Statistical Parametric Speech Processing

Solving problems with the model-based approach

June 2 2016

Mads Græsbøll Christensen

Audio Analysis Lab, AD:MT
Aalborg University
Denmark

AALBORG UNIVERSITY
DENMARK

Preface

- Looking back, there are certain ideas that permeate my research.
- Briefly put, my research has revolved around the ideas of
  - describing and analyzing audio signals using parametric and statistical models.
  - posing and solving engineering problems in audio and acoustics using optimization, linear algebra, and statistics.
- In this talk, I would like to tell more about those ideas and what can be achieved with them.
- I will do this in the context of speech processing.

Outline

Section 1

Introduction

Motivation

- Parametric speech models have been around for many years (e.g., linear prediction in the 70s, sinusoidal model in the 80s).
- Skeptics argue that the models are (always) wrong and that it is not possible to estimate the model parameters well enough under adverse conditions.
- Parametric models can, however, be used for many things and in different ways.
- The harmonic model, for example, is a good model of quasi-periodic signals, like voiced speech. It describes the signal in terms of pitch, amplitudes, and phases.

Motivation

All models are wrong; some models are useful. (G. Box)

Motivation

Everything should be made as simple as possible, but no simpler. (A. Einstein)

Motivation

For every complex problem there is an answer that is clear, simple, and wrong. (H. L. Mencken)

Motivation

Methodology:

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

Why parametric methods?

- They lead to robust, tractable methods whose properties can be analyzed and understood.
- A full parametrization of the signal of interest is obtained.
- Back to basics... how can we hope to solve complicated problems if we cannot solve the simple ones?

Motivation

Some interesting questions:

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

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.

Harmonic Model

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

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

Definitions:

- $s(n)$ is voiced speech
- $e(n)$ is the observation noise
- ω_0 is the fundamental frequency
- $\psi_l = \omega_0 l$ 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 also be written as (with $\mathbf{x}(n)$ being a snapshot)

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{s}(n) + \mathbf{e}(n) && (2) \\ &= \mathbf{Z}(n)\mathbf{a} + \mathbf{e}(n) && (3) \\ &= \mathbf{Z}\mathbf{D}^n\mathbf{a} + \mathbf{e}(n) && (4) \\ &= \mathbf{Z}\mathbf{a}(n) + \mathbf{e}(n), && (5) \end{aligned}$$

with the following definitions:

$$\mathbf{x}(n) = [\ x(n)\ \cdots\ x(n+M-1)\ ]^T$$
$$\mathbf{z}(\omega) = [\ 1\ e^{j\omega}\ \cdots\ e^{j\omega(M-1)}\ ]^T$$
$$\mathbf{Z} = [\ \mathbf{z}(\omega_0)\ \cdots\ \mathbf{z}(\omega_0 L)\ ]$$
$$\mathbf{D} = \mathrm{diag}(e^{j\omega_0}, e^{j\omega_0 2}, \ldots, e^{j\omega_0 L})$$
$$\mathbf{a} = [\ a_1\ \cdots\ a_L\ ]^T$$

Harmonic Model

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

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\}. \tag{6}$$

Written in terms of the harmonic model, we get

$$\begin{aligned}\mathbf{R} &= \mathbf{Z}\mathrm{E}\left\{\mathbf{a}(n)\mathbf{a}^H(n)\right\}\mathbf{Z}^H + \mathrm{E}\left\{\mathbf{e}(n)\mathbf{e}^H(n)\right\} &(7)\\ &= \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \mathbf{Q}, &(8)\end{aligned}$$

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

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

$$\mathbf{P} \approx \mathrm{diag}\left(\left[\, A_1^2 \;\cdots\; A_L^2 \,\right]\right). \tag{9}$$

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, signals are not perfectly periodic
- The model does not differentiate between background noise and unvoiced speech
- It is single-channel

Can this be dealt with? Does it matter?

Section 2

Estimating Parameters

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

$$\mathrm{E}\left\{\hat{\theta}_i\right\} = \theta_i \;\forall \theta_i, \tag{10}$$

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

$$\mathrm{var}(\hat{\theta}_i) \geq \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}, \tag{11}$$

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

$$[\mathbf{I}(\boldsymbol{\theta})]_{il} = -\mathrm{E}\left\{\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i \partial\theta_l}\right\}, \tag{12}$$

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):

$$\mathrm{var}(\hat{\omega}_0) \geq \frac{6\sigma^2}{N(N^2-1)\sum_{l=1}^{L} A_l^2 l^2} \tag{13}$$

$$\mathrm{var}(\hat{A}_l) \geq \frac{\sigma^2}{2N} \tag{14}$$

$$\mathrm{var}(\hat{\phi}_l) \geq \frac{\sigma^2}{2N}\left(\frac{1}{A_l^2} + \frac{3l^2(N-1)^2}{\sum_{m=1}^{L} A_m m^2 (N^2-1)}\right). \tag{15}$$

These depend on the following quantity:

$$PSNR = 10\log_{10}\frac{\sum_{l=1}^{L} A_l^2 l^2}{\sigma^2}\ [\mathrm{dB}]. \tag{16}$$

For colored noise, pre-whitening should be employed.

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.
- 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. These predict more complicated phenomena.

Parameter Estimation Bounds

It is possible to relate estimation errors to reconstruction errors. Let the observed signal be given by

$$\mathbf{x} = \mathbf{s}(\theta) + \mathbf{e} \tag{17}$$

Suppose an estimate $\hat{\theta}$ of θ is used to reconstruct the ith sample as $\hat{s}_i = s_i(\hat{\theta})$, which can be approximated as

$$s_i(\theta + \epsilon) \approx s_i(\theta) + \left(\frac{\partial s_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^H \epsilon. \tag{18}$$

The mean squared error (MSE) is then

$$E\left\{(s_i(\theta) - s_i(\theta + \epsilon))^2\right\} = \left(\frac{\partial s_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^H E\left\{\epsilon\epsilon^H\right\}\left(\frac{\partial s_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right). \tag{19}$$

Parameter Estimation Bounds

If a MLE is used (for sufficiently high N), then

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}^{-1}(\theta)), \tag{20}$$

where $\mathbf{I}(\theta)$ is the FIM! For Gaussian signals with $\mathbf{x} \sim \mathcal{N}(\mathbf{s}(\theta), \mathbf{Q})$ where $\mathbf{Q}$ is the noise covariance matrix, the FIM is given by

$$[\mathbf{I}(\theta)]_{nm} = \frac{\partial \mathbf{s}^H(\theta)}{\partial \theta_n} \mathbf{Q}^{-1} \frac{\partial \mathbf{s}(\theta)}{\partial \theta_m}. \tag{21}$$

The MSE can then be seen to be

$$E\left\{(s_i(\theta) - s_i(\theta + \epsilon))^2\right\} = \left(\frac{\partial s_i(\theta)}{\partial \boldsymbol{\theta}}\right)^H \mathbf{I}^{-1}(\theta) \left(\frac{\partial s_i(\theta)}{\partial \boldsymbol{\theta}}\right). \tag{22}$$

Maximum Likelihood Method

For Gaussian signals, the likelihood function is

$$p(\mathbf{x}(n);\theta) = \frac{1}{\pi^M \det(\mathbf{Q})} e^{-(\mathbf{x}(n)-\mathbf{Z}\mathbf{a}(n))^H \mathbf{Q}^{-1}(\mathbf{x}(n)-\mathbf{Z}\mathbf{a}(n))}. \tag{23}$$

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

$$p(\{\mathbf{x}(n)\};\theta) = \prod_{n=0}^{G-1} p(\mathbf{x}(n);\theta). \tag{24}$$

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

$$\hat{\theta} = \arg\max \mathcal{L}(\theta). \tag{25}$$

Maximum Likelihood Method

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

$$\mathcal{L}(\boldsymbol{\theta}) = -N \ln \pi - N \ln \sigma^2 - \frac{1}{\sigma^2}\|\mathbf{x} - \mathbf{Z}\mathbf{a}\|_2^2. \tag{26}$$

The concentrated MLE is given by

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathcal{L}(\omega_0) = \arg\max_{\omega_0} \mathbf{x}^H\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x} \tag{27}$$

$$\approx \arg\max_{\omega_0} \sum_{l=1}^{L} \left| \sum_{n=0}^{N-1} x(n) e^{-j\omega_0 l n} \right|^2. \tag{28}$$

This can be computed using an FFT (i.e., using *harmonic summation*)!

Subspace Method

Recall that the model is

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{a}(n) + \mathbf{e}(n), \tag{29}$$

and that the covariance matrix then is

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \sigma^2\mathbf{I}, \tag{30}$$

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

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

Subspace Method

Let $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ be the EVD of the $\mathbf{R}$, and let $\mathbf{G}$ be formed as

$$\mathbf{G} = \begin{bmatrix} \mathbf{u}_{L+1} & \cdots & \mathbf{u}_M \end{bmatrix} \quad (31)$$

i.e., from the eigenvectors $\mathbf{u}_k$ corresponding to the $M - L$ smallest eigenvalues. Then we have that $\mathbf{Z}^H\mathbf{G} = \mathbf{0}$.

By measuring the angles between subspaces, we can obtain an estimate as

$$\hat{\omega}_0 = \arg\min_{\omega_0} \|\mathbf{Z}^H\mathbf{G}\|_F^2 = \arg\min_{\omega_0} \sum_{l=1}^{L} \|\mathbf{z}^H(\omega_0 l)\mathbf{G}\|_2^2. \quad (32)$$

This maximizes the angles between the subspaces $\mathcal{R}(\mathbf{Z})$ and $\mathcal{R}(\mathbf{G})$.

Filtering Method

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

$$y(n) = \sum_{m=0}^{M-1} h(m)x(n-m) = \mathbf{h}^H \mathbf{x}(n), \tag{33}$$

with $M \leq N$ and where $\mathbf{h}$ is a vector formed from $\{h(n)\}$. The output power is then $\mathrm{E}\left\{|y(n)|^2\right\} = \mathbf{h}^H \mathbf{R} \mathbf{h}$.

The filtered output can be seen to be

$$\mathbf{h}^H \mathbf{x}(n) = \mathbf{h}^H \mathbf{Z} \mathbf{D}^n \mathbf{a} + \mathbf{h}^H \mathbf{e}. \tag{34}$$

If $\mathbf{h}^H \mathbf{Z} = \mathbf{1}^T$ with $\mathbf{1} = [\ 1\ \cdots\ 1\ ]^T$ the voiced speech would pass undistorted and the noise term $\mathbf{h}^H \mathbf{e}$ could be minimized!

Filtering Method

We would thus like to design a filter as

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R} \mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{1}^T. \tag{35}$$

This has the solution

$$\mathbf{h} = \mathbf{R}^{-1} \mathbf{Z} \left(\mathbf{Z}^H \mathbf{R}^{-1} \mathbf{Z} \right)^{-1} \mathbf{1}. \tag{36}$$

We can use this filter to estimate the pitch as

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

Section 3

Solving Problems

Multi-Channel Modeling
Introduction

- A myriad of different pitch estimators exist, but very few have been proposed for multiple channels except a few heuristic ones.
- We will now derive a method for multi-channel pitch estimation based on a parametric model.
- The signals in the various channels share the same fundamental frequency but can have different amplitudes, phases, and noise characteristics.
- This means that the model allows for different conditions in the various channels!

Multi-Channel Modeling
Signal Model

The method operates on snapshots $\mathbf{x}_k(n) \in \mathbb{C}^M$ for the kth channel.

These are modeled as sums of sinusoids in Gaussian noise $\mathbf{e}_k$ with covariance $\mathbf{Q}_k$, i.e.,

$$\mathbf{x}_k(n) = \mathbf{Z}(n)\mathbf{a}_k + \mathbf{e}_k(n), \quad (38)$$

with $\mathbf{a}_k = [\, A_{k,1}e^{j\phi_{k,1}} \;\cdots\; A_{k,L}e^{j\phi_{k,L}} \,]^T$. Let θ_k be the parameter vector for the kth channel. The likelihood function is then

$$p(\mathbf{x}_k(n);\theta_k) = \frac{1}{\pi^M \det(\mathbf{Q}_k)} e^{-\mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}. \quad (39)$$

Multi-Channel Modeling

Signal Model

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

$$p(\{\mathbf{x}_k(n)\};\{\boldsymbol{\theta}_k\}) = \prod_{k=1}^{K} \frac{1}{\pi^{MG}\det(\mathbf{Q}_k)^G} e^{-\sum_{n=0}^{G-1} \mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}. \quad (40)$$

For simplicity, we assume that the noise is white in each channel but has different σ_k^2, i.e., $\mathbf{Q}_k = \sigma_k^2\mathbf{I}$.

The log-likelihood function then reduces to

$$\ln p(\{\mathbf{x}_k(n)\};\{\boldsymbol{\theta}_k\}) = -GM\sum_{k=1}^{K}\ln\left(\pi\sigma_k^2\right) - \sum_{k=1}^{K}\sum_{n=0}^{G-1}\frac{\|\mathbf{e}_k(n)\|^2}{\sigma_k^2}. \quad (41)$$

Multi-Channel Modeling
Estimator

The MLE of the amplitudes for channel k are

$$\hat{\mathbf{a}}_k = \left(\sum_{n=0}^{G-1} \mathbf{Z}^H(n)\mathbf{Z}(n) \right)^{-1} \sum_{n=0}^{G-1} \mathbf{Z}^H(n)\mathbf{x}_k(n). \tag{42}$$

This can be used to form a noise variance estimate as

$$\hat{\sigma}_k^2 = \frac{1}{GM} \sum_{n=0}^{G-1} \|\hat{\mathbf{e}}_k(n)\|^2 = \frac{1}{GM} \sum_{n=0}^{G-1} \|\mathbf{x}_k(n) - \mathbf{Z}(n)\hat{\mathbf{a}}_k\|^2. \tag{43}$$

This yields the following log-likelihood for channel k at time n

$$\ln p(\mathbf{x}_k(n); \omega_0) = -M \ln \pi - M \ln \hat{\sigma}_k^2.$$

Multi-Channel Modeling
Estimator

For all n and k, this yields

$$\ln p(\{\mathbf{x}_k(n)\}; \omega_0) = -GMK \ln \pi - GM \sum_{k=1}^{K} \ln \hat{\sigma}_k^2. \tag{44}$$

The maximum likelihood estimator (MLE) can finally be stated as

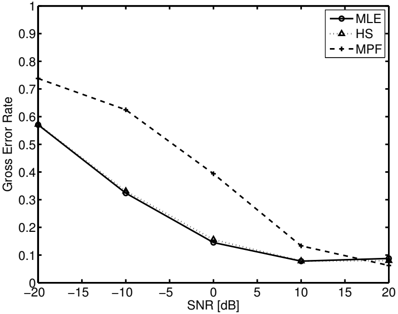
$$\hat{\omega}_0 = \arg\min_{\omega_0} \sum_{k=1}^{K} \ln \hat{\sigma}_k^2. \tag{45}$$

This estimator can then be approximated as

$$\hat{\omega}_0 = \arg\min_{\omega_0} \sum_{k=1}^{K} \ln \left(\|\mathbf{x}_k\|^2 - \frac{1}{N} \|\mathbf{Z}\mathbf{x}_k\|^2 \right), \tag{46}$$

where $\mathbf{x}_k = \mathbf{x}_k(0)$ for $M = N$. This can be computed using FFTs.

Multi-Channel Modeling
Experiments



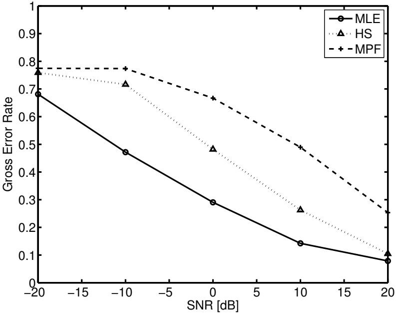


Figure: Gross error rate for (left) symmetrical noise level and (right) asymmetrical noise level (i.e., different noise levels).

Noise Reduction
Introduction

- The harmonic signal model has been used for noise reduction in various ways, like the traditional comb filters.
- We have seen how adaptive and optimal filters can be used for pitch estimation.
- The same principle can be used for finding noise reduction filters.
- Some interesting and well-known special cases can be obtained from these filters.

Noise Reduction
Filter Design

As we saw earlier, we get the following model when a filter **h** is applied to the observed signal $\mathbf{x}(n)$:

$$\hat{s}(n) = \mathbf{h}^H \mathbf{x}(n) = \mathbf{h}^H \mathbf{Z} \mathbf{D}^n \mathbf{a} + \mathbf{h}^H \mathbf{e}. \quad (47)$$

This comprises two terms:

- The filtered voiced speech $\mathbf{h}^H \mathbf{Z} \mathbf{D}^n \mathbf{a}$
- The filtered noise $\mathbf{h}^H \mathbf{e}$

If $\mathbf{h}^H \mathbf{Z} = \mathbf{1}^T$ then $\mathbf{h}^H \mathbf{Z} \mathbf{D}^n \mathbf{a} = \sum_{l=1}^{L} a_l e^{j\omega_0 l n}$ while $E\{|\mathbf{h}^H \mathbf{e}|^2\} = \mathbf{h}^H \mathbf{Q} \mathbf{h}$ is minimized, we have distortionless optimal noise reduction!

Noise Reduction
Filter Design

A distortionless filter should have $\mathbf{h}^H\mathbf{Z} = \mathbf{1}^T$ and should minimize the residual noise, i.e.,

$$\min_{\mathbf{h}} \mathbf{h}^H\widehat{\mathbf{Q}}\mathbf{h} \quad \text{s.t.} \quad \mathbf{Z}^H\mathbf{h} = \mathbf{1} \tag{48}$$

The solution can be shown to be

$$\hat{\mathbf{h}} = \widehat{\mathbf{Q}}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\widehat{\mathbf{Q}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}. \tag{49}$$

with $\widehat{\mathbf{Q}}$ being a particular *noise* covariance matrix estimate.

These filters are adaptive, optimal comb filters! Unlike the normally used Wiener filter, these do not distort the desired signal.

Noise Reduction

Noise Covariance Matrix

We seek to find a filter such that the MSE is minimized:

$$MSE = \frac{1}{G} \sum_{n=M-1}^{N-1} \left| y(n) - \sum_{l=1}^{L} a_l e^{j\omega_0 l n} \right|^2 = \frac{1}{G} \sum_{n=M-1}^{N-1} |\mathbf{h}^H \mathbf{x}(n) - \mathbf{a}^H \mathbf{w}(n)|^2,$$

with $\mathbf{w}(n) = \left[\begin{array}{ccc} e^{j\omega_0 1 n} & \cdots & e^{j\omega_0 L n} \end{array} \right]^T$. Solving for the amplitudes, we get

$$MSE = \mathbf{h}^H \left(\widehat{\mathbf{R}} - \mathbf{G}^H \mathbf{W}^{-1} \mathbf{G} \right) \mathbf{h} \triangleq \mathbf{h}^H \widehat{\mathbf{Q}} \mathbf{h}, \tag{50}$$

where $\mathbf{G} = \frac{1}{G} \sum_{n=M-1}^{N-1} \mathbf{w}(n) \mathbf{x}^H(n)$ and $\mathbf{W} = \frac{1}{G} \sum_{n=M-1}^{N-1} \mathbf{w}(n) \mathbf{w}^H(n)$.

Thus we can estimate $\mathbf{Q}$ as $\widehat{\mathbf{Q}} = \widehat{\mathbf{R}} - \mathbf{G}^H \mathbf{W}^{-1} \mathbf{G}$!

Noise Reduction
Special Cases

Special cases:

- Setting $\mathbf{W} = \mathbf{I}$ yields the usual noise covariance matrix estimate.
- Capon-like filters can be obtained from $\widehat{\mathbf{Q}} = \widehat{\mathbf{R}}$, i.e., $\hat{\mathbf{h}} = \widehat{\mathbf{R}}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\widehat{\mathbf{R}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}$.
- Setting $\widehat{\mathbf{R}} = \sigma^2\mathbf{I}$ yields $\hat{\mathbf{h}} = \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{1}$.
- Noting that $\lim_{M\to\infty} M\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1} = \mathbf{Z}$, we get $\hat{\mathbf{h}} = \frac{1}{M}\mathbf{Z}\mathbf{1}$.
- Binary masking can also be obtained using these principles.

Noise Reduction
Examples

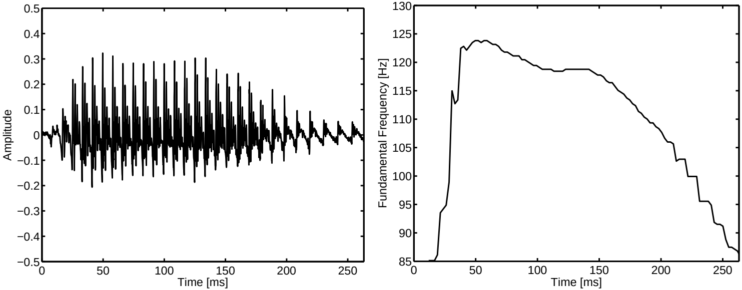


Figure: The original voiced speech signal and the estimated pitch.

Noise Reduction
Examples

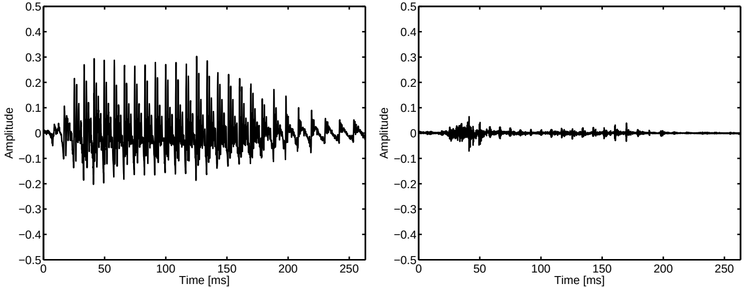


Figure: The extracted signal and the difference between the two signals, i.e., the part of the signal that was not extracted.

Noise Reduction
Examples

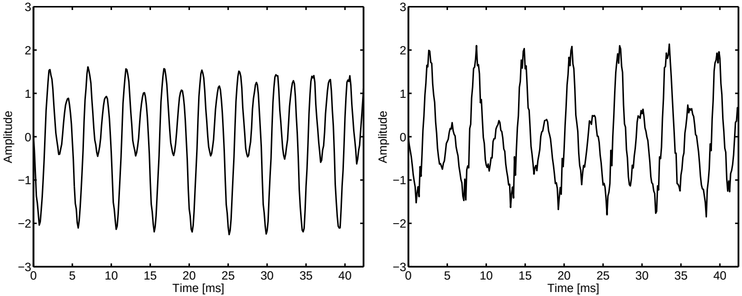


Figure: The voiced speech signal of sources 1 and 2.

Noise Reduction
Examples

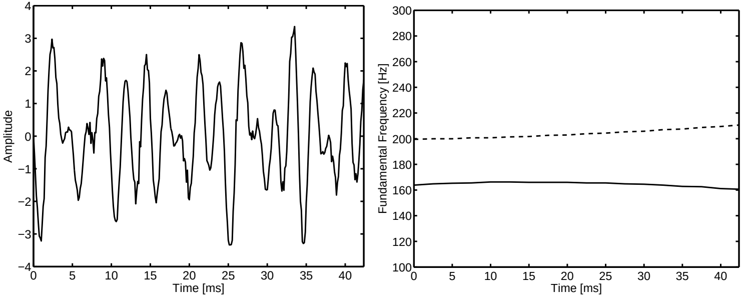


Figure: The mixture of the two signals and the estimated pitch tracks for source 1 (dashed) and 2 (solid).

Noise Reduction
Examples

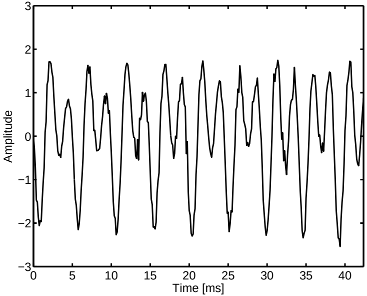


Figure: The estimate of sources 1 and 2 obtained from the mixture.

Non-Stationary Speech

Introduction

- Parametric methods based on the harmonic model have proven to overcoming the problems of correlation-based pitch estimation methods.
- However, as mentioned earlier, there might be concerns the stationarity within segments.
- To investigate whether this is a problem, we will take a closer look at the harmonic chirp model and derive an estimator for determining its parameters.

Non-Stationary Speech
Signal Model

For a segment of a speech signal with $n = n_0, \ldots, n_0 + N - 1$ the new harmonic chirp model is given by

$$x(n) = \sum_{l=1}^{L} A_l e^{j\theta_l(n)} + e(n) \tag{51}$$

where

- L is the number of harmonics (assumed known).
- A_l the lth is the amplitude.
- $\theta_l(n)$ is the instantaneous phase of the lth harmonic.
- $e(n)$ are the stochastic parts of the observed signal.
- n_0 is the start index.

Non-Stationary Speech
Signal Model

The instantaneous phase $\theta_l(\cdot)$ is a continuous function of the continuous variable t. It is given by

$$\theta_l(t) = \int_0^t l\omega_0(\tau)d\tau + \phi_l, \tag{52}$$

where $\omega_0(t)$ is the time-varying pitch and ϕ_l is the phase of the lth harmonic. The instantaneous frequency of the lth harmonic is then

$$\omega_l(t) = \frac{d\theta_l(t)}{dt} = l\omega_0(t). \tag{53}$$

In pitch estimation, it is most often assumed that the pitch is constant, i.e., $\omega_l(t) = l\omega_0$, which results in the harmonic model (HM).

Non-Stationary Speech
Signal Model

If the pitch is slowly and smoothly varying an appropriate model would be $\omega_0(t) = \alpha_0 t + \omega_0$ which yields

$$\theta_l(t) = \frac{1}{2}\alpha_0 l t^2 + \omega_0 l t + \phi_l, \tag{54}$$

where $\alpha_0 l$ is then the chirp rate of the lth harmonic. We term α_0 the fundamental chirp rate.

The resulting model is called the harmonic chirp model (HCM).

We would like to jointly estimate α_0 and ω_0 from $x(n)$.

Non-Stationary Speech
NLS Estimator

Define a vector from the observed signal with $n_0 = -(N-1)/2$ as

$$\mathbf{x} = \begin{bmatrix} x(n_0) & x(n_0+1) & \dots & x(n_0+N-1) \end{bmatrix}. \tag{55}$$

and a matrix as

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}(\omega_0, \alpha_0) & \mathbf{z}(2\omega_0, 2\alpha_0) & \dots & \mathbf{z}(L\omega_0, L\alpha_0) \end{bmatrix}, \tag{56}$$

with columns

$$\mathbf{z}(l\omega_0, l\alpha_0) = \begin{bmatrix} e^{j(\frac{1}{2}\alpha_0 l n_0^2 + \omega_0 l n_0)} & \dots & e^{j(\frac{1}{2}\alpha_0 l (n_0+N-1)^2 + \omega_0 l (n_0+N-1))} \end{bmatrix}^T. \tag{57}$$

For convenience, we introduce $\mathbf{\Pi}_{\omega_0,\alpha_0} = \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H$.

Non-Stationary Speech
NLS Estimator

As before, the nonlinear least squares (NLS) estimator can be used:

$$\{\hat{\alpha}_0, \hat{\omega}_0\} = \arg\min_{\alpha_0,\omega_0} \|\mathbf{x} - \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}\|^2. \tag{58}$$

We solve this iteratively as follows (with i being the iteration index). First obtain an estimate $\hat{\alpha}_0^{(i)}$ from $\hat{\omega}_0^{(i-1)}$ for $i = 1, 2, \ldots$ as

$$\hat{\alpha}_0^{(i)} = \arg\max_{\alpha_0} \left\{\mathbf{x}^H \mathbf{\Pi}_{\hat{\omega}_0^{(i-1)},\alpha_0} \mathbf{x}\right\}, \tag{59}$$

and then update the estimate of the fundamental frequency, ω_0, as

$$\hat{\omega}_0^{(i)} = \arg\max_{\omega_0} \left\{\mathbf{x}^H \mathbf{\Pi}_{\omega_0,\hat{\alpha}_0^{(i)}} \mathbf{x}\right\}. \tag{60}$$

This is then repeated for $i =, 1, 2, \ldots$ until convergence. We initialize with $\alpha_0^{(0)} = 0$.

Non-Stationary Speech
Experiments

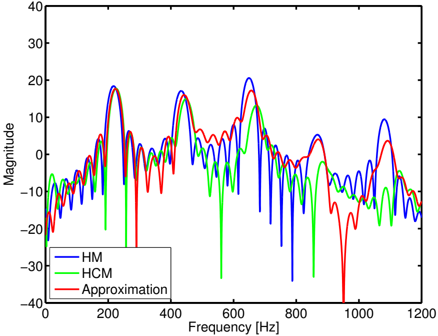


Figure: Spectrum of harmonic model, harmonic chirp model, and an approximation.

Non-Stationary Speech
Experiments

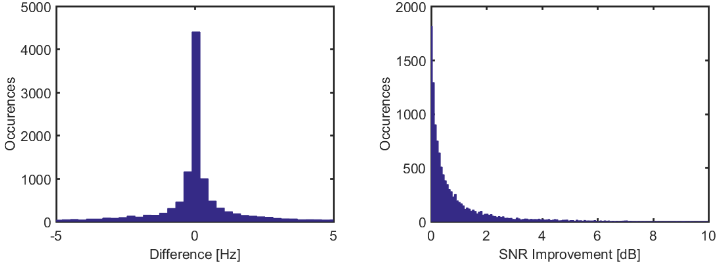


Figure: Histogram of differences in pitch estimates (left) and reconstruction SNRs (right) between HM and HCM.

Section 4

Discussion and Applications

Summary

We have see how

- the problem of finding the parameters of the harmonic model can be analyzed.
- the parameters of the harmonic model can be found in various ways.
- the harmonic model and its estimators can be extended to multiple channels under quite general conditions.
- the harmonic model can be used for designing optimal and distortionless filters that do not require knowledge of noise statistics.
- it is fairly straighforward to take the non-stationary nature of speech into account.

Other Problems

Some problems that can be solved with the approach are:

- Unknown model order: Can be solved with the MAP/BIC principles and angles between subspaces.
- Model selection: Can be solved with the MAP/BIC principles.
- Colored noise: Can be handled with pre-whitening if we know or can find the noise PSD/covariance matrix.
- Missing data: Can be solved with model-based interpolation.
- Segmentation: The MAP optimal segmentation can be found using dynamic programming.
- Detection: Can be solved with GLRT with the harmonic model.

Applications

These ideas are/can be used in many applications, including:

- Hearing aids
- Voice over IP
- Telecommuncation
- Reproduction systems
- Voice analysis
- Intelligence, law enforcement, defense
- Music equipment/software

Some Other Results

- Parametric models can be used for speech/audio compression.
- Feedback cancellation can be improved using a model of the near-end signal.
- Beamforming can be improved with the model-based approach.
- Optimal filters can be designed for the chirp model too.
- Model-based TDOA estimation is better than the state of the art.
- It is possible to take common panning techniques in stereo into account.
- We have recently shown that fast implementations can be found!

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.
- This is because the parametric models lead to mathematically tractable methods.
- 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

I would like to thank the following people:

- The elders: Søren, Andreas, Jacob, Marc, Petre
- The young guns: Jesper, Jesper, Jesper, Daniele, Tobias
- The Ph.D. students: Sam, Sidsel, Vincent, Martin, Knud, Mathew, Charlotte
- My family: Mia, Silke, Freja, Erik