Statistical Parametric Speech Processing
Solving problems with the model-based approach

September 14 2016

Mads Græsbøll Christensen

Audio Analysis Lab, AD:MT
Aalborg University
Denmark
Outline

Introduction
  Motivation
  Harmonic Model

Estimating Parameters
  Parameter Estimation Bounds
  Maximum Likelihood Method
  Subspace Method
  Filtering Method

Some Examples
  Multi-Channel Modeling
  Noise Reduction
  Non-Stationary Speech

Discussion and Applications

References
Motivation

- Parametric speech processing is processing based on parametric models.
- Signal models described in terms of physically meaningful parameters.
- Parametric speech models have been around for many years (e.g., linear prediction in the 70s, sinusoidal model in the 80s).
- Skeptics argue that the models are (always) wrong and that it is not possible to estimate the model parameters well enough under adverse conditions.
- Parametric models can, however, be used for many things and in different ways.
- As an example, we will here take our starting point in the harmonic model.
Motivation

All models are wrong; some models are useful. (G. Box)
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 \( n = 0, \ldots, N - 1 \))

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

(1)

Definitions:

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

The model can also be written as (with $x(n)$ being a snapshot)

$$x(n) = Z(n)a + e(n)$$  \hspace{1cm} (2)

$$= ZD^n a + e(n)$$  \hspace{1cm} (3)

$$= Za(n) + e(n),$$  \hspace{1cm} (4)

with the following definitions:

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

$$z(\omega) = [ 1 \ e^{j\omega} \ \cdots \ e^{j\omega(M-1)} ]^T$$

$$Z = [ z(\omega_0) \ \cdots \ z(\omega_0L) ]$$

$$D = \text{diag}( [ e^{j\omega_0} \ e^{j\omega_0^2} \ \cdots \ e^{j\omega_0L} ])$$

$$a = [ a_1 \ \cdots \ a_L ]^T$$
The covariance matrix of $\mathbf{x}(n)$ is

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

(5)

Written in terms of the harmonic model, we get

$$
\mathbf{R} = \mathbf{Z} \mathbb{E} \left\{ \mathbf{a}(n)\mathbf{a}^H(n) \right\} \mathbf{Z}^H + \mathbb{E} \left\{ \mathbf{e}(n)\mathbf{e}^H(n) \right\}
$$

$$
= \mathbf{ZPZ}^H + \mathbf{Q},
$$

(6)

(7)

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)

$$
\mathbf{P} \approx \text{diag} \left( [ A_1^2 \cdots A_L^2 ] \right).
$$

(8)
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 $\theta_i$ (i.e., the $i$th element of $\theta \in \mathbb{R}^P$) is unbiased if

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

(9)

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

$$\text{var}(\hat{\theta}_i) \geq [I^{-1}(\theta)]_{ii},$$

(10)

where the Fisher Information Matrix (FIM) $I(\theta)$ is given by

$$[I(\theta)]_{ii} = -E \left\{ \frac{\partial^2 \ln p(x; \theta)}{\partial \theta_i \partial \theta_l} \right\},$$

(11)

with $\ln p(x; \theta)$ being the log-likelihood function for $x \in \mathbb{C}^N$. 
The CRLBs can be derived for the harmonic model (for WGN):

\[ \text{var}(\hat{\omega}_0) \geq \frac{6\sigma^2}{N(N^2 - 1) \sum_{l=1}^{L} A_l^2 l^2} \]  \hspace{1cm} (12)

\[ \text{var}(\hat{A}_l) \geq \frac{\sigma^2}{2N} \]  \hspace{1cm} (13)

\[ \text{var}(\hat{\phi}_l) \geq \frac{\sigma^2}{2N} \left( \frac{1}{A_l^2} + \frac{3l^2(N - 1)^2}{\sum_{m=1}^{L} A_m m^2(N^2 - 1)} \right) \]  \hspace{1cm} (14)

These depend on the following quantity:

\[ \text{PSNR} = 10 \log_{10} \frac{\sum_{l=1}^{L} A_l^2 l^2}{\sigma^2} \text{ [dB]} \]  \hspace{1cm} (15)

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

\[ x = s(\theta) + e \]  \hspace{1cm} (16)

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

\[ s_i(\theta + \epsilon) \approx s_i(\theta) + \left( \frac{\partial s_i(\theta)}{\partial \theta} \right)^H \epsilon. \]  \hspace{1cm} (17)

The mean squared error (MSE) is then

\[ E \left\{ (s_i(\theta) - s_i(\theta + \epsilon))^2 \right\} = \left( \frac{\partial s_i(\theta)}{\partial \theta} \right)^H E \left\{ \epsilon \epsilon^H \right\} \left( \frac{\partial s_i(\theta)}{\partial \theta} \right). \]  \hspace{1cm} (18)
Parameter Estimation Bounds

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

$$
\epsilon \sim \mathcal{N}(0, \mathbf{I}^{-1}(\theta)), \quad (19)
$$

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

$$
[I(\theta)]_{nm} = \left( \frac{\partial \mathbf{s}^H(\theta)}{\partial \theta_n} \right) \mathbf{Q}^{-1} \left( \frac{\partial \mathbf{s}(\theta)}{\partial \theta_m} \right). \quad (20)
$$

The MSE can then be seen to be

$$
E \left\{ (s_i(\theta) - s_i(\theta + \epsilon))^2 \right\} = \left( \frac{\partial s_i(\theta)}{\partial \theta} \right)^H \mathbf{I}^{-1}(\theta) \left( \frac{\partial s_i(\theta)}{\partial \theta} \right). \quad (21)
$$
Maximum Likelihood Method

For Gaussian signals, the likelihood function is

\[ p(x(n); \theta) = \frac{1}{\pi^M \det(Q)} e^{- (x(n) - Za(n))^H Q^{-1} (x(n) - Za(n))}. \] (22)

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

\[ p(\{x(n)\}; \theta) = \prod_{n=0}^{G-1} p(x(n); \theta). \] (23)

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

\[ \hat{\theta} = \arg \max \mathcal{L}(\theta). \] (24)
For white Gaussian noise \((Q = \sigma^2 I)\) with \(M = N\) the log-likelihood function is

\[
L(\theta) = -N \ln \pi - N \ln \sigma^2 - \frac{1}{\sigma^2} \|x - Za\|^2_2.
\]  

(25)

The concentrated MLE is given by

\[
\hat{\omega}_0 = \arg \max_{\omega_0} L(\omega_0) = \arg \max_{\omega_0} x^H Z (Z^H Z)^{-1} Z^H x
\]

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

(26)

(27)

This can be computed using an FFT (i.e., using harmonic summation)!
Recall that the model is
\[ x(n) = Za(n) + e(n), \]  
(28)
and that the covariance matrix then is
\[ R = E \{ x(n)x^H(n) \} = ZPZ^H + \sigma^2 I, \]  
(29)
where \( ZPZ^H \) has rank \( L \) and
\[ P = \text{diag} \left( [ A_1^2 \, \cdots \, A_L^2 ] \right). \]
Subspace Method

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

$$\mathbf{G} = \begin{bmatrix} \mathbf{u}_{L+1} & \cdots & \mathbf{u}_M \end{bmatrix}$$  \hspace{1cm} (30)

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

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

$$\hat{\omega}_0 = \arg \min_{\omega_0} \| \mathbf{Z}^H \mathbf{G} \|_F^2 = \arg \min_{\omega_0} \sum_{l=1}^{L} \| \mathbf{z}^H(\omega_0 l) \mathbf{G} \|_2^2.$$  \hspace{1cm} (31)

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

\[
y(n) = \sum_{m=0}^{M-1} h(m)x(n - m) = h^H x(n),
\]

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

The filtered output can be seen to be

\[
h^H x(n) = h^H Z D^\alpha a + h^H e. \tag{33}
\]

If \( h^H Z = 1^T \) with \( 1 = [ 1 \ \cdots \ 1 ]^T \) the voiced speech would pass undistorted and the noise term \( h^H e \) could be minimized!
Filtering Method

We would thus like to design a filter as

$$\min_h h^H R_h \quad \text{s.t.} \quad h^H Z = 1^T.$$  \hspace{1cm} (34)

This has the solution

$$h = R^{-1} Z (Z^H R^{-1} Z)^{-1} 1.$$  \hspace{1cm} (35)

We can use this filter to estimate the pitch as

$$\hat{\omega}_0 = \arg \max_{\omega_0} 1^H (Z^H R^{-1} Z)^{-1} 1.$$  \hspace{1cm} (36)
These methods are more robust to noise than non-parametric methods (YIN stops working below 10 dB, these work for -5 dB).
They are better for low fundamental frequencies too and get better for higher SNR and $N$.
The model order varies and has to be found on a per segment basis.
Fast implementations that make the exact NLS as fast as harmonic summation exist.
Colored noise can be dealt with.
They can be extended to multiple pitches, although not always trivially.
Section 3

Some Examples
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 take a look at 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!
The method operates on snapshots $x_k(n) \in \mathbb{C}^M$ for the $k$th channel.

These are modeled as sums of sinusoids in Gaussian noise $e_k$ with covariance $Q_k$, i.e.,

$$x_k(n) = Z(n)a_k + e_k(n),$$

(37)

with $a_k = [ A_{k,1} e^{j\phi_{k,1}} \cdots A_{k,L} e^{j\phi_{k,L}} ]^T$. Let $\theta_k$ be the parameter vector for the $k$th channel. The likelihood function is then

$$p(x_k(n); \theta_k) = \frac{1}{\pi^M \det(Q_k)} e^{-e_k^H(n)Q_k^{-1}e_k(n)}.$$  

(38)
Multi-Channel Modeling
Signal Model

If the deterministic part is stationary and $e_k(n)$ is i.i.d. over $n$ and independent over $k$, the combined likelihood is

$$p(\{x_k(n)\}; \{\theta_k\}) = \prod_{k=1}^{K} \frac{1}{\pi^{MG} \text{det}(Q_k)^G} e^{-\sum_{n=0}^{G-1} e_k^H(n) Q_k^{-1} e_k(n)}. \quad (39)$$

For simplicity, we assume that the noise is white in each channel but has different $\sigma_k^2$, i.e., $Q_k = \sigma_k^2 I$.

The log-likelihood function then reduces to

$$\ln p(\{x_k(n)\}; \{\theta_k\}) = -GM \sum_{k=1}^{K} \ln (\pi \sigma_k^2) - \sum_{k=1}^{K} \sum_{n=0}^{G-1} \frac{\|e_k(n)\|^2}{\sigma_k^2}. \quad (40)$$
Multi-Channel Modeling

Estimator

The MLE of the amplitudes for channel \( k \) are

\[
\hat{a}_k = \left( \sum_{n=0}^{G-1} Z^H(n)Z(n) \right)^{-1} \sum_{n=0}^{G-1} Z^H(n)x_k(n).
\] (41)

This can be used to form a noise variance estimate as

\[
\hat{\sigma}_k^2 = \frac{1}{GM} \sum_{n=0}^{G-1} \| \hat{e}_k(n) \|^2 = \frac{1}{GM} \sum_{n=0}^{G-1} \| x_k(n) - Z(n)\hat{a}_k \|^2.
\] (42)

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

\[
\ln p(x_k(n); \omega_0) = -M \ln \pi - M \ln \hat{\sigma}_k^2.
\]
For all $n$ and $k$, this yields

$$\ln p(\{x_k(n)\}; \omega_0) = -GMK \ln \pi - GM \sum_{k=1}^{K} \ln \hat{\sigma}_k^2. \tag{43}$$

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

$$\hat{\omega}_0 = \arg \min_{\omega_0} \sum_{k=1}^{K} \ln \hat{\sigma}_k^2. \tag{44}$$

This estimator can then be approximated as

$$\hat{\omega}_0 = \arg \min_{\omega_0} \sum_{k=1}^{K} \ln \left( \|x_k\|^2 - \frac{1}{N} \|Z^H x_k\|^2 \right), \tag{45}$$

where $x_k = x_k(0)$ 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).
As we have seen, it was fairly straightforward to extend the MLE to multiple channels.

It works well and under very general conditions.

It is fast too.

Easy to build in more specific knowledge, like array structure, nearfield, TDOAs, binaural setups.

The multi-channel model contains the usual broadband model as a special case with $\omega_0 = 2\pi/N$.

Can be used for pitch/DOA estimation and model-based beamforming.
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.
As we saw earlier, we get the following model when a filter $h$ is applied to the observed signal $x(n)$:

$$\hat{s}(n) = h^H x(n) = h^H ZD^n a + h^H e.$$  \hfill (46)

This comprises two terms:

- The filtered voiced speech $h^H ZD^n a$
- The filtered noise $h^H e$

If $h^H Z = 1^T$ then $h^H ZD^n a = \sum_{i=1}^{L} a_i e^{i\omega_0 ln}$ while $E\{|h^H e|^2\} = h^H Q h$ is minimized, we have distortionless optimal noise reduction!
Noise Reduction
Filter Design

A distortionless filter should have $h^HZ = 1^T$ and should minimize the residual noise, i.e.,

$$\min_h h^H\hat{Q}h \quad \text{s.t.} \quad Z^Hh = 1 \quad (47)$$

The solution can be shown to be

$$\hat{h} = \hat{Q}^{-1}Z\left(Z^H\hat{Q}^{-1}Z\right)^{-1}1. \quad (48)$$

with $\hat{Q}$ 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:

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

with $w(n) = [e^{j\omega_0 1 n} \ldots e^{j\omega_0 L n}]^T$. Solving for the amplitudes, we get

$$MSE = h^H \left( \hat{R} - G^H W^{-1} G \right) h \triangleq h^H \hat{Q} h,$$  \hspace{1cm} (49)

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

Thus we can estimate $Q$ as $\hat{Q} = \hat{R} - G^H W^{-1} G$!
Special cases:

- Setting $W = I$ yields the usual noise covariance matrix estimate.
- Capon-like filters can be obtained from $\hat{Q} = \hat{R}$, i.e.,
  $$\hat{h} = \hat{R}^{-1}Z \left(Z^H\hat{R}^{-1}Z\right)^{-1}1.$$  
- Setting $\hat{R} = \sigma^2 I$ yields $\hat{h} = Z \left(Z^HZ\right)^{-1}1$.
- Noting that $\lim_{M \to \infty}MZ \left(Z^HZ\right)^{-1} = Z$, we get $\hat{h} = \frac{1}{M}Z1$.
- Binary masking can also be obtained using these principles.
Noise Reduction
Examples

Figure: The original voiced speech signal and the estimated pitch.
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.
We have seen how the harmonic model can be used for designing filters for noise reduction.

The filters are distortionless, i.e., they let the signal of interest pass undistorted.

Meanwhile, the noise is attenuated as much as possible.

The resulting filters are thus optimal in terms of output SNR and minimum distortion!

They do not require a priori knowledge of noise statistics.

They can be generalized to multiple channels.
Parametric methods based on the harmonic model have proven to overcome the problems of correlation-based methods.

However, as mentioned earlier, there might be concerns about 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.
For a segment of a speech signal with \( n = n_0, \ldots, n_0 + N - 1 \) the new harmonic chirp model is given by

\[
x(n) = \sum_{l=1}^{L} A_l e^{i\theta_l(n)} + e(n)
\]

(50)

where

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

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

(51)

where $\omega_0(t)$ is the time-varying pitch and $\phi_l$ is the phase of the $l$th harmonic. In the harmonic model (HM) we have that $\omega_l(t) = l\omega_0$.

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

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

(52)

where $\alpha_0$ is the fundamental chirp rate.

The resulting model is called the harmonic chirp model (HCM).
Define a vector with \( n_0 = -(N - 1)/2 \) as

\[
x = \begin{bmatrix} x(n_0) & x(n_0 + 1) & \ldots & x(n_0 + N - 1) \end{bmatrix}.
\]  
(53)

and a matrix as

\[
Z = \begin{bmatrix} z(\omega_0, \alpha_0) & z(2\omega_0, 2\alpha_0) & \ldots & z(L\omega_0, L\alpha_0) \end{bmatrix},
\]  
(54)

with columns

\[
z(l\omega_0, l\alpha_0) = \begin{bmatrix} e^{i\frac{1}{2} \alpha_0 l n_0^2 + \omega_0 l n_0} & \ldots & e^{i\frac{1}{2} \alpha_0 l (n_0 + N - 1)^2 + \omega_0 l (n_0 + N - 1)} \end{bmatrix}^T.
\]

For convenience, we introduce \( \Pi_{\omega_0, \alpha_0} = Z (Z^H Z)^{-1} Z^H \).
Non-Stationary Speech
NLS Estimator

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

\[
\{ \hat{\alpha}_0, \hat{\omega}_0 \} = \arg \min_{\alpha_0, \omega_0} \| x - Z (Z^H Z)^{-1} Z^H x \|^2.
\] (55)

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

\[
\hat{\alpha}_0^{(i)} = \arg \max_{\alpha_0} \left\{ x^H \Pi_{\hat{\omega}_0^{(i-1)}, \alpha_0} x \right\},
\] (56)

and then update the estimate of the fundamental frequency, \( \omega_0 \), as

\[
\hat{\omega}_0^{(i)} = \arg \max_{\omega_0} \left\{ x^H \Pi_{\omega_0, \hat{\alpha}_0^{(i)}} x \right\}.
\] (57)

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 for 30 sentences.
As we have seen, it is quite easy to account for non-stationarity.

Although the differences in pitch estimates are small, they may matter.

There exists fast implementations for the exact NLS for the harmonic chirp model too!

It is also possible to use HCM with the distortionless filters, meaning we can design filters that account for the non-stationarity of speech.
Section 4

Discussion and Applications
Summary

We have seen 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 straightforward to take the non-stationary nature of speech into account.
Applications

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

- Hearing aids
- Voice over IP
- Telecommunication
- Reproduction systems
- Voice analysis
- Intelligence, law enforcement, defense
- Music equipment/software
Some Other Results

- Parametric models can be used for speech/audio compression.
- Model-based interpolation/extrapolation can be used for packet losses/corrupt data.
- Feedback cancellation can be improved using a model of the near-end signal.
- Beamforming can be improved with the model-based approach.
- Jointly optimal segmentation and parameter estimates can be found with dynamic programming.
- Optimal filters can be designed for the chirp model too.
- 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.
- 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.


S. M. Nørholm, J. R. Jensen, and M. G. Christensen, “Enhancement and noise statistics estimation for non-stationary voiced speech,”
References II


References V


M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, “Joint fundamental frequency and order estimation using optimal
References VI


References VII


References VIII


References X


References XII


