Model-based Analysis and Processing of Speech and Audio Signals

April 29, 2022

Mads Græsbøll Christensen

Audio Analysis Lab, CREATE
Aalborg University, Denmark

AALBORG UNIVERSITY
DENMARK

Preface

- Looking back, there are certain ideas that permeate my research over the past decades.
- Briefly put, my research has revolved around the ideas of
  - describing and analyzing audio signals using parametric and statistical models.
  - posing, analyzing, and solving problems in speech and audio using optimization, linear algebra, and statistics.
- In this presentation, I would like to talk more about the *model-based approach* and what can be achieved with it.
- I will do this primarily in the context of a specific model and our contributions.

Outline

Model-based Approach

Research questions: *What are good models of speech and audio signals recorded in adverse conditions, how do we find their parameters, and how can we use them?*

Model-based Approach

- Processing based on generative signal models described in terms of physically meaningful parameters.
- Speech and audio models have been around for many years (*we tried it in the 70s and it didn't work*).
- Skeptics argue that the models are (always) wrong and that it is not possible to estimate the parameters anyway.
- However, models can be used for many things and in different ways.
- The approach leads to robust, tractable and often fast methods that can be improved and analyzed.

Model-based Approach

Partial models, imperfect as they may be, are the only means developed by science for understanding the universe.

Rosenblueth & Wiener, 1945.

Model-based Approach

What is a good model?

- Fits the data well
- Physically meaningful
- As simple as possible!

We will now explore with an example how we can

- model speech and audio signals
- estimate parameters
- use and improve the model

Harmonic Model

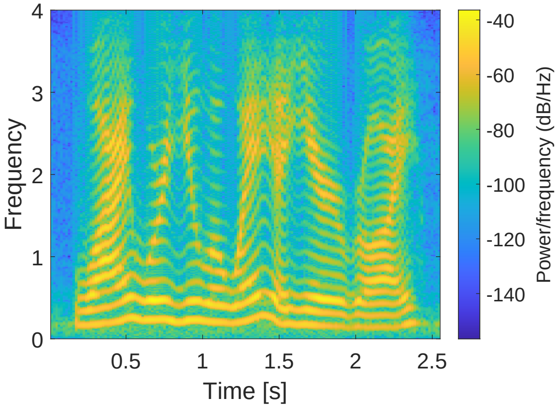


Figure: Spectrogram of speech utterance "why were you away a year, Roy?".

Harmonic Model

Many speech and audio signals are periodic or approximately so. Such signals can be modeled by the harmonic model given by (for $n = 0, \ldots, N-1$)

$$x(n) = s(n) + e(n) \quad (1)$$

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

Definitions:

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

Harmonic Model

The model can be written in matrix-vector notation as

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{s}(n) + \mathbf{e}(n) && (3)\\ &= \mathbf{Z}(n)\mathbf{a} + \mathbf{e}(n) && (4)\\ &= \mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{e}(n) && (5) \end{aligned}$$

with the following definitions:

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

$$\mathbf{z}(n,\omega) = [\ e^{j\omega n}\ e^{j\omega(n+1)}\ \dots\ e^{j\omega(n+M-1)}\ ]^T$$

$$\mathbf{Z}(n) = [\ \mathbf{z}(n,\omega_0)\ \cdots\ \mathbf{z}(n,\omega_0 L)\ ], \mathbf{Z} = \mathbf{Z}(0)$$

$$\mathbf{D}(n) = \mathrm{diag}(\left[e^{j\omega_0 n}\ e^{j\omega_0 2n}\ \dots\ e^{j\omega_0 L n}\ \right])$$

$$\mathbf{a} = [\ a_1\ \cdots\ a_L\ ]^T$$

Harmonic Model

The covariance matrix of $\mathbf{x}(n)$ denoted $\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\}$, can be Written in terms of the model, i.e.,

$$\mathbf{R} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \mathbf{Q}, \tag{6}$$

where $\mathbf{P} \approx \mathrm{diag}\left(\left[\begin{array}{ccc} A_1^2 & \cdots & A_L^2 \end{array}\right]\right)$ and $\mathbf{Q} = \mathrm{E}\left\{\mathbf{e}(n)\mathbf{e}^H(n)\right\}$. Often it is assumed that $\mathbf{Q} = \sigma^2\mathbf{I}$.

Let the output signal $y(n)$ of a filter having coefficients $\mathbf{h} \in \mathbb{C}^M$ be defined as

$$\begin{aligned} y(n) =& \mathbf{h}^H\mathbf{x}(n) & (7) \\ =& \mathbf{h}^H\mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{h}^H\mathbf{e}. & (8) \end{aligned}$$

The output power $\mathbf{h}^H\mathbf{R}\mathbf{h}$ is then

$$\mathrm{E}\left\{|y(n)|^2\right\} = \mathbf{h}^H\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{h} + \mathbf{h}^H\mathbf{Q}\mathbf{h}. \tag{9}$$

Harmonic Model

Let the EVD of $\mathbf{R}$ be

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H, \tag{10}$$

where $\mathbf{U}$ contains the M eigenvectors $\mathbf{u}_k$ of $\mathbf{R}$, i.e., and $\boldsymbol{\Lambda}$ is a diagonal matrix containing the corresponding (sorted) eigenvalues, λ_k.

Let $\mathbf{S}$ and $\mathbf{G}$ be formed as

$$\mathbf{S} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_L \end{bmatrix} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix} \mathbf{u}_{L+1} & \cdots & \mathbf{u}_M \end{bmatrix}. \tag{11}$$

Assuming $\mathbf{Q} = \sigma^2\mathbf{I}$ and observing that $\mathbf{U}\left(\boldsymbol{\Lambda} - \sigma^2\mathbf{I}\right)\mathbf{U}^H = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$ it follows that

$$\mathbf{Z}^H\mathbf{G} = \mathbf{0} \quad \text{and} \quad \mathcal{R}(\mathbf{S}) = \mathcal{R}(\mathbf{Z}). \tag{12}$$

Harmonic Model

Is it possible to

- estimate the nonlinear parameters?
- take non-stationarity into account?
- deal with interference and noise?
- extend the model to arrays?

Let us find out...

Fundamental Frequency Estimation

The variance of an unbiased estimate $\hat{\theta}_i$ of θ_i (i.e., the *i*th element of $\boldsymbol{\theta} \in \mathbb{R}^P$) is bounded by the Cramér-Rao lower bound (CRLB):

$$\mathrm{var}(\hat{\theta}_i) \geq \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}, \quad (13)$$

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

$$[\mathbf{I}(\boldsymbol{\theta})]_{il} = -\mathrm{E}\left\{ \frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_l} \right\}, \quad (14)$$

with $\ln p(\mathbf{x}; \boldsymbol{\theta})$ being the log-likelihood function for $\mathbf{x} \in \mathbb{C}^N$. The asymptotic CRLB for ω_0 (for WGN) is

$$\mathrm{var}(\hat{\omega}_0) \geq \frac{6\sigma^2}{N^3 \sum_{l=1}^{L} A_l^2 l^2}. \quad (15)$$

Fundamental Frequency Estimation

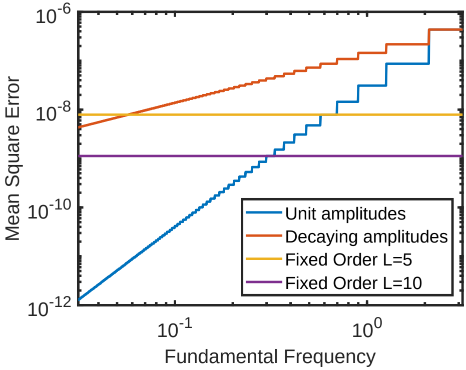


Figure: CRLB as a function of ω_0 for different cases.

Fundamental Frequency Estimation

For white Gaussian noise ($\mathbf{Q} = \sigma^2\mathbf{I}$) with $M = N$ the log-likelihood function is

$$\ln p(\mathbf{x}; \boldsymbol{\theta}) = -N \ln \pi - N \ln \sigma^2 - \frac{1}{\sigma^2} \|\mathbf{x} - \mathbf{Za}\|_2^2. \tag{16}$$

The maximum likelihood estimator is given by (Quinn 1991)

$$\hat{\omega}_0 = \arg\max_{\omega_0} \ln p(\mathbf{x}; \boldsymbol{\theta}) = \arg\max_{\omega_0} \mathbf{x}^H \mathbf{Z} \left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{x} \tag{17}$$

$$\approx \arg\max_{\omega_0} \sum_{l=1}^{L} \left| \sum_{n=0}^{N-1} x(n) e^{-j\omega_0 l n} \right|^2. \tag{18}$$

This can be computed using an FFT, i.e., using *harmonic summation* (Noll 1969) but (17) can also be implemented fast (Nielsen 2017)!

Fundamental Frequency Estimation

The principal angles $\{\xi_k\}$ between the two subspaces with projection matrices $\mathbf{\Pi}_Z$ and $\mathbf{\Pi}_G$, are defined for $k = 1, \ldots, K$ as (with $K = \min\{2L, M - 2L\}$)

$$\cos(\xi_k) = \max_{\mathbf{y}} \max_{\mathbf{z}} \frac{\mathbf{y}^H \mathbf{\Pi}_Z \mathbf{\Pi}_G \mathbf{z}}{\|\mathbf{y}\|_2 \|\mathbf{z}\|_2} \triangleq \mathbf{y}_k^H \mathbf{\Pi}_Z \mathbf{\Pi}_G \mathbf{z}_k = \kappa_k, \quad (19)$$

with $\mathbf{y}^H \mathbf{y}_i = 0$ and $\mathbf{z}^H \mathbf{z}_i = 0$ for $i = 1, \ldots, k-1$. The κ_ks are related to the Frobenius norm as $\|\mathbf{\Pi}_Z \mathbf{\Pi}_G\|_F^2 = \sum_{k=1}^{K} \kappa_k^2$. Thus, we can estimate ω_0 and L as (Christensen 2009)

$$(\hat{\omega}_0, L) = \arg\min_{\omega_0, L} \frac{1}{MK} \operatorname{Tr}\left\{\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{G}\mathbf{G}^H\right\} \quad (20)$$

$$\approx \arg\min_{\omega_0, L} \frac{1}{MK} \|\mathbf{Z}^H\mathbf{G}\|_F^2. \quad (21)$$

Fundamental Frequency Estimation

Recall that the filtered signal is $y(n) = \mathbf{h}^H \mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{h}^H \mathbf{e}$ and that

$$\mathrm{E}\left\{|y(n)|^2\right\} = \mathbf{h}^H \mathbf{Z}\mathbf{P}\mathbf{Z}^H \mathbf{h} + \mathbf{h}^H \mathbf{Q}\mathbf{h}. \quad (22)$$

Idea: design a filter as

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}\mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{1}^T. \quad (23)$$

This has the solution

$$\mathbf{h}^\star = \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}^H \mathbf{R}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}. \quad (24)$$

We can use this filter to estimate the fundamental frequency as (Christensen 2008)

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{h}^H \mathbf{R}\mathbf{h} \quad (25)$$

$$= \arg\max_{\omega_0} \mathbf{1}^T \left(\mathbf{Z}^H \mathbf{R}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}. \quad (26)$$

Pre-Whitening

Many estimators assume that $\mathbf{Q} = \sigma^2\mathbf{I}$. How do we deal with colored noise? Suppose that $\mathbf{e}(n) \sim \mathcal{N}(0, \mathbf{Q})$. We can transform $\mathbf{x}(n)$ as

$$\mathbf{A}^H\mathbf{x}(n) = \mathbf{A}^H\mathbf{s}(n) + \mathbf{A}^H\mathbf{e}(n). \tag{27}$$

Let $\mathbf{A}$ be the Cholesky factor of $\mathbf{Q}^{-1}$, then $\mathbf{A}^H\mathbf{Q}\mathbf{A} = \mathbf{I}$ and the noise is now distributed as $\mathbf{A}^H\mathbf{e}(n) \sim \mathcal{N}(0, \mathbf{I})$. This can be implemented as a filter and is called pre-whitening.

The matrix $\mathbf{Q}$ can be estimated in a number of ways:

- Noise trackers (Gerkmann 2012)
- Parametric NMF (Srinivasan 2007, Jensen 2018)
- Harmonic model (Nørholm 2016, Quinn 2021)

Separation & Noise Reduction

How do we deal with interference and noise? Introducing sources $x_k(n)$ indexed by k, we obtain (Christensen 2008)

$$x(n) = \sum_{k=1}^{K} x_k(n) = \sum_{k=1}^{K} \left(\sum_{l=1}^{L_k} a_{k,l} e^{j\omega_k l n} + e_k(n) \right) \tag{28}$$

$$= \underbrace{\sum_{l=1}^{L} a_l e^{j\omega_0 l n}}_{target} + \underbrace{e(n)}_{interference+noise} . \tag{29}$$

$e(n)$ is no longer Gaussian! The filtered signal $\mathbf{x}(n)$ is

$$\mathbf{h}^H \mathbf{x}(n) = \mathbf{h}^H \mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{h}^H \mathbf{e}. \tag{30}$$

If $\mathbf{h}^H\mathbf{Z} = \mathbf{1}^T$ then $\mathbf{h}^H\mathbf{Z}\mathbf{D}(n)\mathbf{a} = \mathbf{1}^T\mathbf{D}(n)\mathbf{a} = \sum_{l=1} a_l e^{j\omega_0 l n}$.

Separation & Noise Reduction

The output power can be written as $\mathrm{E}\left\{|y(n)|^2\right\} = \mathbf{h}^H\mathbf{ZPZ}^H\mathbf{h} + \mathbf{h}^H\mathbf{Qh}$. Optimal filters can be derived by as:

$$\min_{\mathbf{h}} \mathbf{h}^H\mathbf{Qh} \quad \text{s.t.} \quad \mathbf{h}^H\mathbf{Z} = \mathbf{1}. \tag{31}$$

The solution to this problem is:

$$\mathbf{h}^\star = \mathbf{Q}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Q}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}. \tag{32}$$

These filters attenuate noise and interference optimally!

Simplifications (Christensen 2010):

1. $\mathbf{Q} = \mathbf{R} \rightarrow \mathbf{h}^\star = \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}$.
2. $\mathbf{Q} = \sigma^2\mathbf{I} \rightarrow \mathbf{h}^\star = \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{1}$.
3. $\lim_{M\rightarrow\infty} M\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1} = \mathbf{Z} \rightarrow \mathbf{h}^\star = \frac{1}{M}\mathbf{Z1}$.

Non-Stationarity

Can we deal with a time-varying fundamental frequency? A more general signal model is the following:

$$x(n) = \sum_{l=1}^{L} A_l e^{j\theta_l(n)} + e(n), \tag{33}$$

where $\theta_l(t) = \int_0^t l\omega_0(\tau)d\tau + \phi_l$ is the instantaneous phase and $\omega_0(t)$ is the fundamental frequency. If $\omega_0(t) \approx \omega_0 + \alpha_0 t$, we get

$$\theta_l(t) = \frac{1}{2}\alpha_0 l t^2 + \omega_0 l t + \phi_l, \tag{34}$$

where α_0 is the fundamental chirp rate. The resulting model is called the *harmonic chirp model* (HCM)! α_0 and ω_0 can be estimated with NLS (Christensen 2014, Nørholm 2016).

Optimal filters can be designed for the HCM too (Nørholm 2016)!

Non-Stationarity

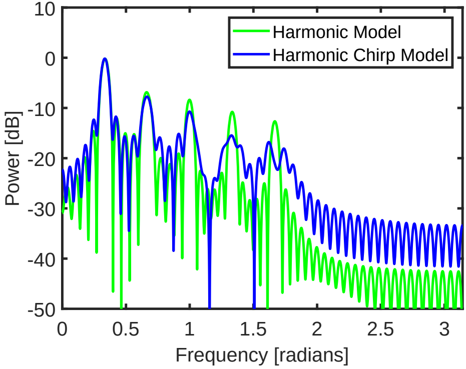
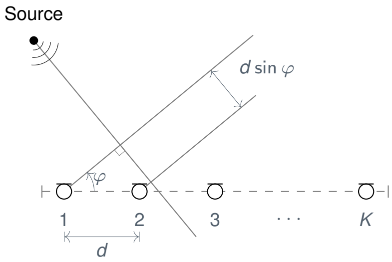


Figure: Spectrum of harmonic model and harmonic chirp model.

Array Processing

Suppose we have a uniform linear array and sources in the farfield:



The delay between microphone 1 and k is then related to the angle φ as $\Delta_k = \frac{d \sin \varphi}{c} f_s(k-1)$ where f_s is the sampling frequency.

Array Processing

Then we have that $\mathbf{s}_k(n) = \mathbf{s}(n - \Delta_k)$ and $\mathbf{x}_k(n) = \mathbf{s}_k(n) + \mathbf{e}_k(n)$ where k denotes the channel, and the signal for the kth channel is

$$\mathbf{x}_k(n) = \mathbf{Z}\mathbf{D}(n - \Delta_k)\mathbf{a} + \mathbf{e}_k(n), \tag{35}$$

with $\mathbf{e}_k$ being its noise. The signal at microphone k is then $\mathbf{s}_k(n) = \mathbf{s}(n - \Delta_k)$ and thus (Jensen 2014)

$$\mathbf{s}_k(n) = \mathbf{Z}\mathbf{D}\left(n - \frac{d \sin\varphi}{c} f_s (k-1)\right)\mathbf{a}. \tag{36}$$

As we can see, it is easy to account for fractional delays and other geometries can easily be incorporated too.

Array Processing

The observed signal can be organized in a matrix $\mathbf{X}(n) \in \mathbb{C}^{K \times M}$ as

$$\mathbf{X}(n) = \begin{bmatrix} x_1(n) & \cdots & x_1(n-M+1) \\ \vdots & \ddots & \vdots \\ x_K(n) & \cdots & x_K(n-M+1) \end{bmatrix}. \tag{37}$$

Defining $\mathbf{i}_k$ as the kth column of $\mathbf{I}_K$, the observed signal can be written as

$$\mathbf{X}^T(n)\mathbf{i}_k = \mathbf{Z}\mathbf{D}(n-\Delta_k)\mathbf{a} + \mathbf{e}_k(n). \tag{38}$$

Define the spatial frequency $\omega_s = \omega_0 f_s \frac{d \sin\varphi}{c}$ and the vectors

$$\mathbf{z}_t(\omega_0 l) = \begin{bmatrix} 1 & e^{-j\omega_0 l} & \cdots & e^{-j\omega_0 l(M-1)} \end{bmatrix}^T \tag{39}$$

$$\mathbf{z}_s(\omega_s l) = \begin{bmatrix} 1 & e^{-j\omega_s l} & \cdots & e^{-j\omega_s l(P-1)} \end{bmatrix}^T. \tag{40}$$

Array Processing

By introducing $\gamma_l(n) = a_l e^{j\omega_0 l n}$, the matrix $\mathbf{X}(n)$ can be modeled as

$$\mathbf{X}(n) = \sum_{l=1}^{L} \gamma_l(n)\mathbf{z}_\mathrm{s}(\omega_\mathrm{s} l)\mathbf{z}_\mathrm{t}^T(\omega_0 l) + \mathbf{E}(n), \tag{41}$$

where $\mathbf{E}(n) \in \mathbb{C}^{K\times M}$ is defined similarly to $\mathbf{X}(n)$. Defining $\bar{\mathbf{x}}(n) = \mathrm{vec}\{\mathbf{X}(n)\}$ where $\mathrm{vec}\{\cdot\}$ is the vectorization operator, the model can be written as

$$\bar{\mathbf{x}}(n) = \sum_{l=1}^{L} \gamma_l(n)\bar{\mathbf{z}}_l + \bar{\mathbf{w}}(n), \tag{42}$$

where $\bar{\mathbf{z}}_l$ is the vectorized version of the spatio-temporal model, i.e.,

$$\bar{\mathbf{z}}_l = \mathrm{vec}\{\mathbf{z}_s(\omega_\mathrm{s} l)\mathbf{z}_\mathrm{t}^T(\omega_0 l)\} \tag{43}$$

$$= \mathbf{z}_\mathrm{s}(\omega_\mathrm{s} l) \otimes \mathbf{z}_\mathrm{t}(\omega_0 l). \tag{44}$$

Array Processing

Defining the matrix $\bar{\mathbf{Z}} = \begin{bmatrix} \bar{\mathbf{z}}_1 & \cdots & \bar{\mathbf{z}}_L \end{bmatrix}$ we can impose the constraint $\bar{\mathbf{h}}^H \bar{\mathbf{Z}} = \mathbf{1}^T$, leading to the filter design problem:

$$\min_{\bar{\mathbf{h}}} \bar{\mathbf{h}}^H \bar{\mathbf{R}} \bar{\mathbf{h}} \quad \text{s.t.} \quad \bar{\mathbf{Z}}^H(n) \bar{\mathbf{h}} = \mathbf{1}, \tag{45}$$

where $\bar{\mathbf{R}}$ is the covariance matrix of $\bar{\mathbf{x}}(n)$. The solution to the above optimization problem is given by:

$$\bar{\mathbf{h}}^{\star} = \bar{\mathbf{R}}^{-1} \bar{\mathbf{Z}} \left(\mathbf{Z}^H \bar{\mathbf{R}}^{-1} \bar{\mathbf{Z}} \right)^{-1} \mathbf{1}. \tag{46}$$

This yields the following estimator of ω_0 and φ (Jensen 2015):

$$\{\hat{\omega}_0, \hat{\varphi}\} = \arg \max_{\omega_0, \varphi} \mathbf{1}^T \left(\mathbf{Z}^H \bar{\mathbf{R}}^{-1} \bar{\mathbf{Z}} \right)^T \mathbf{1}. \tag{47}$$

The filter can also be used for separation and noise reduction!

Contributions

Key contributions of thesis:

- Methods for order estimation (Paper A)
- Optimal filters for periodic signals (Paper B, E)
- Fundamental frequency estimators (Paper C)
- Models, estimators, and filters for non-stationary signals (Paper D, E)
- Sparse linear prediction (Paper F)
- Model-based array processing (Paper G)

Contributions

Related contributions:

- The NLS fundamental frequency estimator can be implemented fast (Nielsen 2017) and in a Bayesian framework (Shi 2019)
- Subspace and optimal filtering methods can be unified (Jensen 2016)
- Model-based enhancement can improve speech intelligibility in babble noise (Kavelekalam 2019)
- Signal models can be used to detect string, fret, picking position, etc. (Hjerrild 2017)
- Parametric NMF can estimate statistics for pre-whitening (Nielsen 2018, Esquivel 2019)
- The model selection problem can be solved in a number of ways (Nielsen 2014)
- Real-time and stable sparse linear prediction possible (Jensen 2013, Giacobello 2014)

Contributions

What can this all be used for?

- Hearing aids
- Voice analysis
- Telecommunication
- Reproduction systems
- Fault detection
- Music equipment

And many other things...

Conclusions

- As we have seen, there are a number of advantages to the model-based approach.
- Critical assumptions can easily be identified and can be mitigated, if necessary.
- The harmonic model can be used for (approximately) periodic signals, such as speech and audio.
- It is possible to estimate its parameters in adverse conditions and computationally efficient implementations exist.
- It is possible to deal with noise, interference, and non-stationarity and to extend the principles to arrays.
- There are many more problems that could probably benefit from this approach!
- These include applications with multiple channels, adverse conditions, and when the fine details and the physics matter.

Thanks to

Current/former students and collaborators
Jesper R. Jensen and Jesper K. Nielsen
Family and friends
Erik
Ane

AALBORG UNIVERSITY
DENMARK

AUDIO ANALYSIS LAB

Audio for good health and well-being