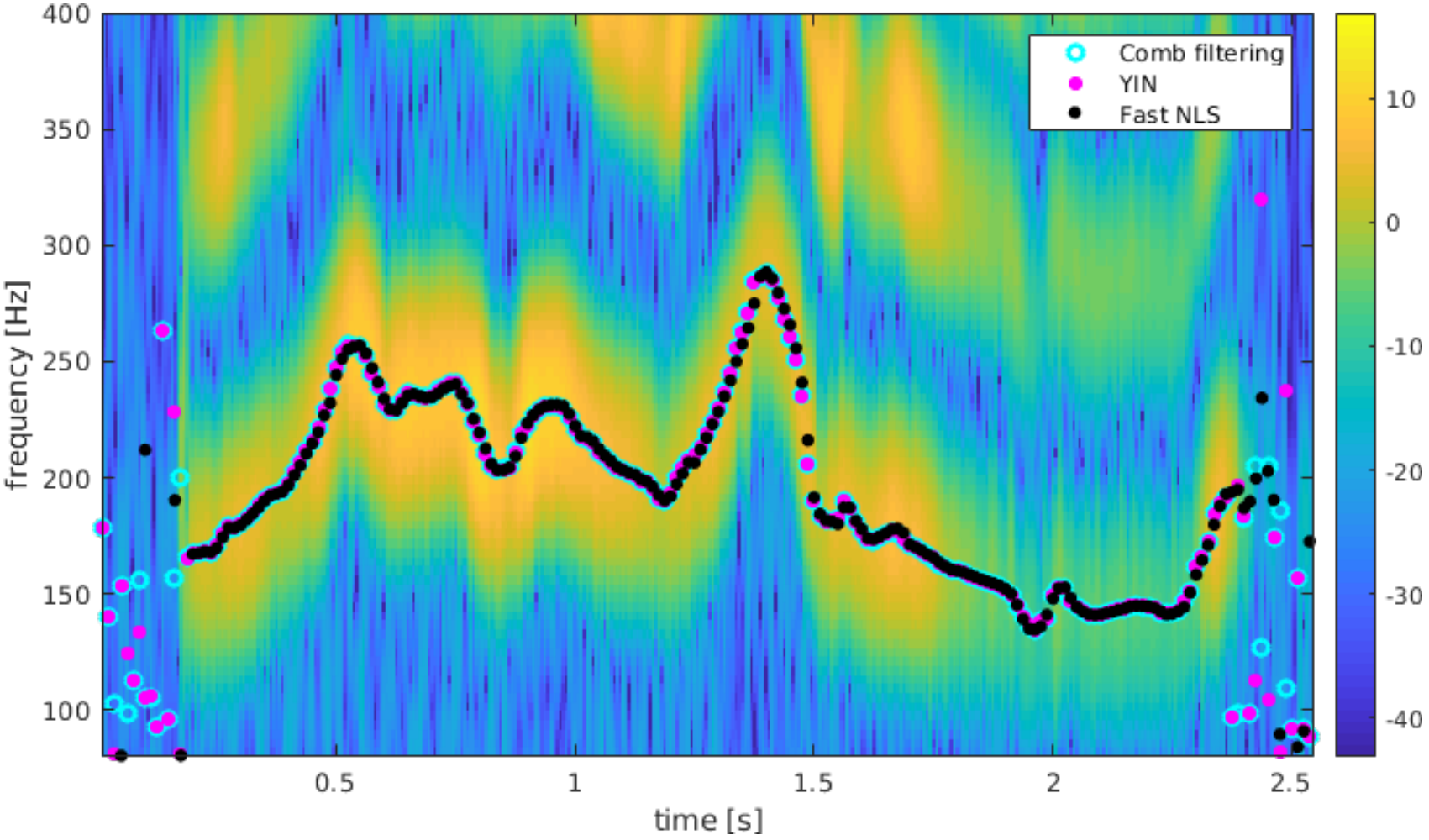
No noise and window size of 25 ms.



Comparison of Methods

Robustness to noise

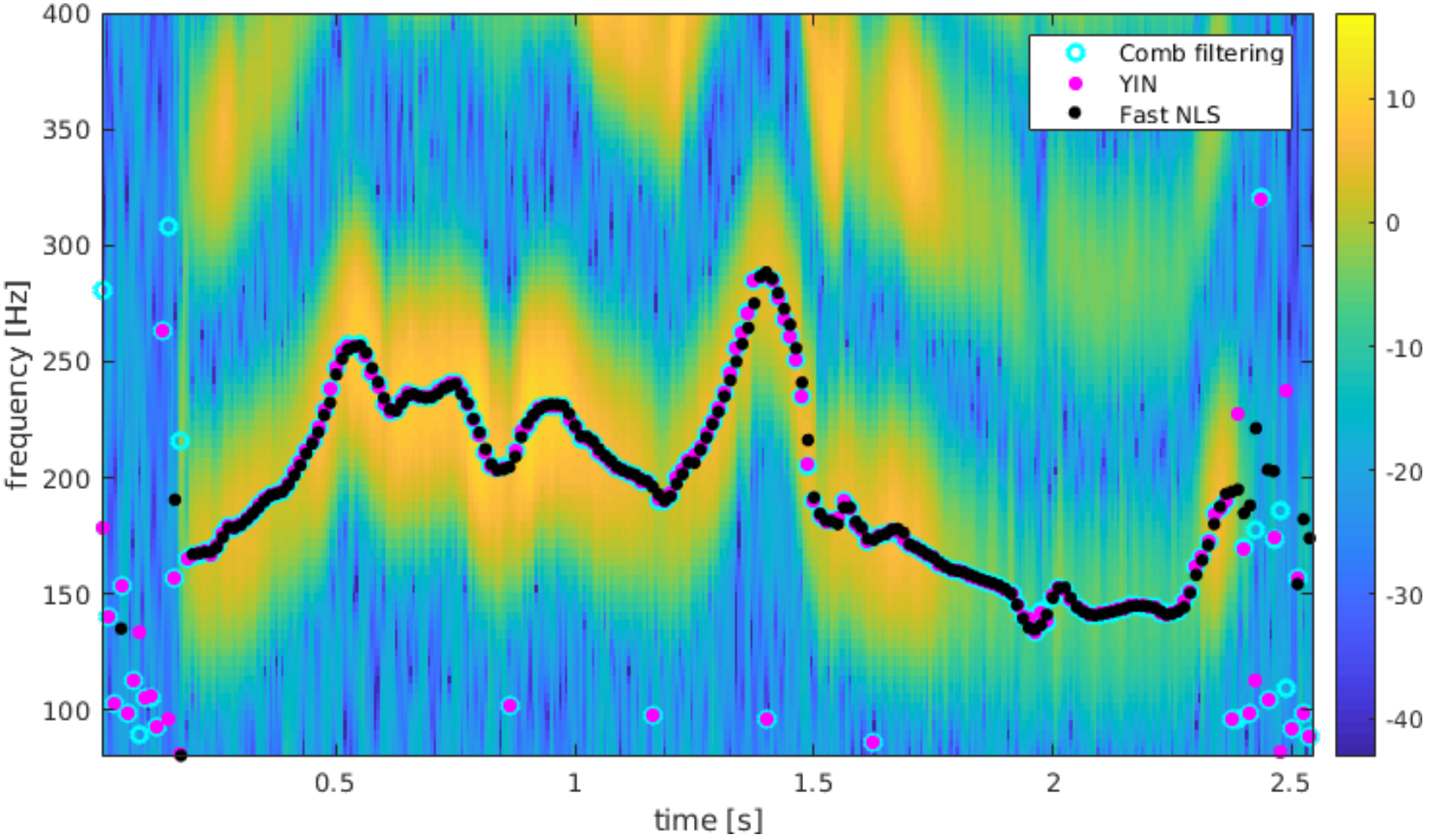
20 dB SNR and window size of 25 ms.



Comparison of Methods

Robustness to noise

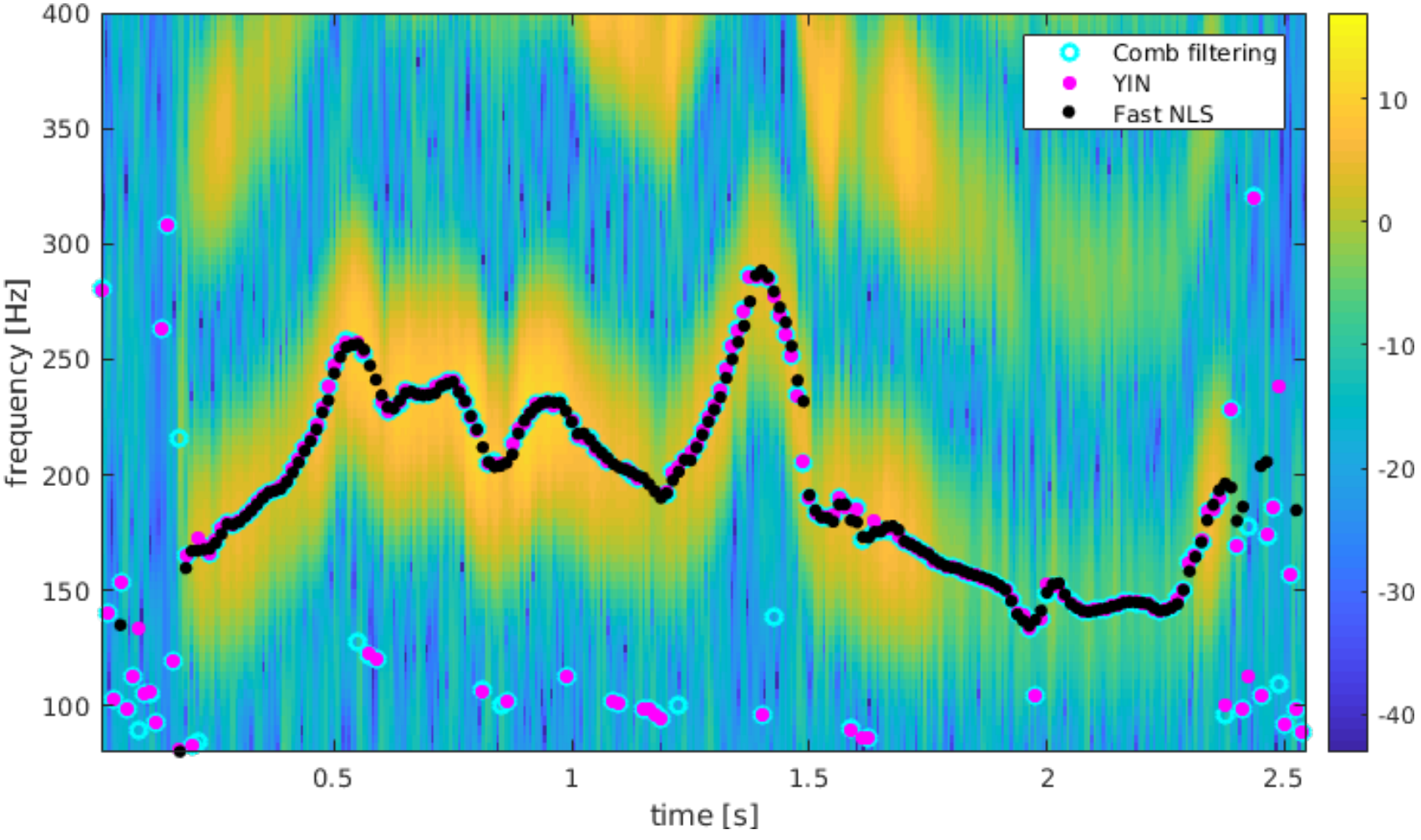
15 dB SNR and window size of 25 ms.



Comparison of Methods

Robustness to noise

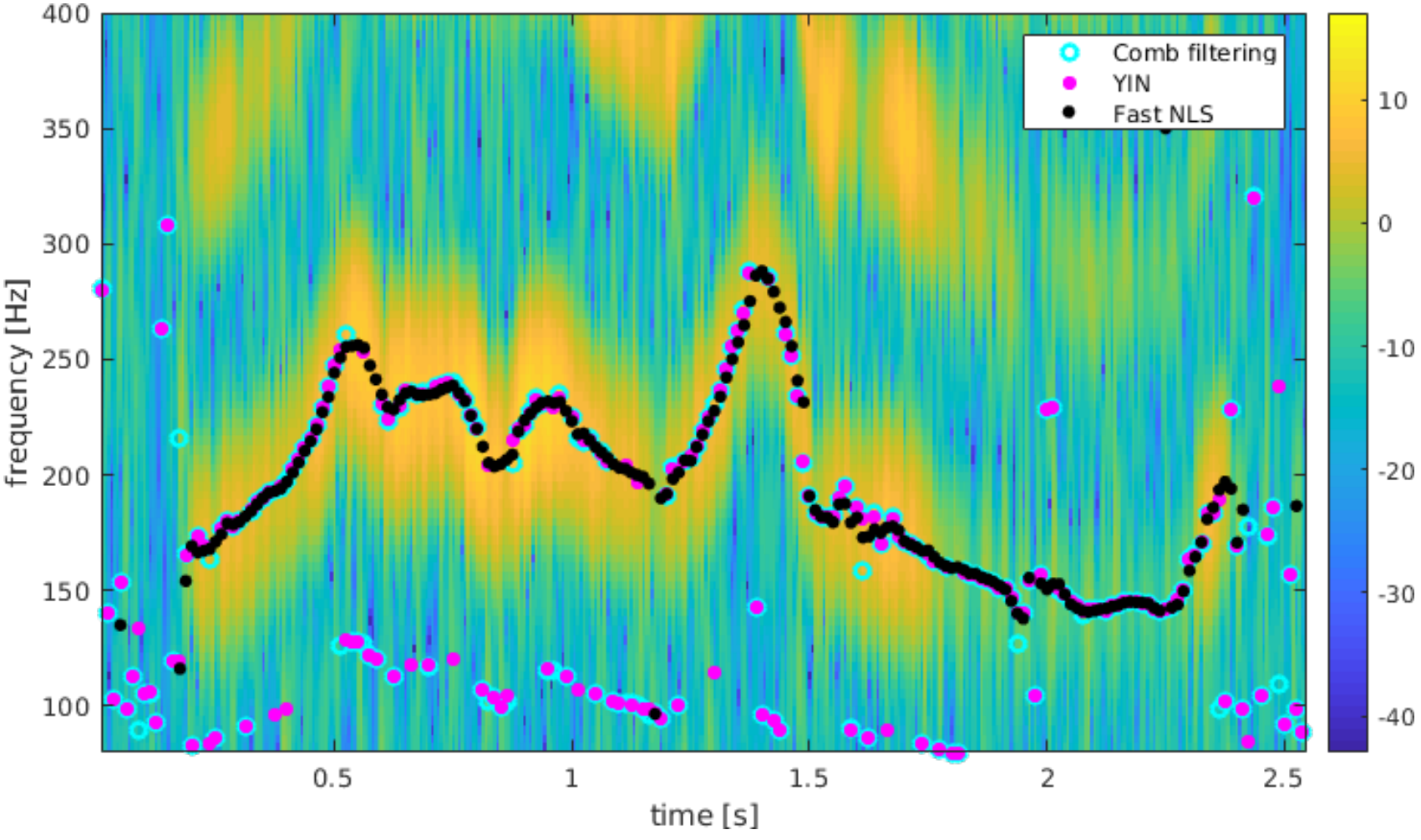
10 dB SNR and window size of 25 ms.



Comparison of Methods

Robustness to noise

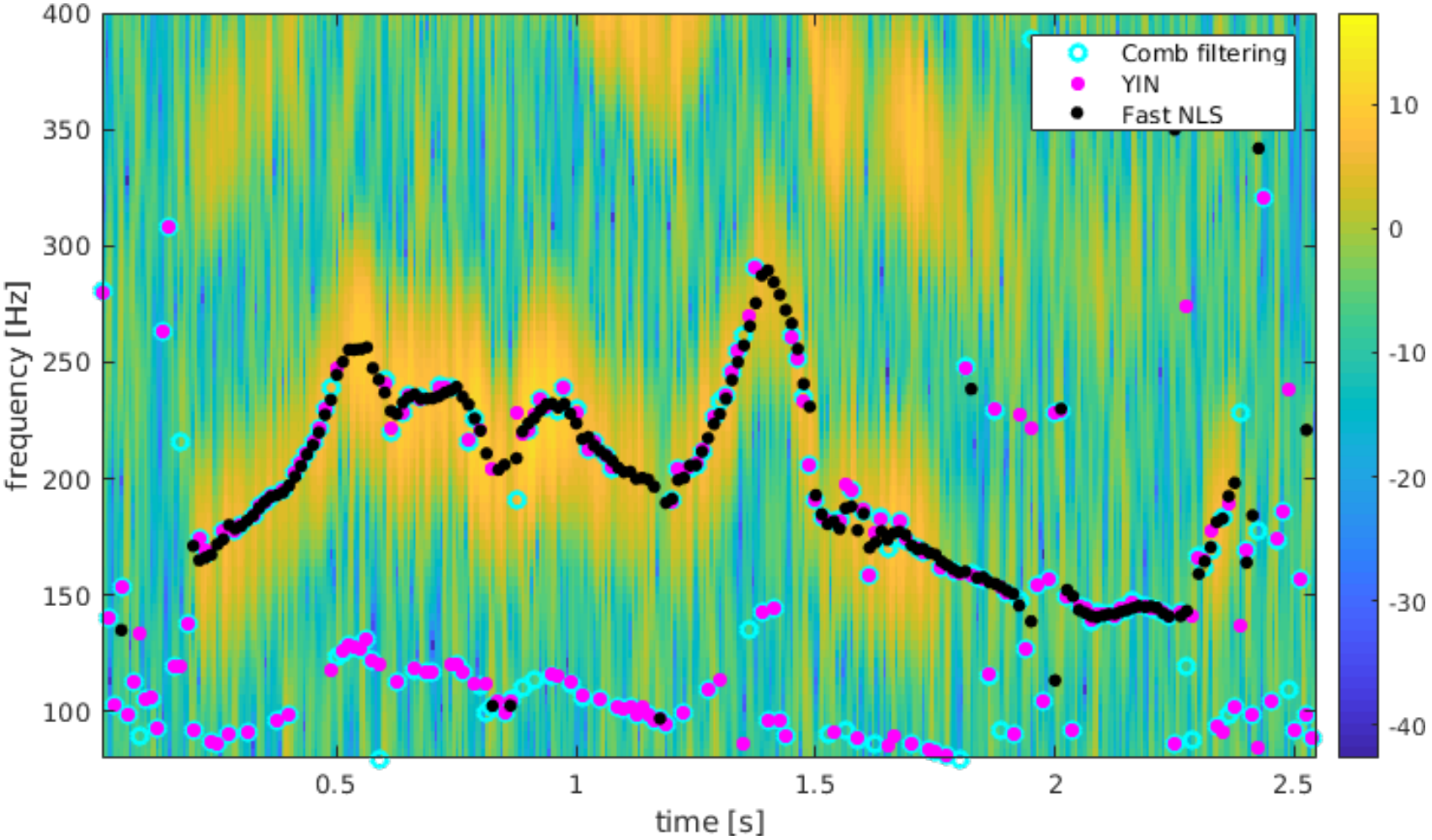
5 dB SNR and window size of 25 ms.



Comparison of Methods

Robustness to noise

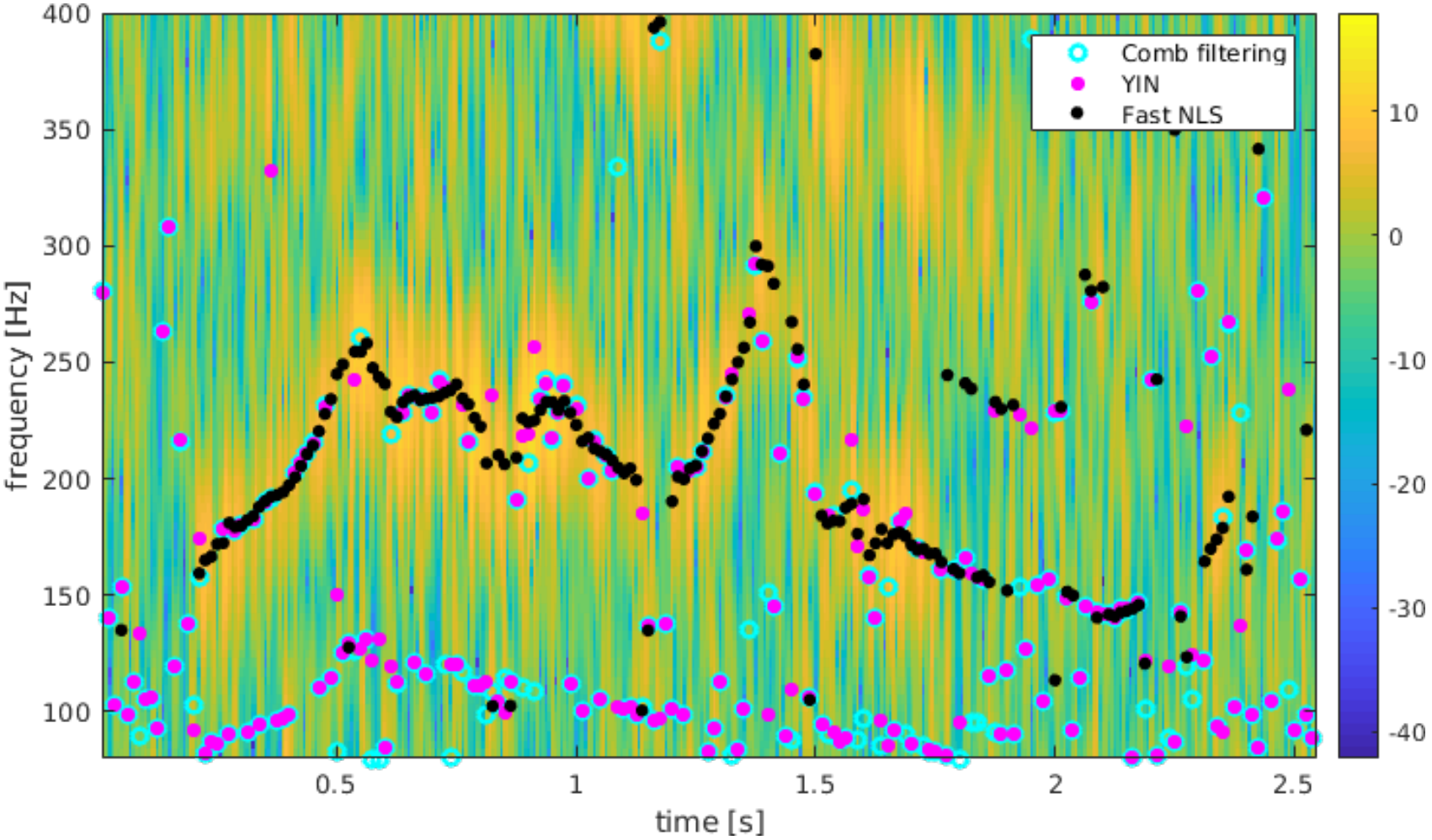
0 dB SNR and window size of 25 ms.



Comparison of Methods

Robustness to noise

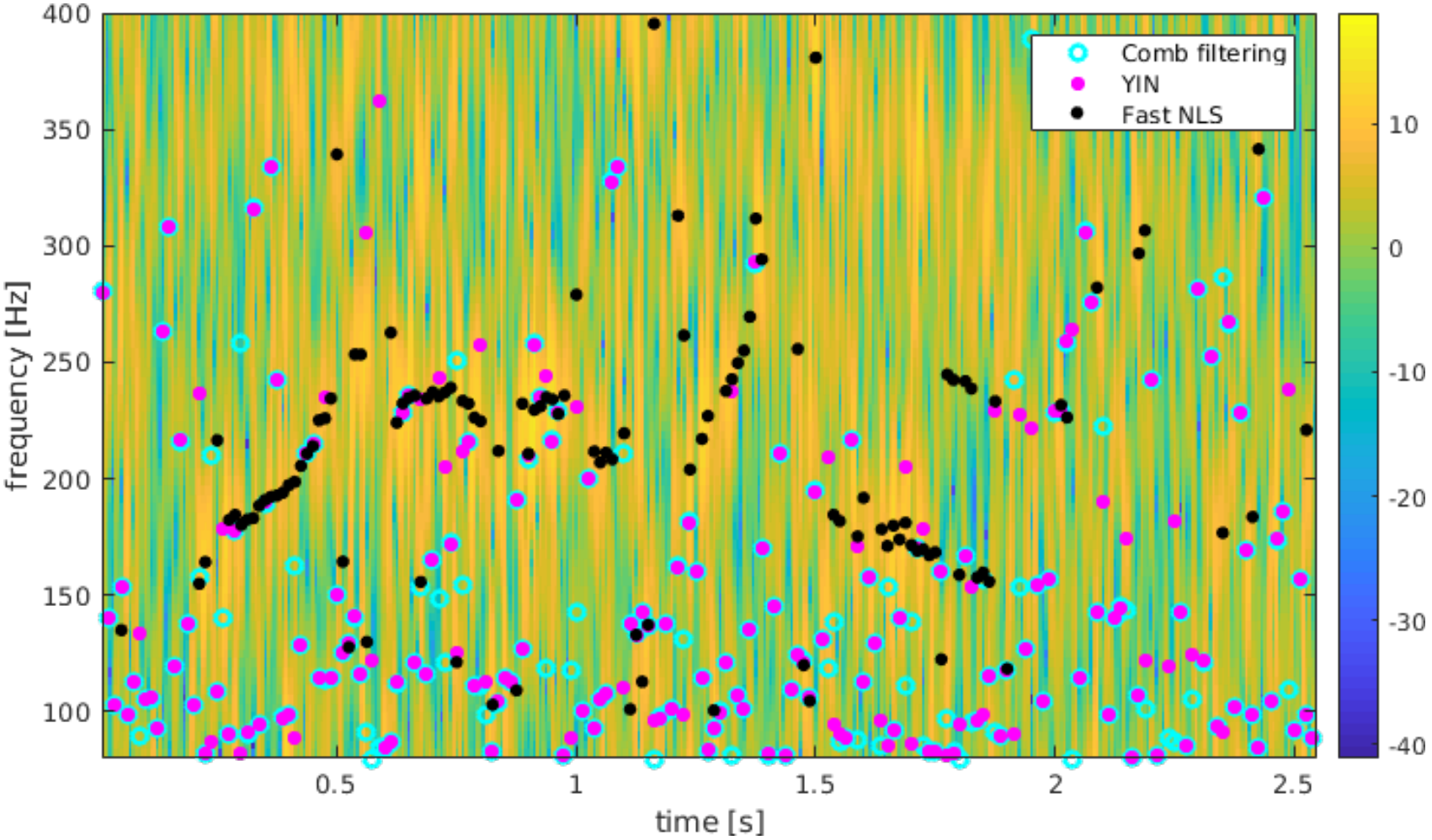
-5 dB SNR and window size of 25 ms.



Comparison of Methods

Robustness to noise

-10 dB SNR and window size of 25 ms.



Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

- Correlation-based Methods
- Nonlinear Least Squares Methods
- Comparison of Methods
  - Robustness to noise
  - Time-frequency resolution
  - Summary
- Non-stationary Pitch Estimation
- Multi-channel Pitch Estimation
- Summary

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

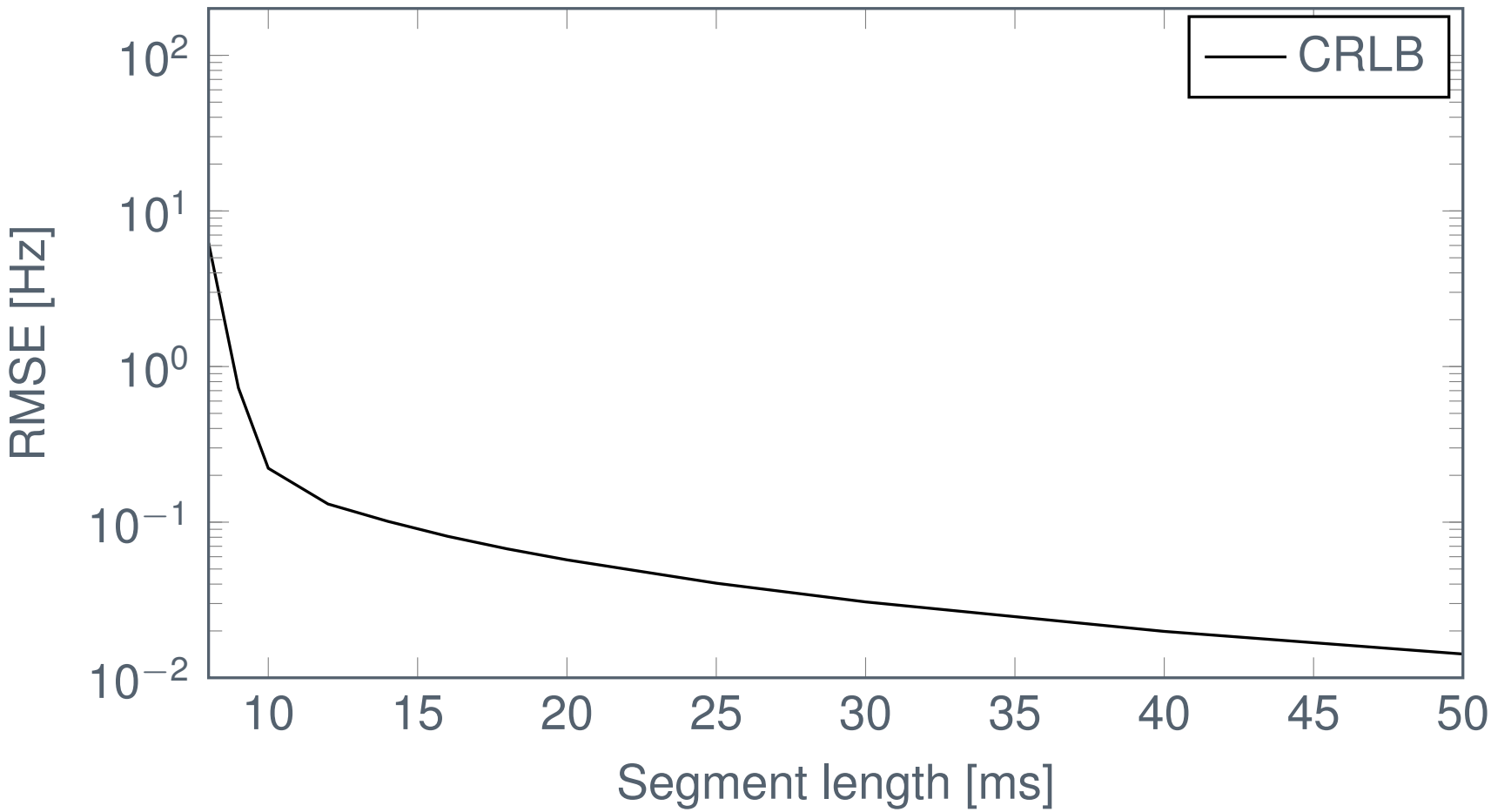
Comparison of Methods
Time-frequency resolution

Simulation setup

- SNR of 30 dB at a sampling frequency of 8000 Hz.
- Estimate the pitch from 1000 Monte Carlo runs for every segment time.
- In each run, the true pitch is randomly selected from $[90, 380]$ Hz and the true phases are also generated at random.
- The true amplitudes are exponentially decreasing.
- The true model order is 7.
- Each method searches for a pitch in the range $[80, 400]$ Hz.
- The maximum model order in NLS is set to 15.
- The noise is white and Gaussian.
- No pitch tracking used in any method.

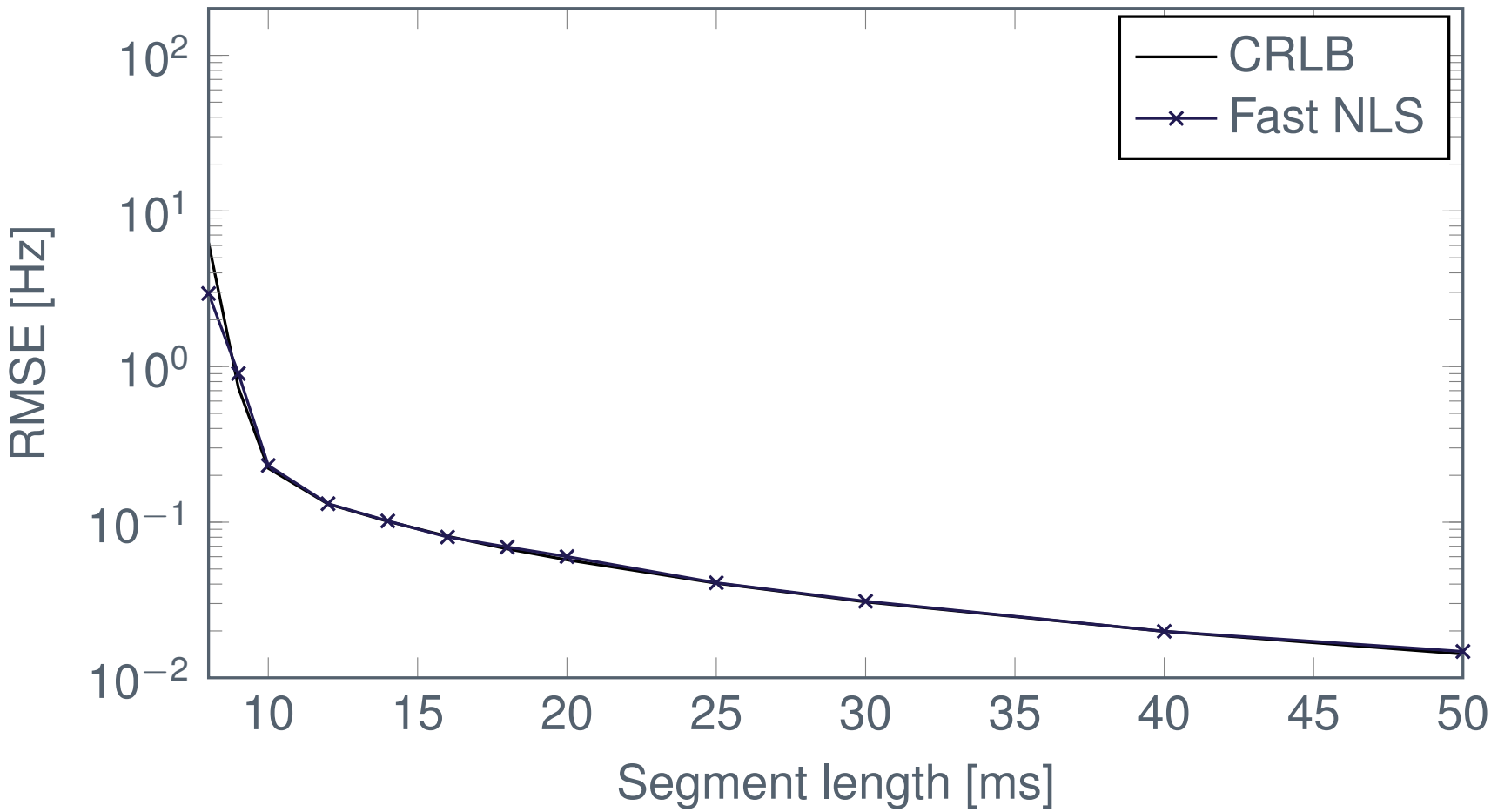
Comparison of Methods

Time-frequency resolution



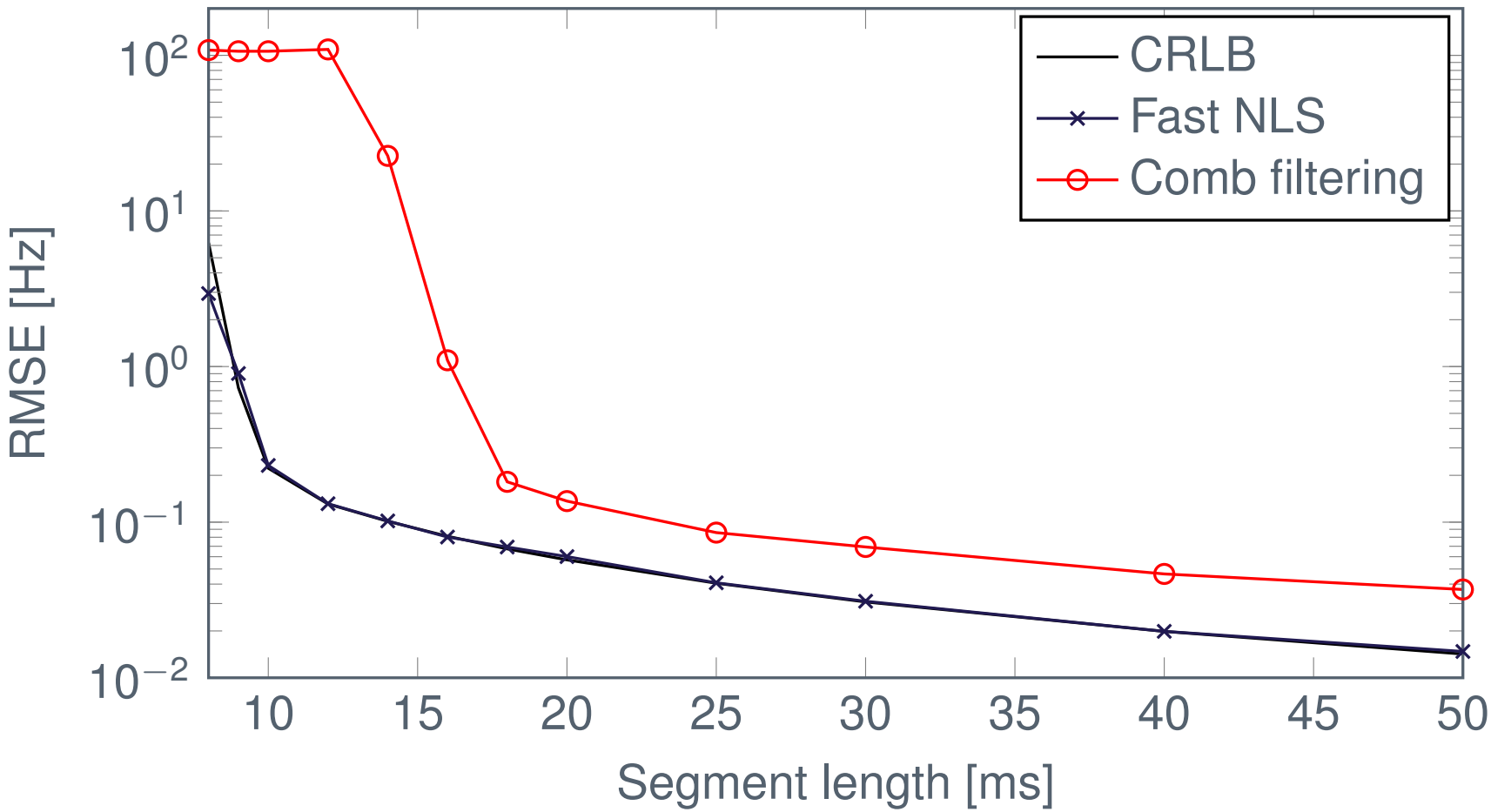
Comparison of Methods

Time-frequency resolution



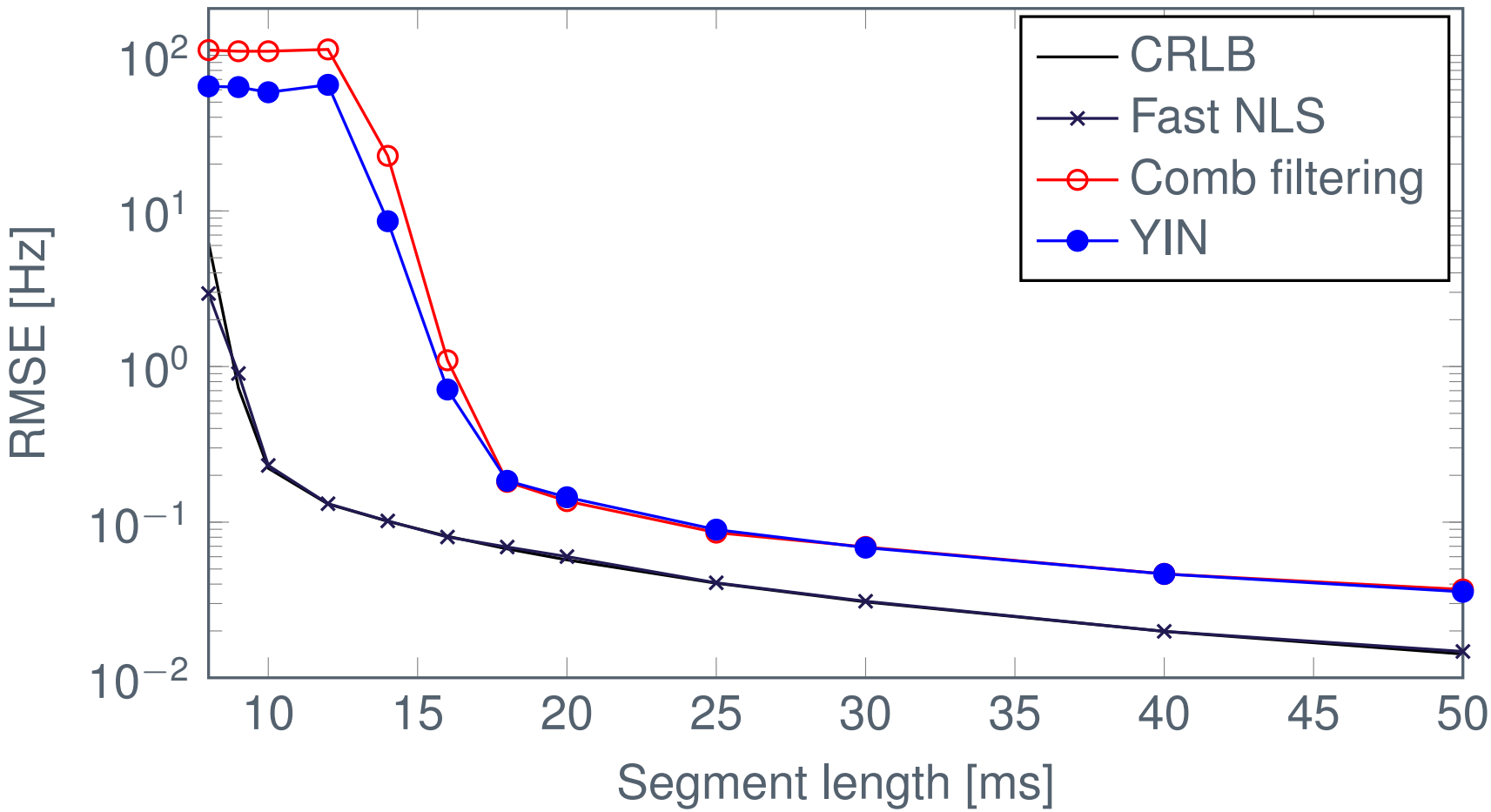
Comparison of Methods

Time-frequency resolution



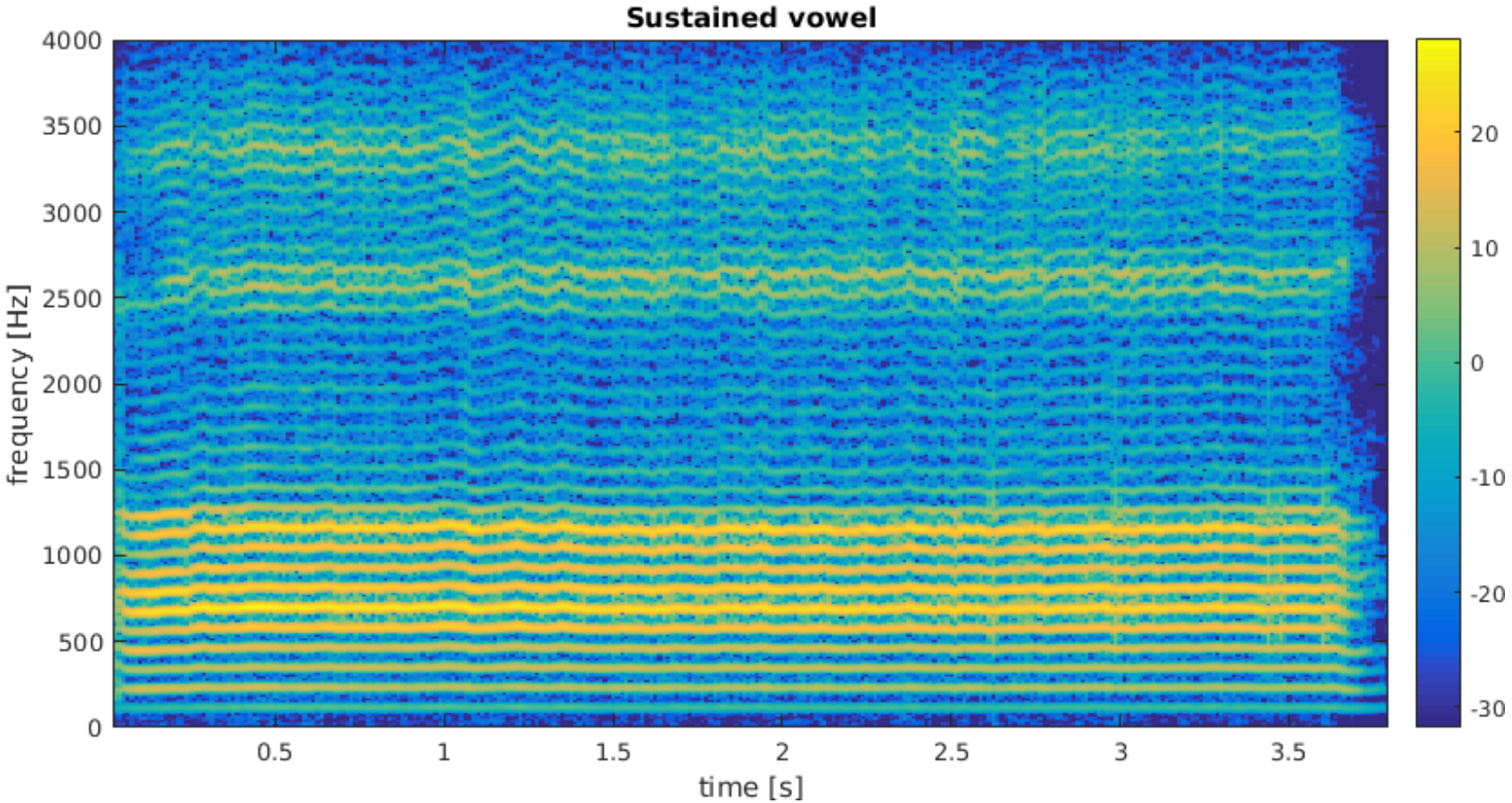
Comparison of Methods

Time-frequency resolution



Comparison of Methods

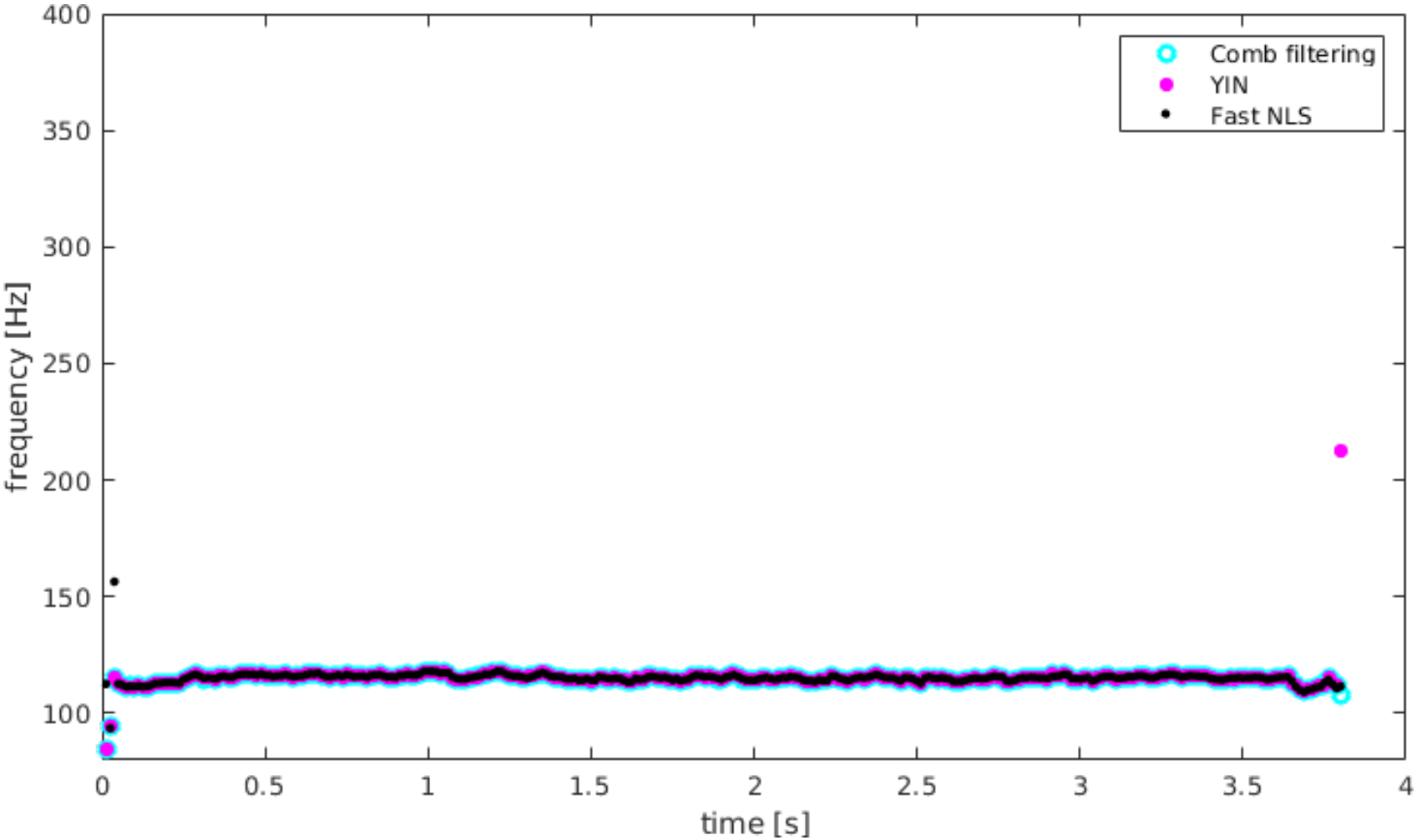
Time-frequency resolution



Comparison of Methods

Time-frequency resolution

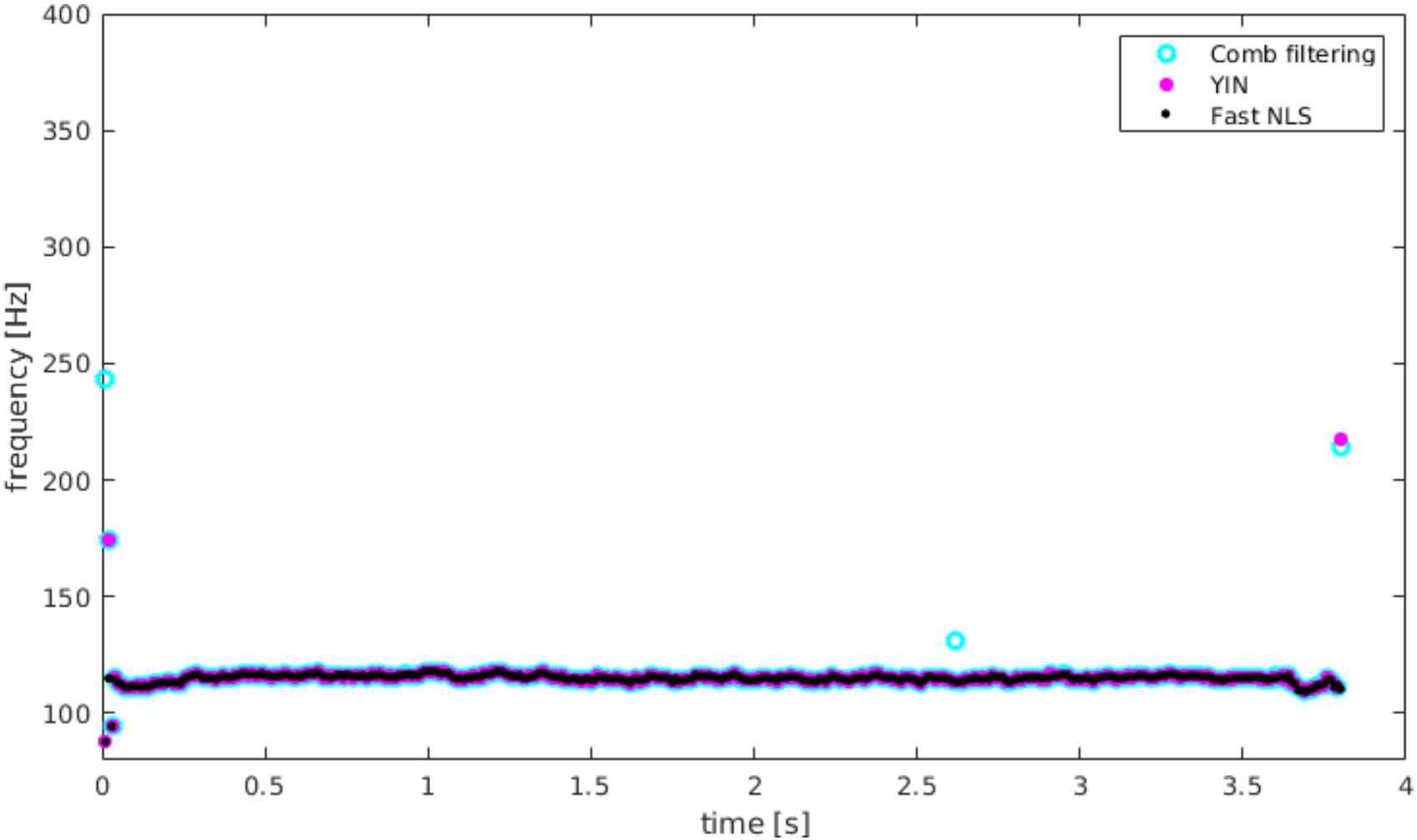
Window size of 25 ms and no noise.



Comparison of Methods

Time-frequency resolution

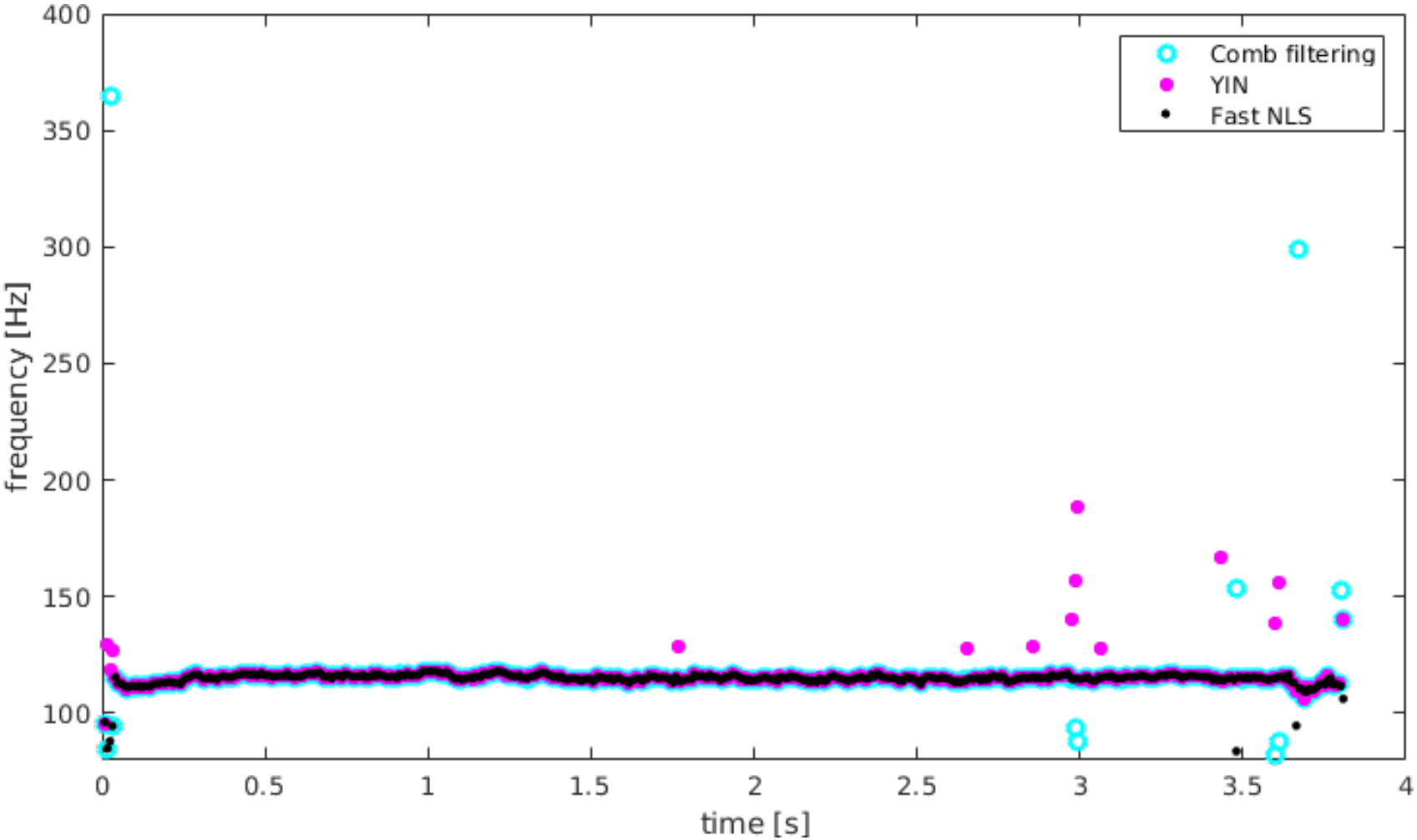
Window size of 20 ms and no noise.



Comparison of Methods

Time-frequency resolution

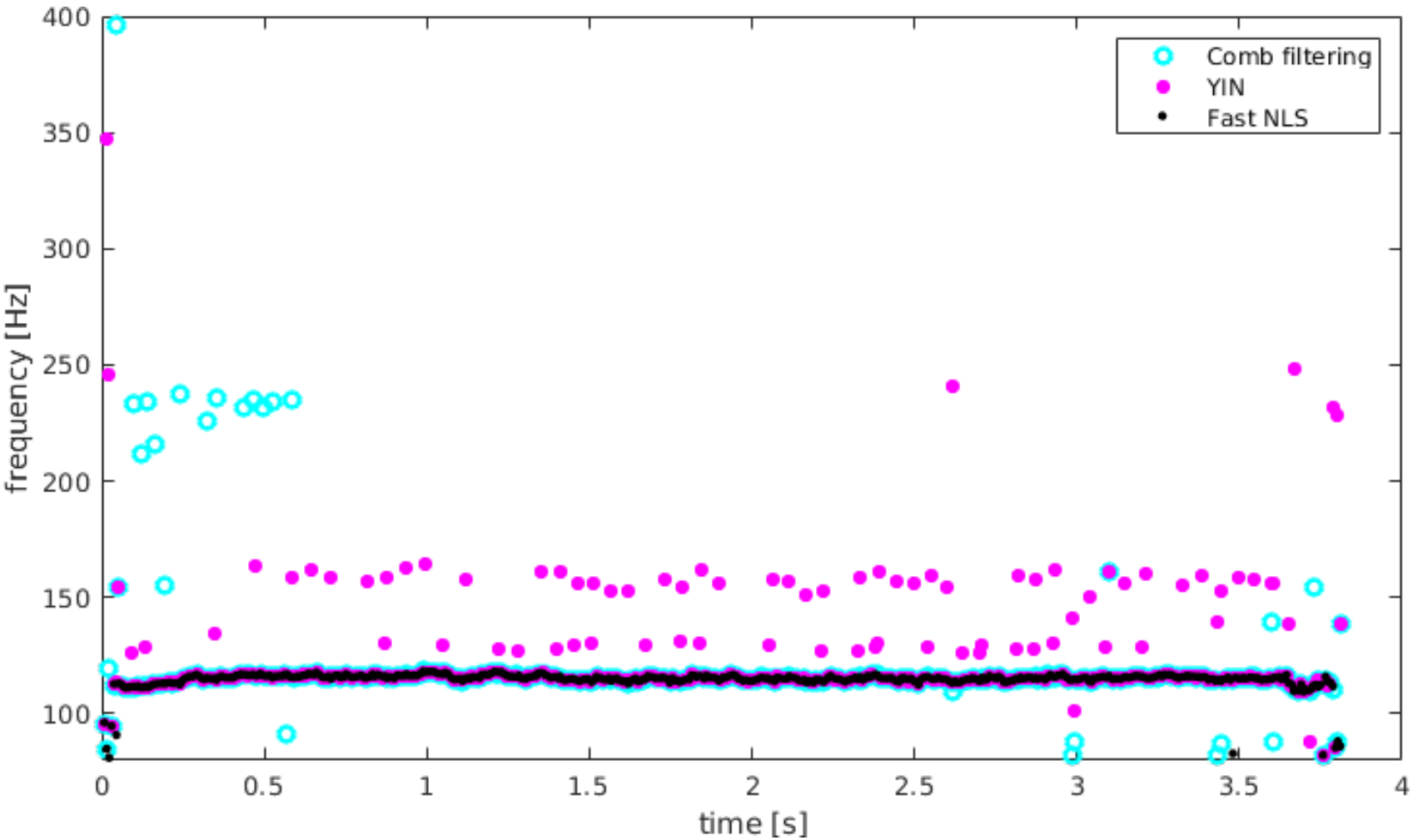
Window size of 16 ms and no noise.



Comparison of Methods

Time-frequency resolution

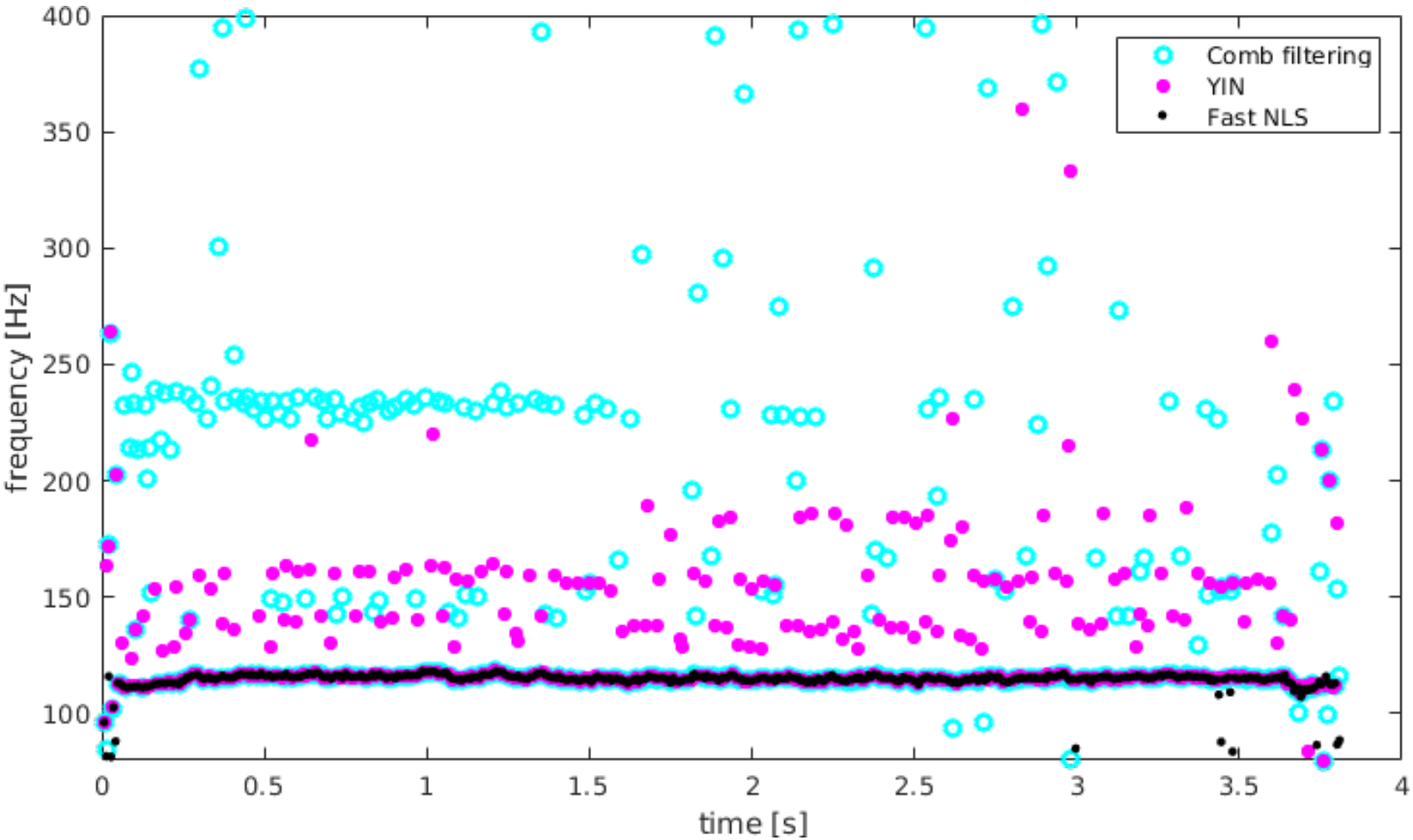
Window size of 15 ms and no noise.



Comparison of Methods

Time-frequency resolution

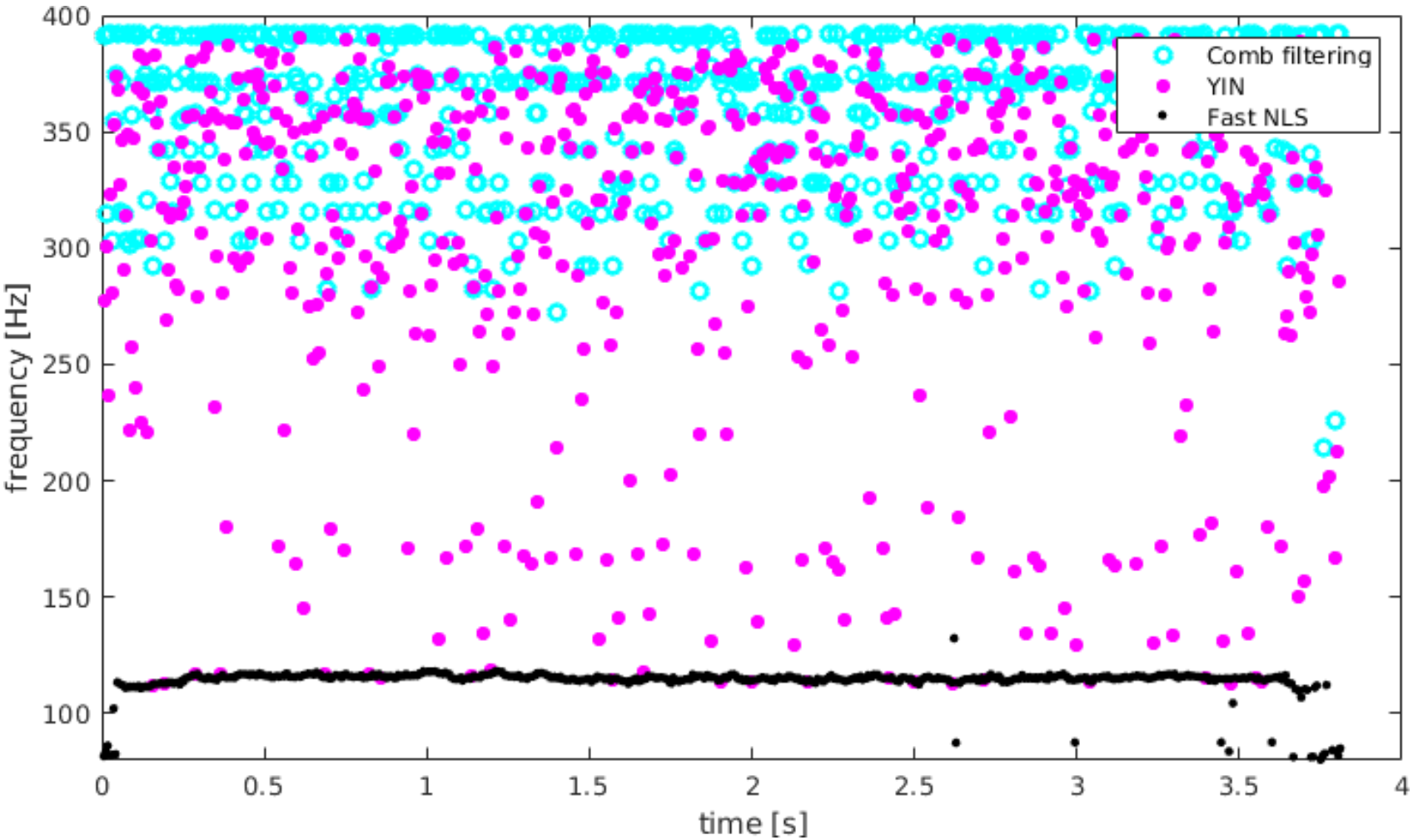
Window size of 14 ms and no noise.



Comparison of Methods

Time-frequency resolution

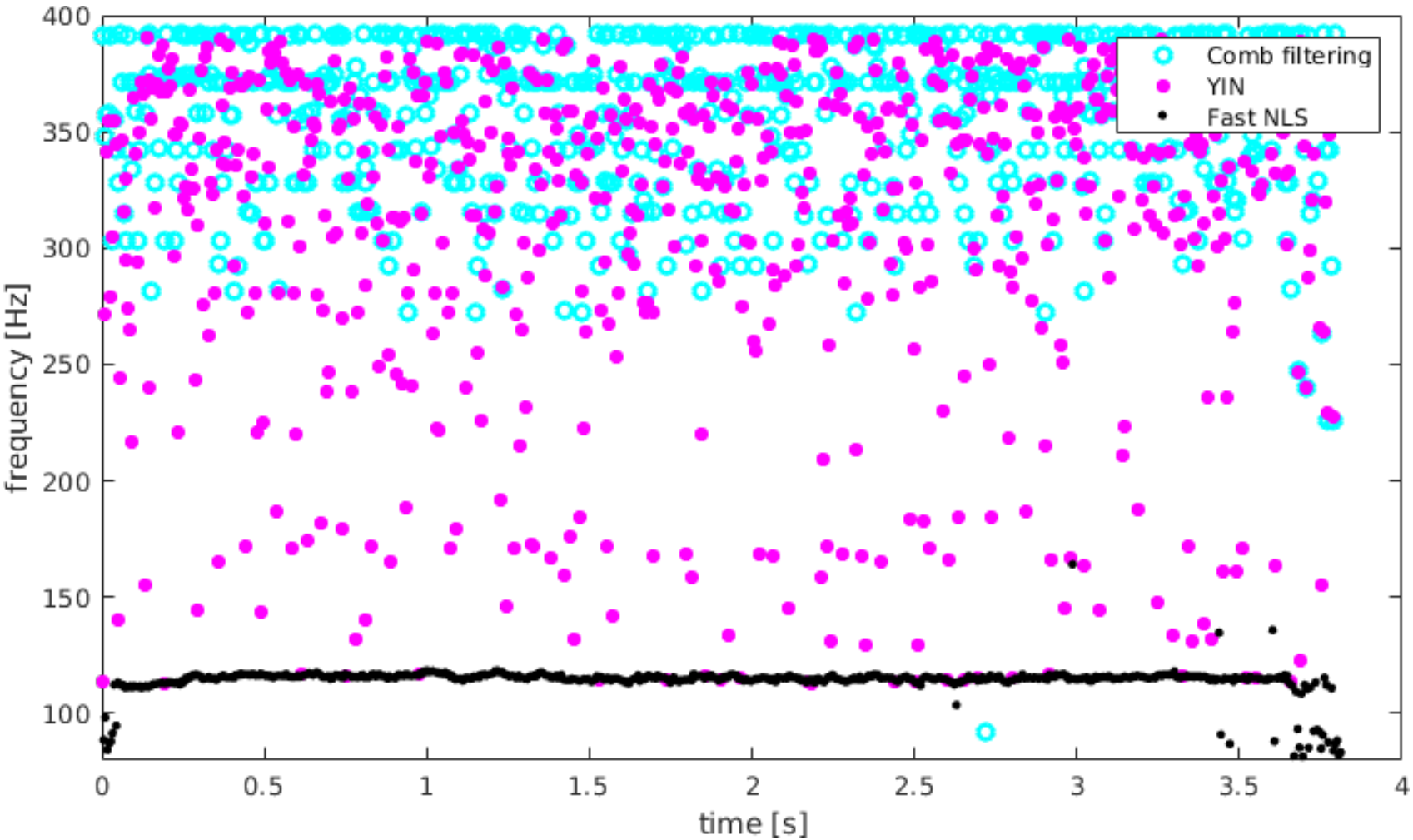
Window size of 12 ms and no noise.



Comparison of Methods

Time-frequency resolution

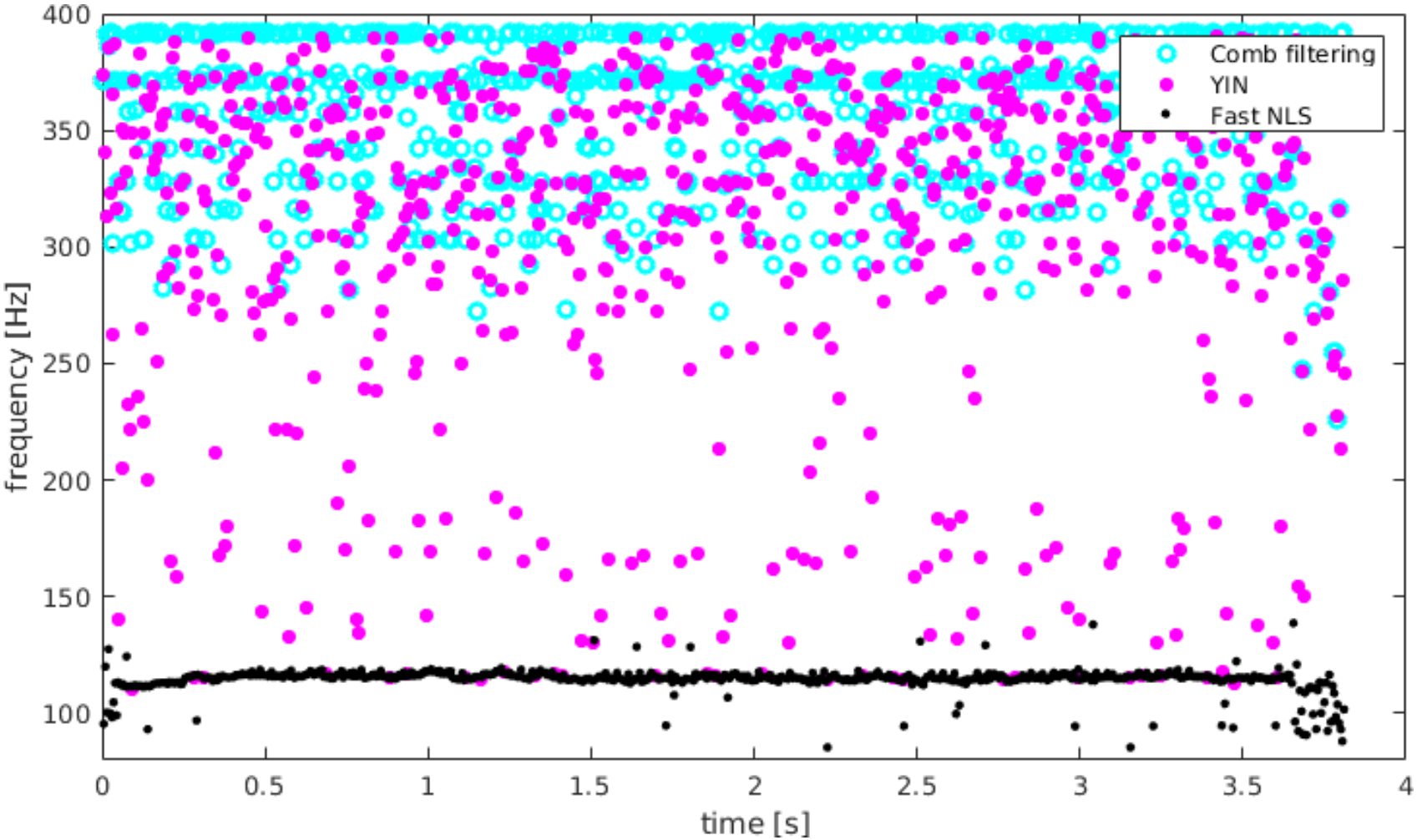
Window size of 11 ms and no noise.



Comparison of Methods

Time-frequency resolution

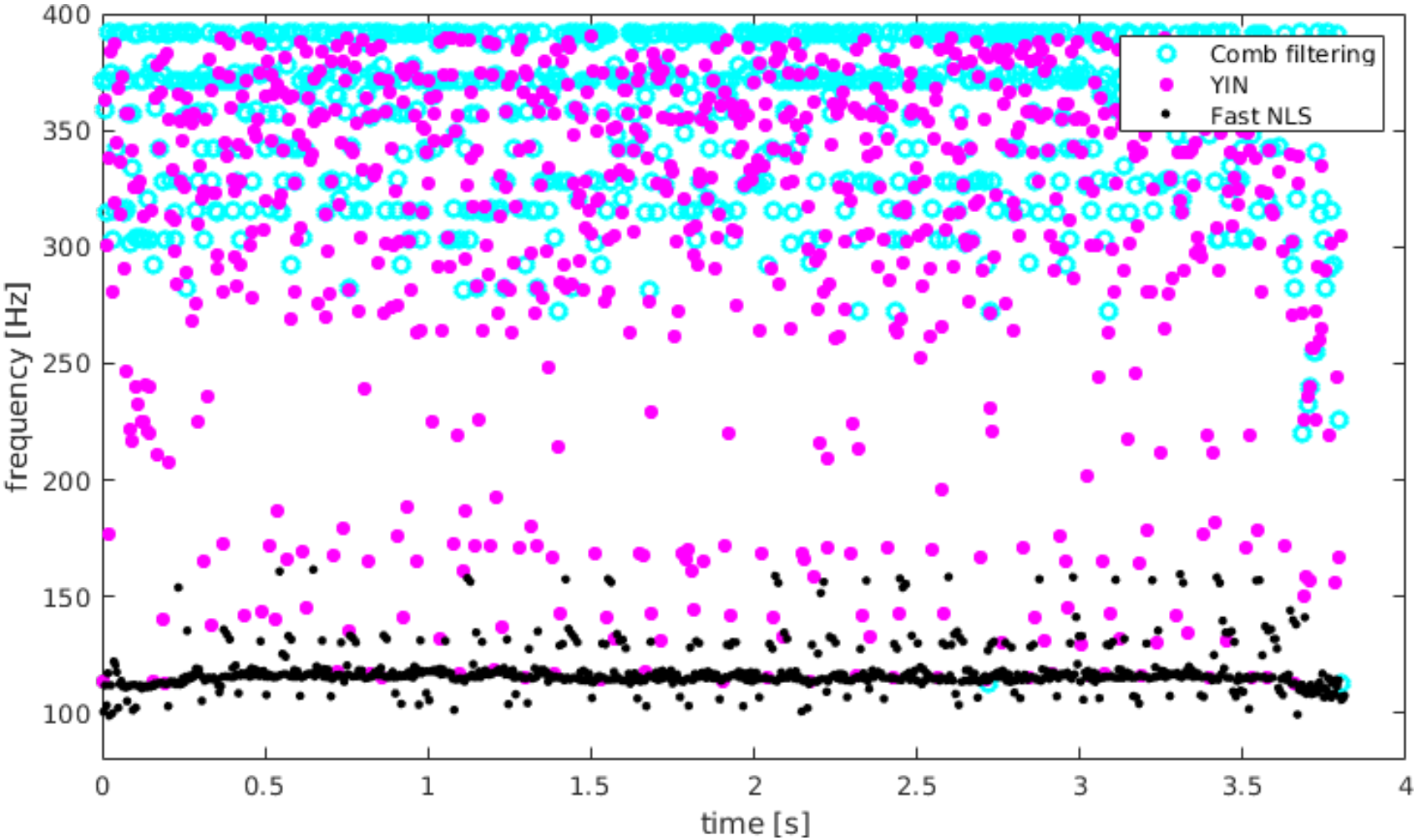
Window size of 10 ms and no noise.



Comparison of Methods

Time-frequency resolution

Window size of 9 ms and no noise.



Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

- Correlation-based Methods
- Nonlinear Least Squares Methods
- Comparison of Methods
  - Robustness to noise
  - Time-frequency resolution
  - Summary
- Non-stationary Pitch Estimation
- Multi-channel Pitch Estimation
- Summary

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Comparison of Methods
Summary

Correlation-based Methods
A periodic signal satisfies that

$$x(n) = x(n - \tau) \tag{115}$$

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

- \+ Intuitive and simple
- \+ Low computational complexity
- \+ Mature and refined set of methods
- +/- No need to estimate the model order
- \- Interpolation needed for fractional delay estimation
- \- Poor time-frequency resolution
- \- Are sensitive to noise

Comparison of Methods
Summary

Parametric Methods
Estimate the parameters in

$$x(n) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l) + e(n) \tag{116}$$

- \+ High estimation accuracy
- \+ Work very well in even noisy conditions
- \+ Good time-frequency resolution
- +/- The model order has to be estimated
- \- Higher computational complexity
- \- Early stage methods without fine tuning (yet)
- \- Might produce over-optimistic results (e.g., due to non-stationarity)

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

- Correlation-based Methods
- Nonlinear Least Squares Methods
- Comparison of Methods
- Non-stationary Pitch Estimation
- Multi-channel Pitch Estimation
- Summary

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Non-stationary Pitch Estimation

- Real-world signals are 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

$$\theta_l(n) \approx \phi_l + l\omega_0 n + l\beta_0 n^2/2 \tag{117}$$

  where β_0 is the fundamental chirp rate.
- We refer to this model as the harmonic chirp model

$$s(n) = \sum_{l=1}^{L} A_l \cos(l\beta_0 n^2/2 + l\omega_0 n + \phi_l) \tag{118}$$

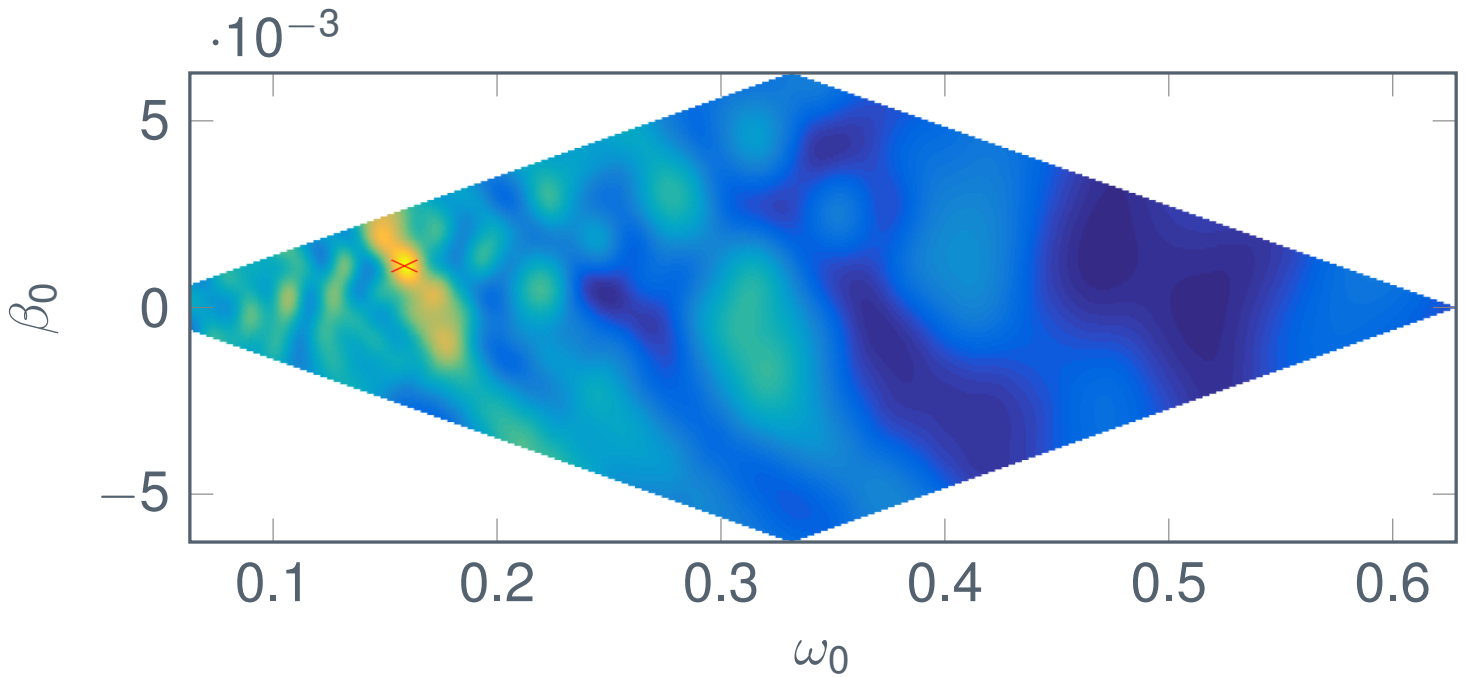
Non-stationary Pitch Estimation

Nonlinear least squares (NLS) objective

$$J_L(\omega_0, \beta_0) = \boldsymbol{x}^T \boldsymbol{Z}_L(\omega_0, \beta_0) \left[\boldsymbol{Z}_L^T(\omega_0, \beta_0) \boldsymbol{Z}_L(\omega_0, \beta_0) \right]^{-1} \boldsymbol{Z}_L^T(\omega_0, \beta_0) \boldsymbol{x} \tag{119}$$

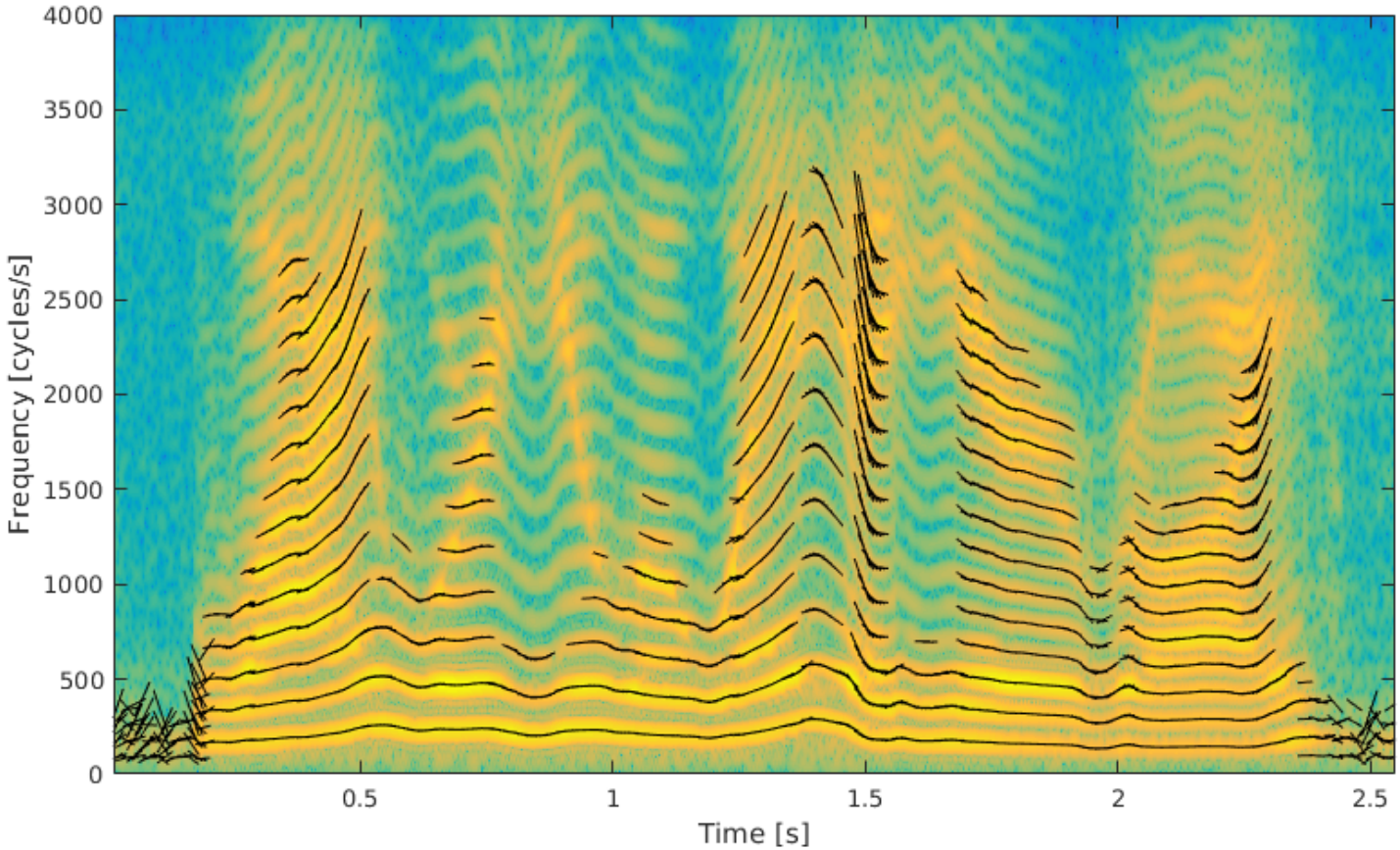
Harmonic chirp summation objective:

$$J_L(\omega_0, \beta_0) = \boldsymbol{x}^T \boldsymbol{Z}_L(\omega_0, \beta_0) \boldsymbol{Z}_L^T(\omega_0, \beta_0) \boldsymbol{x} \tag{120}$$



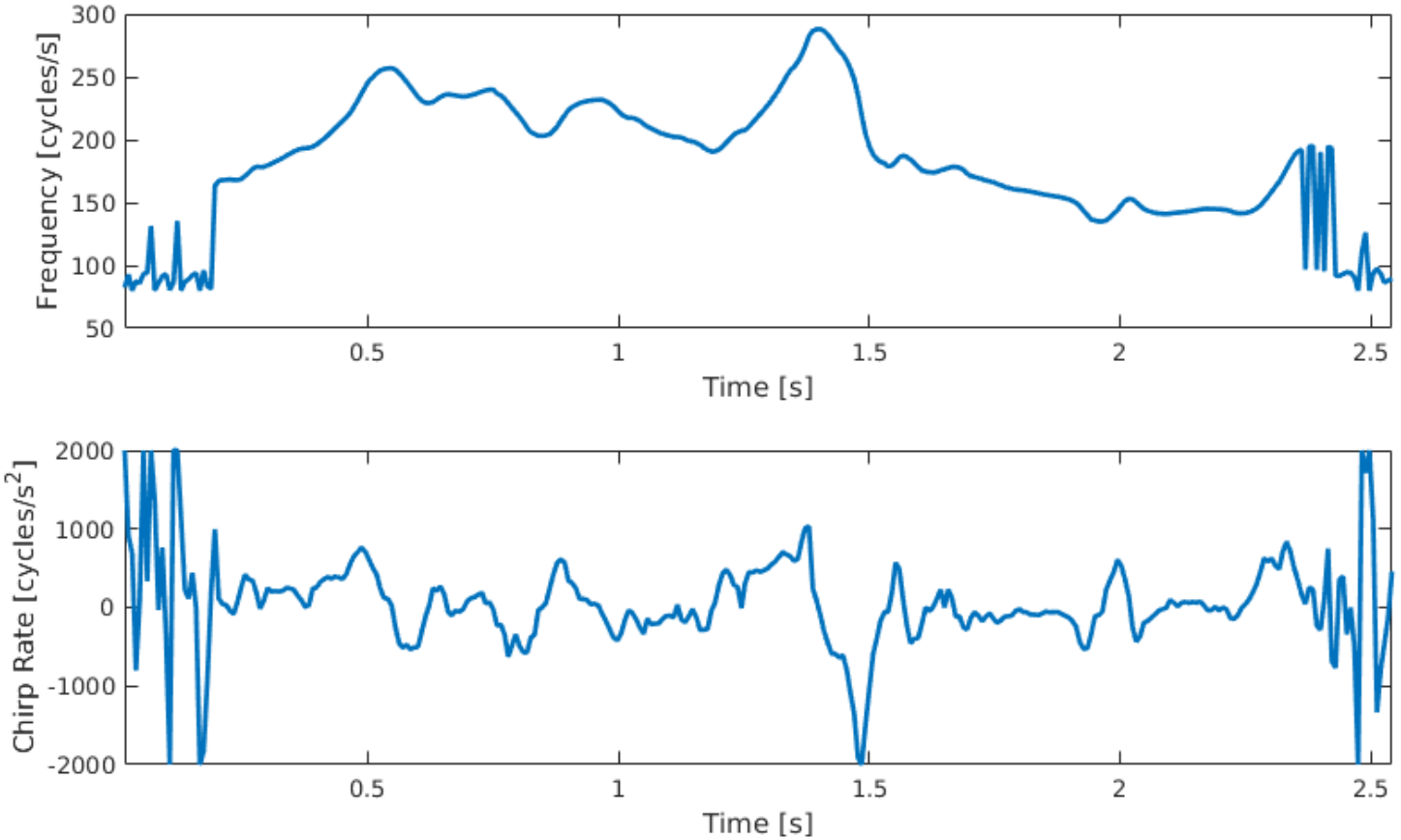
Non-stationary Pitch Estimation

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



Non-stationary Pitch Estimation

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



Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

- Correlation-based Methods
- Nonlinear Least Squares Methods
- Comparison of Methods
- Non-stationary Pitch Estimation
- Multi-channel Pitch Estimation
- Summary

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

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

$$x_k(n) = \sum_{l=1}^{L} A_{l,k} \cos(l\omega_0 n + \phi_{l,k}) + e_k(n) \tag{121}$$

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

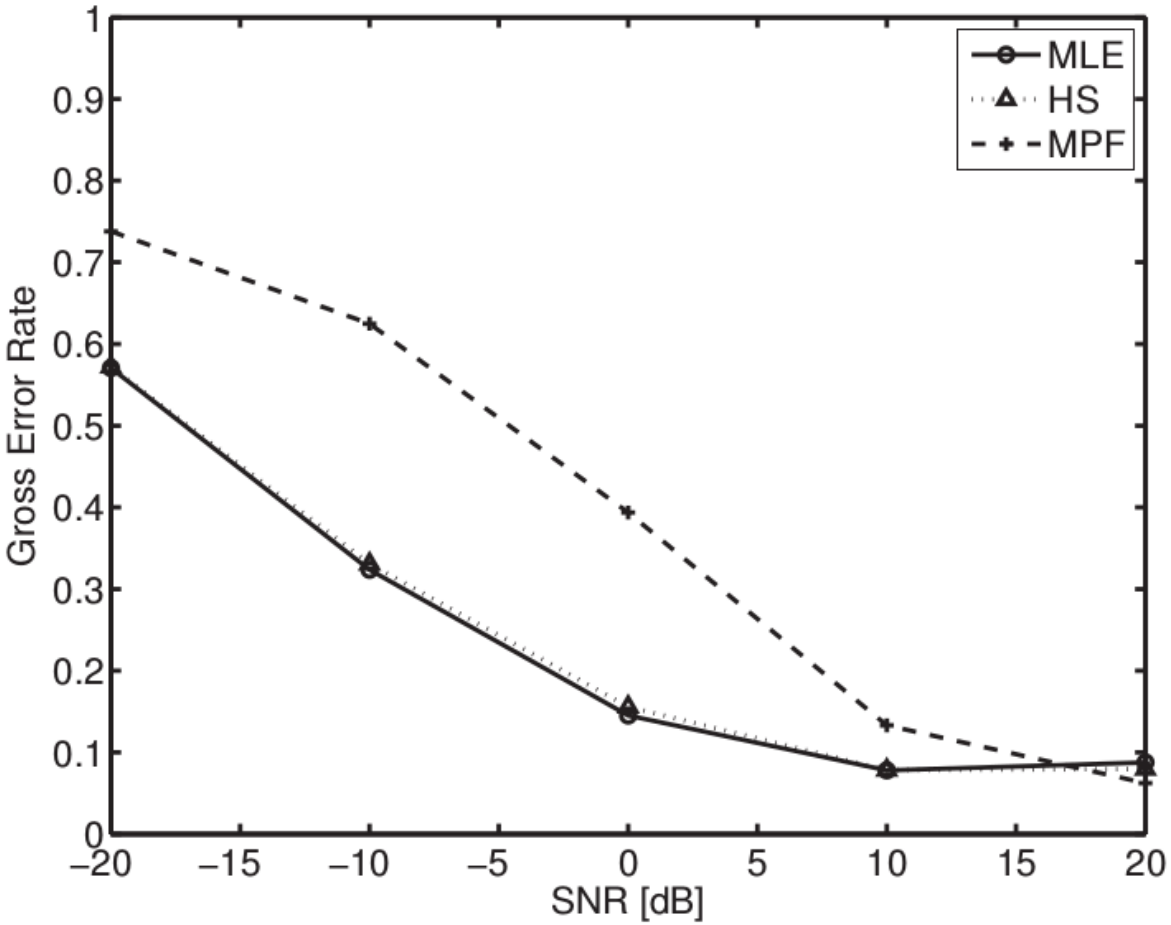
$$J_L(\omega_0) = \sum_{k=1}^{K} \boldsymbol{x}_k^T \boldsymbol{Z}_L(\omega_0) \left[\boldsymbol{Z}_L^T(\omega_0)\boldsymbol{Z}_L(\omega_0)\right]^{-1} \boldsymbol{Z}_L^T(\omega_0)\boldsymbol{x}_k$$

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

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

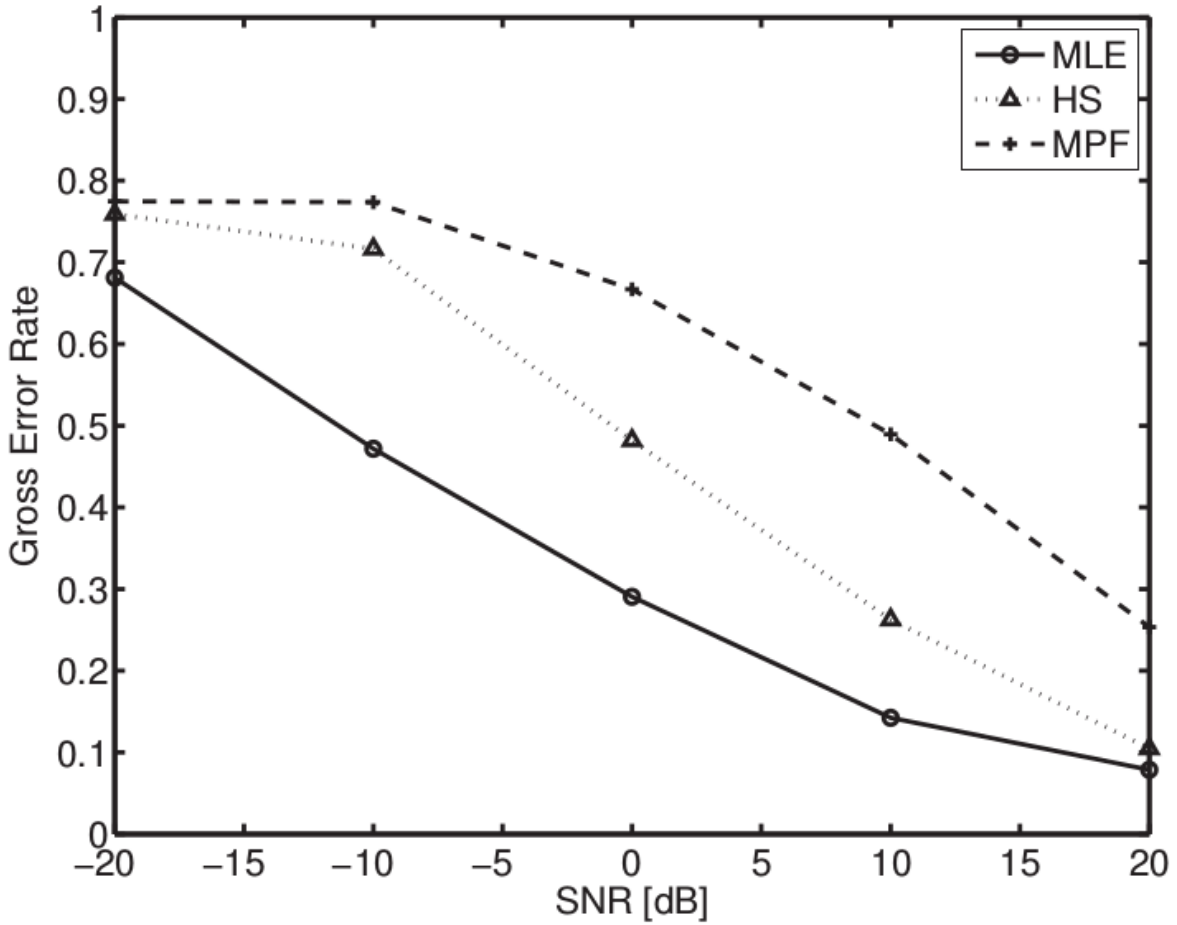
Multi-channel pitch estimation

Same noise variance on every channel.



Multi-channel pitch estimation

Different noise variances on every channel.



Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

- Correlation-based Methods
- Nonlinear Least Squares Methods
- Comparison of Methods
- Non-stationary Pitch Estimation
- Multi-channel Pitch Estimation
- Summary

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

Summary and Conclusion

Summary

- Published correlation-based methods are more mature than published parametric methods in that they tend to include everything (pitch detection, estimation, and tracking) and are less computationally costly.
- However, parametric pitch estimation methods typically outperform correlation-based methods in terms of estimation accuracy, noise robustness, and time-frequency resolution.
- The modelling assumptions are explicit in parametric methods.
- Consequently, we can easily extend the model to take more complex phenomena into account.
- Besides NLS, examples of other parametric methods are subspace and filtering methods (Christensen and Jakobsson, 2009).

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement
- Classical Optimal Filtering
- Model-based Speech Enhancement
- Enhancement Example
- Summary

Model-based Array Processing and Enhancement

Summary and Conclusion

Speech Enhancement

The speech enhancement problem

We observe a noisy speech signal

$$x(n) = s(n) + e(n) \tag{122}$$

where

$s(n)$ is the clean speech, and

$e(n)$ is the noise.

We wish to improve the speech intelligibility and quality, but

- we have two unknowns for every observation so
- we need prior information about the speech and/or noise to solve the problem

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement
- Classical Optimal Filtering
- Model-based Speech Enhancement
- Enhancement Example
- Summary

Model-based Array Processing and Enhancement

Summary and Conclusion

Classical optimal filtering

Estimate the clean speech by designing a filter $\boldsymbol{h} \in \mathbb{R}^M$ so that

$$\hat{s}(n) = \boldsymbol{h}^T \boldsymbol{x}(n) = \boldsymbol{h}^T \boldsymbol{s}(n) + \boldsymbol{h}^T \boldsymbol{e}(n) \,. \tag{123}$$

Two conflicting requirements:

Speech distortion $\boldsymbol{h}^T \boldsymbol{s}(n) \approx s(n)$

Noise suppression $\boldsymbol{h}^T \boldsymbol{e}(n) \approx 0$

Many different ways of designing the filter, e.g.,

- Wiener filter
- LCMV-filter
- Variable-span linear filter

They all assume that the second-order statistics is available.

Classical optimal filtering

Wiener filter

$$\boldsymbol{h}_{\mathrm{W}} = \boldsymbol{R}_x^{-1}\boldsymbol{R}_s\boldsymbol{i}_{1,M} = (\boldsymbol{I}_M - \boldsymbol{R}_x^{-1}\boldsymbol{R}_e)\boldsymbol{i}_{1,M} \tag{124}$$

where $\boldsymbol{i}_{1,M} = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}^T \in \mathbb{R}^M$ and

- $\boldsymbol{R}_x$ covariance matrix of the noisy speech
- $\boldsymbol{R}_s$ covariance matrix of the clean speech
- $\boldsymbol{R}_e$ covariance matrix of the noise.

The second-order statistics is usually estimated using

- a VAD to signal when the noise statistics can be updated,
- noise tracking (MS, IMCRA, or MMSE), or
- nonnegative matrix factorisation (NMF).

However, no method so far which consistently improves speech intelligibility and quality for nonstationary noise.

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

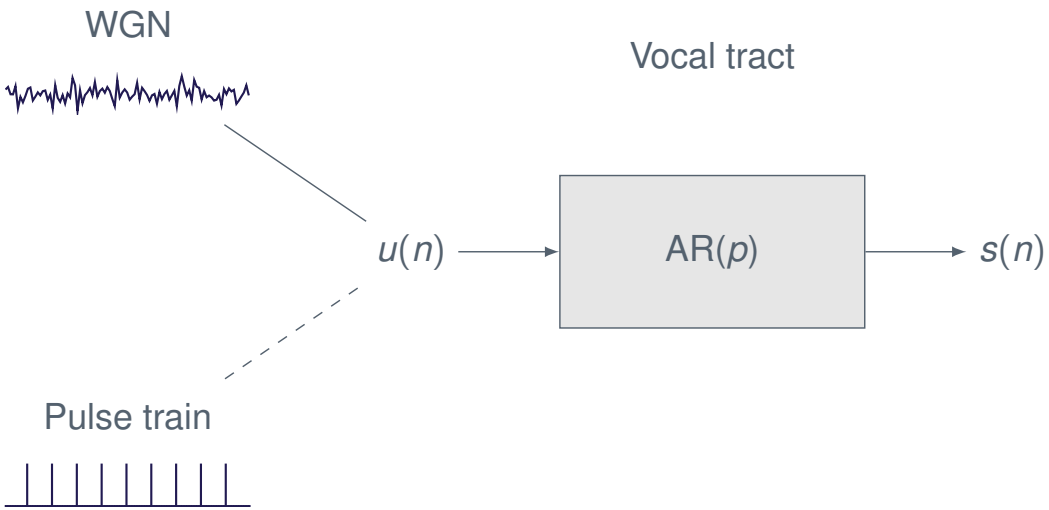
- Classical Optimal Filtering
- Model-based Speech Enhancement
- Enhancement Example
- Summary

Model-based Array Processing and Enhancement

Summary and Conclusion

Model-based Speech Enhancement

Speech model



We initially assume that the excitation signal is white Gaussian noise (WGN) so that

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{125}$$

where $\{a_i\}_{i=1}^{p}$ and $u(n)$ are the AR-parameters and the excitation noise, respectively.

Model-based Speech Enhancement

Speech model

For $n = 0, 1, \cdots, N-1$, a stationary AR-process can be written as

$$p(\boldsymbol{s}|\sigma_{\mathrm{s}}^2, \boldsymbol{a}) = \mathcal{N}(\boldsymbol{0}, \sigma_{\mathrm{s}}^2 \boldsymbol{Q}(\boldsymbol{a})) \tag{126}$$

where σ_{s}^2 and $\boldsymbol{Q}(\boldsymbol{a})$ are the excitation variance and normalised covariance matrix, respectively.

- Asymptotically (or if $\boldsymbol{s}$ is periodic in N), we have that (Gray, 2006)

$$\boldsymbol{Q}(\boldsymbol{a}) = N^{-1} \boldsymbol{F} \boldsymbol{D}(\boldsymbol{a}) \boldsymbol{F}^H \tag{127}$$

where $\boldsymbol{F} = \{\exp(j2\pi nk/N)\}$ is the DFT matrix and

$$\boldsymbol{D}(\boldsymbol{a}) = \left[\boldsymbol{\Lambda}^H(\boldsymbol{a})\boldsymbol{\Lambda}(\boldsymbol{a})\right]^{-1} \tag{128}$$

$$\boldsymbol{\Lambda}(\boldsymbol{a}) = \mathrm{diag}\left(\boldsymbol{F}^H \begin{bmatrix}\boldsymbol{a} \\ \boldsymbol{0}\end{bmatrix}\right) \tag{129}$$

Model-based Speech Enhancement

Signal model

- We also assume a stationary AR-model for the noise vector $\boldsymbol{e}$ with excitation variance σ_{e}^2 and AR parameters $\boldsymbol{b}$.

Thus, the signal model is

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{e} \tag{130}$$

where

$$p(\boldsymbol{s}|\sigma_{\mathrm{s}}^2, \boldsymbol{a}) = \mathcal{N}(\boldsymbol{0}, \sigma_{\mathrm{s}}^2 \boldsymbol{Q}(\boldsymbol{a})) \tag{131}$$

$$p(\boldsymbol{e}|\sigma_{\mathrm{e}}^2, \boldsymbol{b}) = \mathcal{N}(\boldsymbol{0}, \sigma_{\mathrm{e}}^2 \boldsymbol{Q}(\boldsymbol{b})) \tag{132}$$

so that the observation model is

$$p(\boldsymbol{x}|\boldsymbol{s}, \sigma_{\mathrm{e}}^2, \boldsymbol{b}) = \mathcal{N}(\boldsymbol{s}, \sigma_{\mathrm{e}}^2 \boldsymbol{Q}(\boldsymbol{b})) \ . \tag{133}$$

Model-based Speech Enhancement

Inference

The posterior is

$$p(\boldsymbol{s}|\boldsymbol{x}, \sigma_{\mathrm{s}}^2, \boldsymbol{a}, \sigma_{\mathrm{e}}^2, \boldsymbol{b}) \propto p(\boldsymbol{x}|\boldsymbol{s}, \sigma_{\mathrm{e}}^2, \boldsymbol{b}) p(\boldsymbol{s}|\sigma_{\mathrm{s}}^2, \boldsymbol{a}) = \mathcal{N}(\hat{\boldsymbol{s}}, \boldsymbol{\Sigma}) \tag{134}$$

where

$$\boldsymbol{\Sigma} = \sigma_{\mathrm{s}}^2 \boldsymbol{Q}(\boldsymbol{a}) \left[\sigma_{\mathrm{s}}^2 \boldsymbol{Q}(\boldsymbol{a}) + \sigma_{\mathrm{e}}^2 \boldsymbol{Q}(\boldsymbol{b})\right]^{-1} \sigma_{\mathrm{e}}^2 \boldsymbol{Q}(\boldsymbol{b}) = \boldsymbol{R}_{\mathrm{s}} \boldsymbol{R}_{\mathrm{x}}^{-1} \boldsymbol{R}_{\mathrm{e}} \tag{135}$$

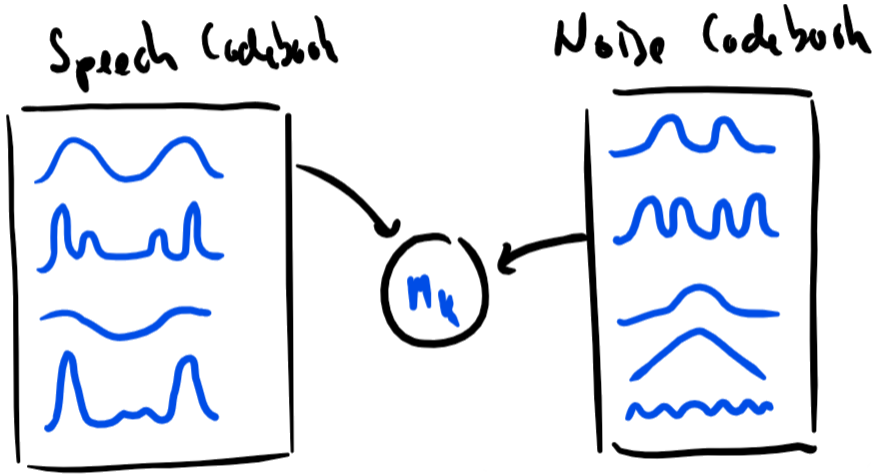
$$\hat{\boldsymbol{s}} = \sigma_{\mathrm{s}}^2 \boldsymbol{Q}(\boldsymbol{a}) \left[\sigma_{\mathrm{s}}^2 \boldsymbol{Q}(\boldsymbol{a}) + \sigma_{\mathrm{e}}^2 \boldsymbol{Q}(\boldsymbol{b})\right]^{-1} \boldsymbol{y} = \boldsymbol{R}_{\mathrm{s}} \boldsymbol{R}_{\mathrm{x}}^{-1} \boldsymbol{y} \tag{136}$$

$$= N^{-1} \boldsymbol{F} \left\{ \sigma_{\mathrm{s}}^2 \boldsymbol{D}(\boldsymbol{a}) \left[\sigma_{\mathrm{s}}^2 \boldsymbol{D}(\boldsymbol{a}) + \sigma_{\mathrm{e}}^2 \boldsymbol{D}(\boldsymbol{b})\right]^{-1} \boldsymbol{F}^H \boldsymbol{y} \right\} \tag{137}$$

- Time- and frequency-domain Wiener filtering in the same formulation.
- For an AR-order of $p = N - 1$, we get the traditional Wiener filter.
- Easier and more robust to estimate a few AR-parameters.

Model-based Speech Enhancement

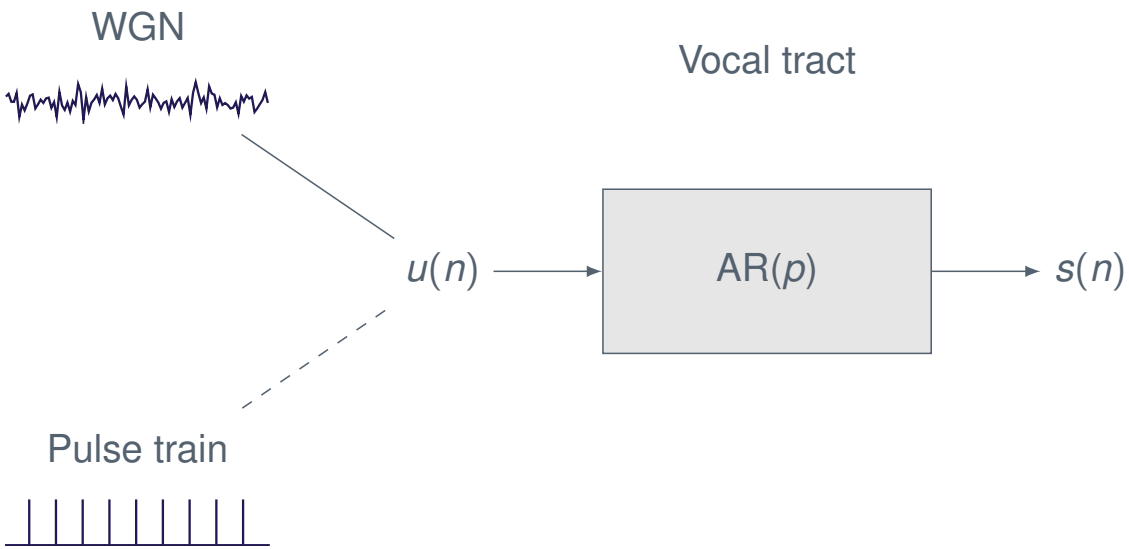
Estimating the model parameters



- Typical AR-parameters for speech and noise can be obtained from training and stored in codebooks (Srinivasan et al., 2006, 2007). Both specific and general codebooks can be trained.
- For every combination of speech and noise codebook vectors, the excitation variances and model probabilities are estimated.
- Model-averaged estimates of σ_{s}^2, σ_{e}^2, $\boldsymbol{a}$, and $\boldsymbol{b}$ are computed.

Model-based Speech Enhancement

Speech model



But . . .

- speech is non-stationary, and
- the WGN excitation is not a very good model for voiced speech.

Model-based Speech Enhancement

State-space model

An AR-process can be written as a state-transition equation

$$\boldsymbol{s}(n) = \boldsymbol{A}(\boldsymbol{a})\boldsymbol{s}(n-1) + \boldsymbol{i}_{1,M}\sqrt{\sigma_{\mathrm{s}}^2}u(n) \tag{138}$$

where

$$\boldsymbol{s}(n) = \begin{bmatrix} s(n) & \cdots & s(n-p+1) & \cdots & s(n-M+1) \end{bmatrix}^T \tag{139}$$

$$\boldsymbol{A}(\boldsymbol{a}) = \begin{bmatrix} \boldsymbol{a}^T & \boldsymbol{0}^T & 0 \\ \boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{M-p-1} & \boldsymbol{0} \end{bmatrix} \tag{140}$$

- $M \geq p$ is set to control the delay in a fixed-lag Kalman smoother.
- We use the estimated model parameters from the previous slides. These are updated regularly (e.g., every 25 ms).

Model-based Speech Enhancement

State-space model

If we also rewrite the AR-noise model using a state-transition equation (with $M = q$), we obtain the state space model

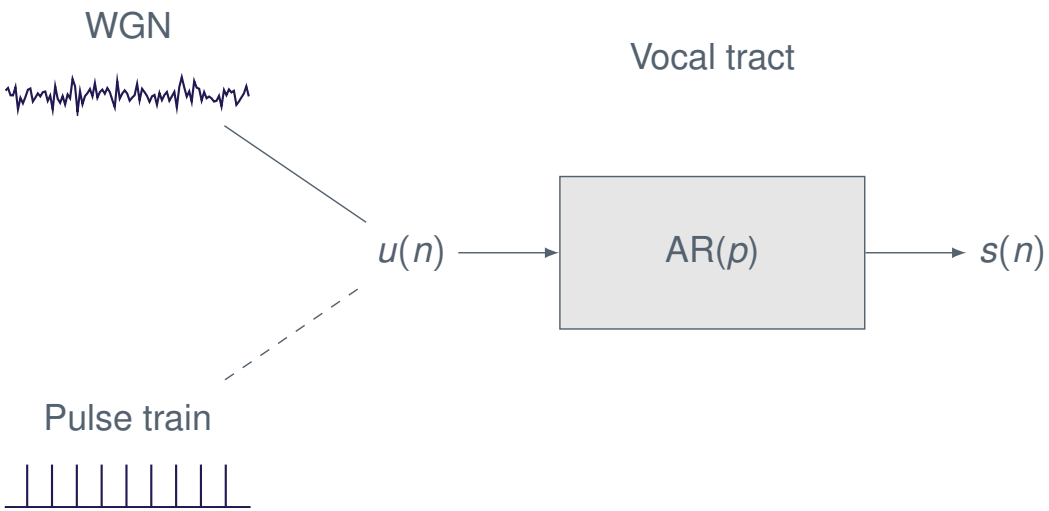
$$x(n) = \begin{bmatrix} \boldsymbol{i}_{1,M}^T & \boldsymbol{i}_{1,q}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{s}(n) \\ \boldsymbol{e}(n) \end{bmatrix}$$

$$\begin{bmatrix} \boldsymbol{s}(n) \\ \boldsymbol{e}(n) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}(\boldsymbol{a}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}(\boldsymbol{b}) \end{bmatrix} \begin{bmatrix} \boldsymbol{s}(n-1) \\ \boldsymbol{e}(n-1) \end{bmatrix} + \begin{bmatrix} \sqrt{\sigma_s^2}\boldsymbol{i}_{1,M} & \boldsymbol{0} \\ \boldsymbol{0} & \sqrt{\sigma_e^2}\boldsymbol{i}_{1,q} \end{bmatrix} \begin{bmatrix} u(n) \\ \epsilon(n) \end{bmatrix}$$

- Can be implemented as a traditional Kalman filter.
- The output of the fixed-lag Kalman smoother is the M'th sample of the estimated state vector.

Model-based Speech Enhancement

Speech model with pulse train excitation



$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{141}$$

$$u(n) = vu(n-\tau) + w(n) \tag{142}$$

where $v \in [0, 1]$ and $\tau > 0$ are the degree of voicing and pitch period, respectively.

Speech model with pulse train excitation

Speech model

We can rewrite the speech model as the state-transition equation

$$\begin{bmatrix} \boldsymbol{s}(n) \\ \boldsymbol{u}(n) \end{bmatrix} = \boldsymbol{F}(\boldsymbol{a}, v, \tau) \begin{bmatrix} \boldsymbol{s}(n-1) \\ \boldsymbol{u}(n-1) \end{bmatrix} + \begin{bmatrix} \boldsymbol{i}_{1,M} \\ \boldsymbol{i}_{\tau,\tau_{\text{MAX}}} \end{bmatrix} \sqrt{\sigma_s^2} w(n) \tag{143}$$

where

$$\boldsymbol{u}(n) = \begin{bmatrix} u(n) & u(n-1) & \cdots & u(n-\tau_{\text{MAX}}) \end{bmatrix}^T \tag{144}$$

$$\boldsymbol{F}(\boldsymbol{a}, v, \tau) = \begin{bmatrix} \boldsymbol{A}(\boldsymbol{a}) & v\boldsymbol{i}_{1,M}\boldsymbol{i}_{\tau,\tau_{\text{MAX}}}^T \\ \boldsymbol{0} & \begin{bmatrix} v\boldsymbol{i}_{\tau,\tau_{\text{MAX}}}^T \\ \boldsymbol{I}_{\tau_{\text{MAX}}-1} \quad \boldsymbol{0} \end{bmatrix} \end{bmatrix} \tag{145}$$

Model-based Speech Enhancement

Signal model

In total, we get the state-space model

$$x(n) = \begin{bmatrix} \boldsymbol{i}_{1,M}^T & \boldsymbol{0}^T & \boldsymbol{i}_{1,q}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{s}(n) \\ \boldsymbol{u}(n) \\ \boldsymbol{e}(n) \end{bmatrix}$$

$$\begin{bmatrix} \begin{bmatrix} \boldsymbol{s}(n) \\ \boldsymbol{u}(n) \end{bmatrix} \\ \boldsymbol{e}(n) \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}(\boldsymbol{a}, v, \tau) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}(\boldsymbol{b}) \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \boldsymbol{s}(n-1) \\ \boldsymbol{u}(n-1) \end{bmatrix} \\ \boldsymbol{e}(n-1) \end{bmatrix} + \begin{bmatrix} \begin{bmatrix} \boldsymbol{i}_{1,M} \\ \boldsymbol{i}_{\tau,\tau_{\text{MAX}}} \end{bmatrix} \sqrt{\sigma_s^2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{i}_{1,q}\sqrt{\sigma_e^2} \end{bmatrix} \begin{bmatrix} w(n) \\ \epsilon(n) \end{bmatrix}$$

This can again be implemented using a fixed-lag Kalman smoother.

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

- Classical Optimal Filtering
- Model-based Speech Enhancement
- Enhancement Example
- Summary

Model-based Array Processing and Enhancement

Summary and Conclusion

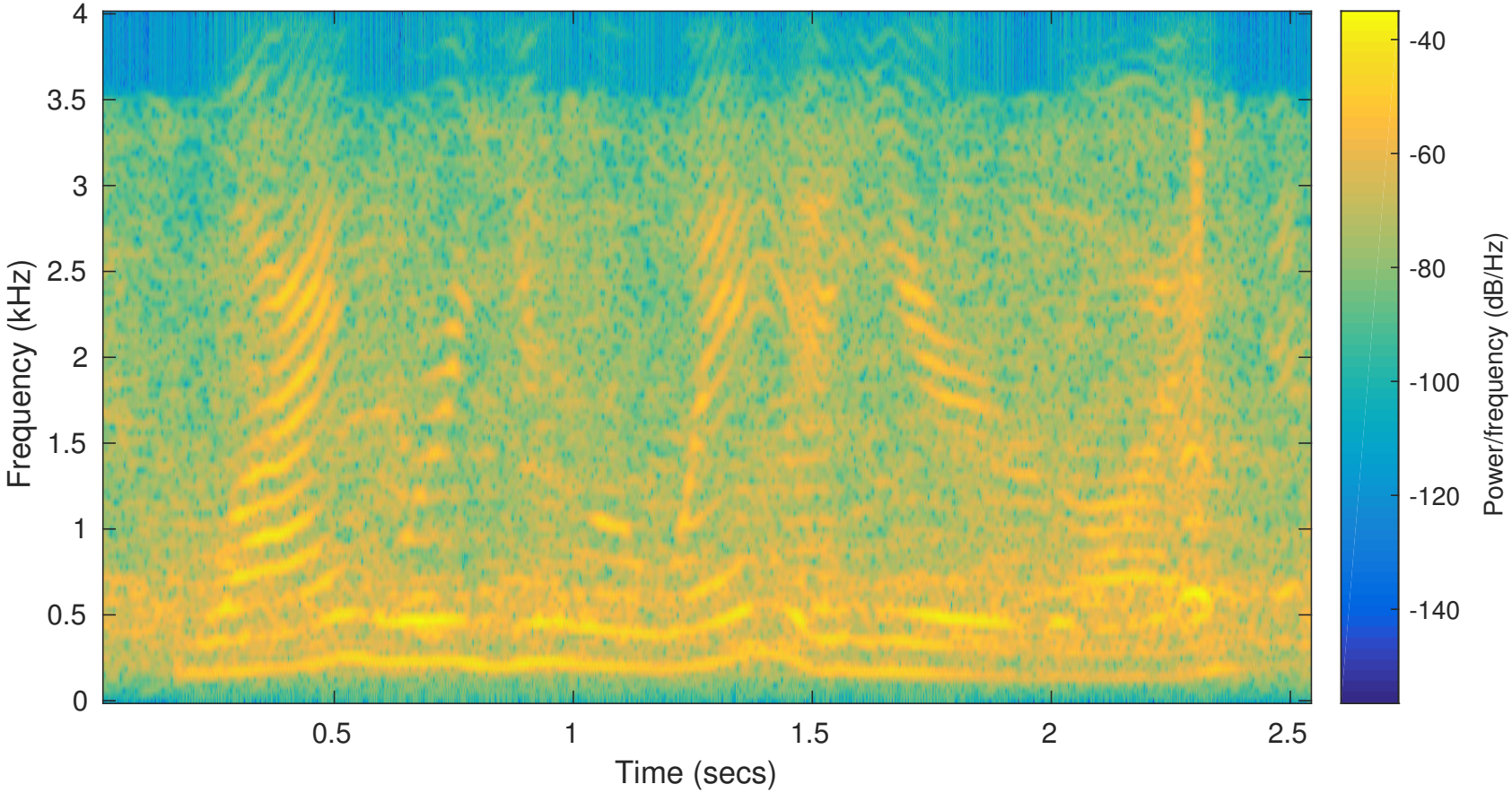
Enhancement Example

Simulation setup

- Model parameters re-estimated every 25 ms.
- Pitch estimated in the range [80,400] Hz.
- AR-order of 14 for both the speech and noise spectra.
- Speaker specific codebook of 64 entries.
- Babble noise codebook of 8 entries.
- Codebooks trained using standard vector quantisation technique from speech coding.

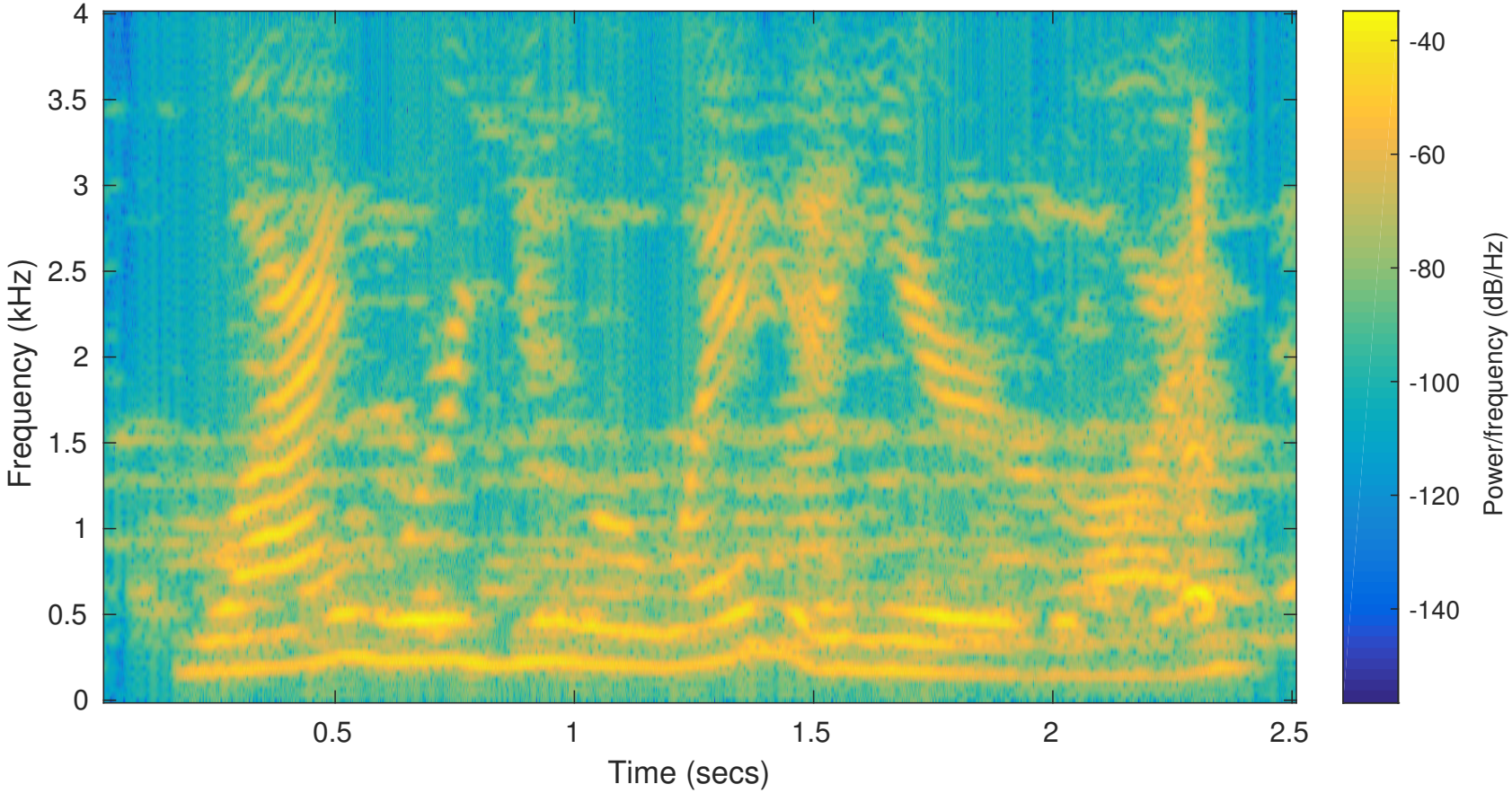
Enhancement Example

Noisy speech (SNR of 5 dB)



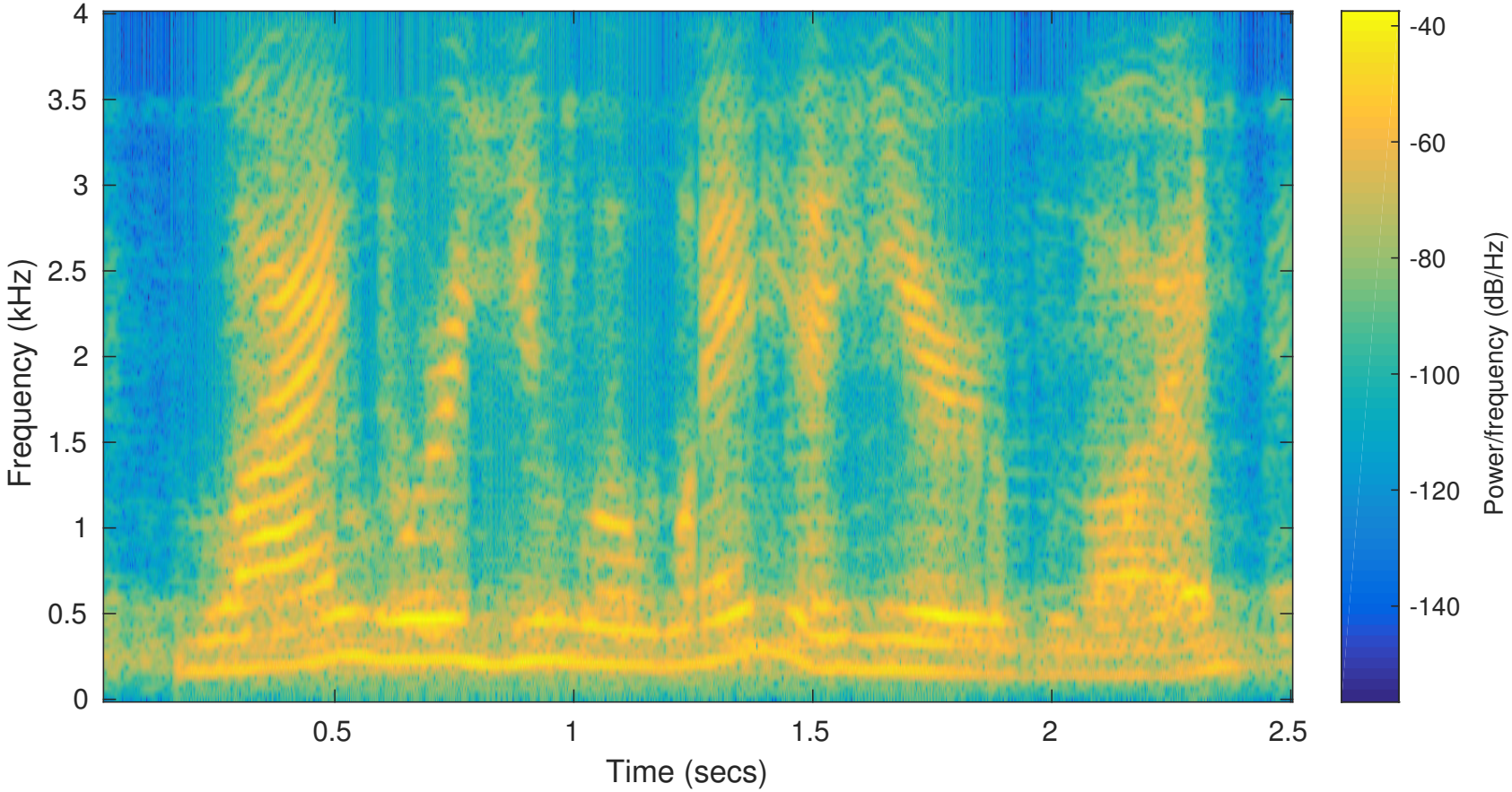
Enhancement Example

Enhanced speech (Wiener filter with IMCRA)



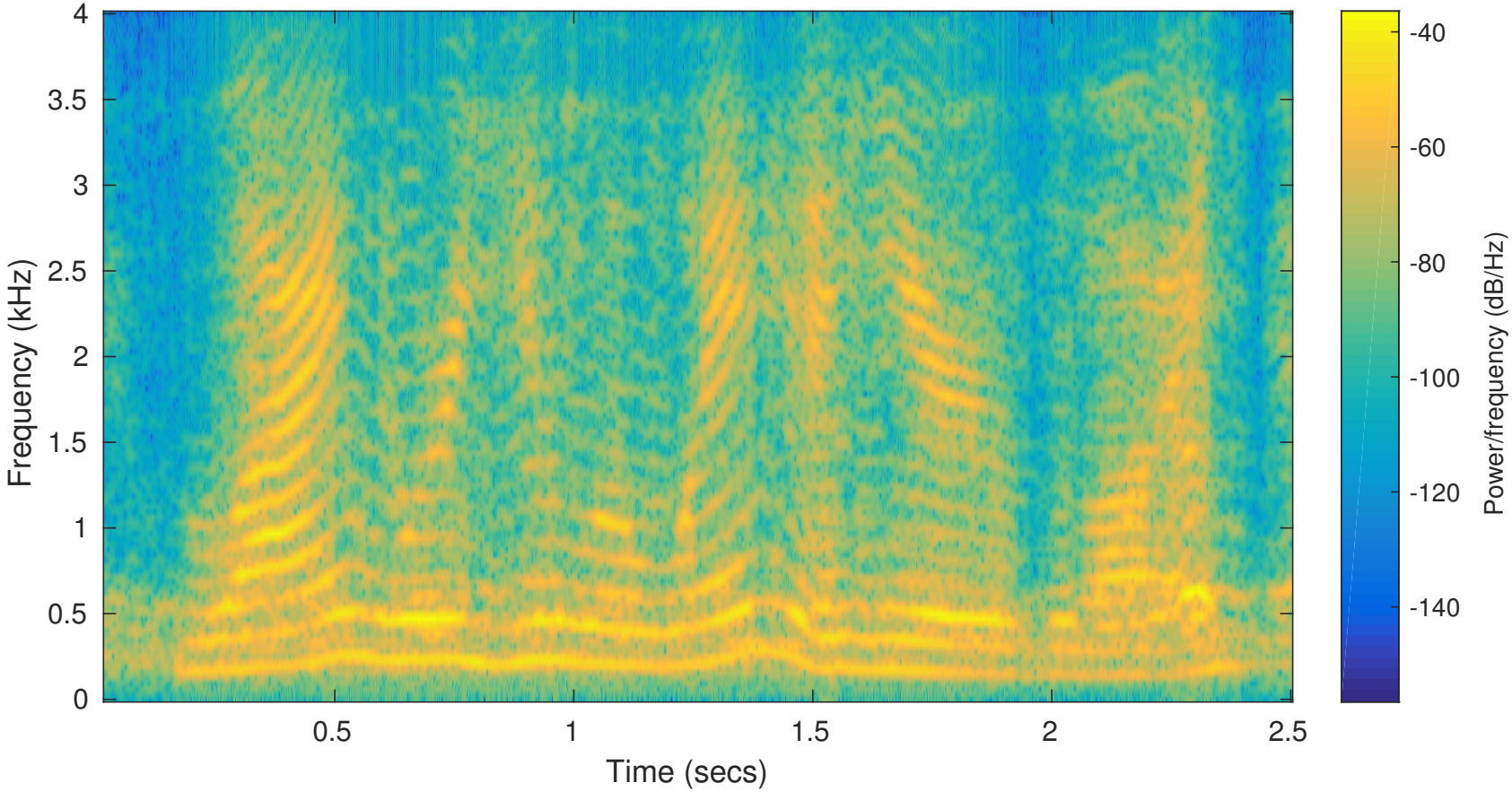
Enhancement Example

Enhanced speech (Kalman filter without voiced model)



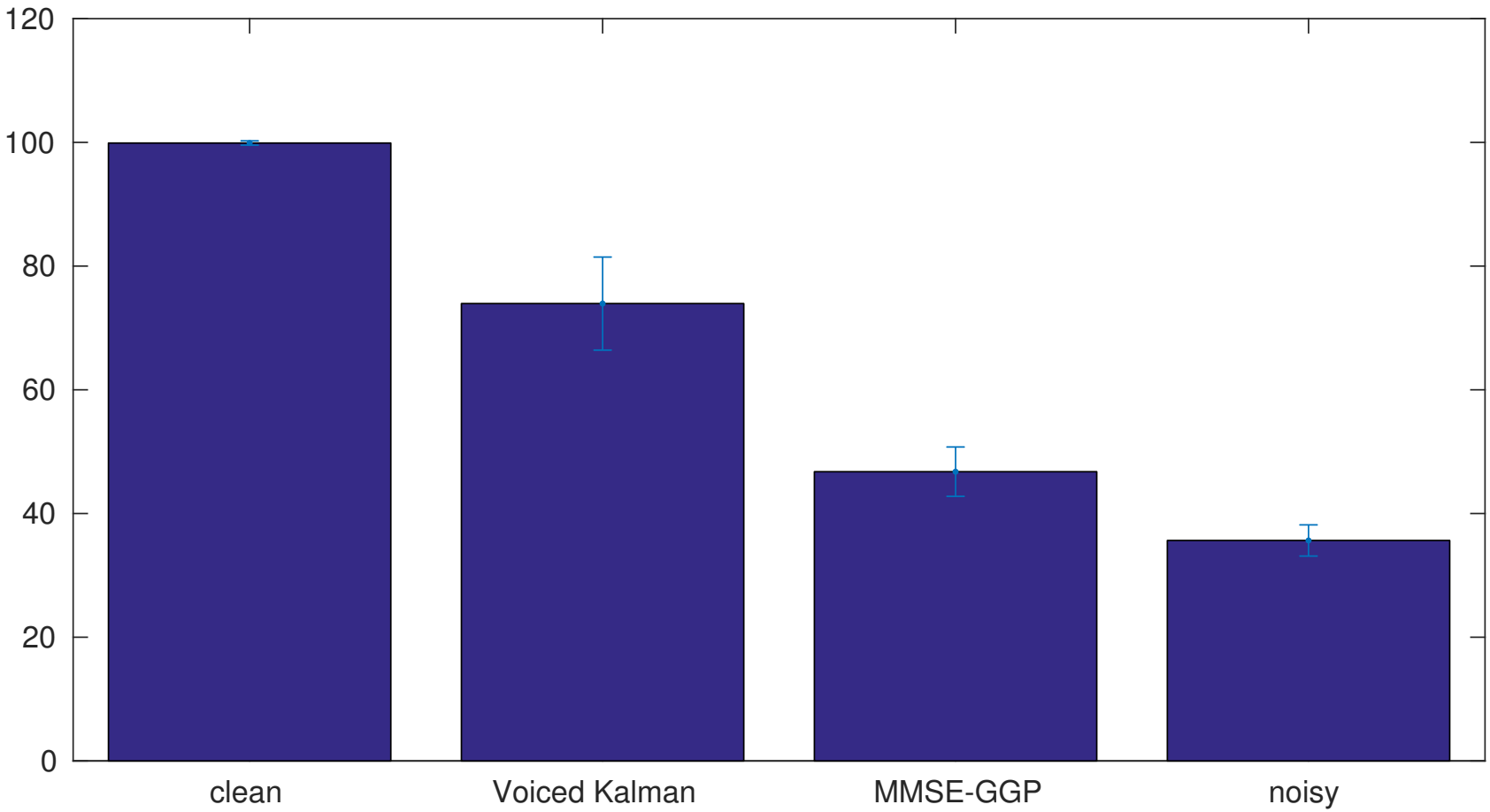
Enhancement Example

Enhanced speech (Kalman filter with voiced model)



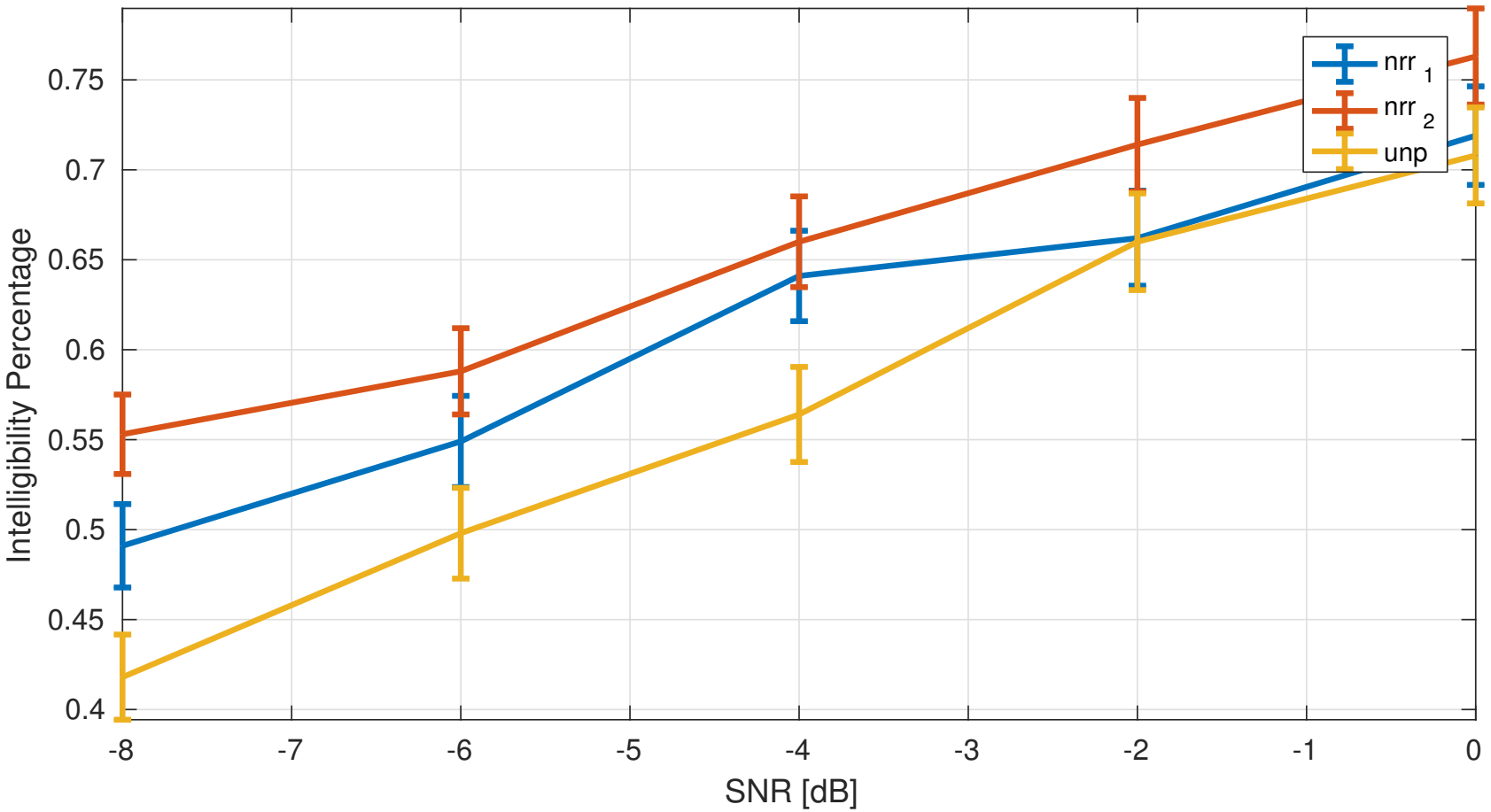
Enhancement Example

Speech Quality (MUSHRA test @ SNR of 10 dB)



Enhancement Example

Speech Intelligibility



Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

- Classical Optimal Filtering
- Model-based Speech Enhancement
- Enhancement Example
- Summary

Model-based Array Processing and Enhancement

Summary and Conclusion

Summary

- The traditional Wiener filter can be viewed a special case of a more general, but very simple model-based approach.
- This model-based approach allows us to incorporate prior spectral information in the form of AR-parameters.
- Including specific prior information about the speaker and/or noise can give a very good performance . . .
- . . . but also result in a poor performance if the speaker and noise are not what was assumed.
- We can easily modify the model so that we better can handle non-stationary speech (Kalman filter) and voiced speech (pulse train model).
- The single-channel approach presented here can improve both speech quality and intelligibility in non-stationary noise such as babble noise.
- The model can also easily be extended to the binaural/multichannel case (Kavalekalam, 2017).

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement
- TDOA and DOA based Models
- Model-based Estimation of Fractional TDOAs
- Joint Estimation of DOA and Pitch
- Model Extensions for Robust Estimation
- Distortionless Enhancement of Audio
- Summary

Summary and Conclusion

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

- TDOA and DOA based Models
  - Model-based Estimation of Fractional TDOAs
  - Joint Estimation of DOA and Pitch
  - Model Extensions for Robust Estimation
  - Distortionless Enhancement of Audio
  - Summary

Summary and Conclusion

Model-Based Array Processing

- Taking the signal model into account can be of great benefit in microphone array processing methods.
- This includes TDOA estimators, DOA estimators, localization methods, beamforming/enhancement methods, etc.
- Examples of benefits offered by such parametric methods includes:
  - Estimation of fractional TDOAs does not require any additional heuristics.
  - Facilitates robust TDOA/DOA estimation in scenarios with multiple sources, even with overlap in frequency or space.
  - Facilitates joint DOA and pitch estimation, which can yield more accurate estimates compared to separate estimators and non-parametric methods.
  - Enables distortionless enhancement of periodic signals, e.g., audio and speech (voiced).

Array Model
Two channels

Let us start by considering a simple array with 2 microphones.

In the time-domain, the observations are modeled as:

$$\begin{aligned} x_1(n) &= s(n) + e_1(n), \\ x_2(n) &= \beta_2 s(n - \eta_2) + e_2(n), \quad n = 0, \ldots, N-1. \end{aligned} \tag{146}$$

where

- $s(n)$: desired deterministic signal,
- $e_{\{1,2\}}(n)$: background noise observed on microphone {1,2},
- β_2: level difference between ref. (mic 1) and mic 2 (assumed frequency independent),
- η_2: TDOA between ref. and mic 2.

Periodic Model

Assuming periodic signal:

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_0 l n + \phi_l) = \sum_{l=-L}^{L} \alpha_l e^{j\omega_0 l n}, \tag{147}$$

with

A_l/α_l: real/complex amplitude ($A_l > 0$, $\alpha_l = \frac{A_l}{2} e^{j\phi_l}$),

ϕ_l: phase ($\phi_l \in [-\pi, \pi[$),

ω_0: fundamental frequency.

L: model order.

Complex Model

Also possible to use a complex model by using the Hilbert transform. Can ease computational complexity and mathematical notation.

Clean desired signal modeled as

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n}, \tag{148}$$

Important observation:

Widely used broadband model is a special case of (147), i.e., for

$$\omega_0 = 2\pi/N \qquad \wedge \qquad L = \lfloor N/2 \rfloor. \tag{149}$$

Array Model
Extension to multiple channels

Inserting the periodic signal in the observation models yields

$$x_1(n) = \sum_{l=-L}^{L} \alpha_l e^{j\omega_0 l n} + e_1(n), \tag{150}$$

$$x_2(n) = \beta_2 \sum_{l=-L}^{L} \alpha_l e^{j\omega_0 l n} e^{-j\omega_0 \eta_2 l} + e_2(n). \tag{151}$$

With more microphones, the model is easily extended as

$$x_k(n) = \beta_k \sum_{l=-L}^{L} \alpha_l e^{j\omega_0 l n} e^{-j\omega_0 \eta_k l} + e_k(n). \tag{152}$$

Modelling Array Structure

Example: uniform linear array

With a high number of microphones, K, the array structure can be exploited to reduce dimensionality.

Instead of having an unknown TDOA for each microphone pair, we only have an unknown range, r_c, and DOA, θ!

Assuming array center to be ref, the source-to-mic-k range is:

$$r_k(r_c, \theta) = \sqrt{g_k^2 d^2 + r_c^2 - 2 g_k d r_c \sin\theta} \tag{153}$$

with

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

Modelling Array Structure

Example: uniform linear array

We can use this to further specify our model,

$$x_k(n) = \frac{r_c}{r_k} \sum_{l=1}^{L} \gamma_l e^{jl\omega_0 n} e^{-jf_s l\omega_0 \frac{r_k - r_c}{c}} + e_k(n). \tag{154}$$

where γ_l's are harmonic amplitudes in the reference point.

In the far-field, we have that

$$\frac{r_c}{r_k} \approx 1, \quad \text{and} \quad \tau_{c,k} = \frac{r_k - r_c}{c} \approx g_k \frac{d \sin\theta}{c}, \tag{155}$$

which results in the simplified model

$$x_k(n) = \sum_{l=1}^{L} \gamma_l e^{jl\omega_0 n} e^{-jf_s l\omega_0 g_k \frac{d \sin\theta}{c}} + e_k(n). \tag{156}$$

Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

- TDOA and DOA based Models
- Model-based Estimation of Fractional TDOAs
- Joint Estimation of DOA and Pitch
- Model Extensions for Robust Estimation
- Distortionless Enhancement of Audio
- Summary

Summary and Conclusion

TDOA Estimation for Audio Applications

- The models can be used to derive TDOA estimators for microphone pairs (Jensen 2015).
- TDOA estimation important in many microphone array applications:
  - array calibration
  - room geometry estimation
  - noise reduction
  - source localization, etc.
- Typically, solved using cross-correlation based methods in the frequency domain.
- Can be shown to be special case of more general and accurate method based on the models.

Matrix-vector model

Returning to the dual mic scenario, the model can be written in matrix-vector notation:

$$\mathbf{x} = \mathbf{H}(\beta_2, \eta_2, \omega_0)\boldsymbol{\alpha} + \mathbf{e} \tag{157}$$

with

$$\begin{aligned}
\mathbf{H} &= \begin{bmatrix} \mathbf{Z}(\omega_0) \\ \beta_2 \mathbf{Z}(\omega_0)\mathbf{D}(\omega_0, \eta_2) \end{bmatrix}, \\
\mathbf{z}(\omega) &= \begin{bmatrix} 1 & e^{j\omega} & \cdots & e^{j\omega(N-1)} \end{bmatrix}^T, \\
\mathbf{Z}(\omega_0) &= [\mathbf{z}(-L\omega_0) \ \cdots \ \mathbf{z}(-\omega_0) \ \mathbf{z}(\omega_0) \ \mathbf{z}(L\omega_0)], \\
\mathbf{D}(\omega_0, \eta_2) &= \text{diag}\left(e^{jL\omega_0\eta_2}, \ldots, e^{j\omega_0\eta_2}, e^{-j\omega_0\eta_2}, \ldots, e^{-jL\omega_0\eta_2}\right), \\
\boldsymbol{\alpha} &= [\alpha_{-L} \ \cdots \ \alpha_{-1} \ \alpha_1 \ \cdots \ \alpha_L]^T, \\
\mathbf{e}&\text{: white Gaussian with pdf } \mathcal{N}(\mathbf{H}(\beta, \xi, \omega_0)\boldsymbol{\alpha}, \sigma^2 \mathbf{I}_{2N}).
\end{aligned}$$

Maximum Likelihood Estimator

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

$$(\widehat{\beta}, \widehat{\xi}, \widehat{\omega}_0) = \arg\max_{\beta,\xi,\omega_0} J(\beta, \xi, \omega_0), \tag{158}$$

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.

Important Special Case

For $\omega_0 = 2\pi/N$, $L = \lfloor N/2 \rfloor$, and $N \to \infty$, the cost function becomes:

$$J(\eta_2) = \mathbf{x}_1^H \mathbf{Z}(\omega_0)\mathbf{D}^*(\omega_0, \eta_2)\mathbf{Z}^H(\omega_0)\mathbf{x}_2 \tag{159}$$

$$= \sum_{k=-\lceil N/2 \rceil + 1}^{\lceil N/2 \rceil - 1} X_1^*(k) X_2(k) e^{jk\omega_0\eta_2}. \tag{160}$$

For N being even and η_2 being an integer:

$$J(\eta_2) = \sum_{k=0}^{N-1} X_1^*(k) X_2(k) e^{jk\omega_0\eta_2}. \tag{161}$$

This is 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 $\lfloor N/2 \rfloor$.
4. Delay is integer valued.

Fractional TDOA Estimation

Assume no noise and η_0 being true delay, then:

$$X_2(k) = X_1(k)e^{-jk\omega_0\eta_0}, \quad k = 0, \ldots, N-1. \tag{162}$$

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

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

Problem avoided by using:

$$J(\eta) = \sum_{k=-\lceil N/2\rceil+1}^{\lceil N/2\rceil-1} X_1^*(k)X_2(k)e^{jk\omega_0\eta_2}, \tag{163}$$

which is real-valued also for fractional delays.

Experiments
Setup

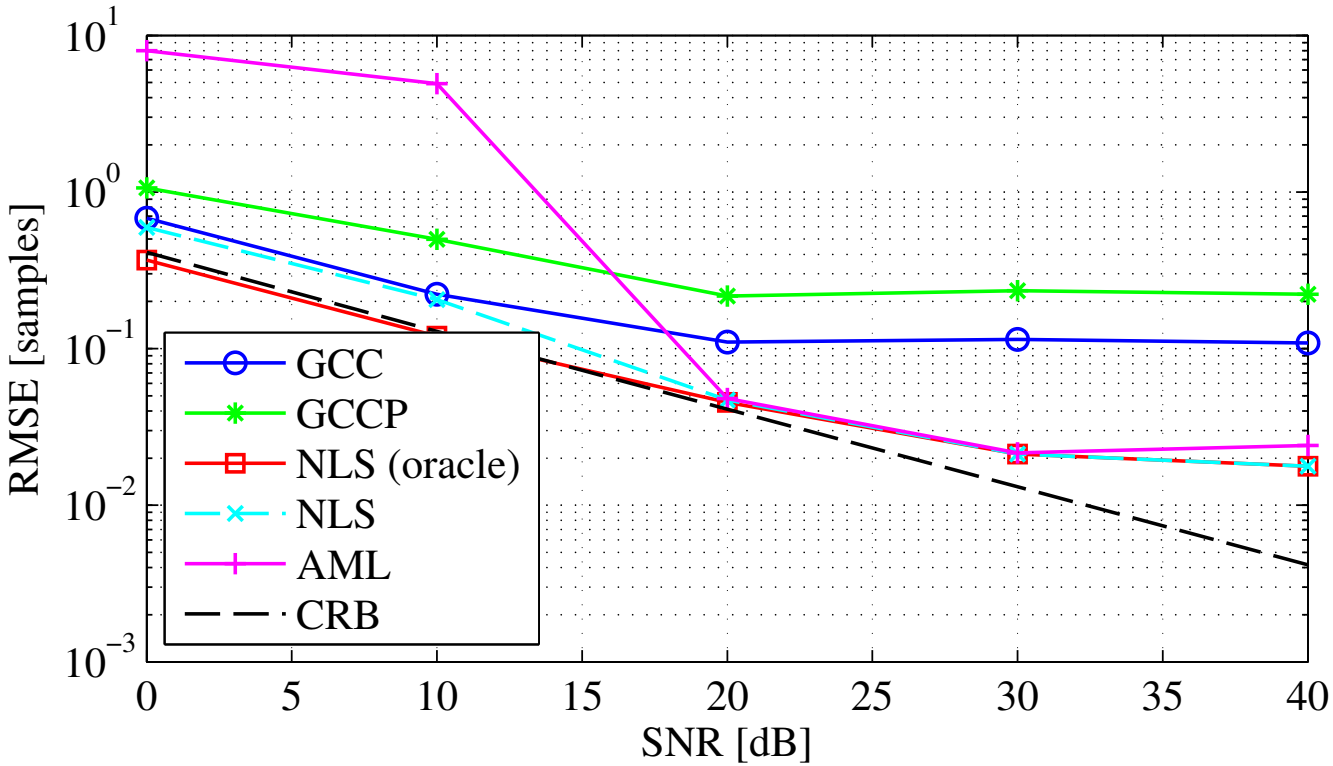
Synthetic data experiments:

- 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.

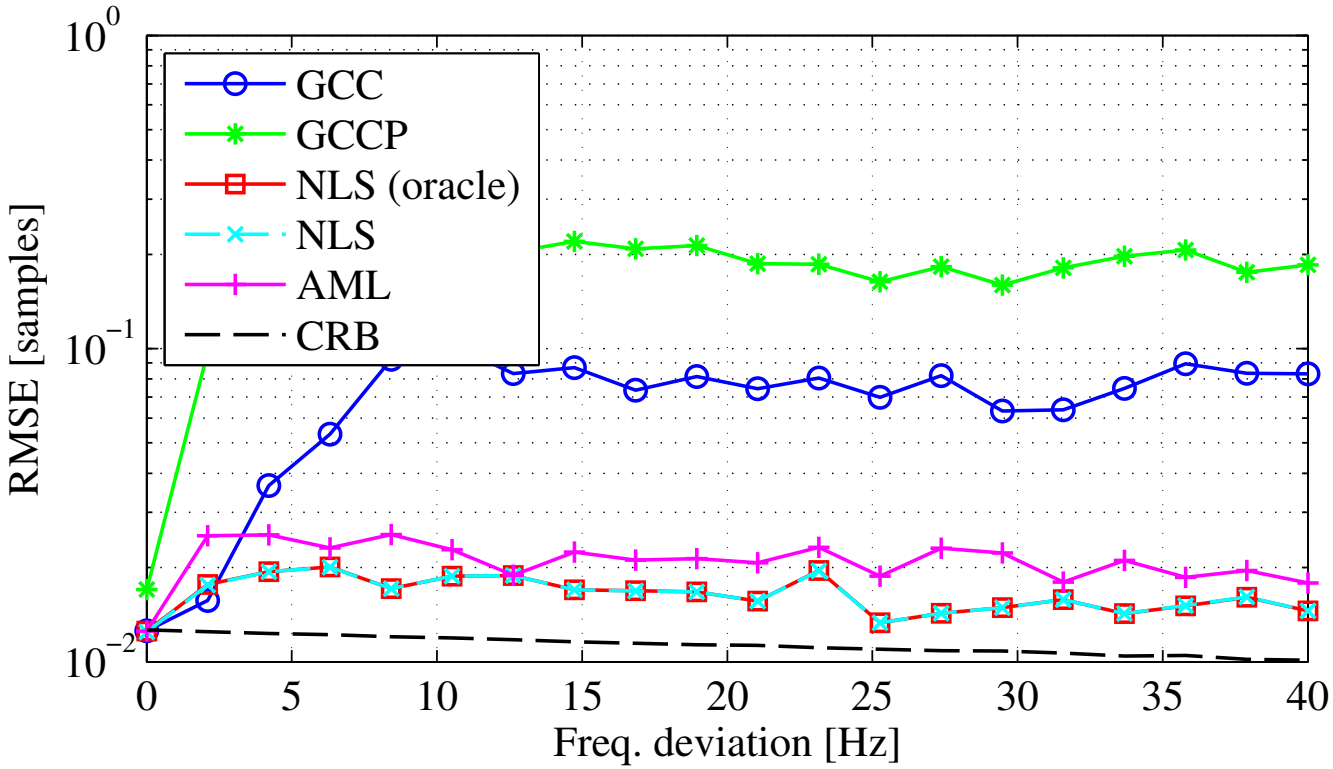
Speech data experiments

- $\sim$2.2 s of female speech (mainly voiced),
- stereo recording made using RIR generator,
- $\eta = 0.75$ samples, no reverb.,
- $N = 100$, $f_s = 8$ kHz.

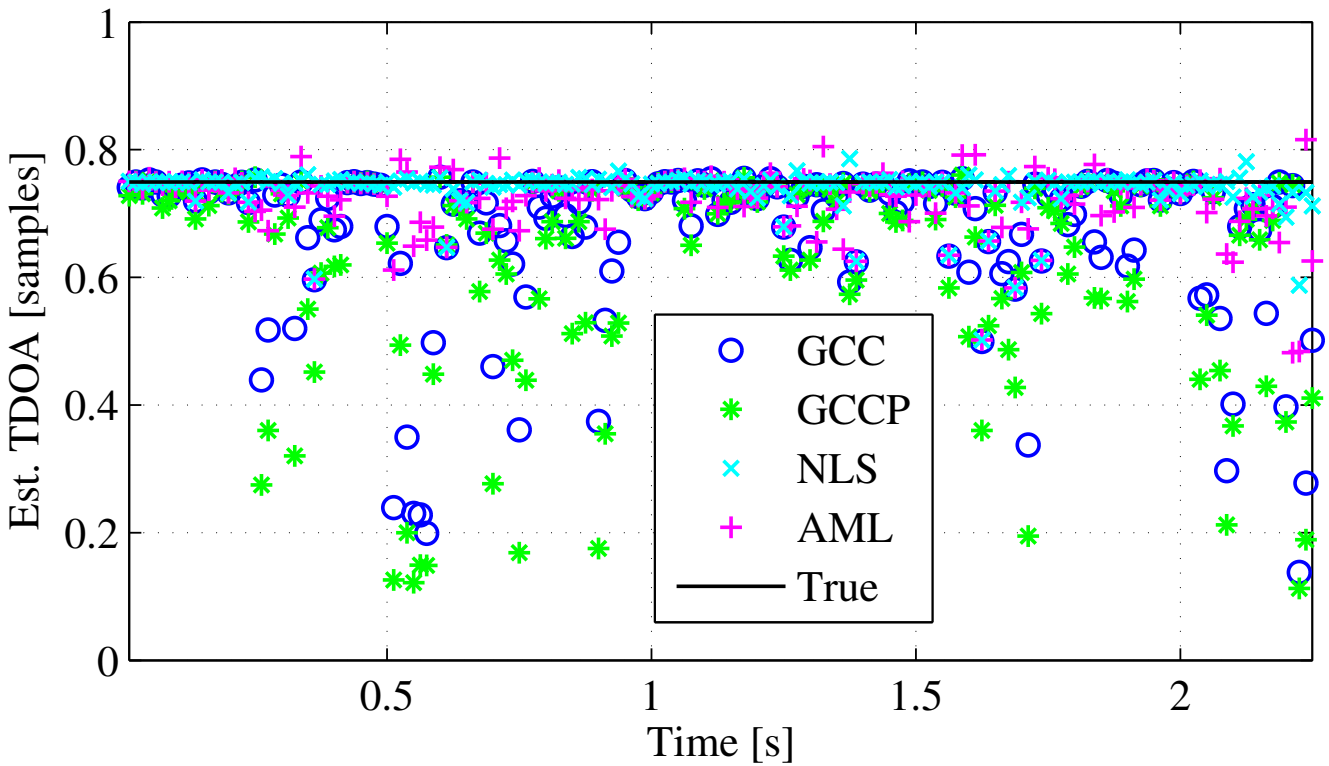
Experiments
Results on synthetic data



Experiments
Results on synthetic data



Experiments
Results on speech data



Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

- TDOA and DOA based Models
- Model-based Estimation of Fractional TDOAs
- Joint Estimation of DOA and Pitch
- Model Extensions for Robust Estimation
- Distortionless Enhancement of Audio
- Summary

Summary and Conclusion

DOA/Pitch Estimation for Multichannel Audio

- With $K \geq 2$ microphones, the models can be utilized for DOA estimation.
- Has applications in beamforming, autonomous steering, surveillance, etc.
- In audio applications, traditional DOA estimators are based on generic broadband model.
- Examples are: steered response power, TDOA-based, and subspace-based.
- More accurate estimates obtainable by exploiting the signal model.
- Periodic signal model can be used for, e.g., musical instruments and voiced speech.

Why Model-based DOA Estimation?

- For periodic signals, 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.

Matrix-Vector Model

Consider vector of N samples from mic k:

$$\begin{aligned}\mathbf{x}_k &= \begin{bmatrix} x_k(0) & x_k(1) & \cdots & x_k(N-1) \end{bmatrix}^T, \\ &= \mathbf{Z}(\omega_0)\mathbf{D}_k(\omega_0, \theta, r_c)\boldsymbol{\gamma} + \mathbf{e}_k, \end{aligned} \tag{164}$$

where

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

Likelihood Function

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

Assuming WGN which is not correlated across mics:

$$\mathcal{L} = \ln p(\{\mathbf{x}_k\}; \boldsymbol{\nu}) = -N\left(K \ln \pi + \sum_{k=0}^{K-1} \ln \sigma_k^2\right) - \sum_{k=0}^{K-1} \frac{\|\mathbf{e}_k\|^2}{\sigma_k^2}, \quad (165)$$

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:

$$\mathrm{CRB}(\omega_0) \approx \frac{6}{N^3 K}\mathrm{PSNR}^{-1}, \tag{166}$$

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

$$\mathrm{PSNR} = \frac{\sum_{l=1}^{L} l^2 A_l^2}{\sigma^2}. \tag{167}$$

Observations

- ω_0 CRB decreases with both N and K but independent of 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 $\boldsymbol{\gamma}$ and σ_k^2 's minimizing $\mathcal{L}$ can be found:

$$\widehat{\boldsymbol{\gamma}} = \left(\sum_{k=0}^{K-1} \frac{\mathbf{D}_k^H \mathbf{Z}^H \mathbf{Z} \mathbf{D}_k}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{\mathbf{D}_k^H \mathbf{Z}^H \mathbf{x}_k}{\sigma_k^2}, \tag{168}$$

$$\widehat{\sigma}_k^2 = \frac{\|\mathbf{x}_k - \mathbf{Z}\mathbf{D}_k \boldsymbol{\gamma}\|^2}{N}. \tag{169}$$

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

Resulting estimator after inserting solutions:

$$\{\widehat{\omega_0}, \widehat{r}_c, \widehat{\theta}\} = \arg\min \sum_{k=0}^{K-1} \ln \|\mathbf{x}_k - \mathbf{Z}\mathbf{D}_k \widehat{\boldsymbol{\gamma}}\|^2. \tag{170}$$

Approximate ML Estimator

For large sample sizes, it holds that

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{Z}^H\mathbf{Z} = \mathbf{I}. \tag{171}$$

With this approximation:

$$\widehat{\gamma} = \left(\sum_{k=0}^{K-1} \frac{r_{\mathrm{c}}^2}{r_k^2}\frac{N}{\sigma_k^2}\right)^{-1} \sum_{k=0}^{K-1} \frac{\mathbf{D}_k\mathbf{Z}^H\mathbf{x}_k}{\sigma_k^2}. \tag{172}$$

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:

$$\widehat{\gamma} = \left(\sum_{k=0}^{K-1} \mathbf{D}_k^H \mathbf{Z}^H \mathbf{Z} \mathbf{D}_k \right)^{-1} \sum_{k=0}^{K-1} \mathbf{D}_k^H \mathbf{Z}^H \mathbf{x}_k. \tag{173}$$

With the large sample approximation, it reduces to

$$\widehat{\gamma} = \left(\sum_{k=0}^{K-1} \frac{r_c^2}{r_k^2} N \right)^{-1} \sum_{k=0}^{K-1} \mathbf{D}_k \mathbf{Z}^H \mathbf{x}_k. \tag{174}$$

ML Estimator

Special case: far-field scenarios

In the far-field, the following approximations are made

$$\frac{r_{\mathrm{c}}}{r_k} \approx 1, \quad \text{and} \quad \tau_{\mathrm{c},k} \approx g_k \frac{d \sin\theta}{c}. \tag{175}$$

Amplitude and noise estimates are then:

$$\widehat{\boldsymbol{\gamma}} = \left(\sum_{k=0}^{K-1} \frac{\widetilde{\mathbf{D}}_k^H \mathbf{Z}^H \mathbf{Z} \widetilde{\mathbf{D}}_k}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{\widetilde{\mathbf{D}}_k^H \mathbf{Z}^H \mathbf{x}_k}{\sigma_k^2}, \tag{176}$$

$$\widehat{\sigma}_k^2 = \frac{\|\mathbf{x}_k - \mathbf{Z}\widetilde{\mathbf{D}}_k \boldsymbol{\gamma}\|}{N}, \tag{177}$$

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 the equal noise variance assumption:

$$\widehat{\gamma} = \left(\sum_{k=0}^{K-1} \widetilde{\mathbf{D}}_k^H \mathbf{Z}^H \mathbf{Z} \widetilde{\mathbf{D}}_k \right)^{-1} \sum_{k=0}^{K-1} \widetilde{\mathbf{D}}_k^H \mathbf{Z}^H \mathbf{x}_k. \tag{178}$$

the large sample approximation:

$$\widehat{\gamma} = \left(\sum_{k=0}^{K-1} \frac{N}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{\widetilde{\mathbf{D}}_k \mathbf{Z}^H \mathbf{x}_k}{\sigma_k^2}. \tag{179}$$

or both:

$$\widehat{\gamma} = \frac{1}{NK} \sum_{k=0}^{K-1} \mathbf{D}_k \mathbf{Z}^H \mathbf{x}_k. \tag{180}$$

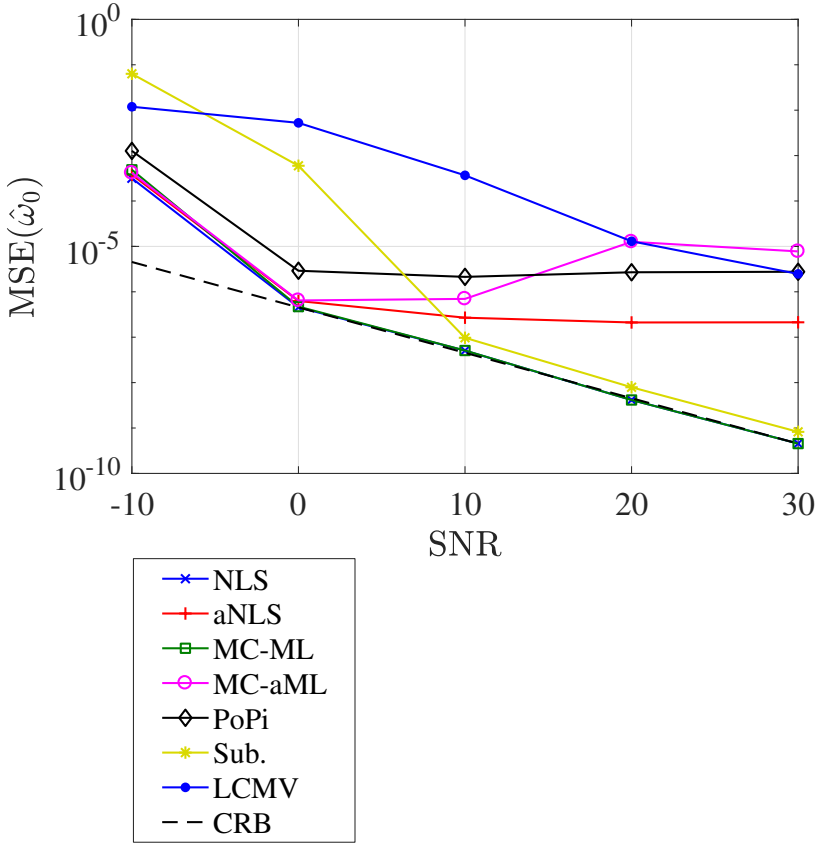
Other Model-Based Estimators

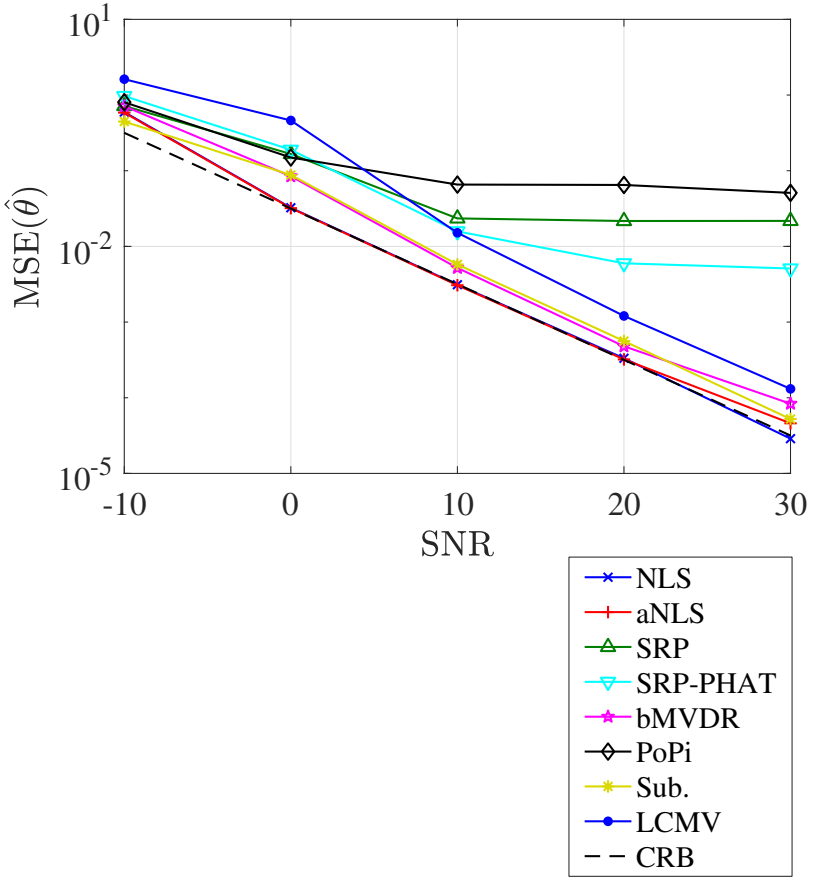
The models have been used as foundation in many more methods for localization/DOA estimation/TDOA estimation, including

- Closed-form DOA and pitch estimator based on weighted least squares (Jensen 2013, Karimian-Azari 2014).
- Optimal filtering-based DOA and pitch estimators (Jensen 2010, Zhou 2013, Karimian-Azari 2013).
- A subspace method for sequentiel pitch and DOA estimation (Wu 2015).
- Maximum likelihood method for source localization with ad-hoc microphone arrays (Hansen 2015).

Experimental Results

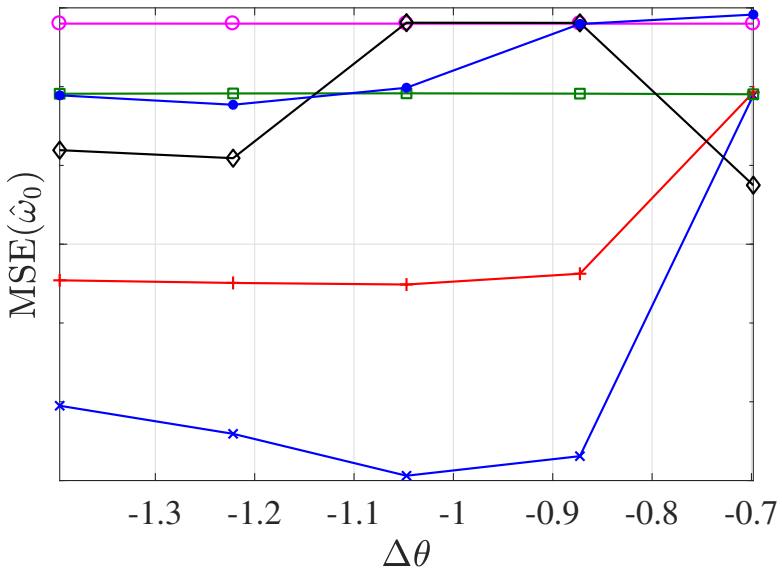
Synthetic source in far-field

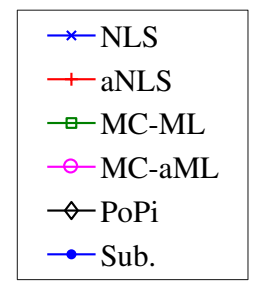


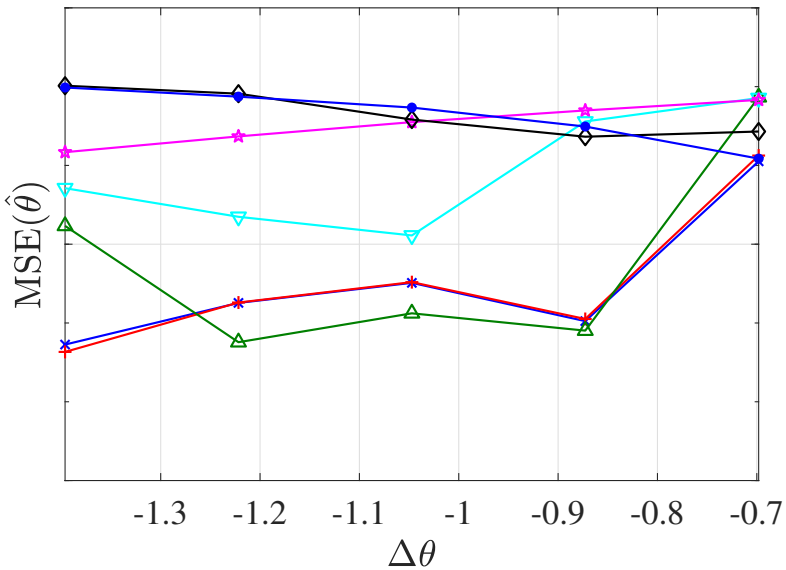


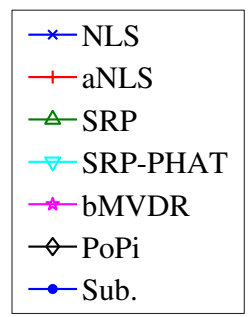
Experimental Results

Synthetic source in far-field



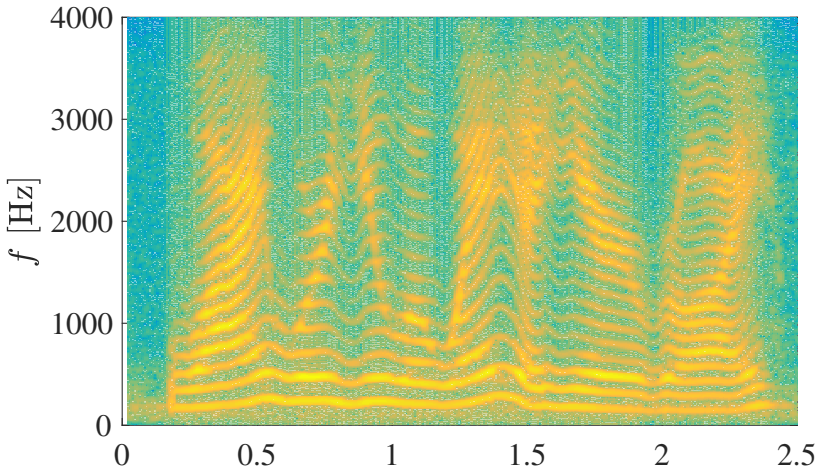


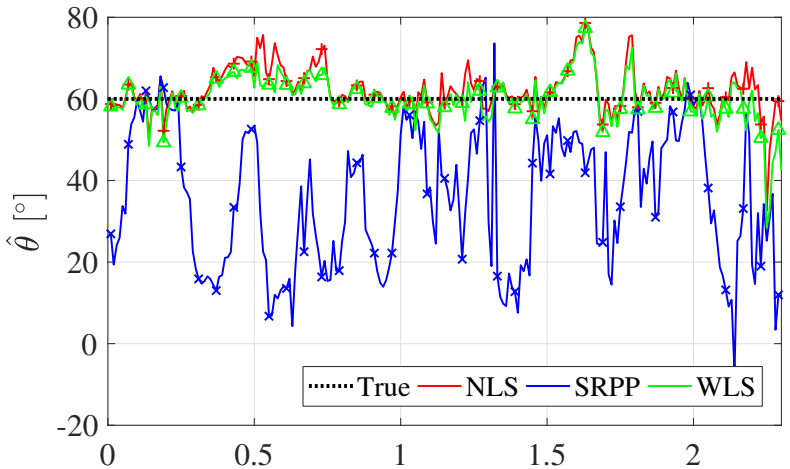


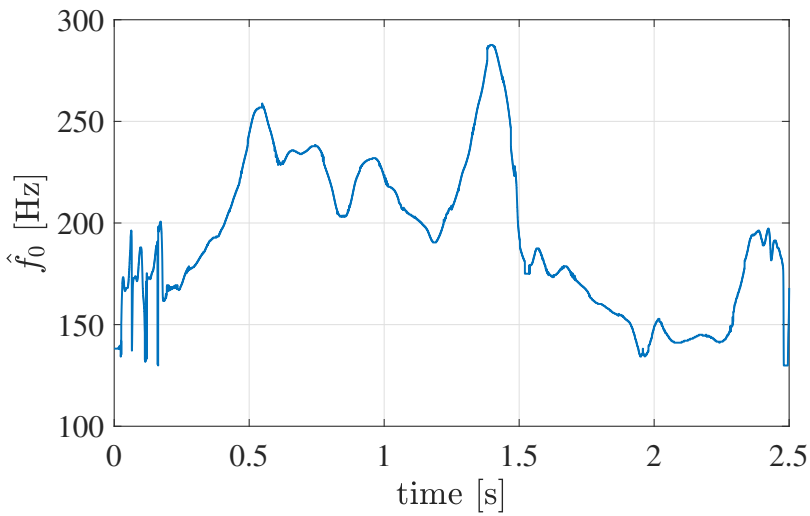


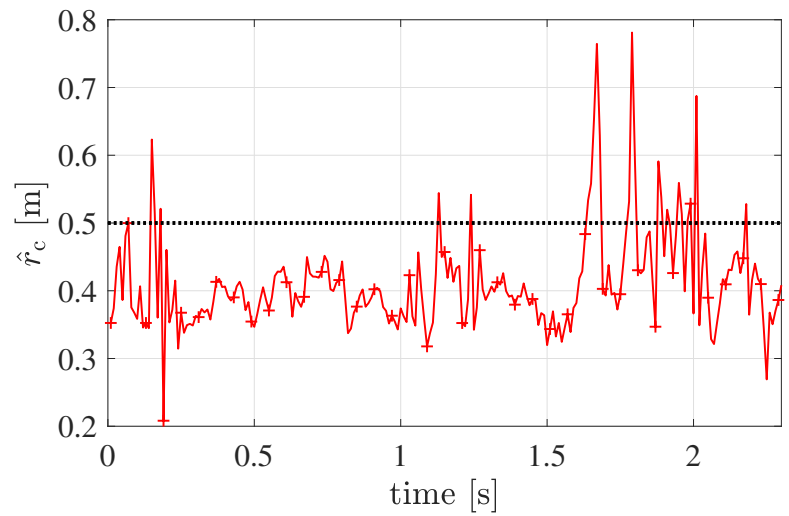
Experimental Results

Real speech in near-field









Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

- TDOA and DOA based Models
- Model-based Estimation of Fractional TDOAs
- Joint Estimation of DOA and Pitch
- Model Extensions for Robust Estimation
- Distortionless Enhancement of Audio
- Summary

Summary and Conclusion

Other Model Extensions

- The multichannel signal models can be extended in other ways to increase robustness.
- Reverberation (early reflections) can be included in the model, which it is not in most existing DOA estimators.
- Non-stationarity (movement, pitch changes, etc.) can be included, e.g., using chirp models.
- Uncertainty about DOA estimates can be included for robust beamforming (Zhao 2017).

Signal Model with Reverberation

An acoustic source is sampled using a microphone array:

$$y_k(n) = (s' * g_k)(n) + v'_k(n) = s_k(n) + v'_k(n), \tag{181}$$

where

$s'(n)$: clean source signal

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

$v'_k(n)$: additive noise (assumed white Gaussian).

Complete Signal Model

Observation modeled as multiple early reflections in noise:

$$\mathbf{y} = \sum_{r=1}^{R} \mathbf{H}(\eta_r)\boldsymbol{\alpha}_r + \mathbf{v}, \tag{182}$$

where

R: number of early reflections

$$\mathbf{H}(\eta_r) = [\mathbf{Z}^T \ (\mathbf{Z}\mathbf{D}_2(\eta_r))^T \ \cdots \ (\mathbf{Z}\mathbf{D}_K(\eta_r))^T]^T$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \cdots & \mathbf{z}_L & \mathbf{z}_1^* & \cdots & \mathbf{z}_L^* \end{bmatrix}$$

$$\mathbf{z}_l = \begin{bmatrix} 1 & e^{jl\omega_0} & \cdots & e^{j(N-1)l\omega_0} \end{bmatrix}^T$$

$$\mathbf{D}_k(\eta_r) = \text{diag}\left(\begin{bmatrix} \mathbf{d}_k^T(\eta_r) & \mathbf{d}_k^H(\eta_r) \end{bmatrix}\right)$$

$$\mathbf{d}_k(\eta_r) = \begin{bmatrix} e^{-j\omega_0 k\eta_r} & \cdots & e^{-jL\omega_0 k\eta_r} \end{bmatrix}^T$$

$\mathbf{v}$: late reverb + noise (assumed white Gaussian).

Reverb Robust DOA Estimation

- The model facilitates ML estimation of the DOA with reverb.
- Idea is to estimate DOAs of both direct-path and early reflections.
- Bias of direct-path estimate reduced in this way.
- Two methods based on nonlinear least squares were proposed:
  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 using an iterative approach.

Nonlinear Least Squares
Unstructured amplitudes

With unstructured amplitudes, the NLS estimator is

$$\{\widehat{\boldsymbol{\eta}}, \widehat{\overline{\boldsymbol{\alpha}}}\} = \arg \min_{\{\boldsymbol{\eta}, \overline{\boldsymbol{\alpha}}\}} \|\mathbf{y} - \overline{\mathbf{H}}(\boldsymbol{\eta})\overline{\boldsymbol{\alpha}}\|_2^2, \tag{183}$$

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 $\overline{\boldsymbol{\alpha}}$ yields

$$\widehat{\boldsymbol{\eta}} = \arg \max_{\boldsymbol{\eta}} \mathbf{y}^H \overline{\mathbf{H}} \left(\overline{\mathbf{H}}^H \overline{\mathbf{H}} \right)^{-1} \overline{\mathbf{H}}^H \mathbf{y}. \tag{184}$$

Iterative Procedure
Unstructured amplitudes

Consider a modified observed signal model:

$$\mathbf{y}_r = \mathbf{y} - \sum_{q=1, q\neq r}^{R} \mathbf{H}(\widehat{\eta}_q)\widehat{\boldsymbol{\alpha}}_q, \tag{185}$$

This suggests:

$$\widehat{\boldsymbol{\alpha}}_r = \left[\mathbf{H}^H(\eta_r)\mathbf{H}(\eta_r)\right]^{-1}\mathbf{H}(\eta_r)^H\mathbf{y}_r, \tag{186}$$

$$\widehat{\eta}_r = \arg\min_{\eta_r} \|\mathbf{P}^{\perp}_{\mathbf{H}(\eta_r)}\mathbf{y}_r\|_2^2. \tag{187}$$

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".

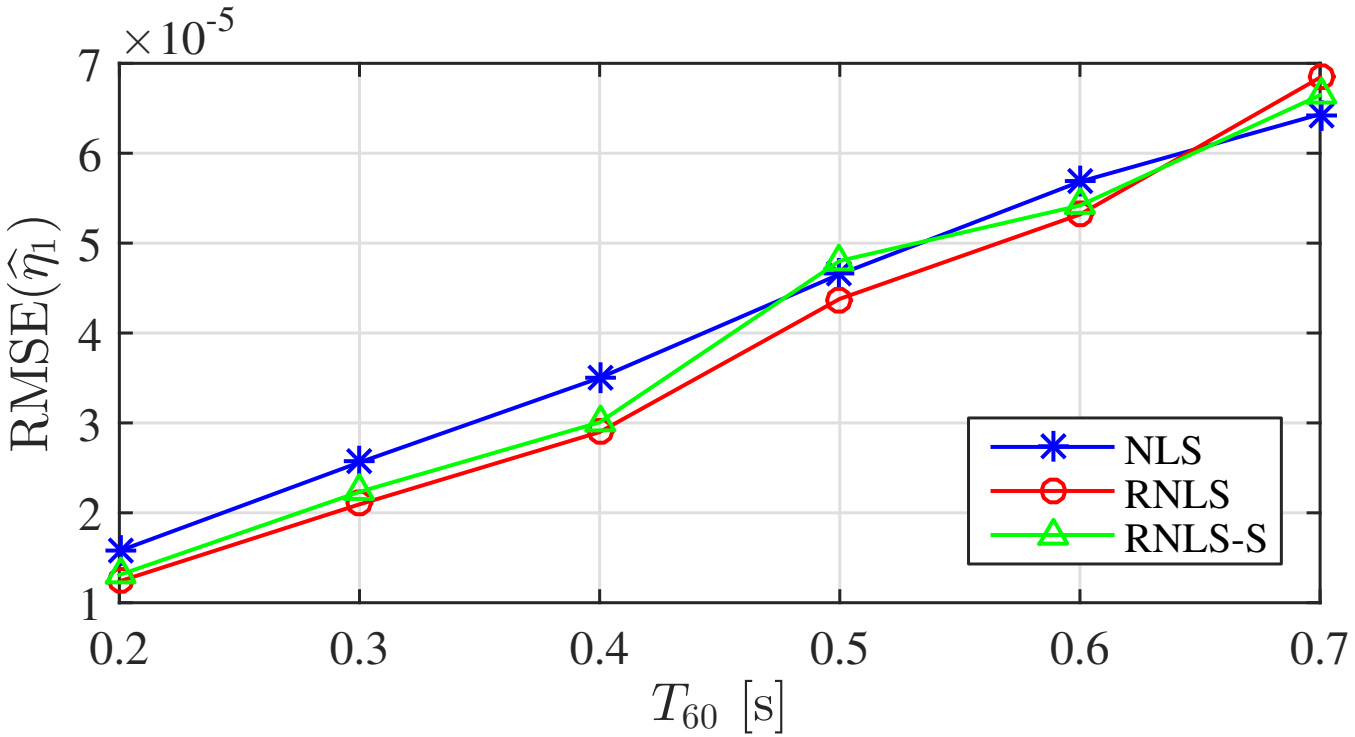
Remaining steps: Continue similarly to the previous steps until R is equal to the number of early reflections.

A similar method can be derived if we also model the dependency between early reflections as scaled and delayed versions of each other (termed RNLS-S).

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.

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

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

- TDOA and DOA based Models
- Model-based Estimation of Fractional TDOAs
- Joint Estimation of DOA and Pitch
- Model Extensions for Robust Estimation
- Distortionless Enhancement of Audio
- Summary

Summary and Conclusion

Distortionless Audio Enhancement

- Model-based approach can be used to derive distortionless filters for the enhancement of periodic signals (musical instruments, voiced speech, etc.).
- Compared to traditional speech enhancement method, these can be guaranteed distortionless!
- Also, do not require noise statistics estimates, but pitch and model order estimates instead.
- While noise statistics are difficult to estimate in practice, pitch can be estimated robustly at low SNRs.
- The filters can even extract nonstationary signal without distortion when using chirp model.
- Traditional STFT based method assumes stationarity within analysis window.

MSE-based Filters

Enhancement filters for periodic audio segments derived from MSE between filter output, $y(n)$ and desired output, $\widehat{y}(n)$,

$$P = \sum_{n=M-1}^{N-1} \frac{|y(n) - \hat{y}(n)|^2}{N - M + 1} = \sum_{n=M-1}^{N-1} \frac{|\mathbf{h}^H \mathbf{x}(n) - \boldsymbol{\alpha}^H \mathbf{w}(n)|^2}{N - M + 1}, \qquad (188)$$

with

$\mathbf{h}$: FIR filter coefficients,

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \cdots & \alpha_L \end{bmatrix}^T,$$

$$\mathbf{w}(n) = \begin{bmatrix} e^{j\omega_0 n} & \ldots & e^{j\omega_0 L n} \end{bmatrix}^T.$$

MSE-based Filters

MSE can be rewritten as:

$$P = \mathbf{h}^H \left(\widehat{\mathbf{R}} - \mathbf{G}^H \mathbf{W}^{-1} \mathbf{G} \right) \mathbf{h} \triangleq \mathbf{h}^H \widehat{\mathbf{Q}} \mathbf{h}. \tag{189}$$

where

- $\mathbf{G}$: time average of $\mathbf{w}(n)\mathbf{x}^H(n)$ outer products,
- $\mathbf{W}$: time average of $\mathbf{w}(n)\mathbf{w}^H(n)$ outer products.
- $\widehat{\mathbf{R}}$: sample mean covariance estimate of $\mathbf{x}(n)$.

Distortionless filter then obtained by solving

$$\min_{\mathbf{h}} \mathbf{h}^H \widehat{\mathbf{Q}} \mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{1}^T. \tag{190}$$

That is, $\widehat{\mathbf{h}} = \widehat{\mathbf{Q}}^{-1} \mathbf{Z} \left(\mathbf{Z}^H \widehat{\mathbf{Q}}^{-1} \mathbf{Z} \right)^{-1} \mathbf{1}$ (SF-APES).

Simplifications

- The design can also be done for a filterbank, with a filter for each harmonic (FB-*).
- The design can be simplified in different ways:
  1. Make the approximation that $\mathbf{W} = \mathbf{I}$ *-APES (appx).
  2. Replace $\widehat{\mathbf{Q}}$ with $\widehat{\mathbf{R}}$ *-Capon.
  3. Assume white Gaussian noise *-WNC.
  4. Exploit asymptoically orthogonal harmonics *-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

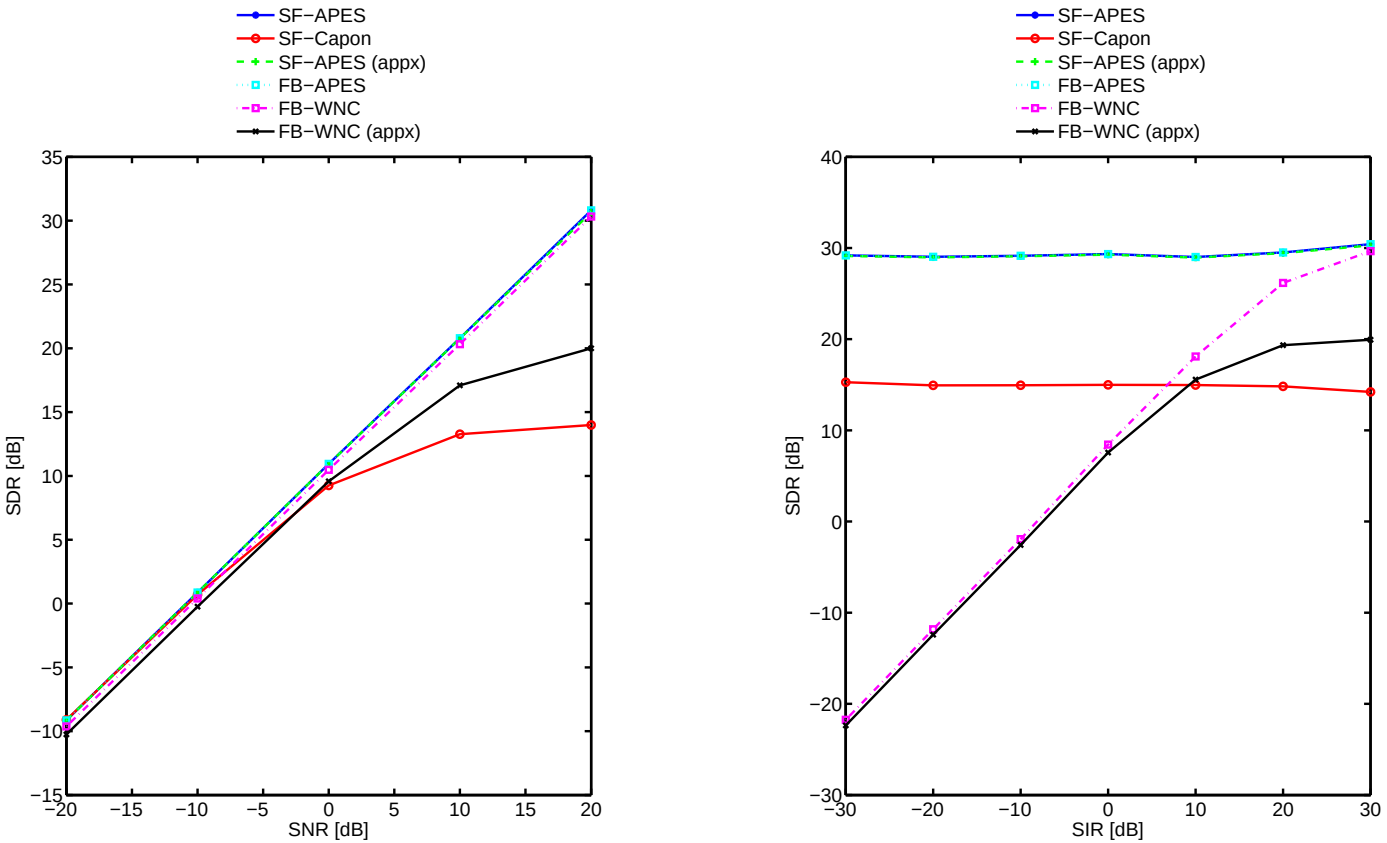


Figure: SDR versus (left) SNR and (right) SIR with an interfering source present (SNR of 10 dB).

Experiments on Synthetic Data

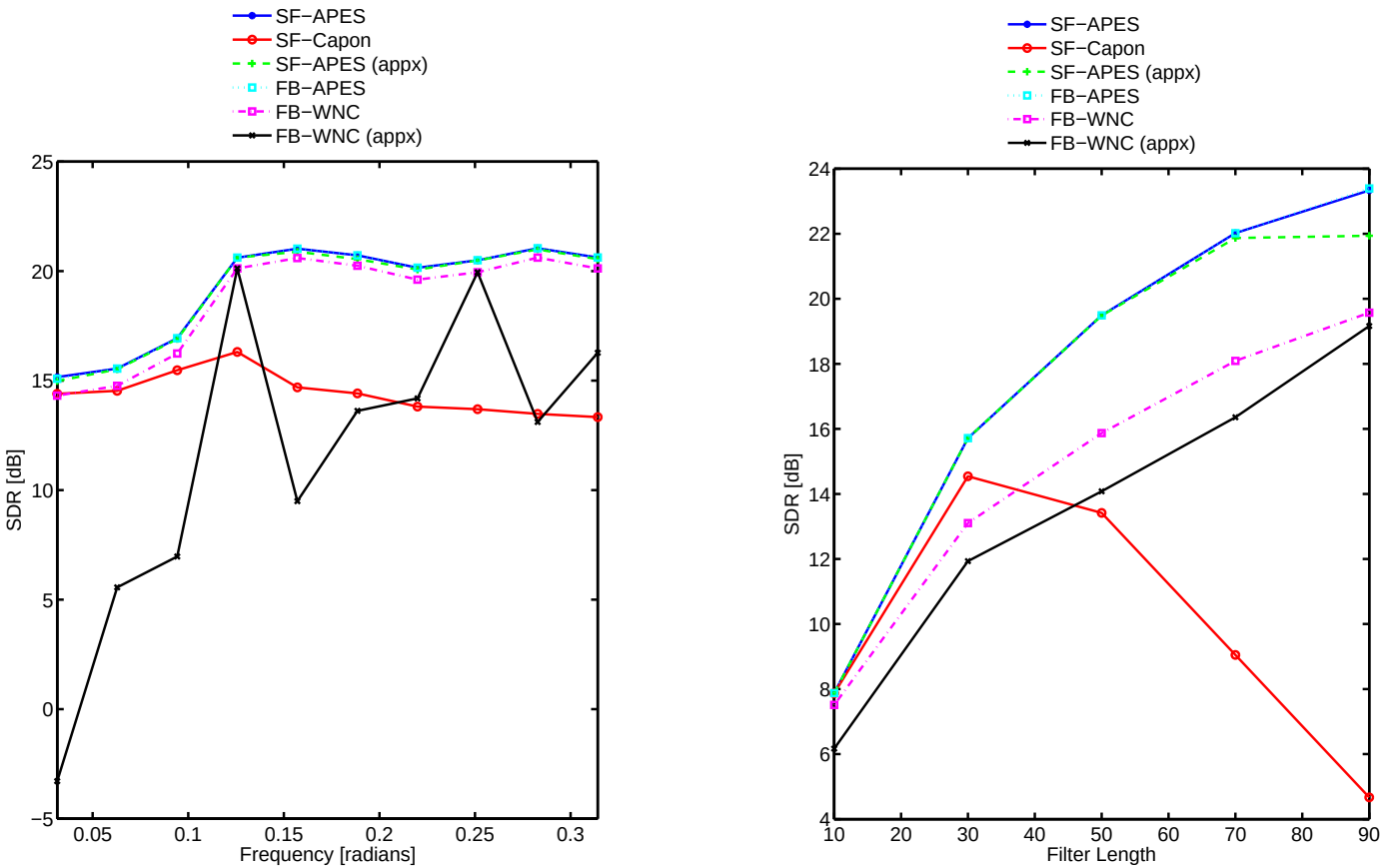


Figure: SDR versus (left) fundamental frequency, and (right) filter length with an interfering source present.

Experiments on Speech Data

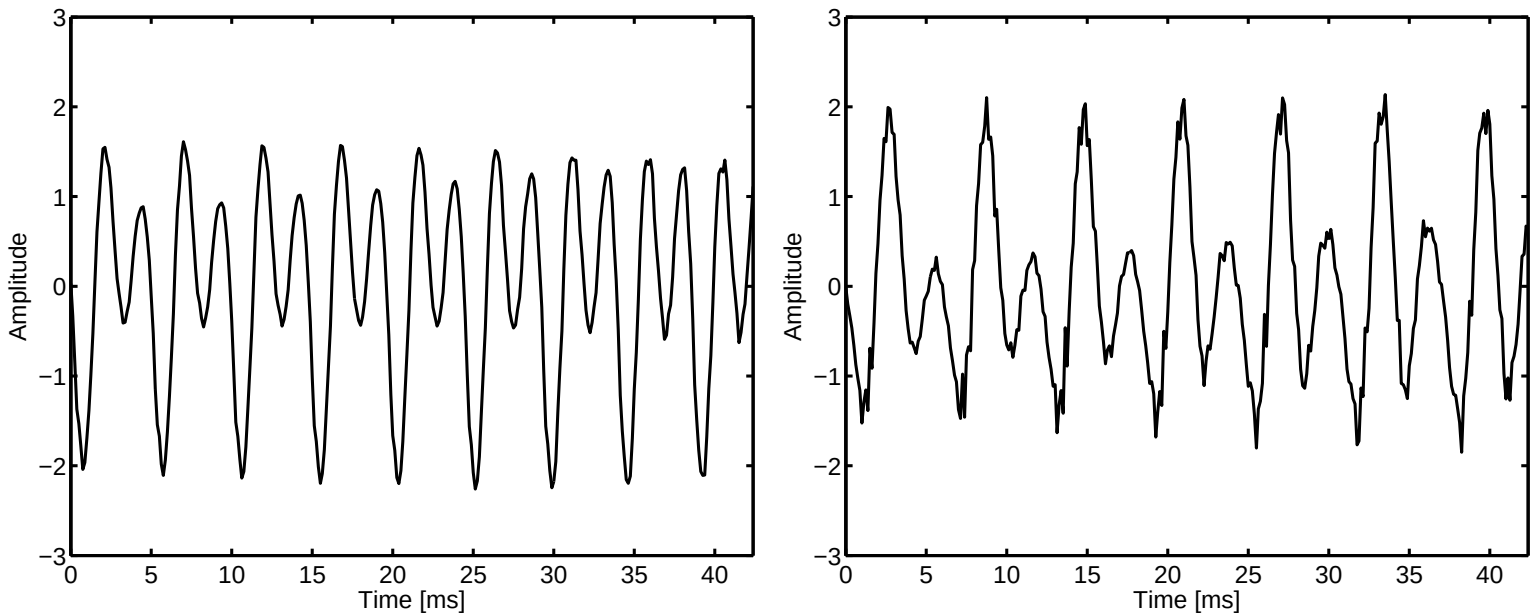


Figure: Plots of: voiced speech signal of sources (left) 1 and (right) 2.

Experiments on Speech Data

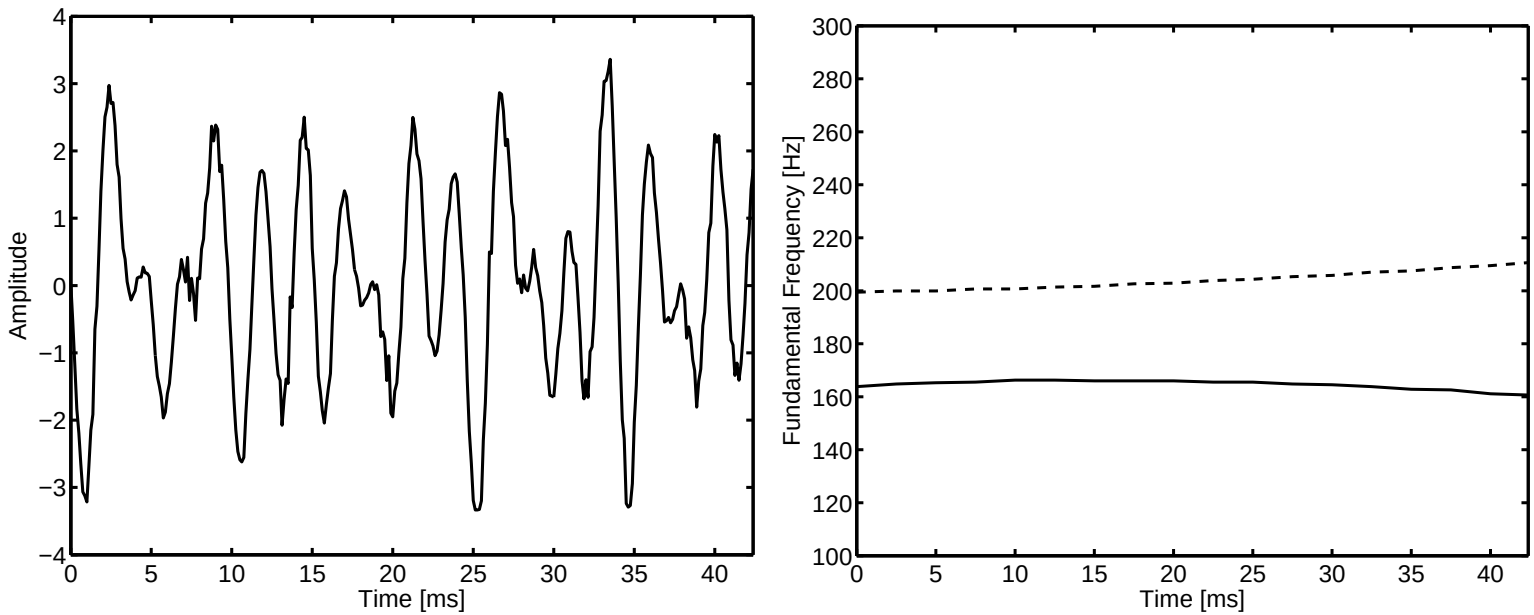


Figure: Plots of: (left) mixture of the two signals and (right) estimated pitch tracks for source 1 (dashed) and 2 (solid).

Experiments on Speech Data

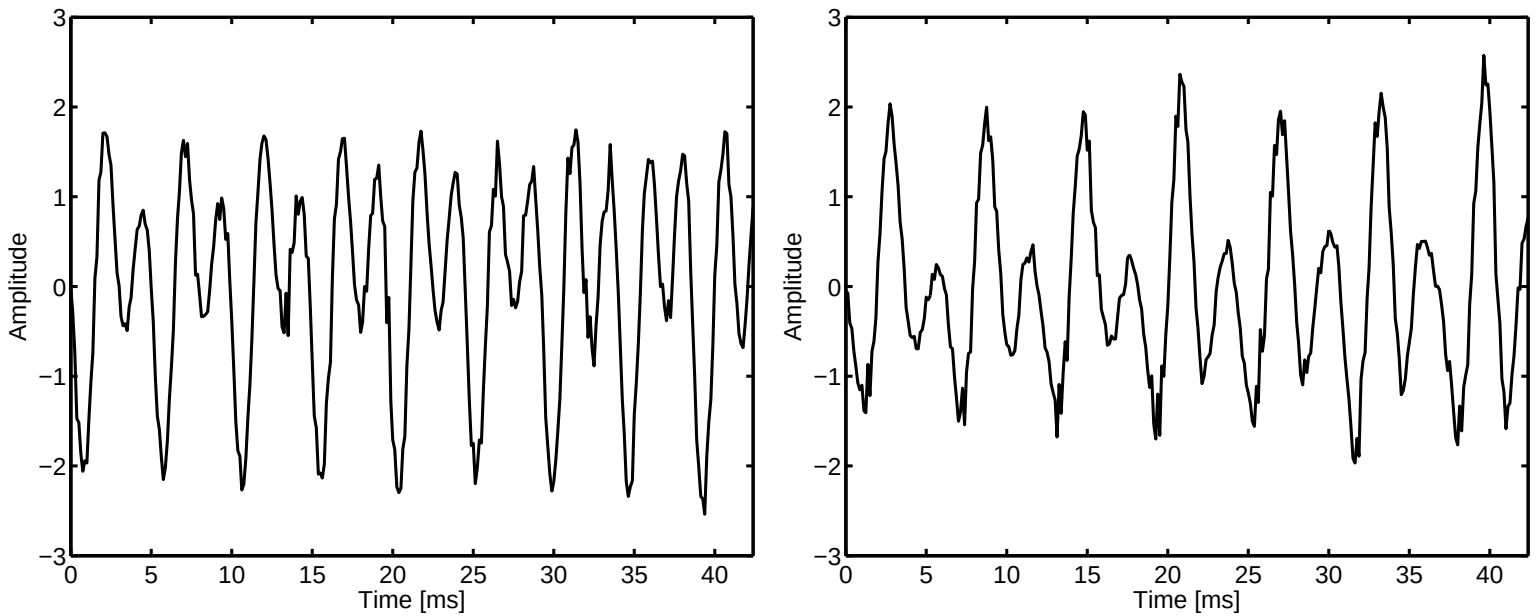
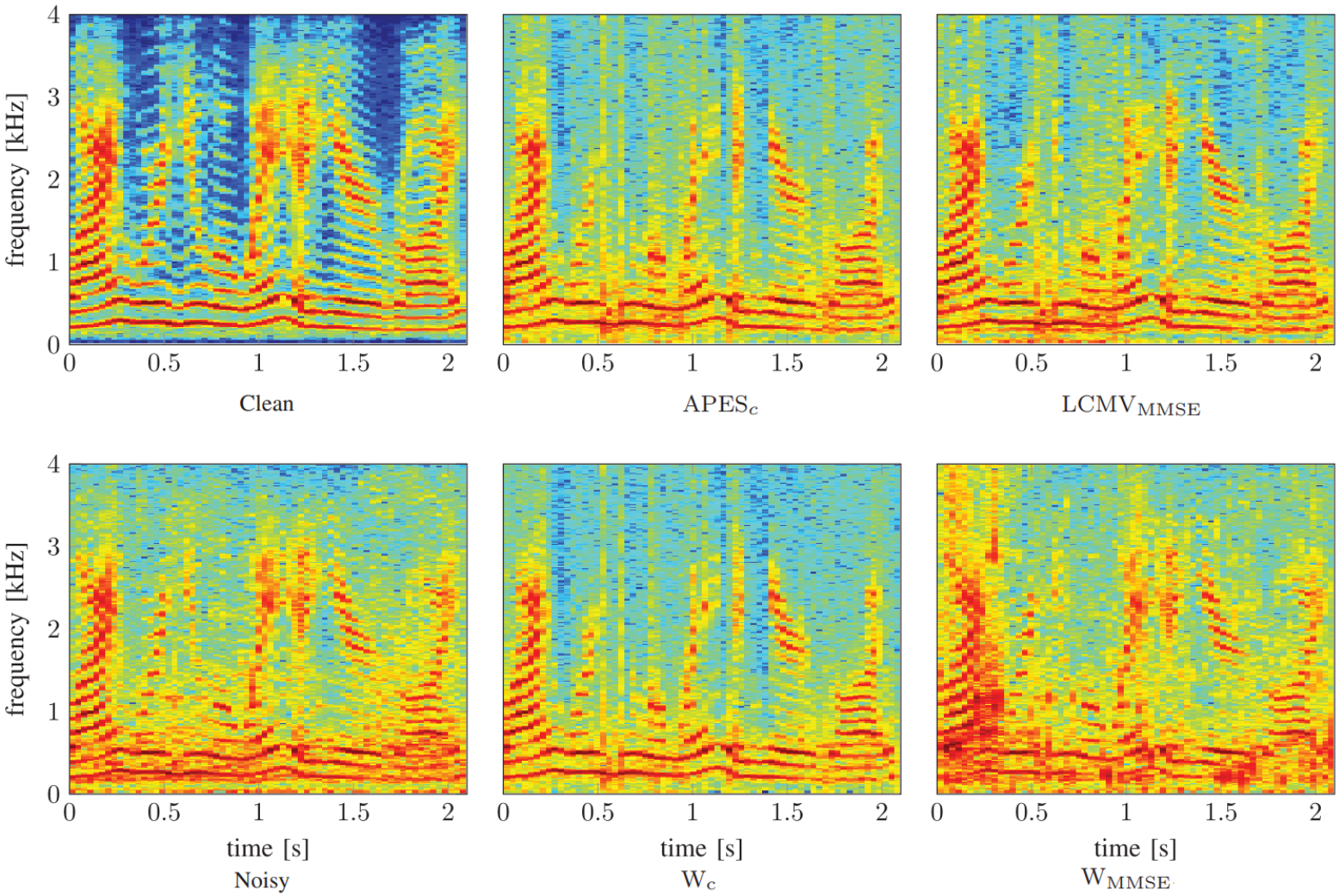


Figure: Plots of: estimate of sources (left) 1 and (right) 2 obtained from mixture.

Extension to Nonstationary Signals

- The idea of distortionless filtering was applied to enhancement of non-stationary signals (e.g., voiced speech) in (Nørholm 2016).
- Based on chirp model, that accounts for a linearly changing pitch.
- That is, non-stationarity is taken into account on a frame level as opposed to in traditional STFT based methods.
- Estimates noise statistics as a by-product, which can be used in other traditional enhancement methods.
- Results showed improvements (SNR, distortion, PESQ) over Wiener filtering based on an MMSE based noise statistics estimator (Gerkmann 2012).

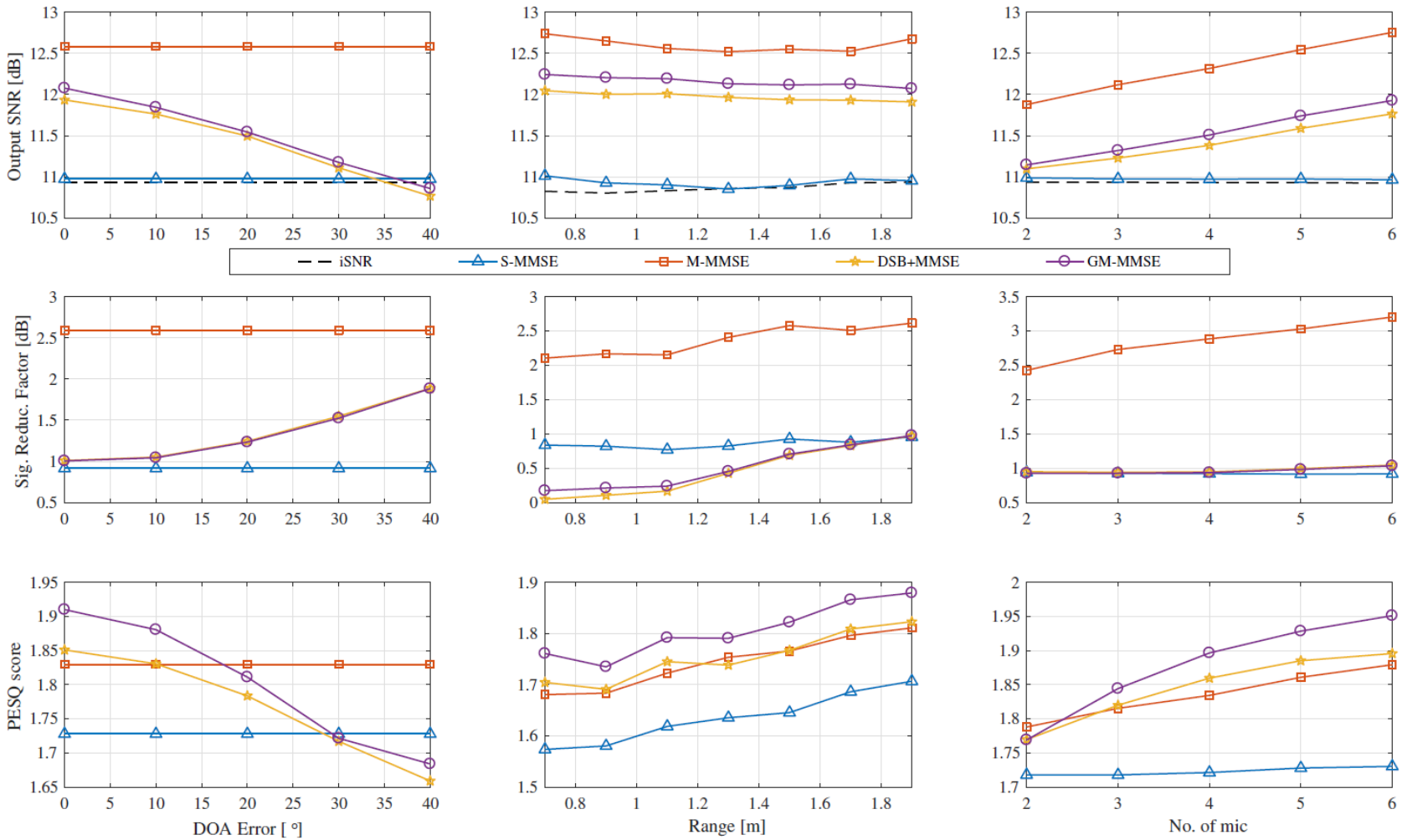
Extension to Nonstationary Signals



Extension to Multichannel

- The filtering method was extended to the multichannel case in (Jensen 2017).
- 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 (GM-MMSE),
  - a method being independent on geometry (M-MMSE).
- Results, in terms of PESQ scores, showed that geometry-based approach is best for larger arrays, but worse when significant DOA errors are present.

Extension to Multichannel



Outline

Introduction

Statistical Speech and Audio Models

Model-based Pitch Estimation

Model-based Single-Channel Enhancement

Model-based Array Processing and Enhancement

- TDOA and DOA based Models
- Model-based Estimation of Fractional TDOAs
- Joint Estimation of DOA and Pitch
- Model Extensions for Robust Estimation
- Distortionless Enhancement of Audio
- Summary

Summary and Conclusion

Summary

- The parametric methods typically provide better performance than their nonparametric counterparts.
- TDOA estimation using two microphones can be done more accurately with a parametric method compared to, e.g., GCC-based methods.
- Also, fractional TDOA's can be estimated without heuristics.
- Pitch can be estimated more accurately when using multiple microphones, and be used for more accurate DOA estimation.
- Joint DOA and pitch estimation yields better source resolvability in multisource scenarios.
- The models, and thus the methods, can be extended to account for, e.g., reverberation and nonstationarity.
- Distortionless filters for enhancement of single- and multichannel audio can be derived from the models.
- These do not require hard-to-obtain noise statistic estimates.

Outline

Introduction
Statistical Speech and Audio Models
Model-based Pitch Estimation
Model-based Single-Channel Enhancement
Model-based Array Processing and Enhancement
Summary and Conclusion

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
- 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
- Petre Stoica
- Tobias Lindstrøm Jensen
- Mathew Shaji Kavalekalam
- Martin Weiss Hansen
- Charlotte Sørensen
- Jesper Boldt
- Sidsel Marie Nørholm
- Sam Karimian-Azari
- Liming Shi