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

August 20, 2017

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

Audio Analysis Lab, AD:MT
Aalborg University
Denmark
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement

Summary and Conclusion
Introduction

Motivation
Who are we?
Parametric speech processing

Statistical Speech Models

Model-based Pitch Estimation of Speech

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.
Motivation

How is the model wrong?

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

How is a (parametric) model **useful**?

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

WGN

Vocal tract

$e(n) \rightarrow \text{AR}(p) \rightarrow x(n)$

Pulse train

Example

Signal model

$x = Xa + e$ (1)

- Estimate $a$ by minimising the 2-norm (Gaussian noise)
- Estimate $a$ by minimising the 1-norm (Laplacian noise) (Giacobello 2012)
Motivation

Example
Motivation

Example
Introduction
   Motivation
   Who are we?
   Parametric speech processing

Statistical Speech Models

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement

Summary and Conclusion
Who are we?

The Audio Analysis Lab

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

The Audio Analysis Lab

People

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

Current Collaborators

- GN ReSound
- Bang & Olufsen
- Brue & Kjær
- Parkinson’s Voice Initiative
- Richard Heusdens, TU Delft (former Guest Professor)
- Andreas Jakobsson, Lund University
- Jingdong Chen, Northwestern Polytechnical University
Who are we?

The Audio Analysis Lab

Current Speech-related Research Projects

- Signal processing for detecting the cocktail party problem and methods for enhancing listening in noisy situations (COCKTAIL), 2014–2018
- Signal Processing for Diagnosis of Parkinson’s Disease from Noisy Speech, 2015–2019

Website: http://audio.create.aau.dk
Youtube: http://tinyurl.com/yd8mo55z
Outline

Introduction
  Motivation
  Who are we?
  Parametric speech processing

Statistical Speech Models

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement

Summary and Conclusion
Introduction

Parametric speech processing

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

Methodology

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

Why parametric methods?

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

Some interesting questions:

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

Observations:

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

Introduction

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

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
Outline

Introduction

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

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
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 explore how we can model speech and how we can manipulate the models.
Harmonic Model

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

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

(2)

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

The model can be written in matrix-vector notation as

\[ x(n) = Za + e(n) \]  
\[ = s(\theta) + e(n) \]

with the following definitions:

\[ x(n) = \begin{bmatrix} x(n) & \cdots & x(n+M-1) \end{bmatrix}^T \]
\[ z(\omega) = \begin{bmatrix} 1 & e^{j\omega} & \cdots & e^{j(M-1)\omega} \end{bmatrix}^T \]
\[ Z = \begin{bmatrix} z(\omega_0) & \cdots & z(\omega_0L) \end{bmatrix} \]
\[ a = \begin{bmatrix} a_1 & \cdots & a_L \end{bmatrix}^T \]

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

The model can be written in different ways:

\[ x(n) = Z(n)a + e(n) \]  \hspace{1cm} (5)
\[ = ZD(n)a + e(n) \]  \hspace{1cm} (6)
\[ = Za(n) + e(n), \]  \hspace{1cm} (7)

where \( D(n) = D^n \) with \( D = \text{diag}([e^{j\omega_0} \ e^{j\omega_0^2} \ ... \ e^{j\omega_0^L}] ) \). Notice that \( D(n)a = \sum_{l=1}^{L} a_l e^{j\omega_0 ln} \).

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

\[ x = Za + e, \] \hspace{1cm} (8)

which is a special case of the model above with \( M = N \).
Harmonic Model

The covariance matrix of $x(n)$ is

$$R = E \{ x(n)x^H(n) \} . \quad (9)$$

Written in terms of the harmonic model, we get

$$R = ZE \{ a(n)a^H(n) \} Z^H + E \{ e(n)e^H(n) \}$$

$$= ZPZ^H + Q, \quad (11)$$

which is called the covariance matrix model.

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

$$P \approx \text{diag} \left( \begin{bmatrix} A_1^2 & \cdots & A_L^2 \end{bmatrix} \right) . \quad (12)$$
Filtering

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

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

(13)

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

$$E\{|y(n)|^2\} = h^H R h.$$  

(14)

Recall that the signal model was

$$x = ZD(n)a + e.$$  

(15)

The filtered output can thus be seen to be

$$h^H x(n) = h^H Z D(n)a + h^H e.$$  

(16)
Filtering

The filtered observed signal $x$ could be written as

$$h^H x(n) = h^H ZD(n)a + h^H e. \quad (17)$$

This comprises two terms:

1. The speech passed through the filter $h^H ZD(n)a$.
2. The residual noise $h^H e$.

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

$$E \{ |y(n)|^2 \} = h^H R h \quad (18)$$

$$= h^H ZPZ^H h + h^H Q h, \quad (19)$$

where $h^H ZPZ^H h$ is the power of the filtered speech and $h^H Q h$ is the residual noise.
Recall that the model is

\[ x(n) = Z a(n) + e(n), \]  

and that the covariance matrix then is given by

\[ R = \mathbb{E} \{ x(n) x^H(n) \} = Z P Z^H + \sigma^2 I, \]

where \( Z P Z^H \) has rank \( L \) and

\[ P = \text{diag} \left( [ A_1^2 \, \cdots \, A_L^2 ] \right). \]
Let
\[ R = U \Lambda U^H \]  \hspace{1cm} (22)
be the eigenvalue decomposition (EVD) of the covariance matrix. \( U \) contains the \( M \) orthonormal eigenvectors of \( R \), i.e.,
\[ U = [ u_1 \, \cdots \, u_M ] , \]  \hspace{1cm} (23)
and \( \Lambda \) is a diagonal matrix containing the corresponding (sorted) positive eigenvalues, \( \lambda_k \). Let \( S \) be formed as
\[ S = [ u_1 \, \cdots \, u_V ] . \]  \hspace{1cm} (24)

The subspace that is spanned by the columns of \( S \) we denote \( R(S) \),
Subspace Model

Similarly, let $G$ be formed as

$$G = \begin{bmatrix} u_{V+1} & \cdots & u_M \end{bmatrix},$$  \hspace{1cm} (25)

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

$$U (\Lambda - \sigma^2 I) U^H = \sum_{k=1}^K Z_k P_k Z_k^H.$$  \hspace{1cm} (26)

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

$$Z^H G = 0$$  \hspace{1cm} (27)

and 2) the matrices $S$ and $Z$ span the same space, i.e.,

$$\mathcal{R}(S) = \mathcal{R}(Z).$$  \hspace{1cm} (28)
Harmonic Model

What’s wrong with this model?

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

Can this be dealt with? Does it matter?
On The Complex Signal Model

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

- Simpler math
- Faster algorithms

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

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

Introduction

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

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
Likelihood Function

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

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

then the likelihood function is given by

$$p(x(n); \theta) = \frac{1}{\sqrt{\pi^M \det(Q)}} e^{-(x(n) - za(n))^H Q^{-1}(x(n) - za(n))} \quad (30)$$

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

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

In the above, \( Q \) could represent the covariance of unvoiced speech, noise or both combined.
The log-likelihood function is

\[ L(\theta) = \ln p(\{x(n)\}; \theta) \]
\[ = \sum_{n=0}^{G} \ln p(x(n); \theta). \]

The maximum likelihood estimator (MLE) is then given by

\[ \hat{\theta} = \arg \max_\theta L(\theta) \]
\[ = \arg \max_\theta \sum_{n=0}^{G} \ln p(x(n); \theta). \]

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 \((Q = \sigma^2 I)\) with \(M = N\) the log-likelihood function is

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

where \(\theta = [\omega_0 \ A_1 \ \ldots \ A_L \ \phi_1 \ \ldots \ \phi_L]\). The concentrated MLE is given by (Quinn 1991)

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

This means that we must find the \(\omega_0\) that results in the largest projection energy!
Outline

Introduction

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

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
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

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

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

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

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

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

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

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

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

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

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

These depend on the following quantity:

\[
\text{PSNR} = 10 \log_{10} \frac{\sum_{l=1}^{L} A_l^2 l^2}{\sigma^2} \text{ [dB]}.
\]
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., $\omega_0$).
- 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 \textit{exact} CRLBs, where no asymptotic approximations are used!

An estimator attaining the bound is said to be \textit{efficient}. A more fundamental property is \textit{consistency}. 
Amplitude Estimation

Figure: CRLB as a function of $\omega_0$ for different cases.
Outline

Introduction

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

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
General Multi-Channel Model

Define \( \mathbf{x}_k(n) \in \mathbb{C}^M \) as the snapshot for the \( k \)th channel.

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

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

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

Interpretation:

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

Let $\theta_k$ be the parameter vector for the $k$th channel. The likelihood function is then

$$p(x_k(n); \theta_k) = \frac{1}{\pi^M \det(Q_k)} e^{-e_k^H(n)Q_k^{-1}e_k(n)}. \quad (46)$$

If the deterministic part is stationary and $e_k(n)$ is i.i.d. over $n$, we get

$$p(\{x_k(n)\}; \theta_k) = \frac{1}{\pi^{MG} \det(Q_k)^G} e^{-\sum_{n=0}^{G-1} e_k^H(n)Q_k^{-1}e_k(n)}. \quad (47)$$

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

$$p(\{x_k(n)\}; \{\theta_k\}) = \prod_{k=1}^{K} \frac{1}{\pi^{MG} \det(Q_k)^G} e^{-\sum_{n=0}^{G-1} e_k^H(n)Q_k^{-1}e_k(n)}. \quad (48)$$
General Multi-Channel Model

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

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

The model ignores noise correlation across channels and array geometry.
Linear Array

For a linear array and sources in the farfield:

\[ \theta \]

\[ d \sin \theta \]

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

$$s_k(n) = s(n - \Delta_k)$$

(49)

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

(50)

Recall that $s(n)$ can be written as $s(n) = ZD(n)a$ and hence

$$s_k \left( n - \frac{d \sin \theta}{c} f_s(k - 1) \right) = ZD \left( n - \frac{d \sin \theta}{c} f_s(k - 1) \right) a.$$  

(51)

As we can see, it is easy to account for fractional delays in the parametric model. Other geometries can easily be incorporated too.
Recall that the matrix $D(n)$ is given by

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

and thus $D(n - \Delta)$ is

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

We can modify the model to account for reverberation.

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

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

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

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

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

Introduction

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

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
Unvoiced Speech

So far, we modeled the observed signal as

\[ x(n) = s(n) + e(n), \]  \hspace{1cm} (56)

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

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

Modified model:

\[ x(n) = s(n) + u(n) + w(n). \]  \hspace{1cm} (57)

\text{voiced} \hspace{1cm} \text{unvoiced} \hspace{1cm} \text{noise}
Unvoiced Speech

What’s a good model for unvoiced speech then?

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

\[ u(n) = \sum_{i=1}^{l} \gamma_i u(n - i) + \eta(n). \] (58)

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

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

In speech applications, the background noise is rarely white.

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

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

Let us examine the following signal model:

\[ x(n) = s(n) + e(n). \]  \( \text{(59)} \)

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

We can transform the observed signal by a matrix \( A \) as

\[ A^H x(n) = A^H s(n) + A^H e(n). \]  \( \text{(60)} \)

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

From the above, we can deduce that \( A \) must be the Cholesky factor of \( Q^{-1} \), i.e., \( AA^H = Q^{-1} \), since \( A^H QA = I \).
Colored Noise

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

Instead, consider the signal model

$$x(n) = s(n) + e(n). \quad (61)$$

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

$$h(n) \ast x(n) = h(n) \ast s(n) + v(n), \quad (62)$$

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

54 / 287
Colored Noise

Since

\[ s(n) = \sum_{l=1}^{L} a_l e^{j\omega_0 ln} \]  \hspace{1cm} (63)

we have that

\[ h(n) * s(n) = h(n) * \sum_{l=1}^{L} a_l e^{j\omega_0 ln} \approx \sum_{l=1}^{L} \tilde{a}_l e^{j\omega_0 ln}. \]  \hspace{1cm} (64)

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

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

How to estimate the noise covariance matrix then?

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

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

For a segment of a speech signal it is given by

$$x(n) = \sum_{l=1}^{L} A_l e^{i\theta_l(n)} + e(n)$$  \hspace{1cm} (65)

where

- $\theta_l(n)$ is the instantaneous phase of the $l$th harmonic.
- everything is as before.
Non-Stationary Speech

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

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

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

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

where $\alpha_0$ is the fundamental chirp rate.

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

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

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

(68)

and a matrix as

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

(69)

with columns

$$
\mathbf{z}(l\omega_0, l\alpha_0) = \begin{bmatrix} e^{j(\frac{1}{2} \alpha_0 ln_0^2 + \omega_0 ln_0)} & \ldots & e^{j(\frac{1}{2} \alpha_0 l(n_0+N-1)^2 + \omega_0 l(n_0+N-1))} \end{bmatrix}^T.
$$

The model can now be written as before:

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

(70)

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

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

Introduction

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

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
Amplitude Estimation

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

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

This can be done in a number of ways including (Stoica 2000):
- Least-squares based estimators, and
- Capon- and APES-based estimators
- Combined using WLS.
Amplitude Estimation

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

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

where

(i) $L$ as well as $\{\psi_l\}_{l=1}^{L}$ are assumed known.

(ii) $\psi_k \neq \psi_l$ for $k \neq l$.

(iii) $e(n)$ denotes a zero mean, complex-valued, and assumed stationary (and possibly colored) additive noise.

How should one proceed to estimate $\{a_l\}_{l=1}^{L}$?
Least-Squares Amplitude Estimation

Form

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

or, using a vector-matrix notation,

\[ x = Za + e. \] (72)

Then, the LS estimator is found as

\[ \hat{a} = (Z^HZ)^{-1}Z^Hx, \] (73)

which is an efficient estimator for all \( N \geq L \) for white Gaussian noise.
Amplitude Estimation

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

$$\text{CRLB}(\hat{a}) = (Z^H Q^{-1} Z)^{-1},$$  \hspace{1cm} (74)

where $Q = \mathbb{E}\{ee^H\}$, which for an additive unit variance white noise implies that $Q = I$.

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

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

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

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

$$x(n) = \begin{bmatrix} x(n) & \ldots & x(n + M - 1) \end{bmatrix}^T$$

$$= \begin{bmatrix} 1 & \ldots & 1 \\ e^{j\psi_1} & \ldots & e^{j\psi_L} \\ \vdots & \ddots & \vdots \\ e^{j\psi_1(M-1)} & \ldots & e^{j\psi_L(M-1)} \end{bmatrix} \begin{bmatrix} a_1 e^{j\psi_1 n} \\ \vdots \\ a_L e^{j\psi_L n} \end{bmatrix} + \begin{bmatrix} e(n) \\ \vdots \\ e(n + M - 1) \end{bmatrix}$$

$$= Z(n)a + e(n), \quad (76)$$

where

$$Z(n) = Z \begin{bmatrix} e^{j\psi_1 n} \\ \vdots \\ e^{j\psi_L n} \end{bmatrix} = ZD(n). \quad (77)$$
Amplitude Estimation

The squared amplitude may be estimated by applying a filter \( h_l \) as

\[
\hat{A}_l^2 = E \{ |h_l^H x(n)|^2 \} = h_l^H E \{ x(n)x(n)^H \} h_l = h_l^H R h_l, \tag{78}
\]

where the filter, \( h_l \), is given by

\[
h_l = \arg \min_{h_l} h_l^H R h_l \quad \text{s.t.} \quad h_l^H z(\psi_l) = 1 \tag{79}
\]

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

with

\[
z(\psi_l) = [1 \ e^{j\psi_l} \ ... \ e^{j\psi_l(M-1)}]^T \tag{81}
\]

This is the classical Capon amplitude (CCA) estimator

\[
\hat{A}_l = \sqrt{h_l^H R h_l} = \left( z^H(\psi_l) R^{-1} z(\psi_l) \right)^{-1/2} \tag{82}
\]
Amplitude Estimation

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

$$h_l^H Z = \begin{bmatrix} 0 & ... & 0 & 1 & 0 & ... & 0 \end{bmatrix} = b_l, \quad (83)$$

which means that

$$h_l^H x(n) = h_l^H [ZD(n)a + e(n)] = a_l e^{i\psi_l n} + h_l^H e(n). \quad (84)$$

This constraint yields the filter

$$h_l = R^{-1} Z (Z^H R^{-1} Z)^{-1} b_l, \quad (86)$$

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

$$\hat{A}_l = \sqrt{h_l^H R h_l} = \sqrt{b_l^T (Z^H R^{-1} Z)^{-1} b_l}. \quad (87)$$
Amplitude Estimation

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

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

where $\hat{Q}$ denotes an estimate of the noise covariance matrix. For sufficiently large $N$ and $M$, one may approximate $\hat{Q} \approx \hat{R}$, where

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

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

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

$$x(n) = Z(n)a + e(n) = \sum_{k=1}^{L} \left[ a_k z(\psi_k) \right] e^{j\psi_k n} + e(n)$$  \hspace{1cm} (90)

suggesting the unstructured LS estimate of $\beta_k$

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

and the covariance matrix estimate

$$\hat{Q} = \hat{R} - \sum_{k=1}^{L} \hat{\beta}_k \hat{\beta}_k^H$$ \hspace{1cm} (92)

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

Finally, one may form a matched filterbank (MAFI) estimator using the matrix filter \( H = [ h_1 \ldots h_L ] \), and express the design criteria as

\[
H = \min_H \text{Tr} \left\{ H^H R H \right\} \quad \text{subject to} \quad H^H Z = I \tag{93}
\]

\[
= R^{-1} Z (Z^H R^{-1} Z)^{-1} \tag{94}
\]

Then,

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

with the \( l \)th index being

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

from which we get the MAFI amplitude estimate as

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

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

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

Introduction

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

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
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

$$p(\mathcal{M}_m | \mathbf{x}) = \frac{p(\mathbf{x} | \mathcal{M}_m)p(\mathcal{M}_m)}{p(\mathbf{x})}.$$  \hspace{1cm} (98)

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

$$\hat{\mathcal{M}}_k = \arg \max_{\mathcal{M}_m, m \in \mathbb{Z}_q} p(\mathcal{M}_m | \mathbf{x}) = \arg \max_{\mathcal{M}_m, m \in \mathbb{Z}_q} \frac{p(\mathbf{x} | \mathcal{M}_m)p(\mathcal{M}_m)}{p(\mathbf{x})}. \hspace{1cm} (99)$$

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

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

\[ p(\mathcal{M}) = \frac{1}{q} \]  

(100)

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

\[ \hat{\mathcal{M}} = \arg \max_{\mathcal{M}_m, m \in \mathbb{Z}_q} p(x|\mathcal{M}_m), \]  

(101)

which is the likelihood function.

The models also depend on \( \theta \), so those have to be integrated out, i.e.,

\[ p(x|\mathcal{M}_m) = \int_{\Theta} p(x|\theta, \mathcal{M}_m) p(\theta|\mathcal{M}_m) d\theta. \]  

(102)
Using Laplace integration, we can write (Djuric 1998)

\[
\int_{\Theta} p(x|\theta, M_m) p(\theta|M_m) d\theta = \pi^{D/2} \text{det} (\hat{H})^{-1/2} p(x|\hat{\theta}, M_m) p(\hat{\theta}|M_m),
\]

where \(D\) is the number of parameters and

\[
\hat{H} = - \frac{\partial^2 \ln p(x|\theta, M_m)}{\partial \theta \partial \theta^T} \bigg|_{\theta=\hat{\theta}}
\]

is the Hessian of the log-likelihood function evaluated at \(\hat{\theta}\) (i.e., the observed information matrix).
Taking the logarithm and ignoring constant terms, we get

\[
\hat{M} = \arg \min_{M_m, m \in \mathbb{Z}_q} \left\{ -\ln p(x | \hat{\theta}, M_m) + \frac{1}{2} \ln \det (\hat{H}) \right\},
\]

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

Note that \( \hat{H} \) is the Fischer information matrix evaluated in \( \hat{\theta} \).

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

\[
\ln \det (\hat{H}) = \ln \det (K^{-2}) + \ln \det (K\hat{H}K).
\]
Posterior Probabilities

For the harmonic model, we introduce

\[ K = \begin{bmatrix} N^{-3/2} & 0 \\ 0 & N^{-1/2}I \end{bmatrix} \] (107)

where \( I \) is an \( 2L_k \times 2L_k \) identity matrix. From this we obtain

\[
\ln \det \left( \hat{H} \right) = \ln \det \left( K^{-2} \right) + \ln \det \left( K\hat{HK} \right)
\]

\[ = 3 \ln N + 2L \ln N + O(1). \] (109)

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

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

Model:

\[ x = Za + e \]  \hspace{1cm} (110)

Hypotheses:

\[ H_0 : a = 0 \]  \hspace{1cm} (111)

\[ H_1 : a \neq 0 \]  \hspace{1cm} (112)

Test statistic:

\[
T(x) = \frac{N - L}{L} \frac{x^H Z (Z^H Z)^{-1} Z^H x}{x^H \left( I - Z (Z^H Z)^{-1} Z^H \right) x}
\]  \hspace{1cm} (113)
The detection rule is then to choose $\mathcal{H}_1$ when
\[ T(x) > \gamma' \] (114)
and $\mathcal{H}_0$ otherwise. The threshold, $\gamma'$, is then chosen according to a desired false alarm (FA) rate as
\[ P_{FA} = Q_{F_L,N-L}(\gamma') \] (115)
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.
Optimal Segmentation

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

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

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

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

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

Costs (white Gaussian noise):

- **Chirp**: \( N \ln \pi + N \ln \hat{\sigma}^2 + N + \frac{3}{2} \ln N + \frac{5}{2} \ln N + L \ln N \)
- **Harmonic**: \( N \ln \pi + N \ln \hat{\sigma}^2 + N + \frac{3}{2} \ln N + L \ln N \)
- **Noise**: \( N \ln \pi + N \ln \hat{\sigma}^2 + N \)

where \( \hat{\sigma}^2 \) is the noise variance estimate for the particular model. The voiced/unvoiced detection can also be done by use of the generalised likelihood ratio test (GLRT).
Optimal Segmentation

\[ J_1 = J_{11} \]
\[ J_2 = J_{11} + J_{22} \]
\[ m = 1 \]

\[ J_1 = J_{11} \]
\[ J_2 = J_{11} + J_{22} \]
\[ m = 2 \]

\[ J_1 = J_{11} \]
\[ J_2 = J_{11} + J_{23} \]
\[ J_3 = J_{12} + J_{33} \]
\[ J_4 = J_{11} + J_{22} + J_{33} \]
\[ m = 3 \]
Conclusion

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

Introduction

Statistical Speech Models

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

Model-based Array Processing and Enhancement

Summary and Conclusion
Periodic signals

A periodic signal repeats itself after some period $\tau$ or, equivalently with some frequency $\omega_0$.

\[ x(n) = x(n - \tau) = x(n - 2\pi/\omega_0) \] (116)
Periodic Signals

Some examples of periodic signals and applications:

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

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech
Correlation-based Methods
The Autocorrelation Method
The Comb Filtering Method
The Least Squares Method
Comparison of Methods
Non-stationary Pitch Estimation
Multi-channel Pitch Estimation
Summary

Model-based Array Processing and Enhancement
The Autocorrelation Method

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

$$e(n) = x(n) - x(n - \tau) = 0$$  \hspace{1cm} (117)

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

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

(118)

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

\[ G(a, \tau) = \sigma_x^2 + a^2 \sigma_x^2 - 2ar_x(\tau) \]  

(119)

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

\[ \hat{\tau} = \arg \max_{\tau \in [\tau_{\text{MIN}}, \tau_{\text{MAX}}]} r_x(\tau) \]  

(120)
The Autocorrelation Method

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

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

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

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

\[ \tau = 45 \]

\[ x(n) \]

\[ x(n-\tau) \]

\[ x(n) * x(n-\tau) \]

\[ E[x(n) * x(n-\tau)] \]
The Autocorrelation Method

\[ \tau = 169 \]

- \( x(n) \)
- \( x(n-\tau) \)
- \( x(n) \times x(n-\tau) \)
The Autocorrelation Method

\[ E[x(n)x(n-\tau)] \]

\[ E[x(n)x(n-f_s/f)] \]
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

Correlation-based Methods
  The Autocorrelation Method
  The Comb Filtering Method

The Least Squares Method

Comparison of Methods

Non-stationary Pitch Estimation

Multi-channel Pitch Estimation

Summary

Model-based Array Processing and Enhancement
The Comb Filtering Method

In the **comb filtering method**, we consider the objective

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

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

$$e(n) = x(n) - ax(n - \tau)$$  \hspace{1cm} (123)
Given $\tau$, the optimal value for $a$ is

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

(124)

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

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

(125)
The Comb Filtering Method
The Comb Filtering Method
The Comb Filtering Method
The Comb Filtering Method
The Comb Filtering Method
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech
  Correlation-based Methods
    The Least Squares Method
      The Harmonic Summation Method
      The Nonlinear Least Squares Estimator
  Comparison of Methods
  Non-stationary Pitch Estimation
  Multi-channel Pitch Estimation
  Summary

Model-based Array Processing and Enhancement
Harmonic Model

\[ x(n) = h_1(n) + h_2(n) + h_3(n) + e(n) \]
Harmonic Model

\[ x(n) = h_1(n) + h_2(n) + h_3(n) + e(n) \]
Harmonic Model

\[ x(n) = h_1(n) + h_2(n) + h_3(n) + e(n) \]
Harmonic Model

Mathematical Model

The signal model for any periodic signal is

\[ s(n) = \sum_{l=1}^{L} h_l(n) = \sum_{l=1}^{L} A_l \cos(\omega_0 ln + \phi_l) \quad (126) \]

where

- \( A_l \) real amplitude of the \( l \)th harmonic
- \( \phi_l \) phase of the \( l \)th harmonic
- \( \omega_0 \) fundamental frequency in radians/sample
- \( L \) the number of harmonics/model order
Method of Least Squares

The method of least-squares

- The vector $\theta$ contains the **model parameters**
- The signal $s(n, \theta)$ is produced by the **signal model**
- The signal $x(n)$ is the **observed data**
- The error consists of **noise** and **model inaccuracies**
From the figure (on the previous slide), we have that

\[ e(n) = x(n) - s(n, \theta) , \quad n = 0, 1, \ldots, N - 1 \quad (127) \]

where \( s(n, \theta) \) is a harmonic model given by

\[ s(n, \theta) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l) \quad (128) \]

\[ \theta = [A_1 \quad \cdots \quad A_L \quad \phi_1 \quad \cdots \quad \phi_L \quad \omega_0]^T \quad (129) \]
Method of Least Squares

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

\[ \hat{\theta} = \arg\min_{\theta} J(\theta) \]  

(130)

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

\[ J(\theta) = \sum_{n=0}^{N-1} |e(n)|^2 \]  

(131)

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

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

Correlation-based Methods

The Least Squares Method

The Harmonic Summation Method

The Nonlinear Least Squares Estimator

Comparison of Methods

Non-stationary Pitch Estimation

Multi-channel Pitch Estimation

Summary

Model-based Array Processing and Enhancement
From Parseval’s theorem, we have that

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

where

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

\[
\alpha_l = A_l \exp(j\phi_l)/2 \tag{134}
\]
Given $\omega_0$, the optimal value for $\alpha_l$ is

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

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

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

The harmonic summation (HS) estimator is

$$\hat{\omega}_0 = \arg\max_{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]} \sum_{l=1}^{L} |X(\omega_0 l)|^2$$  \hspace{1cm} (137)
Some remarks:

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

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech
   Correlation-based Methods
      The Least Squares Method
         The Harmonic Summation Method
         The Nonlinear Least Squares Estimator
   Comparison of Methods
   Non-stationary Pitch Estimation
   Multi-channel Pitch Estimation
   Summary

Model-based Array Processing and Enhancement
NLS Estimator

The harmonic model

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

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

\[ x = Z_L(\omega_0) \alpha_L + e . \]  \hspace{1cm} (139)

where

\[ Z_L(\omega) = \begin{bmatrix} c(\omega) & c(2\omega) & \cdots & c(L\omega) & s(\omega) & s(2\omega) & \cdots & s(L\omega) \end{bmatrix} \]

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

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

\[ \alpha_l = [a_l^T, -b_l^T]^T, \quad a_L = [a_1, \cdots, a_L]^T, \quad b_L = [b_1, \cdots, b_L]^T \]
NLS Estimator

The least squares error is

$$\sum_{n=0}^{N-1} e^2(n) = e^T e = [x - Z_L(\omega_0)\alpha_L]^T [x - Z_L(\omega_0)\alpha_L]$$  \hspace{1cm} (140)

Given \(\omega_0\), the estimate of \(\alpha_L\) is

$$\hat{\alpha}_L = \left[Z_L^T(\omega_0)Z_L(\omega_0)\right]^{-1} Z_L^T(\omega_0)x$$  \hspace{1cm} (141)

Inserting this back into the objective yields the NLS estimator

$$\hat{\omega}_{0,L} = \arg\max_{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]} x^T Z_L(\omega_0) \left[Z_L^T(\omega_0)Z_L(\omega_0)\right]^{-1} Z_L^T(\omega_0)x$$  \hspace{1cm} (142)

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

\[
\hat{\omega}_{0,L} = \arg\max_{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]} \mathbf{x}^T \mathbf{Z}(\omega_0) \left[ \mathbf{Z}^T(\omega_0) \mathbf{Z}(\omega_0) \right]^{-1} \mathbf{Z}^T(\omega_0) \mathbf{x} \tag{143}
\]
NLS Estimator

1. Compute NLS cost function

\[ \hat{\omega}_{0,L} = \arg\max_{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]} x^T Z_L(\omega_0) \left[ Z_L^T(\omega_0) Z_L(\omega_0) \right]^{-1} Z_L^T(\omega_0) x \] (143)

on an \( F/L \)-point uniform grid for all model orders \( L \in \{1, \ldots, L_{\text{MAX}}\} \).

2. Optionally refine the \( L_{\text{MAX}} \) grid estimates.
1. Compute NLS cost function

\[
\hat{\omega}_{0,L} = \arg\max_{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]} x^T Z_L(\omega_0) \left[ Z_L^T(\omega_0) Z_L(\omega_0) \right]^{-1} Z_L^T(\omega_0) x
\]  

(143)

on an \( F/L \)-point uniform grid for all model orders \( L \in \{1, \ldots, L_{\text{MAX}}\} \).

2. Optionally refine the \( L_{\text{MAX}} \) grid estimates.

3. Do model comparison.
NLS Estimator

Nonlinear least squares (NLS) estimator:

\[
\hat{\omega}_{0,L} = \arg\max_{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]} x^T Z_L(\omega_0) \left[ Z_L^T(\omega_0) Z_L(\omega_0) \right]^{-1} Z_L^T(\omega_0) x
\] (144)

Harmonic summation (HS) estimator:

\[
\hat{\omega}_{0,L} = \arg\max_{\omega_0 \in [\omega_{\text{MIN}}, \omega_{\text{MAX}}]} x^T Z_L(\omega_0) Z_L^T(\omega_0) x
\] (145)

Complexities

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

\begin{align*}
\text{NLS} & \quad \mathcal{O}(F \log F) + \mathcal{O}(F L_{\text{MAX}}^3) \\
\text{HS} & \quad \mathcal{O}(F \log F) + \mathcal{O}(F L_{\text{MAX}})
\end{align*}

We have recently decreased the complexity of NLS to that of HS.
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);
A MATLAB implementation

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

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

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

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

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

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech
   Correlation-based Methods
   The Least Squares Method
   Comparison of Methods
      Estimation Accuracy
      Robustness to noise
      Time-frequency resolution
      Summary

Non-stationary Pitch Estimation
Multi-channel Pitch Estimation
Summary
Comparison of Methods

What could be evaluated?

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

How to evaluate a pitch estimator?

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

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

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

Statistical Speech Models

Model-based Pitch Estimation of Speech
  Correlation-based Methods
  The Least Squares Method
  Comparison of Methods
    Estimation Accuracy
    Robustness to noise
    Time-frequency resolution
    Summary

Non-stationary Pitch Estimation
Multi-channel Pitch Estimation
Summary
Comparison of Methods
Estimation Accuracy

Cramér-Rao Lower Bound (CRLB)

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

Asymptotic CRLB for the pitch in WGN

\[
\text{var}(\hat{\omega}_0) \geq \frac{24\sigma^2}{N(N^2 - 1) \sum_{l=1}^{L} A_l^2 l^2}
\]  

(146)
Comparison of Methods
Estimation Accuracy

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

![Graph showing the Relationship Between SNR and RMS Error](image)
Comparison of Methods
Estimation Accuracy

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

![Graph showing comparison between Asymptotic CRLB and fast NLS in terms of RMS error against SNR in dB. The graph illustrates the decrease in RMS error as SNR increases, with the Asymptotic CRLB line approaching the lower end of the x-axis at higher SNRs.]
Setup: $N = 500$, $L = 10$, 1000 repetitions, random, but not low, fundamental frequency, random phases, and constant amplitudes.
Comparison of Methods
Estimation Accuracy

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

The graph shows the RMS difference between the true fundamental frequency \( \omega_0 \) and the estimated frequency \( \hat{\omega}_0 \) as a function of SNR in dB. The lines represent different estimation methods:
- Black line: Asymptotic CRLB
- Grey dashed line: fast NLS
- Red circle line: Harmonic summation
- Blue dot line: YIN

The RMS difference decreases as the SNR increases, indicating better frequency estimation accuracy at higher SNR levels.
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech
  Correlation-based Methods
  The Least Squares Method
  Comparison of Methods
    Estimation Accuracy
    Robustness to noise
    Time-frequency resolution
    Summary

Non-stationary Pitch Estimation
Multi-channel Pitch Estimation
Summary
Comparison of Methods
Robustness to noise

Why were you away a year, Roy?
Comparison of Methods

Robustness to noise

No noise and window size of 25 ms.
Comparison of Methods
Robustness to noise

30 dB SNR and window size of 25 ms.
Comparison of Methods
Robustness to noise

20 dB SNR and window size of 25 ms.
Comparison of Methods
Robustness to noise

15 dB SNR and window size of 25 ms.
Comparison of Methods
Robustness to noise

10 dB SNR and window size of 25 ms.
Comparison of Methods

Robustness to noise

5 dB SNR and window size of 25 ms.
Comparison of Methods
Robustness to noise

0 dB SNR and window size of 25 ms.
Comparison of Methods
Robustness to noise

-5 dB SNR and window size of 25 ms.
Comparison of Methods
Robustness to noise

-10 dB SNR and window size of 25 ms.
Comparison of Methods
Robustness to noise

-15 dB SNR and window size of 25 ms.
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech
  Correlation-based Methods
  The Least Squares Method
  Comparison of Methods
    Estimation Accuracy
    Robustness to noise
    Time-frequency resolution
    Summary

Non-stationary Pitch Estimation
Multi-channel Pitch Estimation
Summary
Comparison of Methods
Time-frequency resolution

Sustained vowel

frequency [Hz]

0 500 1000 1500 2000 2500 3000 3500 4000

time [s]

0 0.5 1 1.5 2 2.5 3 3.5

-30 -20 -10 0 10 20
Comparison of Methods
Time-frequency resolution

Window size of **25 ms** and no noise.
Comparison of Methods
Time-frequency resolution

Window size of 20 ms and no noise.
Comparison of Methods
Time-frequency resolution

Window size of 15 ms and no noise.
Comparison of Methods
Time-frequency resolution

Window size of 12 ms and no noise.

Sustained vowel

frequency [Hz]

0 0.5 1 1.5 2 2.5 3 3.5 4
time [s]

NLS
Yin
Comparison of Methods

Time-frequency resolution

Window size of 11 ms and no noise.
Comparison of Methods

Time-frequency resolution

Window size of 10 ms and no noise.
Comparison of Methods

Time-frequency resolution

Window size of 9 ms and no noise.
Comparison of Methods
Time-frequency resolution

Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 15 dB
Comparison of Methods
Time-frequency resolution

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

![Graph of RMS error against \( \omega_0 \)]
Comparison of Methods
Time-frequency resolution

Setup: \( N = 200, \ L = 10, \) 1000 repetitions, random phases, and constant amplitudes, and SNR of 15 dB
Comparison of Methods
Time-frequency resolution

Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 15 dB
Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 0 dB

![Graph showing the RMS of the difference between $\omega_0$ and $\hat{\omega}_0$ as a function of $\omega_0 \frac{N}{2\pi}$ for different values of $\omega_0$. The x-axis represents $\omega_0 \frac{N}{2\pi}$ in cycles per segment, ranging from 0.8 to 3. The y-axis is a log scale from $10^{-5}$ to $10^{-1}$, showing the RMS values. The graph includes a line labeled "Asymp. CRLB." ]
Setup: $N = 200$, $L = 10$, 1000 repetitions, random phases, and constant amplitudes, and SNR of 0 dB
Comparison of Methods
Time-frequency resolution

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

\[ \text{RMS}(\omega_0 - \hat{\omega}_0) \]

- Asymp. CRLB
- NLS
- HS

\[ \omega_0 \frac{N}{2\pi} \text{ [cycles/segment]} \]
Comparison of Methods
Time-frequency resolution

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

\[
\text{RMS}(\omega_0 - \hat{\omega}_0) \quad \text{Asymp. CRLB}
\]

\[
\text{NLS}
\]

\[
\text{HS}
\]

\[
\text{YIN}
\]
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech
  Correlation-based Methods
  The Least Squares Method

Comparison of Methods
  Estimation Accuracy
  Robustness to noise
  Time-frequency resolution

Summary

Non-stationary Pitch Estimation

Multi-channel Pitch Estimation

Summary
Comparison of Methods
Summary

Correlation-based Methods
A periodic signal satisfies that

\[ x(n) = x(n - \tau) \] (147)

where \( \tau = \frac{2\pi}{\omega_0} \) is the period.

+ Intuitive and simple
+ Low computational complexity
+/- No need to estimate the model order
- Poor time-frequency resolution
- Are very sensitive to noise
- Interpolation needed for fractional delay estimation
Comparison of Methods

Summary

Parametric Methods

Estimate the parameters in

\[ x(n) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l) + e(n) \]  

(148)

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

Introduction

Statistical Speech Models

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

Model-based Array Processing and Enhancement

Summary and Conclusion
Speech is non-stationary since the fundamental frequency is continuously changing.

The harmonic model assumes that the fundamental frequency is constant in a segment of data.

We can extend the model of the phase of the $l$th harmonic component to

$$\theta_l(n) \approx \phi_l + l\omega_0 n + l\beta_0 n^2 / 2 \quad (149)$$

where $\beta_0$ is the fundamental chirp rate.

We refer to this model as the harmonic chirp model.

$$s(n) = \sum_{l=1}^{L} A_l \cos(l\beta_0 n^2 / 2 + l\omega_0 n + \phi_l) \quad (150)$$
Non-stationary Pitch Estimation

Nonlinear least squares (NLS) objective

\[
J_L(\omega_0, \beta_0) = x^T Z_L(\omega_0, \beta_0) \left[ Z_L^T(\omega_0, \beta_0) Z_L(\omega_0, \beta_0) \right]^{-1} Z_L^T(\omega_0, \beta_0) x
\]  

(151)

Harmonic chirp summation objective:

\[
J_L(\omega_0, \beta_0) = x^T Z_L(\omega_0, \beta_0) Z_L^T(\omega_0, \beta_0) x
\]  

(152)
Non-stationary Pitch Estimation

Window size of 30 ms, 75 % overlap, and no noise
Non-stationary Pitch Estimation

Window size of 30 ms, 75% overlap, and no noise
Non-stationary Pitch Estimation

Window size of 30 ms and no noise
Non-stationary Pitch Estimation

Window size of 30 ms and no noise
Non-stationary Pitch Estimation

(a) Harmonic chirp model

(b) Traditional harmonic model
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech
  Correlation-based Methods
  The Least Squares Method
  Comparison of Methods
  Non-stationary Pitch Estimation

Multi-channel Pitch Estimation
Summary

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

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

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

$$J_L(\omega_0) = \sum_{k=1}^{K} x_k^T Z_L(\omega_0) \left[ Z_L^T(\omega_0) Z_L(\omega_0) \right]^{-1} Z_L^T(\omega_0) x_k$$

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

$$J_L(\omega_0) = \sum_{k=1}^{K} \ln \left\{ x_k^T x_k - x_k^T Z_L(\omega_0) \left[ Z_L^T(\omega_0) Z_L(\omega_0) \right]^{-1} Z_L^T(\omega_0) x_k \right\}$$
Multi-channel pitch estimation

Same noise variance on every channel.
Multi-channel pitch estimation

Different noise variances on every channel.
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech
  Correlation-based Methods
  The Least Squares Method
  Comparison of Methods
  Non-stationary Pitch Estimation
  Multi-channel Pitch Estimation

Summary

Model-based Array Processing and Enhancement

Summary and Conclusion
Summary

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

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

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

Summary and Conclusion
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

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

Summary and Conclusion
Traditional TDOA Estimation

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

Time domain model of 2 microphone recordings:

\[ x_1(n) = s(n) + e_1(n), \quad x_2(n) = \beta s(n - \eta) + e_2(n), \quad n = 0, \ldots, N - 1. \] (154)

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

\[ X_1(k) = S(k) + E_1(k), \]
\[ X_2(k) = \beta S(k)e^{-j2\pi km\eta/N} + E_2(k). \] (155)
Problems with Model

- Frequency domain often too restrictive:
  - Source signal often periodic on short time-scale, but $\omega_0$ assumption not satisfied.
  - Leads to edge effects.
  - Can be reduced through zero-padding, but colors noise spectrum.

- Frequency domain model aren’t suited for fractional TDOA estimation:
  - Fractional delay corresponds to a complex valued sensor signal.
Improvements

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

Assuming periodic signal:

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

with

\[ A_k / \alpha_k: \text{real/complex amplitude (} A_k > 0, \, \alpha_k = \frac{A_k}{2} e^{j\phi_k}\), \]

\[ \phi_k: \text{phase (} \phi_k \in [-\pi, \pi[\)), \]

\[ \omega_0: \text{fundamental frequency}. \]

Signal delay by \( \eta \) gives

\[ s(n - \eta) = \sum_{k=-L}^{L} \alpha_k e^{j\omega_0 kn} e^{-j\omega_0 \eta k} = \sum_{k=-L}^{L} \alpha_k e^{j\omega_0 kn} e^{-j\xi k}. \quad (157) \]
Matrix-vector model

Model in matrix-vector notation:

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

with

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

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

\[
(\hat{\beta}, \hat{\xi}, \hat{\omega}_0) = \arg \max_{\beta, \xi, \omega_0} J(\beta, \xi, \omega_0),
\]

with

\[
J(\beta, \eta, \omega_0) = x^H H (H^H H)^{-1} H^H x.
\]

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

Complexity can be reduced through approximations.
Approximate ML Method

For large $N$ we have

$$Z^H(\omega_0)Z(\omega_0) \approx NI_{2L}. \quad (160)$$

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

$$H^H(\beta, \xi, \omega_0)H(\beta, \xi, \omega_0) \approx (1 + \beta^2)NI_{2L}, \quad (161)$$

and as a result:

$$J(\beta, \xi, \omega_0) = \frac{1}{N(1 + \beta^2)} \left[ x_1^HZ(\omega_0)Z^H(\omega_0)x_1 + \beta^2x_2^Z(\omega_0)Z^H(\omega_0)x_2 
+ 2\beta x_1^H Z(\omega_0) D^*(\xi) Z^H(\omega_0) x_2 \right]. \quad (162)$$
Important Special Case

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

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

$$J(\eta) = x_1^H Z(2\pi/N) D^*(2\pi\eta/N) Z^H(2\pi/N) x_2$$  \hspace{1cm} (163)

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

For $N$ being even:

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

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

Thus, the CC TDOA estimator is statistically efficient when:

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

Assume no noise and \( \eta_0 \) being true delay, then:

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

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

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

Problem avoided by using:

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

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

Speech data experiments
- $\sim2.2$ s of female speech (mainly voiced),
- stereo recording made using RIR generator,
- $\eta = 0$ samples, no reverb.,
- $N = 100$, $f_s = 8$ kHz.
Experiments
Results on synthetic data
Experiments
Results on synthetic data

![Graph showing experimental results on synthetic data with various methods: GCC, GCCP, NLS (oracle), NLS, AML, CRB. The graph plots RMSE against frequency deviation.]
Experiments
Results on speech data

![Graph showing estimated TDOA vs time](image-url)

- **GCC**
- **GCCP**
- **NLS**
- **AML**
- **True**

**Legend:**
- Estimated TDOA [samples]
- Time [s]
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

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

Summary and Conclusion
DOA and Pitch Estimation

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

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

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

$$y_k(n) = \beta_k s(n - f_s \tau_k) + e_k(n)$$  \hspace{1cm} (168)

$$= x_k(n) + e_k(n),$$  \hspace{1cm} (169)

with

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

We choose reference point for which:

\[ s_{\text{ref}}(n) = \beta_{\text{ref}} s(n - f_s \tau_{\text{ref}}). \]  (170)

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

\[ x_k(n) = \frac{r_{\text{ref}}}{r_k} s_{\text{ref}}(n - f_s \tau_{\text{ref},k}) + e_k(n), \]  (171)

\[ = \frac{r_{\text{ref}}}{r_k} s_{\text{ref}} \left( n - f_s \frac{r_k - r_{\text{ref}}}{c} \right) + e_k(n), \]  (172)

where

\[ \tau_{\text{ref},k} : \text{TDOA of source between mic 0 and } k. \]
Clean desired signal modeled as

\[ s_{\text{ref}}(n) = \sum_{l=1}^{L} \gamma_l e^{i\omega_0 n}, \]  

(173)

with

\[ \gamma_l = \beta_{\text{ref}} \alpha_l, \]

\( \alpha_l \): complex amplitude of \( l \)'th harmonic,

\( \omega_0 \): fundamental frequency [rad/sample],

\( L \): model order, i.e., number of harmonics.
Uniform Linear Array Model

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

$$r_k(r_c, \theta) = \sqrt{g_k^2 d^2 + r_c^2 - 2g_k dr_c \sin \theta}$$  \hspace{1cm} (174)

with

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

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

Then,

$$x_k(n) = \frac{r_c}{r_k} \sum_{l=1}^{L} \gamma_l e^{j\omega_0 n} e^{-jfs_l\omega_0} r_{k-l} + e_k(n).$$  \hspace{1cm} (175)
Matrix-Vector Model

Consider vector of $N$ samples from mic $k$:

$$\mathbf{x}_k = [x_k(0) \ x_k(1) \ \cdots \ x_k(N-1)]^T,$$

$$= \mathbf{Z}(\omega_0)\mathbf{D}_k(r_k)\gamma + \mathbf{e}_k, \quad (176)$$

where

$$\mathbf{Z}(\omega_0) = [\mathbf{z}(\omega_0) \ \mathbf{z}(2\omega_0) \ \cdots \ \mathbf{z}(L\omega_0)],$$

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

$$[\mathbf{D}_k(r_k)]_{pq} = \begin{cases} \frac{r_c}{r_k} e^{-jfs_p\omega_0} \frac{r_k-r_c}{c}, & p = q, \\
0, & \text{otherwise}, \end{cases}$$

$$\gamma = [\gamma_1 \ \gamma_2 \ \cdots \ \gamma_L]^T.$$
Likelihood function useful for finding optimal estimators and CRB’s.

Assuming WGN which is not correlated across mics:

\[ \mathcal{L} = \ln p(\{x_k\}; \nu) = -N \left( K \ln \pi + \sum_{k=0}^{K-1} \ln \sigma_k^2 \right) - \sum_{k=0}^{K-1} \frac{\|e_k\|^2}{\sigma_k^2}, \quad (177) \]

with

\[ \nu: \text{ vector containing unknown parameters of interest,} \]
\[ \sigma_k^2: \text{ variance of noise at mic } k. \]
Asymptotic CRBs

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

\[
\text{CRB}(\omega_0) \approx \frac{6}{N^3K} \text{PSNR}^{-1},
\]

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

\[
\text{PSNR} = \frac{\sum_{l=1}^{L} l^2 A_l^2}{\sigma^2}.
\]

Oberservations

- \(\omega_0\) CRB decreases with both \(N\) and \(K\) but independent on \(\theta\).
- \(\theta\) CRB decreases with increasing \(\omega_0\), \(N\) and \(K\).
- both \(\omega_0\) and \(\theta\) CRBs decreases by exploiting harmonic structure.
First, closed-form solutions for $\gamma$ and $\sigma_k^2$’s minimizing $\mathcal{L}$ can be found:

\[
\hat{\gamma} = \left( \sum_{k=0}^{K-1} \frac{D_k^H Z^H ZD_k}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{D_k^H Z^H x_k}{\sigma_k^2}, \quad (180)
\]

\[
\hat{\sigma}_k^2 = \frac{\|x_k - ZD_k \hat{\gamma}\|^2}{N}. \quad (181)
\]

Estimates depend on each other → estimated iteratively!

Resulting estimator after inserting closed-form solutions:

\[
\{\hat{\omega}_0, \hat{r}_c, \hat{\theta}\} = \arg \min \sum_{k=0}^{K-1} \ln \|x_k - ZD_k \hat{\gamma}\|^2. \quad (182)
\]
Approximate ML Estimator

For large sample sizes, it holds that

$$\lim_{N \to \infty} \frac{1}{N} Z^H Z = I.$$  \hspace{1cm} (183)

With this approximation:

$$\hat{\gamma} = \left( \sum_{k=0}^{K-1} \frac{r_c^2}{r_k^2} \frac{N}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{D_k Z^H x_k}{\sigma_k^2}. \hspace{1cm} (184)$$

Main computational complexity is $Z^H x_k$, but replaceable with FFT.
ML Estimator
Special case: equal noise levels

With the same noise level at each mic:

$$\hat{\gamma} = \left( \sum_{k=0}^{K-1} D_k^H Z^H Z D_k \right)^{-1} \sum_{k=0}^{K-1} D_k^H Z^H x_k. \quad (185)$$

With the large sample approximation, it reduces to

$$\hat{\gamma} = \left( \sum_{k=0}^{K-1} \frac{r_c^2}{r_k^2} N \right)^{-1} \sum_{k=0}^{K-1} D_k^H Z^H x_k. \quad (186)$$
ML Estimator
Special case: far-field scenarios

In far-field, following approximations hold

\[
\frac{r_c}{r_k} \approx 1, \quad \text{and} \quad \tau_{c,k} \approx g_k \frac{d \sin \theta}{c}.
\]  
(187)

Amplitude and noise estimates are then:

\[
\hat{\gamma} = \left( \sum_{k=0}^{K-1} \frac{\tilde{D}_k^H Z^H Z \tilde{D}_k}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{\tilde{D}_k^H Z^H x_k}{\sigma_k^2},
\]  
(188)

\[
\hat{\sigma}_k^2 = \frac{\| x_k - Z \tilde{D}_k \gamma \|}{N},
\]  
(189)

with

\[
[\tilde{D}_k]_{pq} = \begin{cases} 
eq -j f_s p \omega_0 \tau_{c,k}, & \text{for } p = q, \\ 0, & \text{otherwise.} \end{cases}
\]
Far-field assumption can be combined with equal noise variance assumption:

\[
\hat{\gamma} = \left( \sum_{k=0}^{K-1} \tilde{D}_k^H \tilde{Z}^H \tilde{D}_k \right)^{-1} \sum_{k=0}^{K-1} \tilde{D}_k^H \tilde{Z}^H x_k. \tag{190}
\]

large sample approximation:

\[
\hat{\gamma} = \left( \sum_{k=0}^{K-1} \frac{N}{\sigma_k^2} \right)^{-1} \sum_{k=0}^{K-1} \frac{\tilde{D}_k \tilde{Z}^H x_k}{\sigma_k^2}. \tag{191}
\]

or both:

\[
\hat{\gamma} = \frac{1}{NK} \sum_{k=0}^{K-1} D_k Z^H x_k. \tag{192}
\]
Experimental Results
Synthetic source in far-field

![Graphs showing experimental results for different methods.](image-url)
Experimental Results
Synthetic source in far-field

-1.3 -1.2 -1.1 -1 -0.9 -0.8 -0.7
NLS
aNLS
MC-ML
MC-aML
PoPi
Sub.

-1.3 -1.2 -1.1 -1 -0.9 -0.8 -0.7
NLS
aNLS
SRP
SRP-PHAT
bMVDR
PoPi
Sub.
Experimental Results
Synthetic source in near-field

![Graphs showing MSE(\hat{\theta}) and MSE(\hat{\phi}) vs SNR for different methods: NLS, SRP-PHAT, WLSWM, CRB.](image-url)
Experimental Results
Synthetic source in near-field

![Graphs showing MSE vs. K for different methods: NLS, SRP-PHAT, WLSWM, and CRB.](image)
Experimental Results
Real speech in near-field

- Frequency spectrum
- Phase estimation
- Fundamental frequency estimation
- Resonance frequency estimation
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

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

Summary and Conclusion
Introduction

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

**Model:** direct-path + early reflections + noise.
Signal Model

An acoustic source is sampled using a microphone array:

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

where

- $s'(n)$: clean source signal
- $g_k(n)$: room impulse response from source to mic $k$
- $v'_k(n)$: additive noise (interferers, sensor noise, etc.)

Remarks:

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

With $K$ microphones recording $N$ samples each, we get

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T & \cdots & \mathbf{y}_K^T \end{bmatrix}^T = \mathbf{s} + \mathbf{v}'.$$  \hspace{1cm} (194)

where

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

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

Further model specifications:

$\blacktriangleright$ desired signal assumed quasi-periodic,

$\blacktriangleright$ a ULA structure is assumed.
Periodic Signal Model

Clean desired signal modeled as

\[ s'(n) = \sum_{l=-L}^{L} \alpha_l e^{j\omega_0 n}, \]  
(195)

with

- \( \alpha_l \): complex amplitude of \( l \)'th harmonic,
- \( \omega_0 \): fundamental frequency [rad/sample],
- \( L \): model order, i.e., number of harmonics.

Important observation:

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

\[ \omega_0 = \frac{2\pi}{N} \quad \land \quad L = \lfloor N/2 \rfloor. \]  
(196)
Array Model

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

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

$$\tau_{r,k} = k \frac{d \sin \theta_r}{c} = k \eta_r,$$  \hspace{1cm} (197)

with

- $d$: spacing between two adjacent mics,
- $\theta_r$: DOA of source $r$,
- $c$: sound propagation speed.
Complete Signal Model

Observation modeled as multiple early reflections in noise:

\[ y = \sum_{r=1}^{R} H(\eta_r) \alpha_r + v, \quad (198) \]

where

- \( R \): number of early reflections
- \( H(\eta_r) = \left[ Z^T \left( ZD_2(\eta_r) \right)^T \cdots \left( ZD_K(\eta_r) \right)^T \right]^T \)
- \( Z = [z_1 \cdots z_L z_1^* \cdots z_L^*]^T \)
- \( z_l = [1 \quad e^{j\omega_0} \cdots e^{j(N-1)\omega_0}]^T \)
- \( D_k(\eta_r) = \text{diag} \left( \left[ d_k^T(\eta_r) \quad d_k^H(\eta_r) \right] \right) \)
- \( d_k(\eta_r) = \left[ e^{-j\omega_0 k \eta_r} \cdots e^{-jL\omega_0 k \eta_r} \right]^T \)

Estimation problem: find \( \eta_1 \) from observations!
Two methods for DOA estimation with reverb were proposed.

Idea is to estimate DOAs of both direct-path and early reflections.

Bias of direct-path estimate reduced in this way.

Both methods are based on nonlinear least squares:

1. a method where amplitudes of direct-path and reflections are assumed independent.
2. a method where the relation between the amplitudes is modeled.

Estimation of multiple DOAs facilitated by an iterative approach.
Nonlinear Least Squares
Unstructured amplitudes

With unstructured amplitudes, the NLS estimator is

\[
\{\hat{\eta}, \hat{\alpha}\} = \arg \min_{\{\eta, \alpha\}} \|y - \overline{H}(\eta)\overline{\alpha}\|_2^2,
\] (199)

with

\[
\eta = [\eta_1 \cdots \eta_R]^T
\]
\[
\overline{H}(\eta) = [H(\eta_1) \cdots H(\eta_R)]
\]
\[
\overline{\alpha} = [\alpha_1^T \cdots \alpha_R^T]^T
\]

Solving for \(\overline{\alpha}\) gives

\[
\hat{\eta} = \arg \min_{\eta} \| \left( I - \overline{H}(\eta)(\overline{H}(\eta)^H\overline{H}(\eta))^{-1}\overline{H}(\eta)^H \right) y \|_2^2. \] (200)
Iterative Procedure

Unstructured amplitudes

Consider a modified observed signal model:

\[ y_r = y - \sum_{q=1, q \neq r}^{R} H(\hat{\eta}_q)\hat{\alpha}_q, \]

This suggests:

\[ \hat{\alpha}_r = (H^H(\eta_r)H(\eta_r))^{-1}H(\eta_r)^H y_r, \]
\[ \hat{\eta}_r = \arg \min_{\eta_r} \| P_{H(\eta_r)} y_r \|_2^2. \]

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

Step (1): Assume \( R = 1 \). Estimate \( \eta_1 \) and \( \alpha_1 \) from \( y_1 = y \) as described before.

Step (2): Assume \( R = 2 \). Estimate \( \eta_2 \) and \( \alpha_2 \) from \( y_2 \) using parameter estimates from Step (1). Re-estimate \( \eta_1 \) and \( \alpha_1 \) from \( y_1 \). Iterate until “practical convergence”.

Step (3): Assume \( R = 3 \). Estimate \( \eta_3 \) and \( \alpha_3 \) from \( y_3 \) using parameters from Step (2). Re-estimate \( \eta_1 \) and \( \alpha_1 \) from \( y_1 \). Re-estimate \( \eta_2 \) and \( \alpha_2 \) from \( y_2 \). Iterate until “practical convergence”.

Remaining steps: Continue similarly to the previous steps until \( R \) is equal to the number of early reflections.
An alternative model with amplitude relations can be formulated

\[ y = \sum_{r=1}^{R} \gamma_r H(\eta_r) T_r \alpha + v, \quad (204) \]

where

- \( \gamma_r \): attenuation of reflection \( r \) (\( \gamma_1 = 1 \))
- \( \eta_r \): delay of reflection \( r \) (\( \eta_1 = 0 \))
- \( \alpha \): direct-path harmonic amplitudes
- \( T_r = \text{diag} \left( \begin{bmatrix} t_r^T & t_r^H \end{bmatrix} \right) \)
- \( t_r = \begin{bmatrix} e^{i\omega_0 \xi_r} & \ldots & e^{iL\omega_0 \xi_r} \end{bmatrix}^T \)
Iterative Procedure
Structured amplitudes

Again, consider a modified observed signal model:

\[ y_r = y - \sum_{q=1, q \neq r}^{R} \gamma_q H(\hat{\eta}_q) \hat{\alpha} \]  \hspace{1cm} (205)

With this, LS amplitudes and attenuations estimates are

\[ \hat{\alpha} = [H^H(\eta_1)H(\eta_1)]^{-1}H^H(\eta_1)y_1 \quad (r = 1) \]  \hspace{1cm} (206)

\[ \hat{\gamma}_r = \frac{\text{Re}\{\hat{\alpha}^H T_r^H H^H(\eta_r)y_r\}}{\hat{\alpha}^H T_r^H H^H(\eta_r)H(\eta_r)T_r \hat{\alpha}} \quad (r = 2, \ldots, R). \]  \hspace{1cm} (207)
Iterative Procedure
Structured amplitudes

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

$$\hat{\eta}_1 = \arg \min_{\eta_1} \| P_{\perp H(\eta_1)} y_1 \|_2^2. \quad (208)$$

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

$$\{\hat{\eta}_r, \hat{\xi}_r\} = \arg \min_{\eta_r, \xi_r} \| y_r - \hat{\gamma}_r H(\eta_r) T_r \hat{\alpha} \|_2^2. \quad (209)$$

This method is termed RNLS-S.

Remarks

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

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

Synthetic data

\[ \text{RMSE}(\hat{\eta}_1) \times 10^{-5} \]

- NLS
- RNLS
- RNLS-S

R
\[ 2 \quad 3 \quad 4 \quad 5 \]

\[
\begin{array}{c}
2.2 \\
2.4 \\
2.6 \\
\end{array}
\times 10^{-5}
\]
Experimental Results
Synthetic data

![Graph showing RMSE(\(\hat{\eta}_1\)) vs. \(T_{60}\) for different methods: NLS, RNLS, RNLS-S.](image)
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: \( R = 4, T_{60} = 0.3 \text{ s}, K = 4 \).

<table>
<thead>
<tr>
<th></th>
<th>NLS</th>
<th>RNLS</th>
<th>RNLS-S</th>
<th>SRP-PHAT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(3.8 \cdot 10^{-5})</td>
<td>(3.6 \cdot 10^{-5})</td>
<td>(3.6 \cdot 10^{-5})</td>
<td>(5.4 \cdot 10^{-5})</td>
</tr>
</tbody>
</table>
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
  Parametric vs. Non-Parametric TDOA Estimation
  Joint DOA and Pitch Estimation
  Reverb Robust Speech Localization

Model-based Speech Enhancement
  Kalman Filtering
  Single-Channel LCMV and APES

Reverb Robust Speech Localization

Model-based Speech Enhancement
  Enhancement of Non-Stationary Speech

Non-Intrusive Speech Intelligibility Prediction

Summary and Conclusion
Model-based Enhancement

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

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

\[ z_{l/r}(n) = s_{l/r}(n) + w_{l/r}(n) \quad \forall n = 0, 1, 2 \ldots \] (210)

with

- \( l/r \): indicates left or right channel,
- \( s_{l,r}, w_{l/r} \): speech and noise signal.

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

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

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

\[ u(n) = b(n, p_n)u(n - p_n) + d(n). \] (212)
Noise Model

Noise signal modelled as an autoregressive process

\[ w(n) = \sum_{i=1}^{Q} c_i(n)w(n - i) + v(n) \]

\[ = c(n)^T w(n - 1) + v(n), \tag{213} \]

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

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

\[ z_l(n) \rightarrow \text{Kalman Smoother} \rightarrow \hat{s}_l(n) \]

\[ z_r(n) \rightarrow \text{Kalman Smoother} \rightarrow \hat{s}_r(n) \]

\[ \text{Codebook Based Approach} \]

\[ \text{Pitch Estimator} \]
Binaural Estimation of STP Parameters

- Usage of a fixed lag Kalman smoother for speech enhancement requires the speech and noise STP parameters to be estimated.
- This approach uses a priori information about the spectral envelopes of speech and noise stored in codebooks.
- Single-channel codebook based estimation of STP parameters can be used [Srinivasan, 2007].
The random variables of parameters to be estimated are \( \theta = [a; c; \sigma_u^2; \sigma_v^2] \).

The MMSE estimate of the parameter is written as

\[
\hat{\theta} = \mathbb{E}(\theta|z_l, z_r),
\]

where \( z_l, z_r \) denotes a frame of noisy samples at left, right ears.

Using the Bayes Theorem:

\[
\hat{\theta} = \int_{\Theta} \theta p(\theta|z_l, z_r) d\theta = \int_{\Theta} \theta \frac{p(z_l, z_r|\theta)p(\theta)}{p(z_l, z_r)} d\theta.
\]
Binaural Estimation of STP Parameters

With $\theta_{ij} = [a_i; \sigma_{u,ij}^2; c_j; \sigma_{v,ij}^2]$, discrete counterpart, (215) is:

$$\hat{\theta} = \sum_{i=1}^{N_s} \sum_{j=1}^{N_w} \theta_{ij} \frac{p(z_l, z_r | \theta_{ij}) p(\theta_{ij})}{p(z_l, z_r)}, \quad (216)$$

where the MMSE estimate is expressed as weighted linear combination of $\theta_{ij}$ with weights $p(z_l, z_r | \theta_{ij})$.

Assuming conditional independence for the left and right noisy signal:

$$p(z_l, z_r | \theta_{ij}) = p(z_l | \theta_{ij}) p(z_r | \theta_{ij}). \quad (217)$$
Log-likelihood can be expressed using negative of Itakura Saito distortion between noisy and modelled noisy spectral envelopes:

\[
p(z_l, z_r | \theta_{ij}) = p(z_l | \theta_{ij}) p(z_r | \theta_{ij}) \\
= e^{-(d_{IS}(P_{z_l}(\omega), \hat{P}^{ij}_z(\omega)) + d_{IS}(P_{z_r}(\omega), \hat{P}^{ij}_z(\omega)))}
\]  

(218)

where

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

\[
\hat{P}^{ij}_z(\omega) = \frac{\sigma^{2,ML}_{u,ij}}{|A^i_s(\omega)|^2} + \frac{\sigma^{2,ML}_{v,ij}}{|A^i_w(\omega)|^2}.
\]
Binaural Estimation of STP Parameters

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

Assumes a homogeneous noise field, i.e.

$$\Phi_{ww}(\omega) = \Phi_{wl}(\omega) = \Phi_{wr}(\omega), \Phi_{wl}(\omega) = 0.$$  (219)

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

$$\hat{\Phi}_{ww}(\omega, k) = \sqrt{\Phi_{zl}(\omega, k)\Phi_{zr}(\omega, k) - |\Phi_{zl}(\omega, k)|}$$  (220)

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

Noisy signals modeled as

\[ y = \begin{bmatrix} z_l \\ z_r \end{bmatrix} = \begin{bmatrix} ZD_l \\ ZD_r \end{bmatrix} \alpha + \begin{bmatrix} w_l \\ w_r \end{bmatrix} = H\alpha + w. \]  

(221)

with

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

ML estimate of amplitude vector **\alpha**:

\[ \hat{\alpha} = (H^TH)^{-1}H^Ty \]  

(222)
ML noise variance estimates for the left and right channels:

$$\hat{\sigma}_{l/r}^2 = \frac{1}{N} \| z_{l/r} - ZD_{l/r} \hat{\alpha} \|^2 $$  \hspace{1cm} (223)

Using amplitude and noise estimates, fundamental frequency estimated as:

$$ \{ \hat{\omega}_0, \hat{L} \} = \arg\min_{\{L, \omega_0\}} N \ln \hat{\sigma}_l^2 \hat{\sigma}_r^2 + L \ln 2N. $$  \hspace{1cm} (224)
Fixed Lag Kalman Smoother (FLKS)

Speech model can be written as concatenated state space equation:

\[
\begin{bmatrix}
    s_{l/r}(n) \\
    u(n+1) \\
    w_{l/r}(n)
\end{bmatrix} =
\begin{bmatrix}
    A(n) & \Gamma_1 \Gamma_2^T & 0 \\
    0 & B(n+1) & 0 \\
    0 & 0 & C(n)
\end{bmatrix}
\begin{bmatrix}
    s_{l/r}(n-1) \\
    u(n) \\
    w_{l/r}(n-1)
\end{bmatrix}
\]

\[
+ \begin{bmatrix}
    0 & 0 \\
    \Gamma_2 & 0 \\
    0 & \Gamma_3
\end{bmatrix}
\begin{bmatrix}
    d(n+1) \\
    0 \\
    v(n)
\end{bmatrix} \Leftrightarrow, \quad (225)
\]

\[
x_{l/r}(n+1) = F(n)x_{l/r}(n) + \Gamma_4 g(n+1)
\]

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

Measurement equation given by

\[
z_{l/r}(n) = \Gamma^T x_{l/r}(n). \quad (227)
\]

Eventually, clean speech is then predicted using Kalman filter.
Experimental Setup

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

Parameters

<table>
<thead>
<tr>
<th>fs</th>
<th>Frame Size</th>
<th>$N_{spk}$</th>
<th>$N_w$</th>
<th>$P$</th>
<th>$Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 Khz</td>
<td>200 (25ms)</td>
<td>64</td>
<td>8</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

Compared Methods

- **UV**: binaural enhancement using the conventional AR model,
- **V-UV**: binaural enhancement using the voiced-unvoiced model (includes pitch information from noisy signal).
# Experimental Results

**STOI scores**

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>-2</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Noisy</td>
<td><strong>male</strong></td>
<td>0.6486</td>
<td>0.7178</td>
<td>0.7827</td>
</tr>
<tr>
<td></td>
<td><strong>female</strong></td>
<td>0.6305</td>
<td>0.7003</td>
<td>0.7668</td>
</tr>
<tr>
<td>UV</td>
<td><strong>male</strong></td>
<td>0.6882</td>
<td>0.7655</td>
<td>0.8334</td>
</tr>
<tr>
<td></td>
<td><strong>female</strong></td>
<td>0.6568</td>
<td>0.7332</td>
<td>0.8035</td>
</tr>
<tr>
<td>V-UV</td>
<td><strong>male</strong></td>
<td>0.7036</td>
<td>0.7803</td>
<td>0.8436</td>
</tr>
<tr>
<td></td>
<td><strong>female</strong></td>
<td>0.6857</td>
<td>0.7635</td>
<td>0.8277</td>
</tr>
</tbody>
</table>
## Experimental Results

PESQ scores

<table>
<thead>
<tr>
<th></th>
<th>SNR (dB)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-5</td>
<td>-2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Noisy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>male</td>
<td>1.6313</td>
<td>1.7754</td>
<td>1.8434</td>
<td>1.9577</td>
</tr>
<tr>
<td>female</td>
<td>1.2924</td>
<td>1.4868</td>
<td>1.6331</td>
<td>1.7941</td>
</tr>
<tr>
<td>UV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>male</td>
<td>1.8666</td>
<td>2.0467</td>
<td>2.2978</td>
<td>2.5326</td>
</tr>
<tr>
<td>female</td>
<td>1.5066</td>
<td>1.7273</td>
<td>1.9487</td>
<td>2.1744</td>
</tr>
<tr>
<td>V-UV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>male</td>
<td>1.8720</td>
<td>2.1006</td>
<td>2.3396</td>
<td>2.5489</td>
</tr>
<tr>
<td>female</td>
<td>1.6088</td>
<td>1.8429</td>
<td>2.0625</td>
<td>2.2626</td>
</tr>
</tbody>
</table>
Experimental Results
Examples

Noisy signal

- Left noisy signal
- Right noisy signal

Noisy signal
Experimental Results
Examples

Enhanced signal (UV)
Experimental Results

Examples

Enhanced signal (V-UV)
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
  Parametric vs. Non-Parametric TDOA Estimation
  Joint DOA and Pitch Estimation
  Reverb Robust Speech Localization

Model-based Speech Enhancement
  Kalman Filtering
  Single-Channel LCMV and APES

Enhancement of Non-Stationary Speech

Non-Intrusive Speech Intelligibility Prediction
Voiced Speech Enhancement
Model-based single-channel methods

- Voiced speech exhibits quasi-periodic structure.
- Has been exploited to derive model- and filtering-based enhancement methods.
- Compared to traditional speech enhancement method, these can be guaranteed distortionless!
- Also, they do not require noise statistics estimates.
First, we introduce a source defined for \( n = 0, \ldots, N - 1 \) as

\[
x_k(n) = \sum_{l=1}^{L_k} \alpha_{k,l} e^{j\omega_k ln} + e_k(n),
\]

(228)

where

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

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

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

(229)

and similarly for \( x_k(n) \). Note that when \( M = N \) we simply write \( x_k(n) = x_k \). The signal model can be written into matrix-vector form as

\[
x(n) = \sum_{k=1}^{K} Z_k \begin{bmatrix}
e^{-j\omega_k n} & 0 \\
0 & \ddots \\
0 & e^{-j\omega_k L_k n}
\end{bmatrix} \alpha_k + e(n)
\]

(230)

(231)

where \( Z_k = [z(\omega_k) \ \cdots \ z(\omega_k L_k)] \), \( z(\omega) = [1 \ e^{-j\omega} \ \cdots \ e^{-j\omega(M-1)}]^T \),

and \( \alpha_k = [\alpha_{k,1} \ \cdots \ \alpha_{k,L_k}]^T \).
Enhancement filters for voiced speech derived from MSE between filter output, $y_k(n)$ and desired output, $\hat{y}_k(n)$,

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

$$P = \frac{1}{G} \sum_{n=M-1}^{N-1} |h_k^H x(n) - \alpha_k^H w_k(n)|^2,$$

where

$$h_k = [h_k(0) \cdots h_k(M-1)]^T,$$

$$\alpha_k = [\alpha_{k,1} \cdots \alpha_{k,L}]^T,$$

$$w_k(n) = [e^{j\omega_k n} \cdots e^{j\omega_k Ln}]^T.$$
MSE-based Filters

MSE can be written out as:

\[ P = h_k^H \hat{R} h_k - \alpha_k^H G_k h_k - h_k^H G_k^H \alpha_k + \alpha_k^H W_k \alpha_k, \quad (233) \]

where

\[ G_k = \frac{1}{G} \sum_{n=M-1}^{N-1} w_k(n)x^H(n), \]

\[ W_k = \frac{1}{G} \sum_{n=M-1}^{N-1} w_k(n)w_k^H(n). \]

Can be solved for the unknown amplitudes:

\[ \hat{\alpha}_k = W_k^{-1} G_k h_k, \quad (G \geq L_k, G \geq M). \quad (234) \]
MSE-based Filters

Inserting amplitude estimates, we get:

\[ P = h_k^H \hat{R}_k h_k - h_k^H G_k^H W_k^{-1} G_k h_k, \quad (235) \]
\[ P = h_k^H \left( \hat{R}_k - G_k^H W_k^{-1} G_k \right) h_k, \quad (236) \]
\[ P \triangleq h_k^H \hat{Q}_k h_k. \quad (237) \]

where

\[ \hat{Q}_k = \hat{R}_k - G_k^H W_k^{-1} G_k \quad (238) \]

can be thought of as a *modified* or *noise* covariance matrix estimate.
MSE-based Filters

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

\[
\min_{h_k} h_k^H \hat{Q}_k h_k \quad \text{s.t.} \quad h_k^H Z_k = 1. \tag{239}
\]

Solution to optimization problem:

\[
\hat{h}_k = \hat{Q}_k^{-1} Z_k \left( Z_k^H \hat{Q}_k^{-1} Z_k \right)^{-1} 1. \tag{240}
\]

These filters are termed SF-APES filters.

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

Replacing $W_k$ by $I$ in (237), the usual noise covariance matrix estimate is obtained. As before:

$$\hat{h}_k = \hat{Q}_k^{-1}Z_k \left( Z_k^H \hat{Q}_k^{-1}Z_k \right)^{-1} 1,$$  

(241)

but the modified covariance matrix estimate is now:

$$\hat{Q}_k = \hat{R} - G_k^H G_k.$$  

(242)

Computationally simpler as it does not require the inversion of $W_k$ for each candidate frequency.

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

#2 and #3

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

$$\hat{h}_k = \hat{R}^{-1}Z_k \left( Z_k^H \hat{R}^{-1}Z_k \right)^{-1} 1,$$

which we refer to as SF-Capon.

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

$$\hat{h}_k = Z_k \left( Z_k^H Z_k \right)^{-1} 1,$$

which is fully specified by $Z_k$. Referred to as SF-WNC.
Further simplification possible through large sample approximation:

\[
\lim_{M \to \infty} M Z_k \left( Z_k^H Z_k \right)^{-1} = Z_k \lim_{M \to \infty} \left( \frac{1}{M} Z_k^H Z_k \right)^{-1} = Z_k.
\] (245)

(246)

Leads to trivial filter design:

\[ \hat{h}_k = \frac{1}{M} Z_k 1, \] (247)

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

Experimental details:

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

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

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

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

Figure: Plots of: (left) mixture of the two signals and (right) estimated pitch tracks for source 1 (dashed) and 2 (solid).
Figure: Plots of: estimate of sources (left) 1 and (right) 2 obtained from mixture.
The filtering method was extended to the multichannel case in [Jensen2017].

Idea is to use APES principle on each channel and do weighted average based on MSEs.

Two approaches were considered:

- a method which is dependent on geometry,
- a method being independent on geometry.

Results, in terms of PESQ scores, showed that geometry-based approach is best for larger arrays, but worse when significant DOA errors are present.
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement
  Parametric vs. Non-Parametric TDOA Estimation
  Joint DOA and Pitch Estimation
  Reverb Robust Speech Localization
  Model-based Speech Enhancement

Enhancement of Non-Stationary Speech
  Non-Intrusive Speech Intelligibility Prediction

Summary and Conclusion
Enhancement of Non-Stationary Speech

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

\[ x(n) = s(n) + v(n) = \sum_{l=1}^{L} \alpha_l e^{i\omega_0 n} + v(n), \quad \alpha_l = A_l e^{i\phi_l}. \]  

\( \omega_0 = \frac{2\pi f_0}{f_s} \): fundamental frequency, \( L \): number of harmonics.
In the 30 ms segment
$\Delta f_0 \approx 8 \text{ Hz} \implies$
Harmonic chirp model
Harmonic Chirp Model

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

\[ \omega_l(n) = l(\omega_0 + kn). \] (249)

where \( k \) is the chirp parameter.

Instantaneous phase is integral of instantaneous frequency:

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

Leads to harmonic chirp model:

\[ x(n) = \sum_{l=1}^{L} \alpha_l e^{il(\omega_0 n + k/2n^2)} + v(n). \] (251)
LCMV Filtering

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

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

where

$$\mathbf{h} = [h(0) \ h(1) \ldots \ h(M-1)]^H,$$

$\mathbf{R}_x$: covariance matrix of $\mathbf{x}(n)$,

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

$$\mathbf{Z} = [\mathbf{z}(\omega_0, k) \ \mathbf{z}(2\omega_0, 2k) \ldots \ \mathbf{z}(L\omega_0, Lk)],$$

$$\mathbf{z}(l\omega_0, lk) = [1 \ \ e^{jl(\omega_0+k/2)} \ldots \ e^{jl(\omega_0(M-1)+k/2(M-1)^2)}]^T,$$

$M$: filter length,

$$\mathbf{1} = [1 \ldots 1].$$
The solution is

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

Often, $R_x$ unknown but can be estimated:

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

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

Assumes a stationary signal within the length $N$ segment!
Amplitude and Phase EStimation (APES) based filter uses signal model to obtain noise covariance matrix estimate.

APES filter minimises the mean square error (MSE)

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

with

$$\alpha = [\alpha_1 \alpha_2 ... \alpha_L]^H,$$

$$w(n) = [e^{j(\omega_0 n + k/2n^2)} \ e^{j(\omega_0 n + k/2n^2)} \ ... \ e^{jL(\omega_0 n + k/2n^2)}]^T.$$
The solution is

\[ h = Q^{-1}Z(Z^HQ^{-1}Z)^{-1}1, \]  

(256)

with

\[ Q = R_x - G^H W^{-1} G, \]

\[ G = \frac{1}{N - M + 1} \sum_{n=0}^{N-M+1} w(n)x^H(n), \]

\[ W = \frac{1}{N - M + 1} \sum_{n=0}^{N-M+1} w(n)w^H(n). \]

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

Filters compared:

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

Filters evaluated as function of input SNR:

\[ \text{iSNR} = \frac{\sigma_s^2}{\sigma_v^2}, \quad (257) \]

Performance measured using output SNR and signal reduction factor:

\[ o\text{SNR}(h) = \frac{\sigma_{s,\text{nr}}^2}{\sigma_{v,\text{nr}}^2} = \frac{h^H R_s h}{h^H R_v h}, \quad (258) \]
\[ \xi_{sr}(h) = \frac{\sigma_s^2}{\sigma_{s,\text{nr}}^2} = \frac{\sigma_s^2}{h^H R_s h}, \quad (259) \]

where
\[ \sigma_s^2, \sigma_v^2, \sigma_{s,\text{nr}}^2, \sigma_{v,\text{nr}}^2: \text{variances of the desired signal and noise before and after noise reduction} \]
\[ R_s, R_v: \text{covariance matrices of desired signal and noise.} \]
Simulations
Speech signal

- Female speaker uttering: "Why were you away a year, Roy?".
- $\omega_0$, $k$ and $L$ estimated with NLS estimator.
- Filter length, $M = 50$.
- Segment length $N = 200$.
- The iSNR is -10 dB to 10 dB in steps of 2.5 dB.
- The noise is white Gaussian.
- 50 Monte Carlo simulations.
Simulations
Results

![Graph showing PESQ score vs. iSNR (dB)]

- LCMV\(_c\)
- LCMV\(_h\)
- APES\(_c\)
- APES\(_h\)
Outline

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

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

Summary and Conclusion
Speech Intelligibility Prediction

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

Obtain multi-channel signal  Estimate parameters  Reconstruct clean speech  Intelligibility prediction

\[ x_0 \rightarrow \text{Short-time segmentation} \rightarrow \text{Pitch estimator} \rightarrow \text{Synthesize speech} \rightarrow \text{STOI estimation} \rightarrow d(n) \]

Noisy speech

Clean speech reconstruction

Desired target DOA

Harmonic model order estimator
Signal Model

Assume $K$ microphones records $N$ samples each:

$$x_k) = \beta_k Z D(k) \alpha + e_k$$

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

(260)

with

$$Z = [z(\omega_0) \cdots z(L\omega_0)],$$

$$z(l\omega_0) = [1 \ e^{jl\omega_0} \ \cdots \ e^{jl\omega_0(N-1)}],$$

$$D(k) = \text{diag}\{e^{-j\omega_0 f_s \tau_k}, \ldots, e^{-jL\omega_0 f_s \tau_k}\},$$

$$\alpha = [\alpha_1 \ \cdots \ \alpha_L]^T$$ (complex harmonic amplitudes),

$\omega_0, f_s$: fundamental and sampling frequencies,

$\tau_k, \beta_k$: delay and attenuation of speech between mics. 0 and $k$,

$L$: number of harmonics.
Maximum Likelihood Pitch estimation

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

$$\ln p(x_k; \psi) = -NK \ln \pi - N \sum_{k=0}^{K-1} \ln \sigma_k^2 - \sum_{k=0}^{K-1} \frac{\|e_k\|^2}{\sigma_k^2}, \quad (262)$$

where $\psi$ contains signal parameters.

Unknown, linear, parameters found iteratively using:

$$\hat{\alpha}_k = \left[ \sum_{k=0}^{K-1} \frac{\beta_k^2}{\sigma_k^2} D^H(k)Z^HZD(k) \right]^{-1} \sum_{k=0}^{K-1} \frac{\beta_k}{\sigma_k^2} D^H(k)Z^Hx_k, \quad (263)$$

$$\hat{\beta}_k = \frac{\text{Re}\{\alpha^H D^H(k)Zx_k\}}{\alpha D^H(k)Z^HZD(k)\alpha}, \quad (264)$$

$$\hat{\sigma}_k^2 = \frac{\|x_k - \beta_k ZD(k)\alpha\|^2}{N}. \quad (265)$$
Maximum Likelihood Pitch estimation

For known $\theta$ (i.e., $\tau_k$’s), the ML pitch estimator is then

$$\hat{\omega}_0 = \arg\min_{\omega_0} \sum_{k=0}^{K-1} \ln \| x_k - \hat{\beta}_k ZD(k) \hat{\alpha} \|^2. \quad (266)$$

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

$$\hat{s}_k(n) = Z\hat{\alpha}_k. \quad (267)$$

A final estimate can be obtained through averaging:

$$\hat{s}(n) = \frac{1}{K} \sum_{k=0}^{K-1} \hat{s}_k(n). \quad (268)$$
Experimental Setup
Experimental Results
Pitch Estimates and Reconstruction
Experimental Results
Intrusive vs non-intrusive STOI estimates

Figure: Results from PB-STOI using a ULA setup.
Experimental Results
Intrusive vs non-intrusive STOI estimates

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

Introduction

Statistical Speech Models

Model-based Pitch Estimation of Speech

Model-based Array Processing and Enhancement

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

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

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

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

Thanks to our collaborators:
- Søren Holdt Jensen
- Andreas Jakobsson
- Petra Stoica
- Tobias Lindstrøm Jensen
- Mathew Shaji Kavalekalam
- Charlotte Sørensen
- Jesper Boldt
- Sidsel Marie Nørholm
- Sam Karimian-Azari
- Liming Shi
I. SINGLE- AND MULTI-PITCH ESTIMATION


II. HARMONIC CHIRP ESTIMATION

III. SPEECH ENHANCEMENT


IV. TDOA, DOA, AND SOURCE LOCALISATION


[22] X. Qian and R. Kumaressan. “Joint Estimation of Time Delay and Pitch of Voiced Speech Signals”. In:


V. SPEECH AND AUDIO CODING


VI. Model Comparison


VII. Speech and Audio Modeling


References


X. SPECTRAL ESTIMATION AND FREQUENCY ESTIMATION


XI. ADAPTIVE FEEDBACK CANCELLATION


XII. PACKET-LOSS CONCEALMENT


XIII. SPEECH INTELLIGIBILITY


XIV. DATABASES