Capturing and Exploiting Multiscale Speech Dynamics
From Mathematical Models to Forensic Tools

Patrick J. Wolfe
(with SISL Members and MIT Lincoln Laboratory)

Statistics and Information Sciences Laboratory (SISL)
School of Engineering and Applied Sciences
Department of Statistics, Harvard University
Harvard-MIT Division of Health Sciences & Technology

TELECOM ParisTech Workshop
27 June 2008
Outline
Capturing and exploiting multiscale speech dynamics

1. Introduction and Research Goals
   - Statistics in Speech Signal Processing
   - Extracting Information from Communications Data
   - Statistical Modeling of Multiscale Speech Dynamics

2. Adaptive Short-Time Analysis-Synthesis for Enhancement
   - Short-Time Analysis of Speech
   - Adaptive Analysis-Synthesis Scheme
   - Statistical Adaptation Criteria
   - Application to Signal Enhancement

3. Conditionally Linear Gaussian Models for Formant Tracking
   - Estimating Resonances in the Vocal Tract
   - Models and Methods for Formant Tracking
   - Tracking Results for the VTR Database
Outline
Capturing and exploiting multiscale speech dynamics

1 Introduction and Research Goals
   - Statistics in Speech Signal Processing
   - Extracting Information from Communications Data
   - Statistical Modeling of Multiscale Speech Dynamics

2 Adaptive Short-Time Analysis-Synthesis for Enhancement
   - Short-Time Analysis of Speech
   - Adaptive Analysis-Synthesis Scheme
   - Statistical Adaptation Criteria
   - Application to Signal Enhancement

3 Conditionally Linear Gaussian Models for Formant Tracking
   - Estimating Resonances in the Vocal Tract
   - Models and Methods for Formant Tracking
   - Tracking Results for the VTR Database
Outline
Capturing and exploiting multiscale speech dynamics

1 Introduction and Research Goals
   - Statistics in Speech Signal Processing
   - Extracting Information from Communications Data
   - Statistical Modeling of Multiscale Speech Dynamics

2 Adaptive Short-Time Analysis-Synthesis for Enhancement
   - Short-Time Analysis of Speech
   - Adaptive Analysis-Synthesis Scheme
   - Statistical Adaptation Criteria
   - Application to Signal Enhancement

3 Conditionally Linear Gaussian Models for Formant Tracking
   - Estimating Resonances in the Vocal Tract
   - Models and Methods for Formant Tracking
   - Tracking Results for the VTR Database
The Role of Statistics in Speech Signal Processing
The Statistics and Information Sciences Laboratory (SISL) at Harvard

- **Statistics** is concerned with modeling, inference, and uncertainty quantification
  - ⇒ The discrepancy between *data* and *information*
- **Signal processing** encompasses many inference problems that arise in engineering and the physical sciences
  - ⇒ Natural connections to coding and complexity.
- Issues of *model elicitation* and *model fitting* are of great relevance to contemporary work in speech processing
  - ⇒ Uncertainty quantification is also key
- At SISL we focus on *effective model-based speech processing*
  - A wealth of information remains to be exploited; advances will come through better *mathematics, models, and algorithms*
Our ability to collect data far outpaces our ability to understand it
- Novel sensing capabilities in engineering
- High-throughput technologies in the sciences

Information sciences must shift from enabling data acquisition to enabling automated analysis and understanding
- Decision analysis and decision support
- Knowledge discovery and management
- Uncertainty quantification and robustness

At SISL we focus on the need for statistical inference the context of large communications data sets
- Innovations will come from cross-pollination between theory and application domains; this talk describes joint work with SISL and MIT Lincoln Laboratory members.
Two key points emerged from a recent yearlong study:
- The volume, variety, and velocity of data require novel and effective techniques to handle modern-day massive, heterogeneous datasets
- Sophisticated automated analysis techniques are required to identify and call attention to potential patterns that may otherwise go undetected in the “data deluge”

These describe both the crux of the challenge facing the community and the basis for our work at SISL

⇒ What is the set of questions that can be answered without word-level information, and what speech processing tasks will these answers allow us to accomplish?
Problem Space
Capturing and exploiting multiscale speech dynamics

- Natural *hierarchy* of problems
- Generative statistical models provide a *coherent framework*
  - Parametric representations
  - Anomaly & change detection
  - Uncertainty quantification
- Model *elicitation* and *fitting*

- **Objective I**: Accurately represent physical parameters that characterize speech signals effectively
- **Objective II**: Incorporate dynamics to capture information-carrying transitions in local stationarity
Recently (and perhaps always!) there has been an interest in characterizing the **temporal dynamics** of speech.

Broadly, we are interested in capturing the slowly time-varying nature of speech simultaneously at multiple scales (e.g., segmental, supra- and sub-segmental).

To **represent** speech dynamics we require a set of time-varying parameters that evolve on the scale of choice.
- Parametric Models: Source-filter and ARMA models, pitch
- Nonparametric Models: Transform-based methods

We first consider speech dynamics through adaptive short-time Fourier analysis and its application to speech enhancement.

We’ll subsequently consider parametric modeling of formants and their estimation from acoustic data.
# Outline

Capturing and exploiting multiscale speech dynamics

## 1 Introduction and Research Goals
- Statistics in Speech Signal Processing
- Extracting Information from Communications Data
- Statistical Modeling of Multiscale Speech Dynamics

## 2 Adaptive Short-Time Analysis-Synthesis for Enhancement
- Short-Time Analysis of Speech
- Adaptive Analysis-Synthesis Scheme
- Statistical Adaptation Criteria
- Application to Signal Enhancement

## 3 Conditionally Linear Gaussian Models for Formant Tracking
- Estimating Resonances in the Vocal Tract
- Models and Methods for Formant Tracking
- Tracking Results for the VTR Database
Information-carrying natural sound signals exhibit a large degree of structure, but some of this is obscured by a view of the time series itself.

Likewise the Fourier transform provides a global view of spectral content, but the timings and durations of individual frequencies are hidden in the complex phase.
Prominent features of information-carrying natural sound signals can be seen in these typical time-frequency surfaces.

Our ultimate goal is to leverage these features to model typical audio signals, in order to accomplish various signal processing tasks of interest.
Motivating Adaptive Short-Time Analysis
An underexplored mantra

- Fixed resolution short-time analysis cannot simultaneously resolve both transient events (e.g., plosives) and vowels well.
- Controlling the time-frequency resolution in a signal-adaptive way is desirable to avoid smearing transients while maximally preserving steady-state harmonic content.
- Adaptive short-time analysis methods have been proposed, but lack synthesis counterparts.

Figure: An example of how a fixed-resolution scheme (top) could be modified to achieve an adaptive time segmentation (bottom).
Adaptive Analysis Scheme

Basic Structure

- Compute fixed-resolution STFT using the shortest desired window, then iteratively merge neighboring windows based on some criterion for measuring change.
- If two short-time segments do not significantly differ from one another according to this metric, then we combine them.
- Combine the left \((w_l)\) and right \((w_r)\) windows according to:

\[
wm[n] = w_l[pL - n] + w_r[(p + 1)L - n]
\]

- The resultant representation is less redundant.
- We discuss two examples of adaptation criteria in the second part of the talk, for now we focus on synthesis.
Visualizing Adaptive Procedure
Iterative clustering of finest resolution scheme

Figure: An example of how a fixed-resolution scheme (top) could be modified to achieve an adaptive time segmentation (bottom).

- Scheme begins with finest temporal resolution desired
- Windows merged according to measure of nonstationarity
- The number of possible combinations is large, so a greedy strategy is employed
- Each window is grown left to right. Other approaches (e.g., tree-based search, dynamic programming) also possible
Synthesis from Adaptive STFT
Rely on well-known overlap-add (OLA) method of synthesis

- Given a set of STFT coefficients $X(pL, k)$, where $L$ is a time-decimation factor, the synthesis of a sequence $y[n]$ through the overlap-add method (OLA) is given by:

$$y[n] = \frac{L}{W(0)} \sum_{p=-\infty}^{\infty} \left( \frac{1}{N} \sum_{k=0}^{N-1} X(pL, k) e^{j \frac{2\pi}{N} kn} \right), \text{ where } W(0) = \sum_{n=-\infty}^{\infty} w[n]$$

- Perfect reconstruction achieved if the OLA constraint is met:

$$\sum_{p=-\infty}^{\infty} w[pL - n] = \frac{W(0)}{L}$$

- After adaptive analysis, each selected window can be decomposed into a summation of windows that were used to compute the initial fixed-resolution STFT and satisfy the OLA constraint.

Thus, the OLA constraint is still satisfied and efficient resynthesis is possible via OLA.
Choosing an Adaptation Criterion
What type of change are we interested in capturing

- **Parametric Approaches**
  - Reliance on the source-filter model of speech
  - Speech presence/absence and voiced/unvoiced detection
  - Detect change in pitch
  - Detect changes in formant locations

- **Nonparametric Approaches**
  - Not specific to speech applications
  - Rely on transform-based methods
  - Connections to various concepts of local stationarity in statistics literature
  - Detect changes in autocovariance function
  - Detect changes in time-frequency concentration
Consider an autoregressive (AR) process of order $p$ with time-varying coefficients (TVAR($p$)) defined by:

$$x[n] = -\sum_{i=1}^{p} a_i[n]x[n - i] + w[n] \quad \text{and} \quad a_i[n] = \sum_{j=0}^{m} \alpha_{ij} f_j[n],$$

where the $f_j[n]$ are deterministic functions with $f_0[n] = 1$ and $w[n]$ is a white innovations sequence.

- If $\alpha_{ij} = 0$ for all $j > 0$, then the above reduces to the usual AR($p$) model.

- Smoothness of functions $f_j[n]$ controls the amount of change that can be captured by this model.

- Model can be used to capture the movement of the vocal tract resonances.
Consider the following hypothesis test:

\[ H_0 : \alpha_{ij} = 0 \quad \text{for all} \quad j > 0 \]
\[ H_1 : \alpha_{ij} \neq 0 \quad \text{for at least one} \quad j > 0 \]

- Under the null hypothesis the AR coefficients are constant. Under the alternate, they vary according to the functions \( f_j[n] \)

- This fits nicely into a generalized likelihood ratio test framework, though this requires parameter estimates under both hypotheses

- Under the null hypothesis, we use any of the classical estimators of the AR coefficients
- Under the alternate hypothesis we use the classical least-squares estimators proposed by Grenier, Hall, and others
Suppose a Formant Suddenly Moved 500 Hz
Can we detect this non-stationarity?

Consider a synthetic signal with the center frequency of a pole changing by 500 Hz halfway through.

Use TVAR(2) model with each trajectory expanded in a 5-element Legendre polynomial basis.

The estimated trajectory can be used to detect formant movement.

Our goal is detection, not estimation.

Figure: Original signal, true and estimated coefficient trajectories along with a pole-zero plot are shown above.
Detection Performance of GLRT

For each frequency step size, we considered 11 different signal lengths from 5 – 50 ms in 5 ms increments.

Resolution: each FFT bin of a 2.5 ms segment of a 16 Khz waveform has width of 400 Hz.

Larger changes and those occurring over longer intervals are (of course) easier to detect.

Methodology can be extended to deal with many formants and a periodic source.

**Figure:** ROC Curves for different size frequency jumps (across plots) and signal lengths (within plot)
Adaptation via TVAR Model
Finding maximally flat formant regions

- We can apply these hypothesis tests to see if two neighboring segments should be merged as follows:
  - Compute the GLRT test statistic on the joined segment
  - If we fail to reject the null, the segments can be merged
  - If we reject the null then the segments are not merged

- Suppose two neighboring segments contained one formant that switched frequencies, as in the previous example

- Then depending on the size of the change and length of the signal, we would decide whether to merge the associated windows

- Why are formant transitions important?
  - Have been shown to be important for intelligibility
  - Rate of transitions different for plosives vs. vowels
  - If used in the context of enhancement (e.g., Wiener filtering), spectral peaks will not be smeared
Adaptation via Time-Frequency Concentration 1/2
We focus on this nonparametric approach from now on

- Consider a signal with STFT, $X_p(t, \omega)$, where the parameter $p$ indexes the length of the underlying analysis window.

- Time-frequency concentration as a function of time is given by:

$$C(t, p) = \frac{\iint |X_p(\tau, \omega)w(\tau - t)|^4 d\tau d\omega}{(\iint |X_p(\tau, \omega)w(\tau - t)|^2 d\tau d\omega)^2},$$

where $w(t)$ is a window centered at 0 that localizes the measure.

- Maximizing $C(t, p)$ favors short-time segments that place most of the energy in the smallest region of the time-frequency plane.

- Shorter windows chosen at time-localized transients (plosives).

- Longer windows are chosen around vowels and voiced consonants which tend to be spread over time but localized in frequency.
Adaptation via Time-Frequency Concentration 2/2
Adapting method of Jones and Baraniuk

- Jones & Baraniuk (1994) adapt the instantaneous window length in order to maximize time-frequency concentration
- In particular, to construct an adaptive STFT from $M$ fixed-resolution STFTs, $M$ spectrograms are computed and interpolated onto the finest time-frequency lattice
- The window length corresponding to the STFT with the maximal local concentration is selected

We provide a synthesis technique for this method as shown earlier. For analysis, we iteratively merge neighboring windows based on a modified concentration measure (where $x_w[n]$ is a short-time segment):

$$C(x_w) = \frac{\sum_m \left| \sum_n x_w[n] e^{-j2\pi nm/N} \right|^4}{\left( \sum_m \left| \sum_n x_w[n] e^{-j2\pi nm/N} \right|^2 \right)^2}$$

- If the time-frequency concentration of the concatenated short-time segment exceeds the maximum of the concentrations of the individual short-time segments, then we combine them
Synthetic Waveform Example 1/2
Start simple to make sense of things

- Left: Fixed-resolution (top) and adaptive (bottom) short-time analysis of a simple synthetic waveform.
- Right: Spectrograms of the synthetic waveform denoised by the fixed (top) and adaptive (bottom) schemes using 512-sample Hamming windows with 50% overlap.

- SNR Gain is 4.7 dB
Synthetic Waveform Example 2/2
Impact that adaptation may have in denoising applications

Figure: Performance comparison of the fixed and adaptive-resolution systems in enhancement of a synthetic waveform

- SNR gains for each scheme are computed over a range of input SNRs (top) panels the length of the analysis window for the STFT is varied in increments of 500 samples (bottom) while input SNR was fixed at 1 dB.

- The adaptive scheme achieves higher gains than the fixed scheme for a range of input SNRs and outperforms it regardless of what fixed window length is used.
Adaptive Segmentation of Speech Data
From synthetic to speech waveforms

- Left: Fixed resolution and adaptive short-time analyses of the phrase “and amazed” extracted from a TIMIT utterance.
- Right: Spectrograms of the utterance “piecemeal” observed in white Gaussian noise at 0 dB SNR and enhanced using the fixed-resolution (top) and adaptive (bottom) schemes.
- Note the smeared onset of the plosive “p”
Reduction of Musical Noise
Evaluating the value of adaptation in enhancement

- Noise reduction is typically achieved through the spectral attenuation of each short-time segment and so selecting the appropriate temporal resolution is crucial.

- It is evident that longer windows are chosen for the voiced segments while shorter segments are chosen for transients.

- As has been previously observed, segmenting stationary and transient regions are properly segmented, then the variance of the estimates of the speech spectrum is reduced, thereby reducing the amount of musical noise present.

- We show that the adaptive scheme can reduce the amount of musical noise without relying on inter-frame smoothing, which also implies great potential to preserve transients.

- This is one of many implied improvements.
Objective measures are not very reliable

- Adaptive system using a baseline 10-ms window is compared to a fixed-resolution STFT system using 20-ms windows with 50% overlap in a series of listening tests

- Speech Data
  - Voiced data (collected at MIT/LL) consisting of 2 male and 2 female speakers uttering two different voiced sentences
  - Four phonetically balanced TIMIT sentences (2 male / 2 female speakers)

- Listening Setup
  - Voiced and TIMIT data presented at 0 and 5 dB SNR
  - Presentation: Clean/Noisy/A/B/A/B
  - Each of 10 listeners were asked which of the two enhanced waveforms (adaptive vs. fixed), if any, had musical noise
Informal Listening Tests
Adaptation reduces musicality

<table>
<thead>
<tr>
<th></th>
<th>Adaptive System</th>
<th>No Preference</th>
<th>STFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced</td>
<td>74.3%</td>
<td>21.3%</td>
<td>4.4%</td>
</tr>
<tr>
<td>TIMIT</td>
<td>33.8%</td>
<td>53.7%</td>
<td>12.5%</td>
</tr>
</tbody>
</table>

Table: Reduction in perceived residual “musical noise”

- Adaptive system using a baseline 10-ms window is compared to a fixed-resolution STFT system using 20-ms windows.
- Table shows summary of listener preferences from tests for the enhancement scheme (material presented at 0 and 5dB SNR).
- Results averaged over 10 expert listeners from MIT Lincoln Laboratory, and 4 voiced utterances (2 male, 2 female).
- Significant majority indicated that less musical noise was present in the waveforms enhanced by the adaptive system.
Summary
Adaptive Short-Time Analysis-Synthesis

- Locally stationary nature of speech motivates adaptive short-time analysis methods
- Previous approaches do not treat the synthesis counterpart to adaptive analysis schemes
- We proposed an adaptive short-time analysis-synthesis procedure based on a measure of time-frequency concentration
- Explored application to enhancement of speech signals
- Reduction in musical noise was observed due to better estimates of the short-time speech magnitude spectrum
Outline
Capturing and exploiting multiscale speech dynamics

1 Introduction and Research Goals
   - Statistics in Speech Signal Processing
   - Extracting Information from Communications Data
   - Statistical Modeling of Multiscale Speech Dynamics

2 Adaptive Short-Time Analysis-Synthesis for Enhancement
   - Short-Time Analysis of Speech
   - Adaptive Analysis-Synthesis Scheme
   - Statistical Adaptation Criteria
   - Application to Signal Enhancement

3 Conditionally Linear Gaussian Models for Formant Tracking
   - Estimating Resonances in the Vocal Tract
   - Models and Methods for Formant Tracking
   - Tracking Results for the VTR Database
Vocal tract resonances (VTRs) characterize a physical system and “always exist” (i.e., have frequency values) both during speech activity and when the mouth is closed.

The word “formant” is typically used to refer to the peaks of a smoothed power spectrum of the acoustic waveform.

During voiced sounds with no constriction of the vocal tract, the formants coincide with the VTRs, but in the presence of constrictions (e.g., consonant production) there may be no spectral peaks corresponding to the VTRs – rendering them unobservable.

Hence we model the (hidden) dynamics of vocal tract resonances and estimate them based on observations of cepstral data.

Key difference between “does not exist” and “is not observed”
Traditional Approaches

Peak picking and root finding

- There is a large literature on VTR/formant tracking, though many algorithms take the following form:
  1. Compute magnitude spectrum for every short-time frame
  2. Compute smooth approximation to spectrum by fitting a low-order all-pole model
  3. Find spectral peaks either by root finding (derivative of AR polynomial) or peak picking
  4. Optionally, connect the dots frame-by-frame using a penalty for connecting peaks that are far apart

- Provocation: This class of methods has no underlying model and expensive and “error-prone” subroutines

- Recently Deng et al. (2006) have constructed a hand-corrected VTR database containing 516 phonetically balanced utterances from TIMIT, providing a common ground for comparison...
A statistical modeling approach is preferable

Three pieces: Acoustic Data, Extracted Parameters, Relating Extracted Parameters to Vocal Tract Resonances.

Vocal tract resonances and observed spectra for short-time frames should be parameterized so that:
  - Dynamics of speech are naturally represented
  - Small perturbations of the VTRs are readily observable
  - Analytic closed-form relationships are desirable

We will show that these criteria are at least reasonably satisfied by our approach, which is based on a model of Deng et al. (2006)

We argue there is something natural about this approach (at least from a mathematical point of view). We use this as a starting point and add two key extensions
State Space Model Definition 1/2

The hidden process

- Each VTR is modeled by a second-order digital resonator parameterized by a frequency $f_k$ and bandwidth $b_k$.
- Assume spectral envelope is characterized by $K$ resonators.
- Then frame at $t$, the spectral envelope is represented by a vector $x_t \in \mathbb{R}^{2K}_+$:
  
  $$x_t = (f_1, \ldots, f_K, b_1, \ldots, b_K)^T.$$

- VTR dynamics are modeled via a discrete-time Gauss-Markov process:
  
  $$x_{t+1} = F x_t + w_t,$$

  where $F \in \mathbb{R}^{2K \times 2K}$ is the state transition matrix and $w_t \in \mathbb{R}^{2K}$ is a white noise sequence satisfying $\mathbb{E}(w_i w_j^T) = Q \delta_{ij}$, with $Q$ denoting the process noise covariance.

- Note: $F$ is typically taken to be diagonal.
At each frame \( t \), we have the first \( N \) LPC cepstral coefficients
\[ y_t = (y_t[1], \ldots, y_t[N])^T \]
observed in additive white Gaussian noise.

The \( n^{th} \) coefficient \( y_t[n] \) and the state vector \( x_t \) are related by:
\[
y_t[n] = \frac{1}{n} \sum_{k=1}^{K} \exp \left( \frac{\pi n}{f_s} x_t[k+K] \right) \cos \left( \frac{2\pi n}{f_s} x_t[k] \right),
\]
where \( f_s \) is the sampling frequency.

This defines a vector-valued nonlinear mapping \( h : \mathbb{R}^{2K} \rightarrow \mathbb{R}^N \) from the vector \( x_t \) to the vector \( y_t \) of LPC cepstral coefficients and the resulting observation equation is given by:
\[
y_t = h(x_t) + v_t,
\]
where \( v_t \in \mathbb{R}^N \) is a white noise sequence satisfying \( \mathbb{E}(v_i v_j^T) = R \delta_{ij} \), the process and observation noise sequences are assumed to be uncorrelated.
Supposing that an acoustic waveform contains $T$ frames, we are interested in two quantities:

$$p(x_t | y_1:t) \text{ and } p(x_t | y_1:T),$$

the filtering and smoothing densities respectively.

Uncertainty of Speech Presence: The model-based approach allows us to censor the likelihood by augmenting the state vector $x_t$ with a binary indicator variable $I_t$ for each formant.

Correlation of Formant Trajectories: Off-diagonal terms in the matrix $F$ can be used to represent correlatedness by fitting an order-one vector autoregressive model.
A telling example from the VTR database

- The VTR trajectories are shown in black, with the extended Kalman smoother (top, average RMSE 107.5 Hz) and WaveSurfer output (bottom, average RMSE 214.3 Hz), both shown in white.

- The regions marked with a purple line in the top plot indicate regions of speech absence in which the tracks were coasted.

“They own a big house in the remote countryside.”
Formant Tracking Results 2/3
Error in speech regions and entire waveform

<table>
<thead>
<tr>
<th>Formant Number</th>
<th>Root Mean Square</th>
<th>Error Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entire Utterance</td>
<td>Speech Presence</td>
</tr>
<tr>
<td></td>
<td>(Hz)</td>
<td>(Hz)</td>
</tr>
<tr>
<td></td>
<td>(%)</td>
<td>(%)</td>
</tr>
<tr>
<td>1</td>
<td>90.93</td>
<td>52.31</td>
</tr>
<tr>
<td></td>
<td>39.32%</td>
<td>22.62%</td>
</tr>
<tr>
<td>2</td>
<td>103.0</td>
<td>59.18</td>
</tr>
<tr>
<td></td>
<td>30.55%</td>
<td>17.56%</td>
</tr>
<tr>
<td>3</td>
<td>147.3</td>
<td>99.79</td>
</tr>
<tr>
<td></td>
<td>32.74%</td>
<td>22.18%</td>
</tr>
</tbody>
</table>

Table: Observed reduction in average root mean square error relative to WaveSurfer, taking the VTR database as ground truth. Errors are computed both for the entire utterance (left) and conditioned on speech presence (right).

Is there enough information in the acoustic signal to extract the hand-corrected trajectories? This motivates the need for additional evaluation through synthesis.
The effect of modeling formant cross-correlation on the percentage reduction of root mean square error relative to a benchmark method, accounting for regions of speech presence.
Summary
Conditionally linear Gaussian models for formant tracking

- We have considered a model-based approach to the problem of vocal tract resonance estimation.
- We have proposed two important extensions to the EKF first described by Deng:
  - Uncertainty of speech presence
  - Correlation of formant trajectories
- Model and methods were evaluated on a recently constructed database of hand-corrected VTR trajectories.
- Overall results look promising relative to those produced by Wavesurfer.
Summary
Capturing and exploiting multiscale speech dynamics

Prabahan Basu, Daniel Rudoy, Daniel Spendley (SISL, Harvard) and Thomas Quatieri (MIT Lincoln Laboratory) contributed to the research presented in this talk, which was supported by DARPA and NSF:

- **Speech Enhancement (ICASSP 2008)**
  - TF Concentration as a measure of local stationarity
  - Noise suppression that adapts to speech dynamics

- **Formant Tracking (Interspeech 2007)**
  - Vector autoregressions to exploit inter-formant correlation
  - State-space modeling to exploit inter-frame correlation

- **If this perspective interests you...**
  - New NSF-DMS Focused Research Group for *Overcomplete Representations with Incomplete Data*; workshops in ’08–’09
  - Special Issue of *Statistica Sinica* (outgrowth of 2006 Graybill Workshop on Multiscale Methods and Statistics)