SUPERVISED SPEECH PROCESSING BASED ON GEOMETRIC ANALYSIS

PhD Seminar

Ronen Talmon

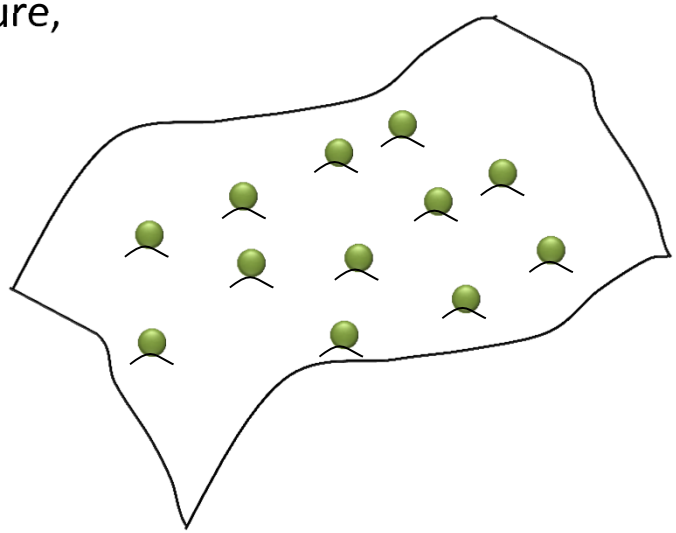
Supervised by Prof. Israel Cohen and Prof. Sharon Gannot

June 29, 2011

Outline

• Modern Data Analysis

- Data points in a high-dimensional space
  - Images, songs, finance and neuroscience data
- The data cannot fill up the high-dimensional space “uniformly”
  - Usually, the space dimensionality is arbitrary chosen by the user or the acquisition system
  - The data lies on a low-dimensional structure, conveying its intrinsic degrees of freedom



• Manifold Learning

- Popular methods
  - Kernel PCA [Schoelkopf et al., 98']
  - ISOMAP [Tenenbaum et al., 00']
  - Locally linear embedding (LLE) [Roweis & Saul, 00']
  - Laplacian eigenmaps [Belkin & Niyogi, 01']
  - Hessian eigenmaps [Donoho & Grimes, 02']
  - **Diffusion maps** [Coifman & Lafon, 04'] and **diffusion geometry**
- Common outline
  - Build a graph from the data using a specially-tailored metric
  - Find a parametric representation of the graph

• Transient Interference Suppression

- Transient is an **abrupt or impulsive** sound followed by **decaying oscillations**, e.g. keyboard typing and door knocking
- Common single-channel speech enhancement algorithms do not suit the abrupt nature of transients
  - For example: spectral subtraction [boll, 79'], decision directed [Ephraim & Malah, 84'], LSA [Ephraim & Malah, 85'], OM-LSA [Cohen & Berdugo, 01']

Problem Formulation

$$Y(l,k) = X(l,k) + T(l,k) + U(l,k)$$

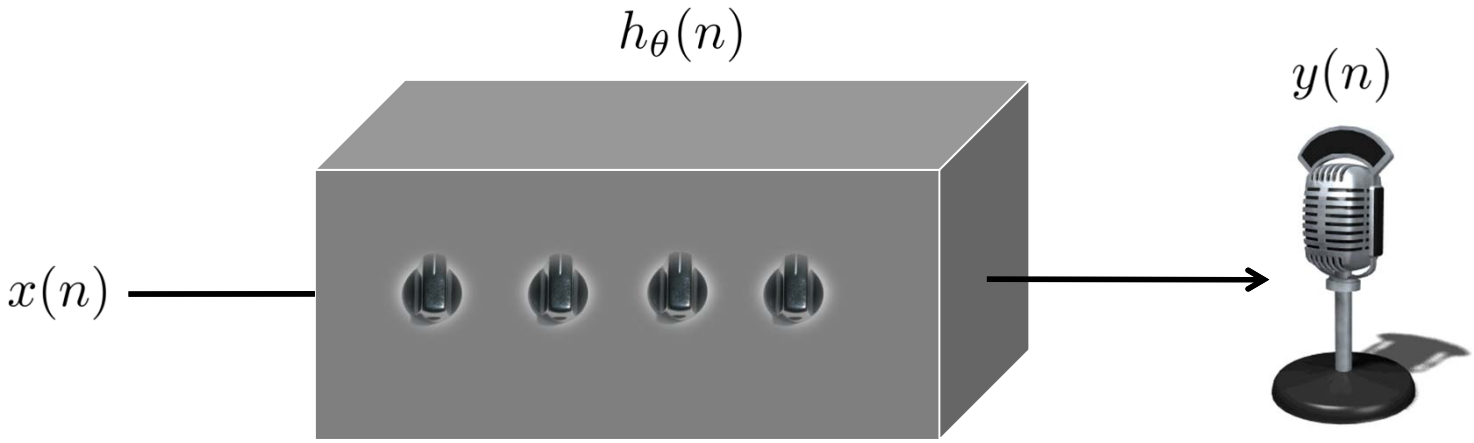
- Aim at estimating the speech given the noisy signal
- Assume the transient interference consists of repeated events

• Linear System Parameterization

- Any musical instrument is controlled by several independent parameters
- For example:
  - A flute is controlled by covering holes
  - A violin is controlled by the length of the strings
  - The speech system is controlled by the position of the tongue, lips, teeth, etc.

• Linear System Parameterization

- Each system can be seen as a “black box” configured by controlling parameters



  - We observe the output of the system
  - The sound depends on the system configuration
  - <u>Our goal:</u> recover the controlling parameters

• Diffusion Maps

Step I:

- Let $\{\mathbf{x}_i\}_i$ be a high-dimensional data set of samples
- Define a pair-wise affinity, e.g.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\varepsilon\right\}$$

Step II:

- We view $\{\mathbf{x}_i\}_i$ as nodes of an undirected symmetric graph
  - Two nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected by an edge with weight $k(\mathbf{x}_i, \mathbf{x}_j)$
- Construct a *random-walk* on the graph by normalizing the kernel

$$p(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)/d(\mathbf{x}_i)$$

with $d(\mathbf{x}_i) = \sum_j k(\mathbf{x}_i, \mathbf{x}_j)$

  - $p(\mathbf{x}_i, \mathbf{x}_j)$ represents the probability of transition from $\mathbf{x}_i$ to $\mathbf{x}_j$ in a single step

• Diffusion Interpretation

- Consider the probability of transition in a few steps
  - Raising the transition matrix of the random-walk P to the power of t
  - The probability of transition is usually larger in few steps than in a single step (brings the samples "closer")
  - Integrates more samples along the transition path
- For large data set and small kernel scale ($\varepsilon \to 0$) the **discrete** random-walk on the **graph** converges to a **continuous** diffusion process on the underlying **manifold** [Singer, 06']
  - **Utilize the analysis of diffusion processes**

• Embedding

- The transition matrix of the random-walk P has
  - A complete sequence of left and right singular vectors $\{\varphi_j, \psi_j\}$
  - Positive singular eigenvalues: $1 = \rho_0 > \rho_1 \geq \rho_2 \geq \ldots$

Diffusion Maps [Coifman & Lafon, '06]

Let $\Psi_t : \mathbb{R}^N \mapsto \mathbb{R}^\ell$ be the diffusion mappings of $\{\mathbf{x}_i\}$ into a new Euclidean space

$$\Psi_t(\mathbf{x}_i) = \left[\rho_1^t \psi_1(i), \cdots, \rho_\ell^t \psi_\ell(i)\right]^T$$

- A fast decay of the eigenvalues enables **dimensionality reduction**
- Notion of feature extraction
  - Ability to **capture the natural parameters of the data**

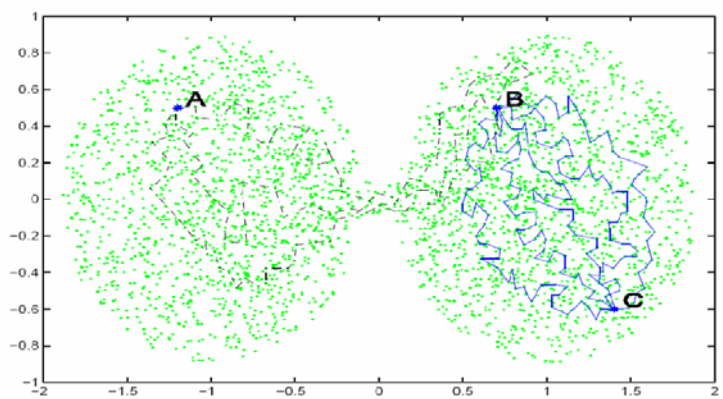
• Diffusion Distance

Diffusion Distance [Coifman & Lafon, '06]

Define a new affinity metric between any two vectors

$$D_t^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_k \left(p_t(\mathbf{x}_i, \mathbf{x}_k) - p_t(\mathbf{x}_j, \mathbf{x}_k)\right)^2 / \varphi_0(k)$$

- Describes the affinity in terms of graph connectivity
- Local structures and rules of transition are integrated into a global metric
- $\boxed{D_t(\mathbf{x}_i, \mathbf{x}_j) = \|\Psi_t(\mathbf{x}_i) - \Psi_t(\mathbf{x}_j)\|}$



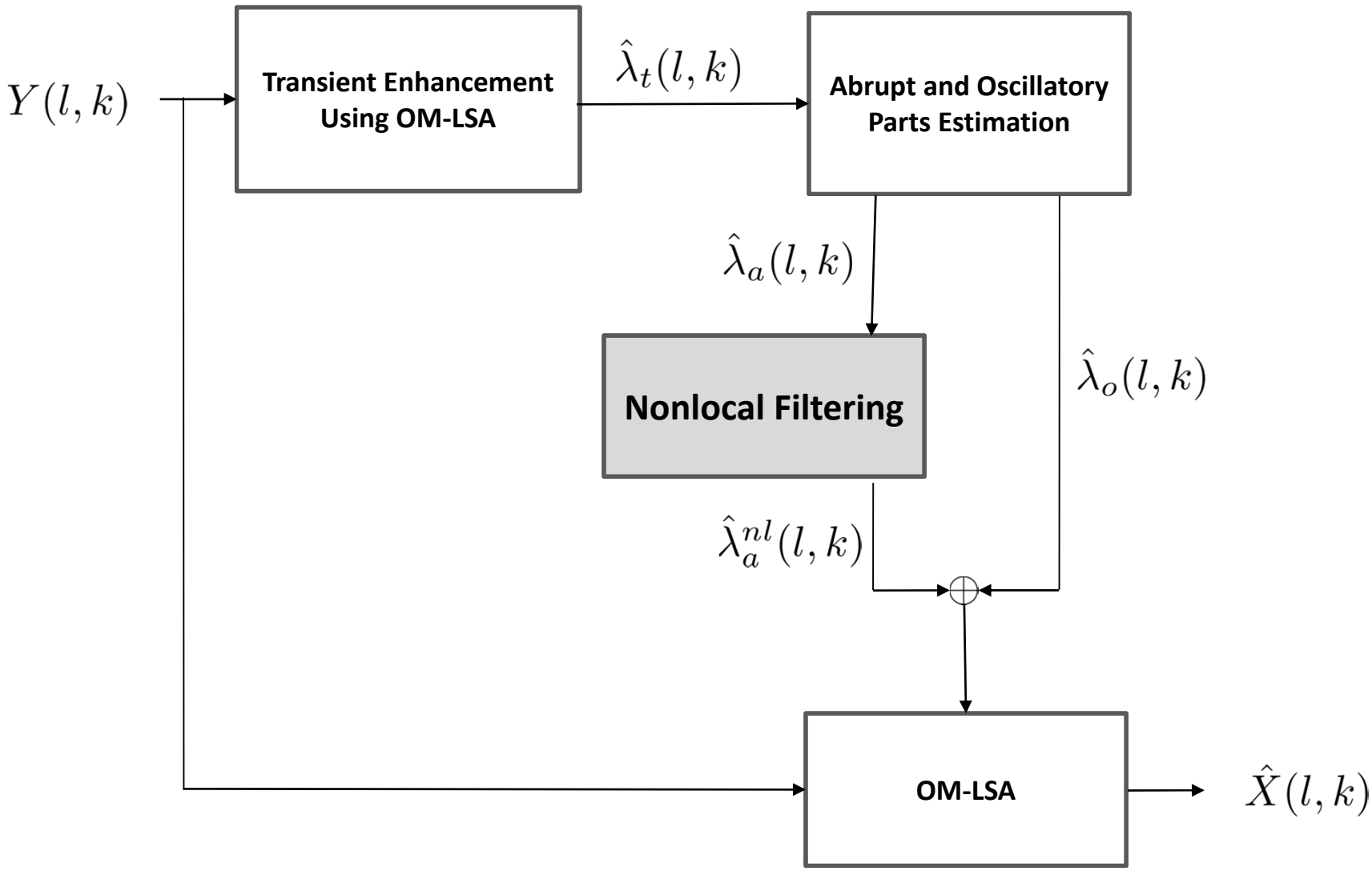
[Coifman & Lafon, 06']

• Overview

Main Components [Talmon, Cohen & Gannot, *submitted to* TASSP]

- **Statistical estimation**
  Exploits the rate of change of signals
- **Manifold learning**
  Capture the unique temporal and spectral features of transients
- **Nonlocal filtering**
  Exploit transient repetitions

• Overview


$Y(l,k)$
Transient Enhancement Using OM-LSA
$\hat{\lambda}_t(l,k)$
Abrupt and Oscillatory Parts Estimation
$\hat{\lambda}_a(l,k)$
$\hat{\lambda}_o(l,k)$
Nonlocal Filtering
$\hat{\lambda}_a^{nl}(l,k)$
OM-LSA
$\hat{X}(l,k)$

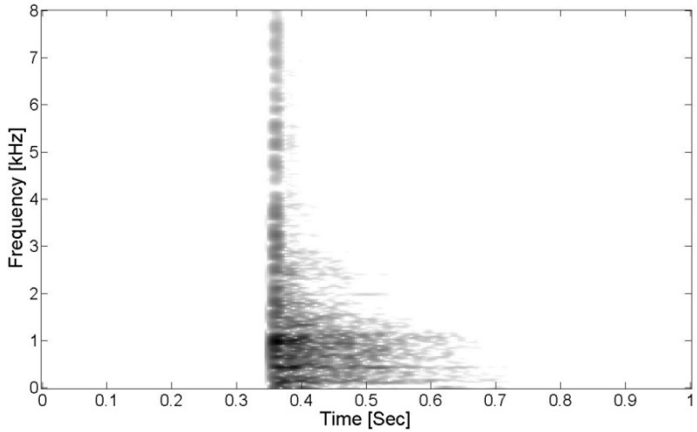
• Nonlocal Filtering

- Notation:
  - $\hat{\boldsymbol{\lambda}}_a(l) = \left[\hat{\lambda}_a(l,0), \ldots, \hat{\lambda}_a(l,N-1)\right]^T$
  - $\bar{p}(l,l')$ is a normalized affinity function between the spectra of frames

Nonlocal Filter Step

$$\boldsymbol{\lambda}_a^{nl}(l) = \sum_{l'} \bar{p}(l,l')\hat{\boldsymbol{\lambda}}_a(l')$$

- Each step can be interpreted as **averaging over similar time frames** according to $\bar{p}(l,l')$
- Let k denote the # of NL steps



• Nonlocal Filtering

- The initial estimate of the abrupt part spectrum

$$\hat{\lambda}_a(l,k) = \lambda_a(l,k) + \varepsilon(l,k) = \begin{cases} \boxed{\beta_a(k)B(l)} + \varepsilon(l,k), & l \in \mathcal{T} \\ \varepsilon(l,k), & l \in \bar{\mathcal{T}} \end{cases}$$

- **Nonlocal filtering:**
  - Temporal averaging which exploits subband dependencies
  - Assume that the kernel separates between transient and non-transient frames

$$\lambda_a^{nl}(l,k) = \begin{cases} \boxed{\beta_a(k)\bar{B}_{\mathcal{T}}} + \bar{\varepsilon}_{\mathcal{T}}(k), & l \in \mathcal{T} \\ \bar{\varepsilon}_{\bar{\mathcal{T}}}(k), & l \in \bar{\mathcal{T}} \end{cases}$$

- A “clean” version of the abrupt spectrum is obtained by:
  - Spectral subtraction suppresses the additive speech residual
  - The speech does not typically span the entire spectrum

• Clustering Using Diffusion Maps

- Graph construction:
  - View the vectors $\left\{\hat{\boldsymbol{\lambda}}_a(l)\right\}$ as nodes
  - Connect the nodes $\hat{\boldsymbol{\lambda}}_a(l)$ and $\hat{\boldsymbol{\lambda}}_a(l')$ by an edge with weight

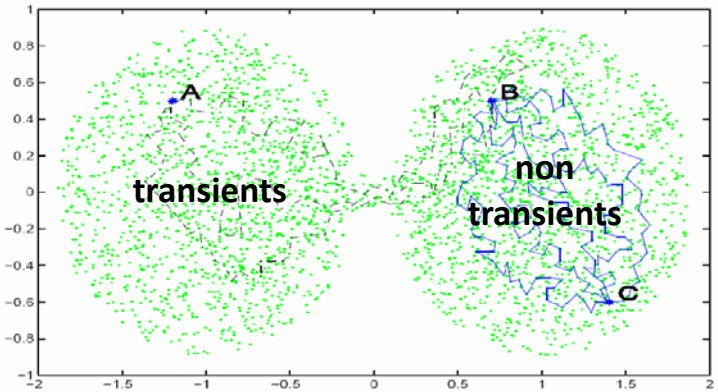
$$k(l,l') = \exp\left\{-\frac{\|\hat{\boldsymbol{\lambda}}_a(l) - \hat{\boldsymbol{\lambda}}_a(l')\|^2}{2\sigma^2}\right\}$$

- **Diffusion mapping**:
  Let $\Psi_t : \mathbb{R}^N \mapsto \mathbb{R}^\ell$ be the diffusion mappings of $\{\hat{\boldsymbol{\lambda}}_a(l)\}$ into a new Euclidean space

$$\Psi_t(\hat{\boldsymbol{\lambda}}_a(l)) = \left[\rho_1^t \psi_1(l), \cdots, \rho_\ell^t \psi_\ell(l)\right]^T$$

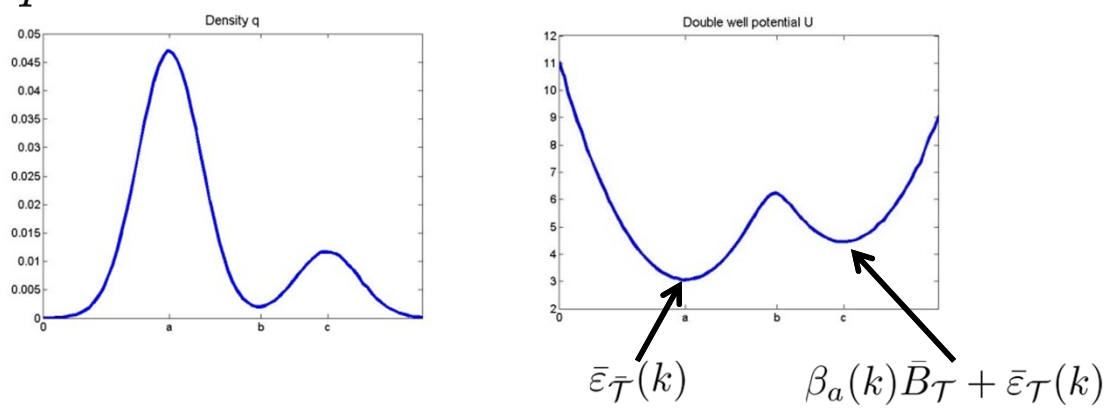
- Define a *new* Gaussian kernel based on the **diffusion distance**

$$\bar{k}(l,l') = \exp\left\{-\frac{\|\Psi_t(\hat{\boldsymbol{\lambda}}_a(l)) - \Psi_t(\hat{\boldsymbol{\lambda}}_a(l'))\|^2}{2\bar{\sigma}^2}\right\}$$



• Statistical Model

- Assume a <u>simple case</u>:
  - No divergence between transient events
  - "Stationary" speech
- The spectral variance *in a single band* has a two Gaussian mixture distribution q :



- Creating a "two-wells" potential $U = -2\log q$:
  - Each well represents the transient presence/absence.
- Full model described in [Talmon, Cohen & Gannot, TASSP 11']

• Diffusion Interpretation

- Using diffusion interpretation of NL filters [Singer et al., 09']
- The diffusion operator (Fokker-Planck):

$$\mathcal{L}f = \Delta f - \nabla f \cdot \nabla U$$

- The filtering depends on the characteristic of the diffusion process in two-wells potential: [Matkowsky & Schuss, 81'], [Gardiner, 04']
  - **Relaxation time** in each well – to properly attenuate the residual speech
  - **Mean first passage time (MFPT)** to exit a well – to approximate wrong identification of transients
- The # of NL steps: $k = \frac{\tau}{\varepsilon}$

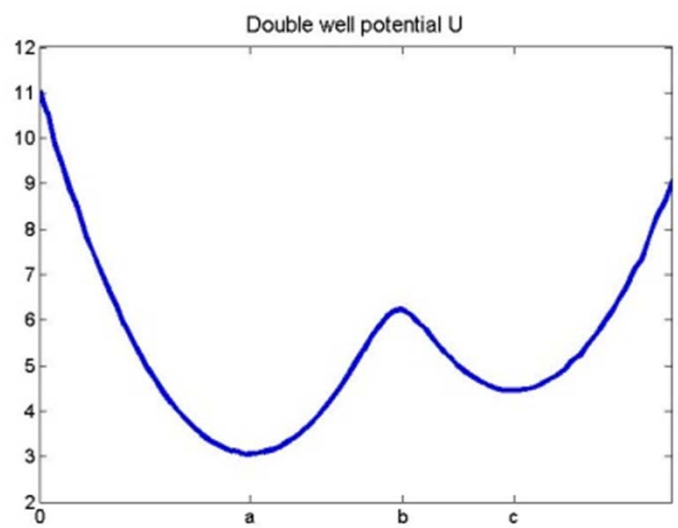
• Relaxation Time

- The characteristic relaxation time inside the well centered at a is given by the curvature of the bottom of the well

$$\tau_r \approx \frac{1}{U''(a)}$$

- The corresponding # of NL steps

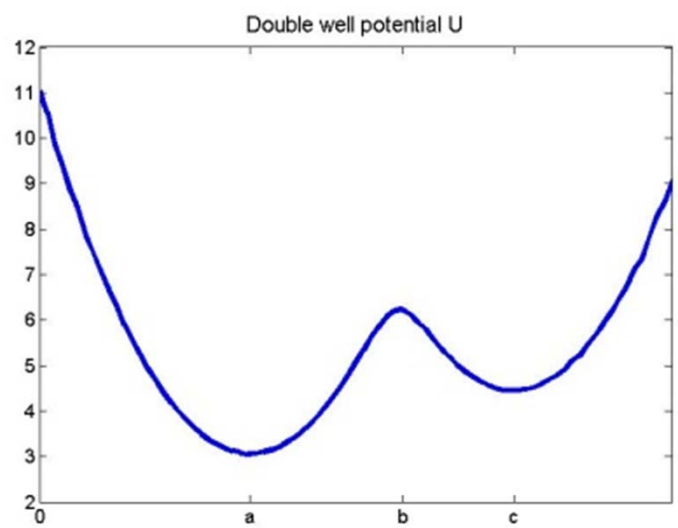
$$k_r = \frac{\tau_r}{\varepsilon} \approx \frac{1}{1-\lambda_1}$$



• Mean First Passage Time

- The mean first passage time from well c to a via the barrier b

$$\tau_{c\to a} \approx \frac{2\pi}{\sqrt{|U''(c)U''(b)|}} \exp\{U(b) - U(c)\}$$



• Results

Limitation

- The wells should be well separated in order to distinguish the presence/absence of transients

$$\tau_r \ll \tau_{c \to a}$$

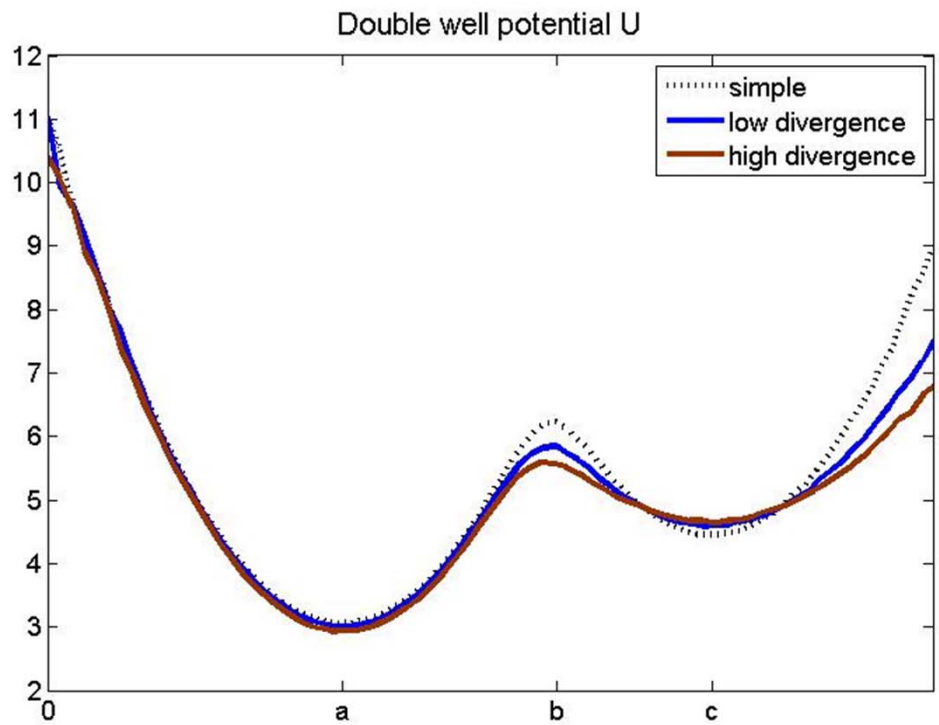
Parameter Setting

- The proper # of NL steps should satisfy

$$\tau_r < \varepsilon k \ll \tau_{c \to a}$$

• Trends

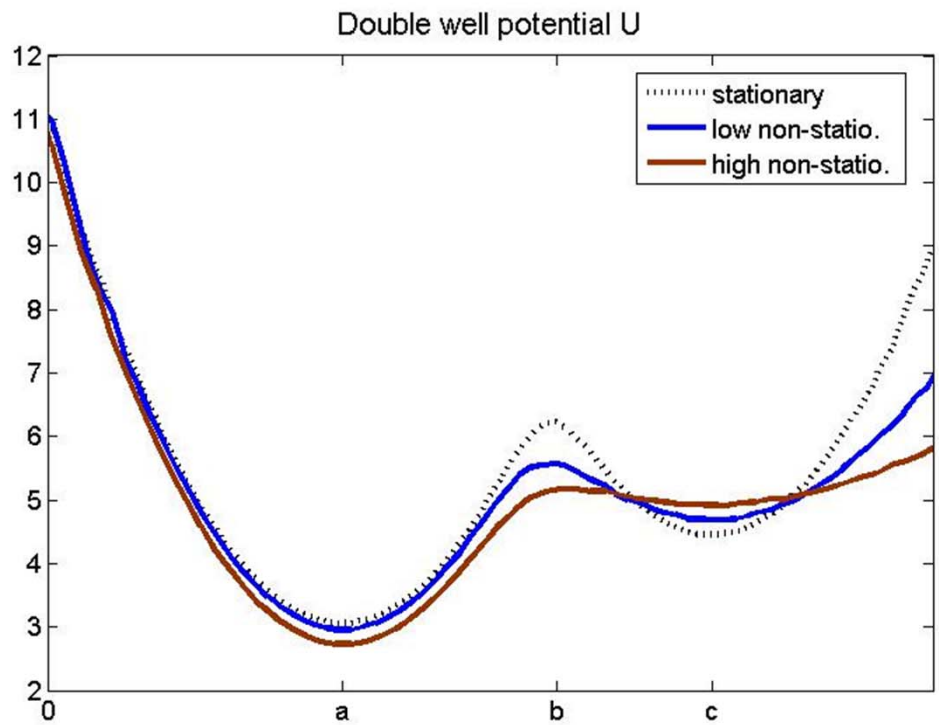
- As the transients become more diverse, the Gaussian distribution is "smeared"



- As illustrated, the right well becomes wider and shallower, and the barrier is lower

• Trends

- As the speech spectral envelope becomes more diverse, we obtain similar trends:



• Results (cont.)

NL Processing

Both cases result in **longer relaxation time** and **shorter mean first passage time**:

- More difficult to distinct the transients
- More steps should be employed

• Misidentification

Claim

Misidentification occurs in case a sample from one hypothesis exists in the opposite well

- In the simple (two Gaussians) case:
  - The probability of misidentifying a transient

$$\Pr(S(l,k) < b | l \in \mathcal{T}) = \Phi\left(\frac{b-c}{\sigma}\right)$$

  - The probability of falsely identifying a transient

$$\Pr(S(l,k) > b | l \in \bar{\mathcal{T}}) = 1 - \Phi\left(\frac{b-a}{\sigma}\right)$$

• High-dimensional Processing

- Intuitively, by comparing the spectral features of all freq. bins, we exploit the unique structure of the transient
- In terms of “diffusion”:
  - The bottoms of the high-dimensional potential wells are more separated
  - The barrier between the wells becomes higher
- As a result:
  - The misidentification probability decreases
  - # of NL steps is less restricted (the MFTP to exit a well increases)

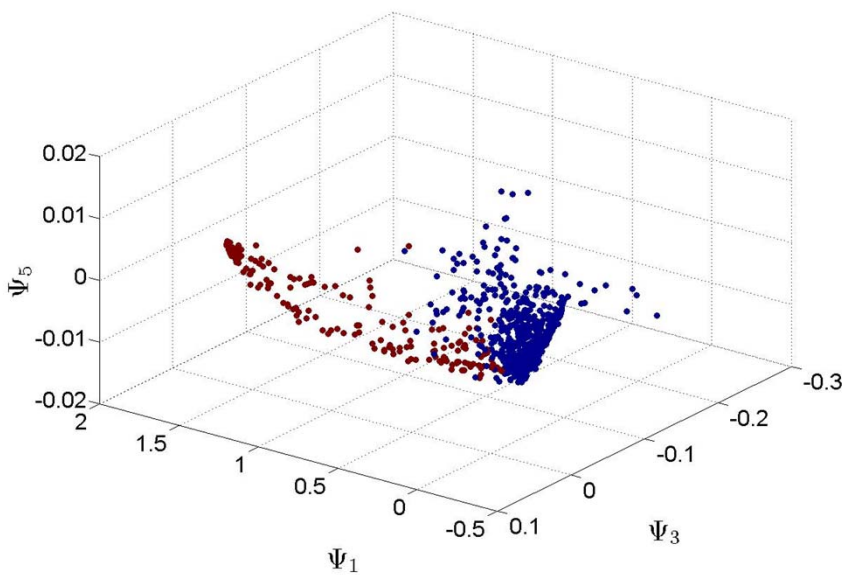
• Experimental Results

- Experimental results on simulated signals supported the results and trends [Talmon, Cohen & Gannot, TASSP 11']
- By-product - **spectral clustering**
  - Further research required for the general problem

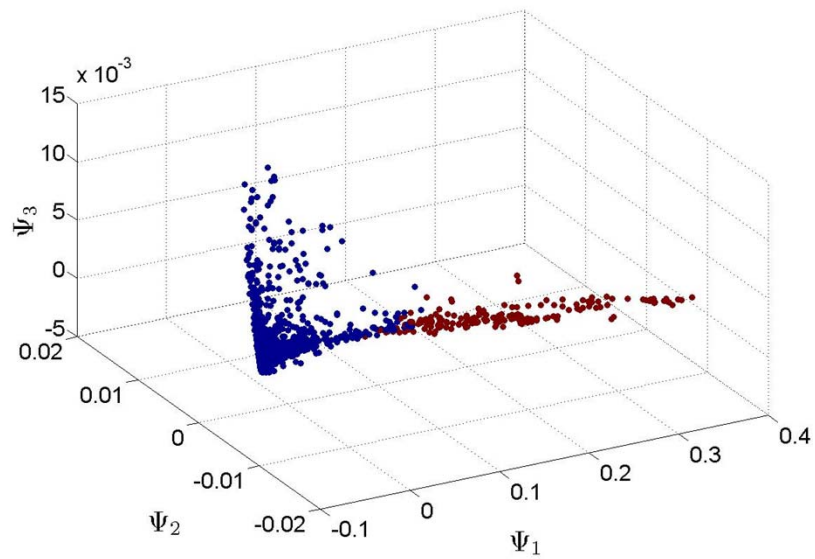
• Experimental Results

- <u>Setup:</u>
  - Recorded speech and transient signals sampled at 16 KHz
  - Speech signals are taken from TIMIT database, and recorded transient interferences are taken from an online free corpus
  - The additive stationary noise part is a computer generated white Gaussian noise with SNR of 20 dB
  - The length of each speech utterance and the corresponding transient interference is 20 sec
  - Such transient interference signal typically consists of 25 to 30 events

• Experimental Results



A scatter plot of the 1st, 3rd, and 5th coordinates of the diffusion map of speech contaminated by **kitchen pocks**

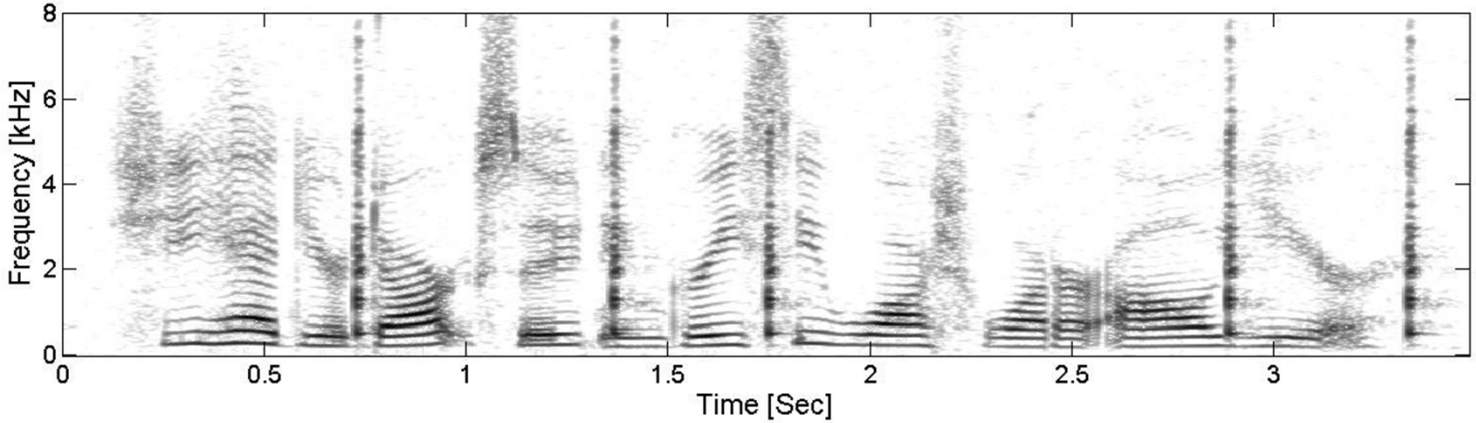


A scatter plot of the 1st, 2nd, and 3rd coordinates of the diffusion map of speech contaminated by **door knocks**

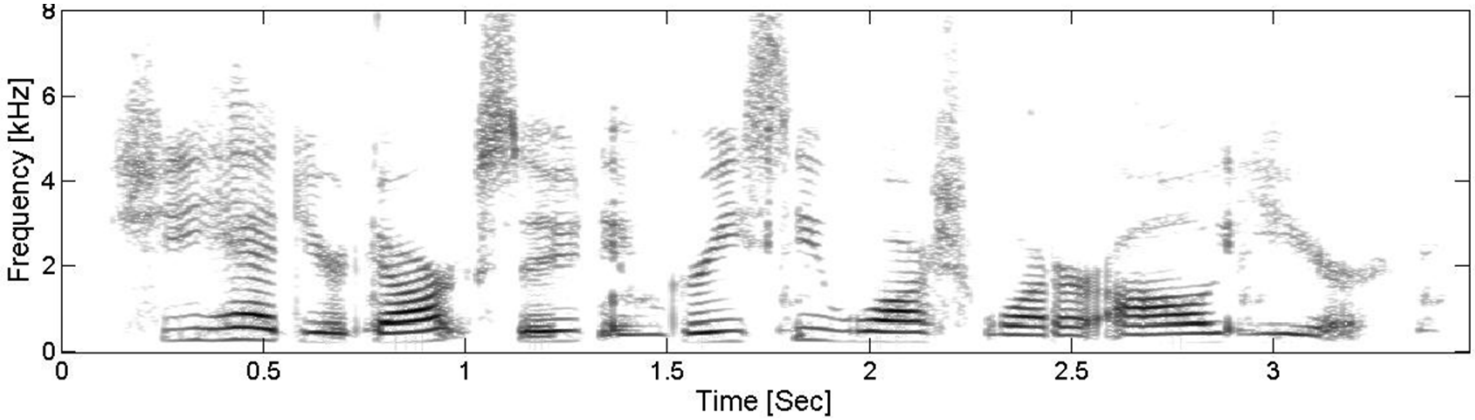
The frame content (either consists of transient interference or not) appears in different color (red and blue)

• Experimental Results

Speech contaminated by **metronome**



Enhanced speech



• Experimental Results

Using **Euclidean distance**:

Transient Type	SNR Improvement [dB]	LSD Improvement [dB]
Metronome	4.10	1.32
Door Knocks	3.89	1.14
Kitchen Pocks	8.05	1.35
Keyboard Typing	7.49	2.34
Household Clicks	4.53	1.25

Using **diffusion distance**:

Transient Type	SNR Improvement [dB]	LSD Improvement [dB]
Metronome	4.69	1.45
Door Knocks	4.83	1.54
Kitchen Pocks	9.25	2.05
Keyboard Typing	8.22	3.22
Household Clicks	4.65	1.42

Audio examples:

• Extensions and Future Work

- **Improved nonlocal operator:** [Talmon, Cohen & Gannot, ICASSP'10]
  - Use the operator $2P - P^2$: Similar eigenvectors and slower decay of spectrum
  - Enables better and more robust enhancement

- **Simultaneous suppression of transients:** [Talmon, Cohen & Gannot, ICASSP'11]
  - Utilize the ability of diffusion maps to distinguish between different transient types

• Extensions and Future Work

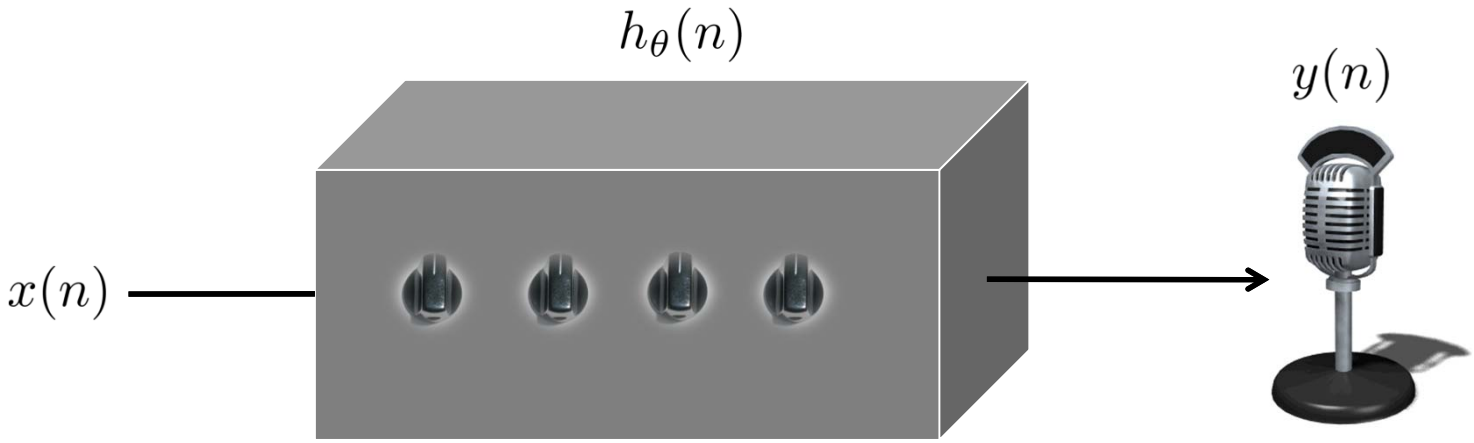
- **Supervised processing:**
  - Use training to capture the structure of the transients and acquire an intrinsic basis
  - Employ nonlinear “graph-based” filtering to enhance the transients by projecting the signal onto the space spanned by the acquired basis

• Discussion

- Combining the classical approach along with a modern data-driven approach
- Capturing the geometric structure of the signals enriches the a-priori assumed statistical model and enables good performance
  - Diffusion maps successfully captures the geometry of transients
- **Up next:**
  - Develop approaches to handle more “complicated” signals (e.g. acoustic, speech and music)

• Concept

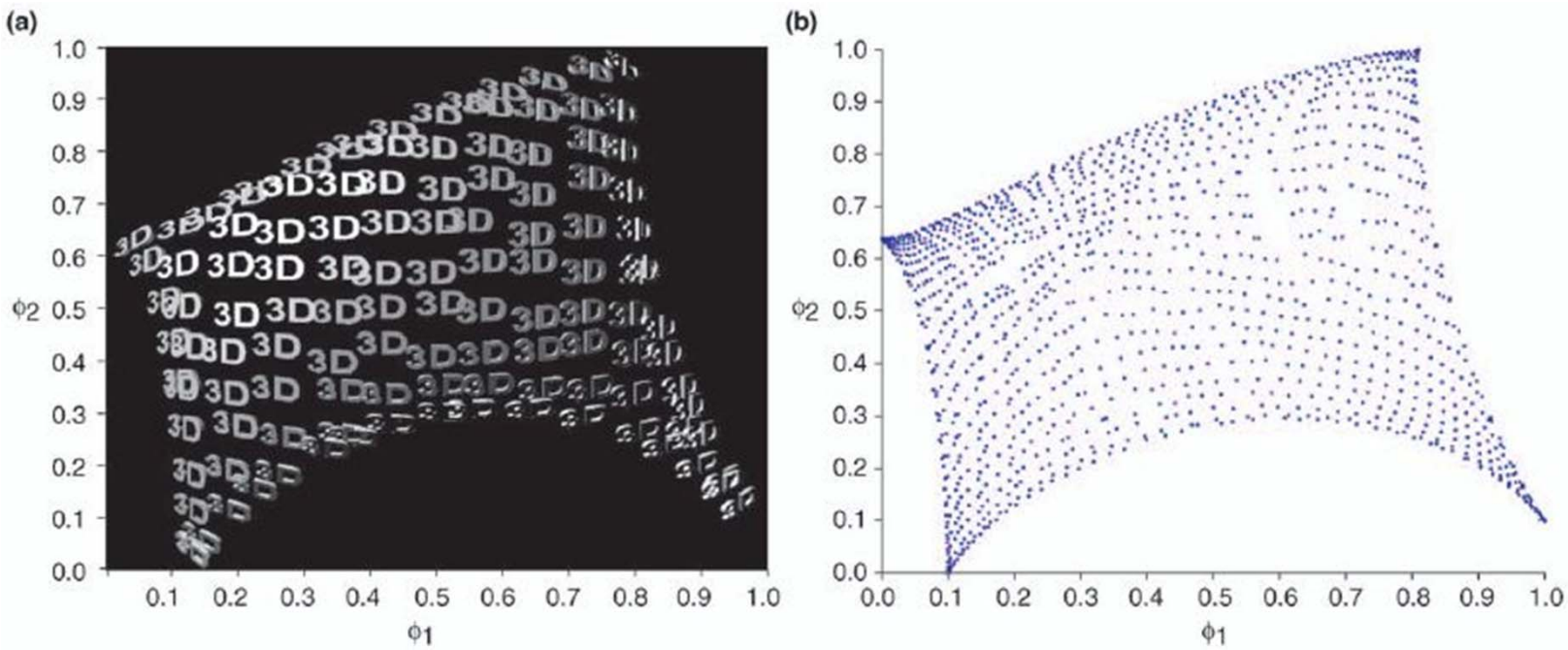
- Each system can be seen as a “black box” configured by controlling parameters



  - Our goal is to recover the controlling parameters of the system
  - Joint work with **R. Coifman and D. Kushnir**

• Parameter Recovery

- We show that the **parameters are conveyed by the eigenvectors** obtained via the diffusion framework [Talmon et al., *submitted to* TSP]
- <u>Example:</u>

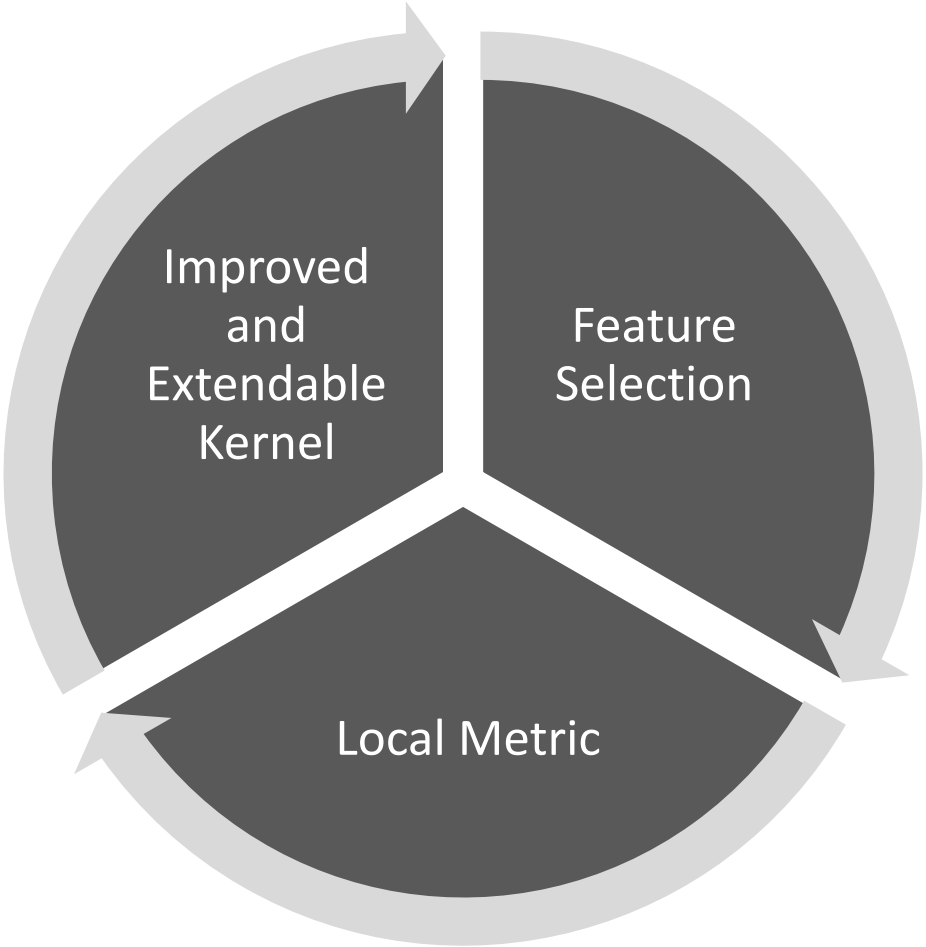


[Coifman & Lafon, 06']

• Temporal Evolution

- Example – violin recordings:
  - Several samples of the same tone
  - Each sample is slightly different due to different finger placements on the string
- Temporal evolution of the controlling parameters
  - Perceptual system variation (large scale)
  - Small fluctuations regime (small scale)
- Formulating the stochastic dynamic as Ito processes [Talmon et al, *submitted to* TSP]

• Main Components



• Feature Selection

- Use the **covariance function** of the observed signal
  - Circumvents the dependency on the realization of the input signal
  - Robust and generic
- Formulated as a nonlinear function of the parameters

  [Talmon et al., *submitted to* TSP]
  - Enables to present the stochastic dynamics using the Ito lemma
- <u>Example – AR process:</u>
  - Observe the AR system of order 1: $y(n) = x(n) - \theta y(n-1)$

    with zero mean white excitation and a controlling parameter (pole) $|\theta| < 1$
  - The covariance of the observable signal nonlinearly depends on the controlling parameter: $c_y(l) = \sigma_x^2 \theta^{-|l|}/(1-\theta^2)$
  - Viewed as a $c : \mathbb{R}^d \to \mathbb{R}^D$ function: $c(\theta) = \mathbf{c}$

• Local Metric

Euclidean distance in the parametric space [Singer & Coifman, 08']

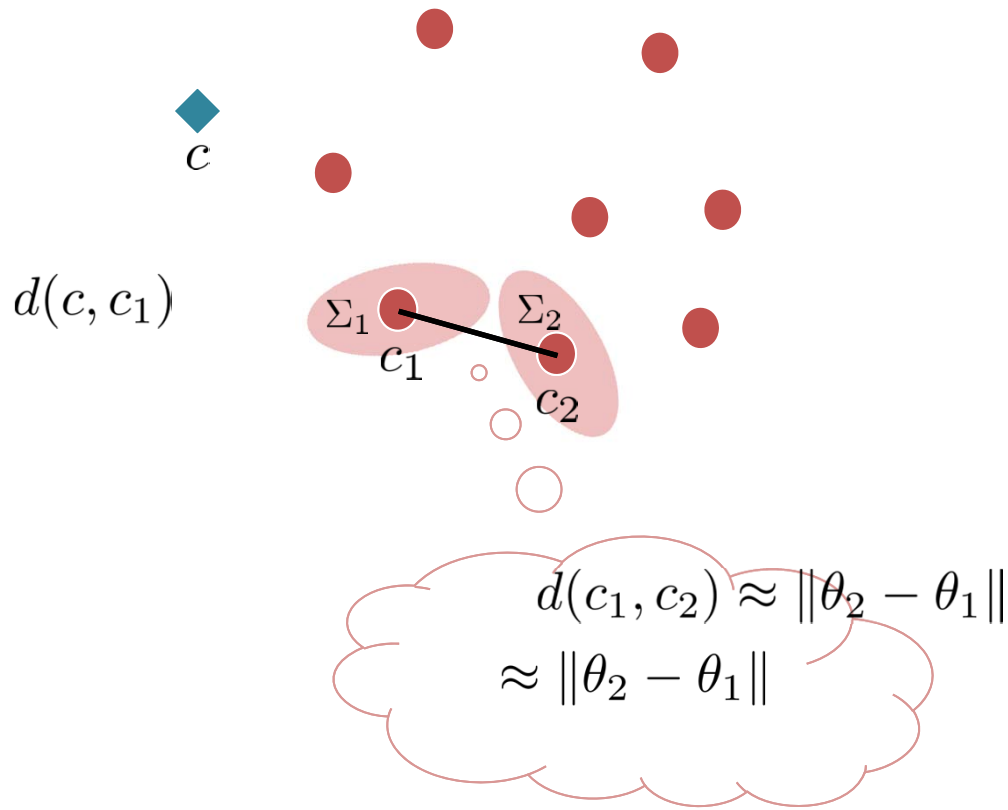
- Let θ_1, θ_2 be two points in the parametric space and let $c_1 = c(\theta_1), c_2 = c(\theta_2)$ be their mapping to the feature space
- The Euclidean distance in the parameter space can be approx. by

$$\|\theta_1 - \theta_2\|^2 = 2(c_2 - c_1)^T \left[\Sigma_1 + \Sigma_2\right]^{-1} (c_2 - c_1) + \mathcal{O}(\|c_2 - c_1\|^4)$$

where Σ_1 and Σ_2 are the local covariance matrices of the features

- The corresponding Gaussian kernel W converges to a diffusion operator
- No natural extension is available

• Local Metric



• Improved and Extendable Kernel

New Kernel [Kushnir at al., 11']

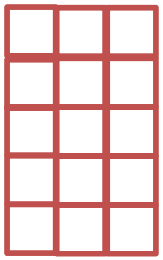
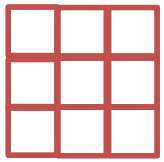
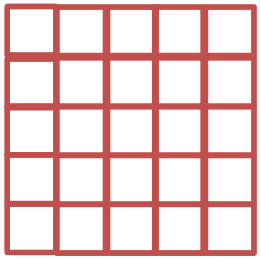
- Use the following metric

$$d(c, c_j) = \|c_j - c\|_j$$

Define affinity

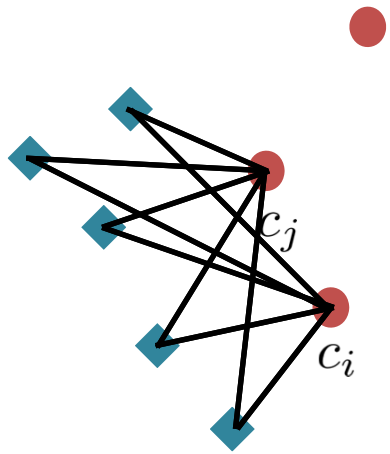
$$a(c, c_j) = \exp\left\{-\|c_j - c\|_j^2/\varepsilon\right\}$$

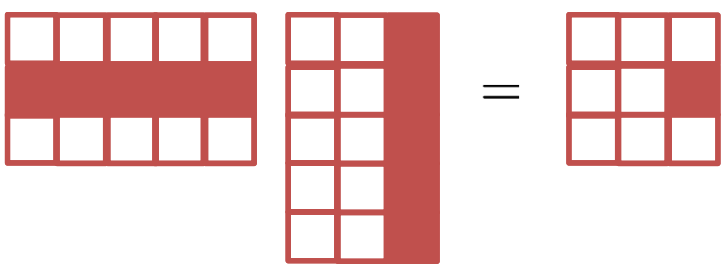
- The original kernel $A^T A = W$
- Extended kernel AA^T

$A =$  $A^T A =$  $AA^T =$ 

• Improved and Extendable Kernel

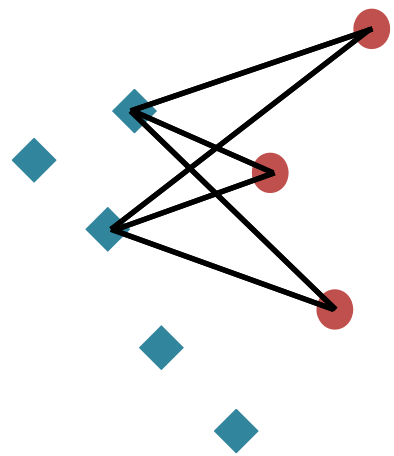
$A_j =$

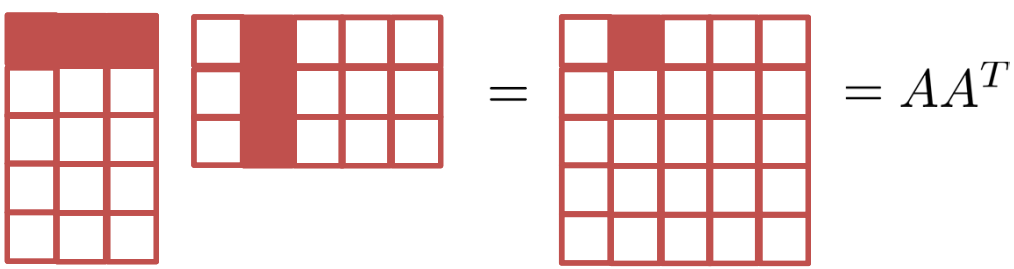




$$(A^T A)_{ij} = W_{ij}$$

• Improved and Extendable Kernel





$= AA^T$

• Improved and Extendable Kernel

- Let $\{\lambda_j, \varphi_j, \psi_j\}$ be the SVD of A
- We have

$$\psi_j = \frac{1}{\lambda_j} A \varphi_j$$

- In addition:
  - $\{\varphi_j\}$ are the eigenvectors of $A^T A = W$
  - $\{\psi_j\}$ are the eigenvectors of $A A^T$

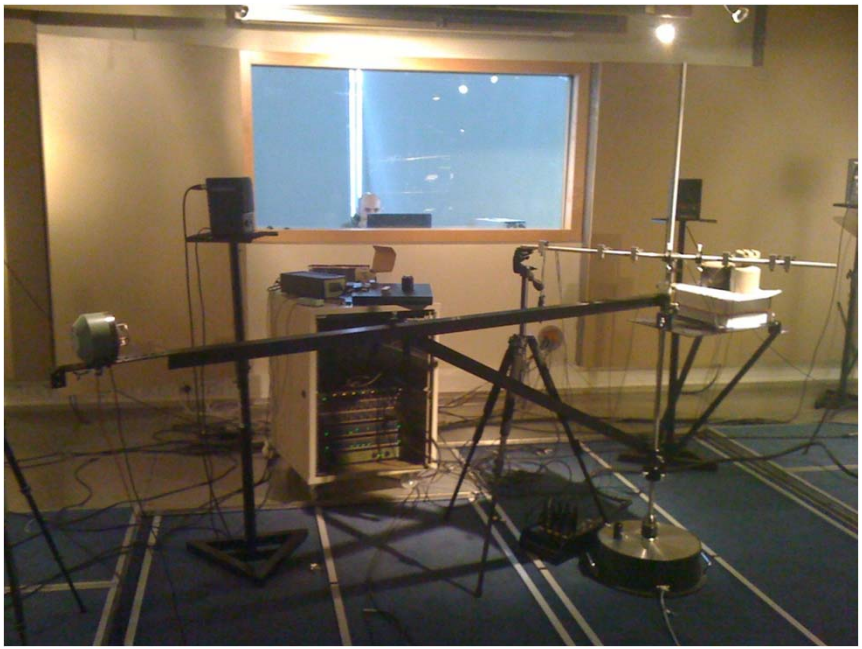
- The eigenvectors represent the independent parameters

• Experimental Results

- **Example I – AR process**
  - Accurate recovery of the poles of AR processes
  - Enables to capture the actual degrees of freedom when the poles are dependent
  - May be viewed as natural parameterization of speech signals
- **Example II – Acoustic channel**
  - Accurate recovery of the acoustic channel parameters: source and microphone positions, wall reflection coefficients, room dimensions
  - May be used for source localization tasks and acoustic system calibration

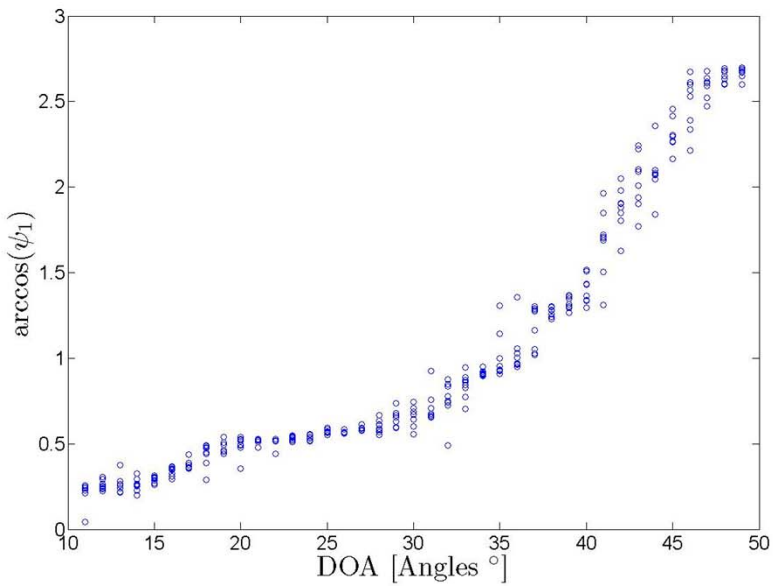
• Experimental Results

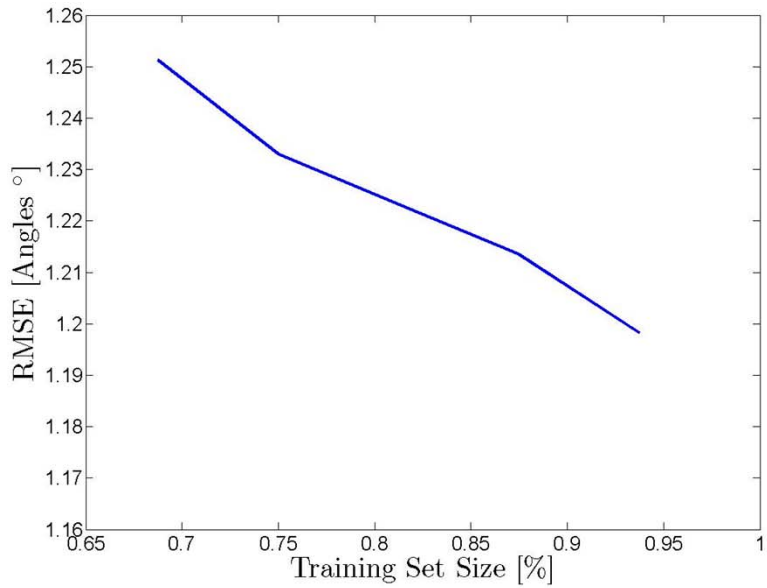
- Direction of arrival estimation



  - 60 DOA angles of 1° spacing
  - Room reverberation time of $T_{60} = 0.3\text{s}$

• Experimental Results





- The eigenvector recovers the DOA
- Accurate estimation of the DOA

• Conclusions

- **Combining geometric information is beneficial**
  - Better performance
  - Efficient and simple
  - Insight and analysis
- Presented methods enable capturing of
  - **Transient interferences**
  - **Acoustic parameters**
  - Artificial signals (**AR processes**)

• Future Work

- Further improve the methods
  - Support more complicated signals (e.g. **speech and music**)
- Develop filtering/processing framework
  - Supervised
  - Nonlinear
  - **Incorporate temporal information**
  - **Handle measurement noise**

THANK YOU