Computer Vision, Speech Communication & Signal Processing Group,
Intelligent Robotics and Automation Laboratory
National Technical University of Athens, Greece (NTUA)
Robot Perception and Interaction Unit,
Athena Research and Innovation Center (Athena RIC)

Nonlinear Aspects of Speech Production: Fractals and Chaotic Dynamics

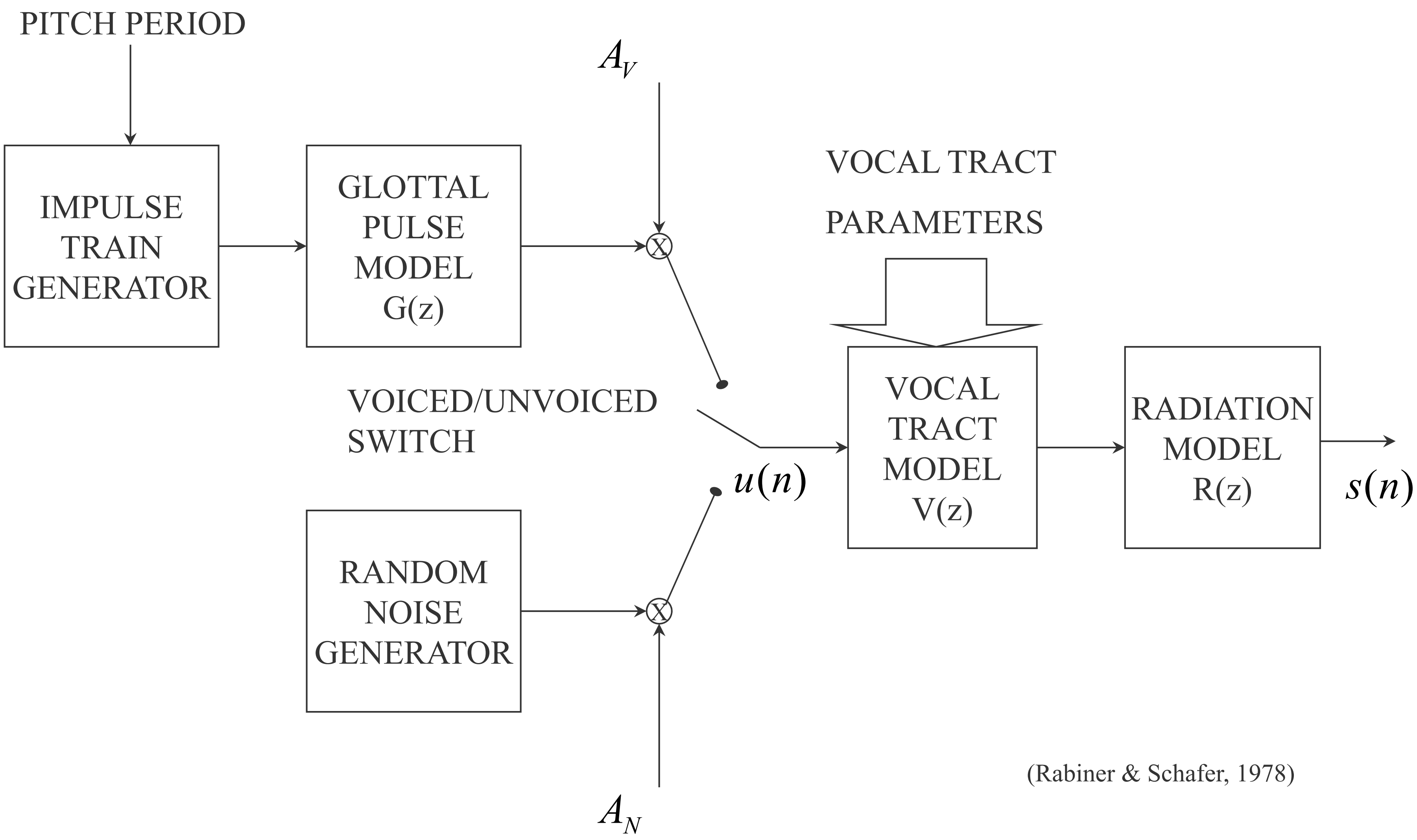
Petros Maragos

Summer School on Speech Signal Processing (S4P)
DA-IICT, Gandhinagar, India, 9-11 Sept. 2018

Outline

- Nonlinear Speech Processing → Turbulence: Fractals, Chaotic Dynamics
- Multiscale Fractal Dimensions of Speech Sounds
- Fractal Modulations for Fricative Sounds
- Chaotic Dynamics of Speech Sounds
- Algorithms for Speech Fractal & Chaos Analysis
- Application to Speech Recognition
- Application to Music Recognition

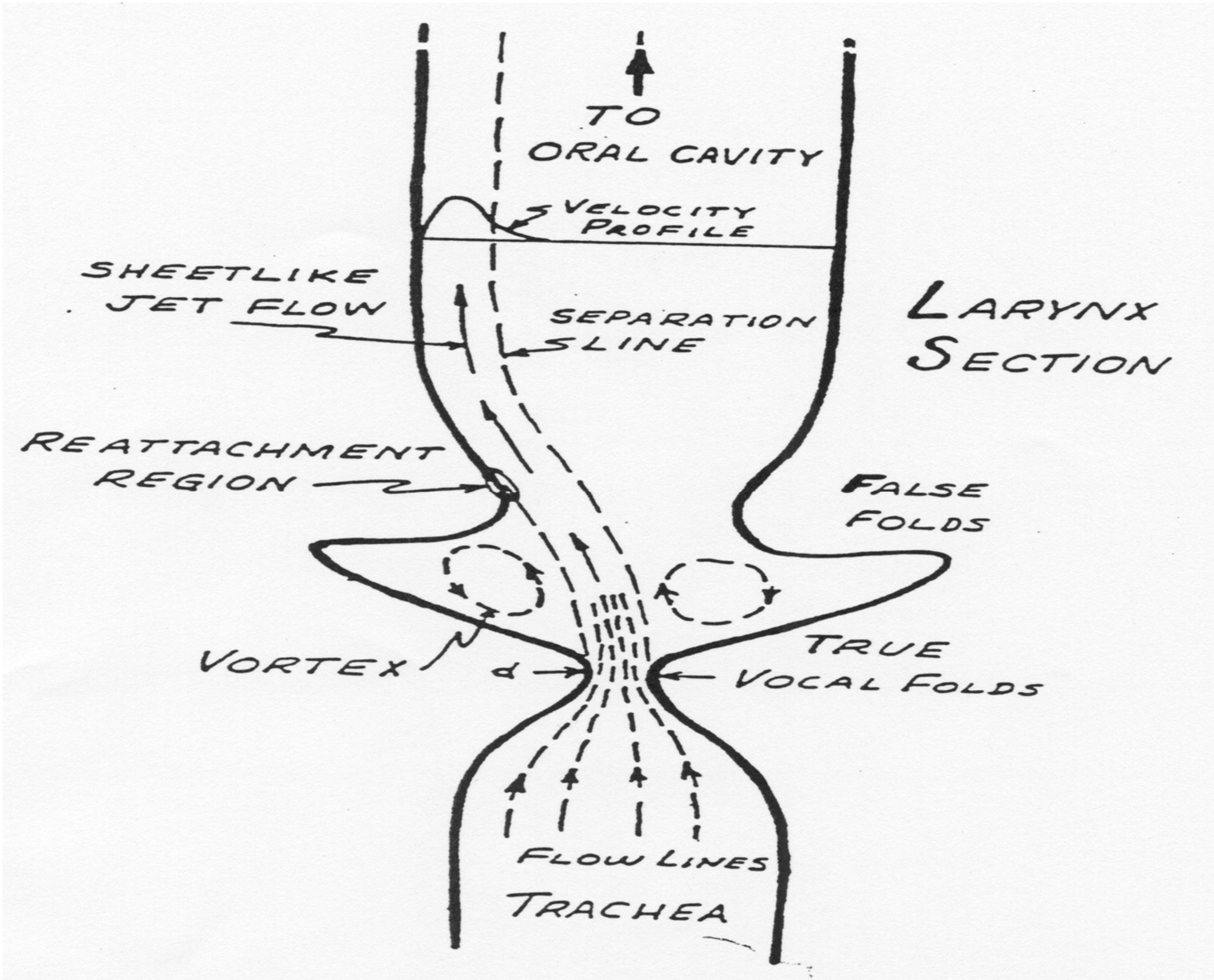
Linear Source-Filter Model



(Rabiner & Schafer, 1978)

Nonlinear Fluid Dynamic of the Vocal Tract

(Kaiser 1993)



Physics of Speech Airflow

- **airflow variables:** ρ **= air density;** p **= pressure**
  $\vec{u}$ **= 3D air particle velocity**

- **governing equations:**

mass conservation (continuity eqn): $$\frac{\partial \rho}{\partial t} + \nabla \cdot \left(\rho \vec{u} \right) = 0$$

momentum conservation (Navier-Stokes eqn):

$$\rho \left(\frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \nabla \vec{u} \right) = -\nabla p + \rho \vec{g} + \mu \left[\nabla^2 \vec{u} + \frac{1}{3} \nabla \left(\nabla \cdot \vec{u} \right) \right]$$

state equation: $p / \rho^{1.4} = \text{const.}$

- **time-varying boundary conditions**

Speech Aerodynamics

- **Reynolds number:** $\mathrm{Re} = \dfrac{\rho \times (velocity\ scale) \times (length\ scale)}{\mu}$

- **low viscosity μ ⇒ high Re**

  ⇒ inertia forces >> viscous forces

- **“aerodynamic” phenomena (Re >>1):**

  air jet, rotational motion, separated airflow,
  boundary layers, vortices, turbulence

- **experimental & theoretical evidence for nonlinear phenomena:**

Teager (1970s–1980s), Kaiser (1983 –), Thomas (1986),

McGowan (1988), Barney, Shadle & Davis (1999), ...

Vortices

- **vorticity:** $\vec{\omega} = \nabla \times \vec{u}$
- **VORTEX is a flow region of similar** $\vec{\omega}$
- **a vortex can be generated by:**
  - velocity gradients in boundary layers
  - separated air flow
  - curved geometry of vocal tract
- **dynamics of vortex propagation:**

$$\frac{\partial \vec{\omega}}{\partial t} + \vec{u} \cdot \nabla \vec{\omega} = \vec{\omega} \cdot \nabla \vec{u} + \left(\mu/\rho\right) \nabla^2 \vec{\omega}$$

$\vec{\omega} \cdot \nabla \vec{u} \rightarrow$ **vorticity twisting & stretching**

$\nabla^2 \vec{\omega} \rightarrow$ **diffusion of vorticity**

Nonlinear Speech Processing

- **Modulations**
- **Turbulence**
  - **Fractals**
  - **Chaos**

Turbulence

- **flow state with broad-spectrum rapidly-varying (in space and time) velocity and vorticity**
- **transition to turbulence is easier for higher Re flows**
- **eddies: vortices of a characteristic size ℓ**
- **Energy Cascade Theory (Richardson,1922)**

  (multiscale hierarchy of eddies)
- **5/3 spectral law (Kolmogorov, 1941):**

$$S(k,r) \propto r^{2/3} k^{-5/3}$$

$k = 2\pi / \ell =$ **wavenumber**

$r =$ **energy dissipation rate**

$S(k,r) =$ **velocity wavenumber spectrum**

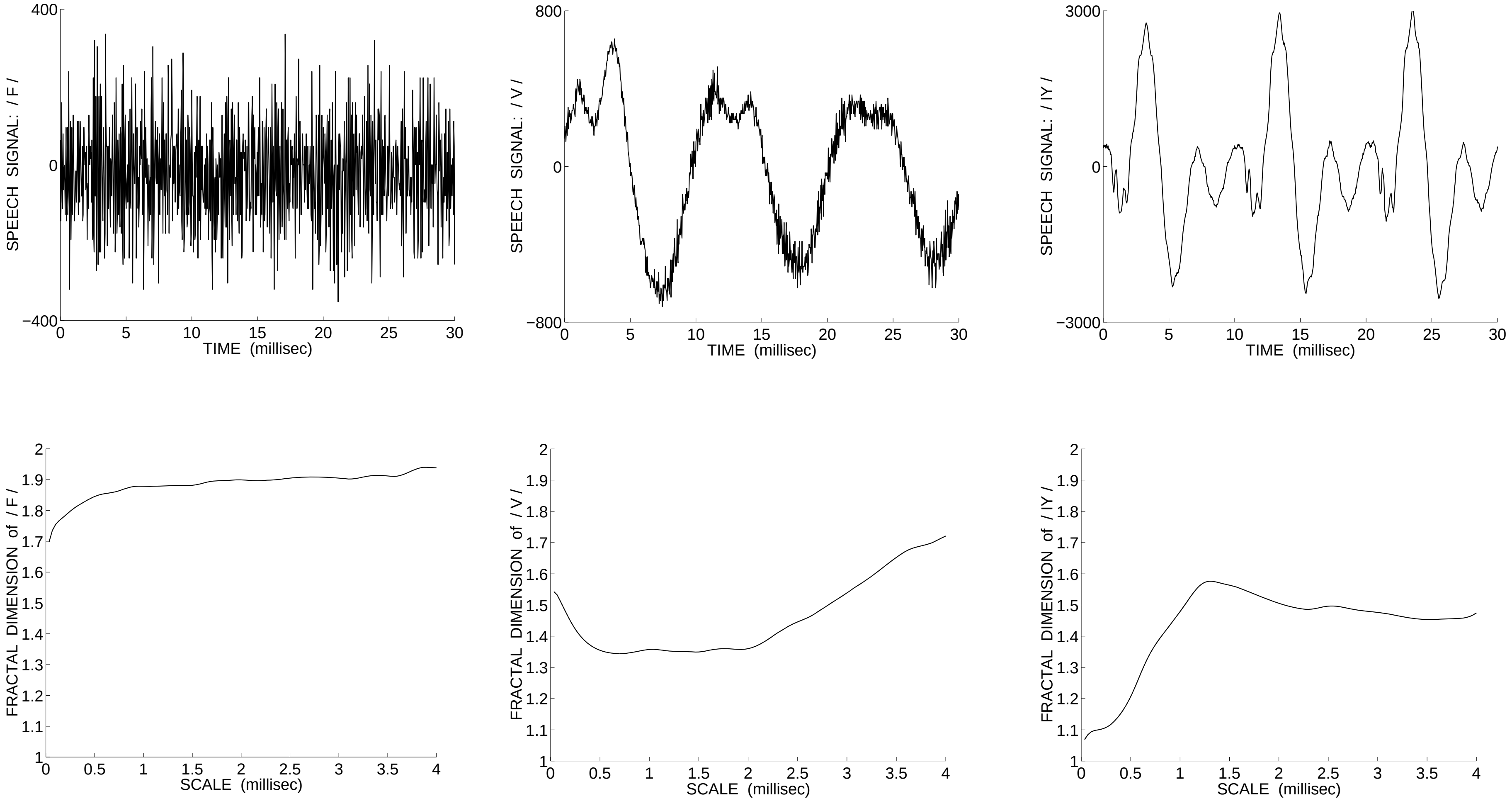
Turbulence, Fractals and Chaos

- **fractal geometry quantifies multiscale structures in turbulence**
- **Kolmogorov's 5/3 law**

$$Var\left[u(x)-u(x+\Delta x)\right] \propto (\Delta x)^{2/3}$$

- **we use fractal dimension to $\approx$ quantify "amount" of turbulence in speech**
- **chaos $\triangleleft \cdots \triangleright$ turbulence**

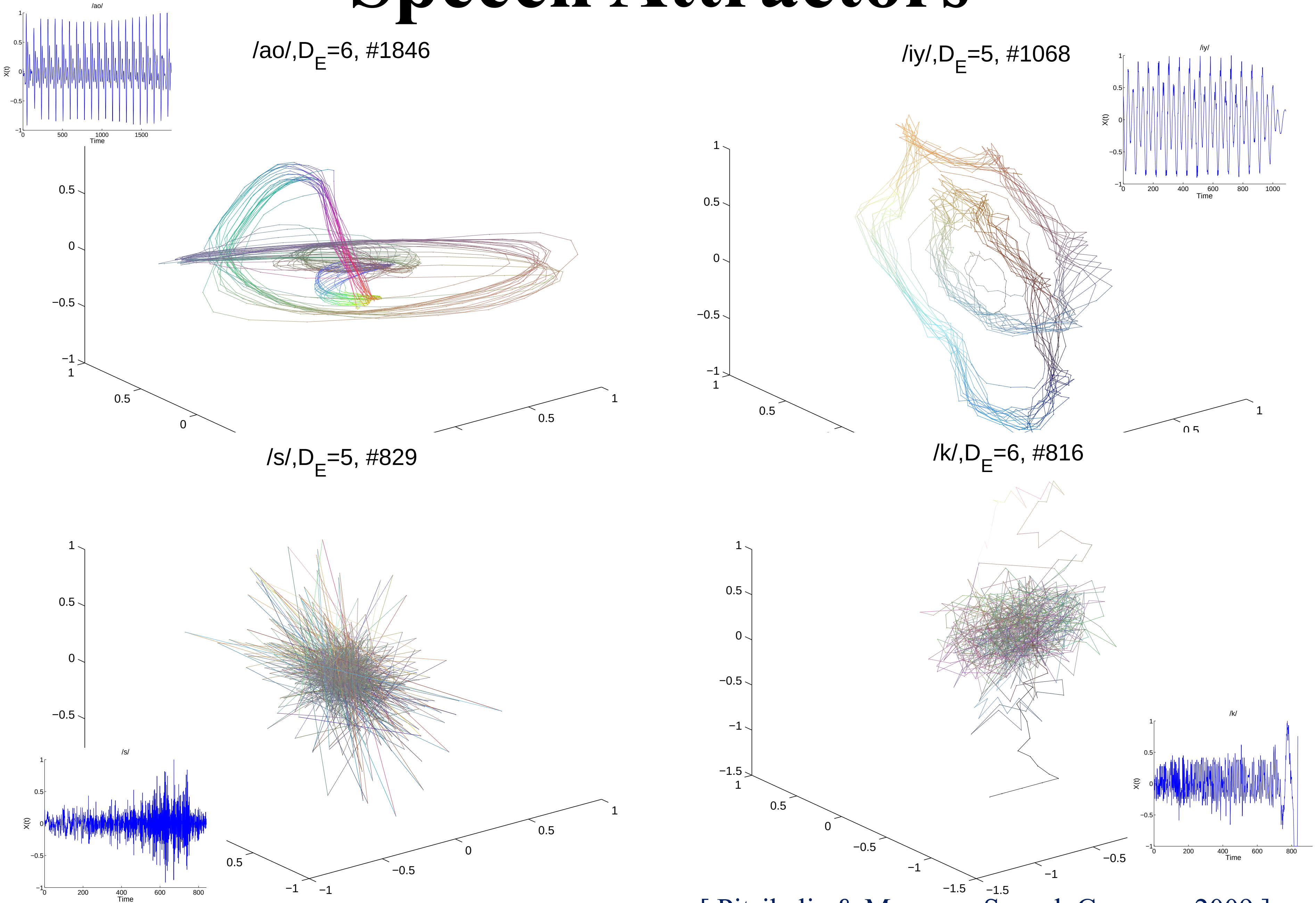
Multiscale Fractal Dimension of Speech Spounds



/f/ **/v/** **/iy/**

[P. Maragos & A. Potamianos, JASA 1999]

Speech Attractors



[Pitsikalis & Maragos, Speech Commun 2009]

Multiscale Fractal Dimensions for Speech Sounds

Refs:

- P. Maragos and A. Potamianos, "*Fractal Dimensions of Speech Sounds: Computation and Application to Automatic Speech Recognition*", Journal of Acoustical Society of America, March 1999.
- P. Maragos, "*Fractal Signal Analysis Using Mathematical Morphology*", in Advances in Electronics and Electron Physics, vol.88, Academic Press, 1994.

FRACTALS: Definitions

- Mandelbrot's definition

  set S is fractal $\Leftrightarrow$

  Hausdorff dim $D_H(S)$ $>$ topological dim $D_T(S)$

- Examples

  $D_T = 0 < D_H \le 1 \;\Rightarrow\; S$ = fractal dust

  $D_T = 1 < D_H \le 2 \;\Rightarrow\; S$ = fractal curve

  $D_T = 2 < D_H \le 3 \;\Rightarrow\; S$ = fractal surface

- Signals

A function $f : \Re^{\nu} \to \Re$ is a fractal if its graph

$Gr(f)$ is a fractal set in $\Re^{\nu+1}$

f is continuous $\Rightarrow \nu = D_T \le D_H[Gr(f)] \le \nu + 1$

‘Fractal’ Dimensions (of sets in $\mathbf{R}^\nu$)

D_H = Hausdorff dimension

D_{MB} = Minkowski-Bouligand dimension

D_{BC} = box counting dimension

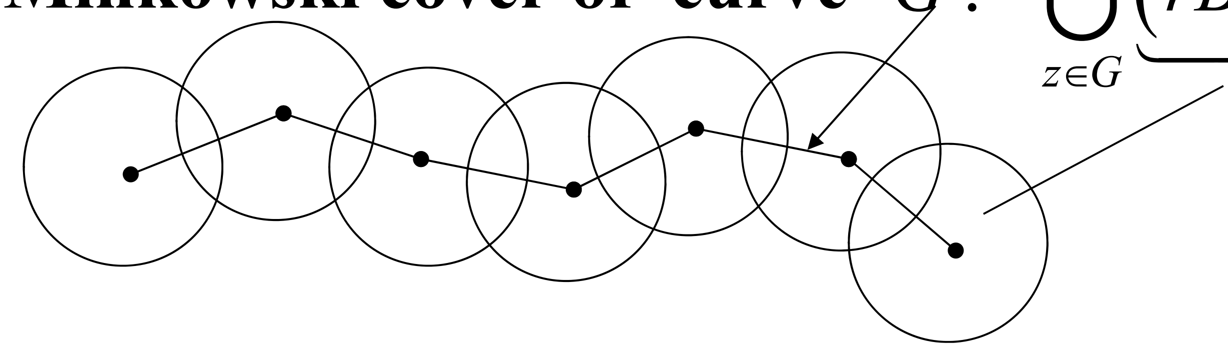
D_S = similarity dimension

$$\boxed{0 \le D_T \le D_H \le D_{MB} = D_{BC} \le \nu}$$

$$\boxed{D_H \le D_S}$$

Morphological Measurement of Fractal Dimension

- **Minkowski cover of curve** $G : \bigcup_{z \in G} \underbrace{(rB) + z} = C_B(r)$



- **Fractal (Minkowski-Bouligand) dimension** $D \in [1, 2]$

$$A_B(r) = area\left[C_B(r)\right];\ length\ of\ G(r) = \frac{A_B(r)}{2r} \propto r^{1-D}$$

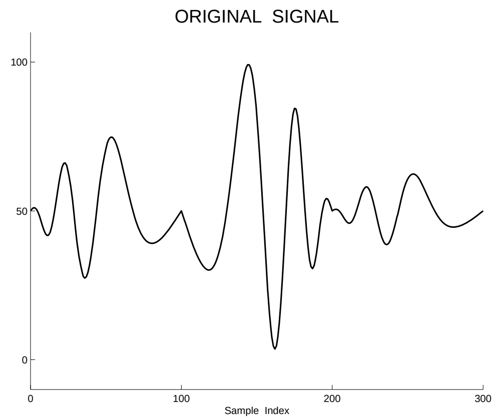
- **Least-Squares line fit to data**

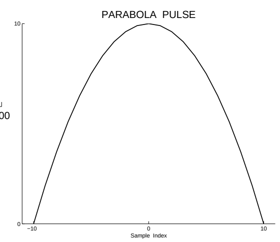
$$\left(\log\left[A_B(r)/r^2\right], \log(1/r)\right) \to D$$

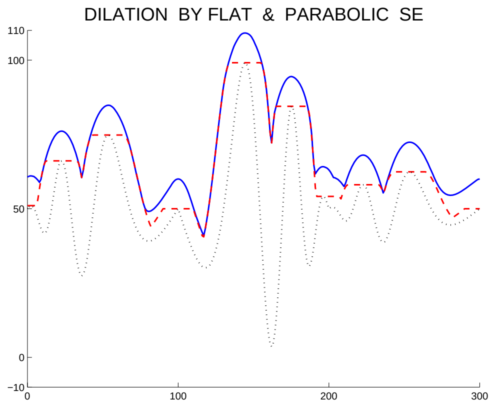
Morphological (Flat & Weighted) Filters

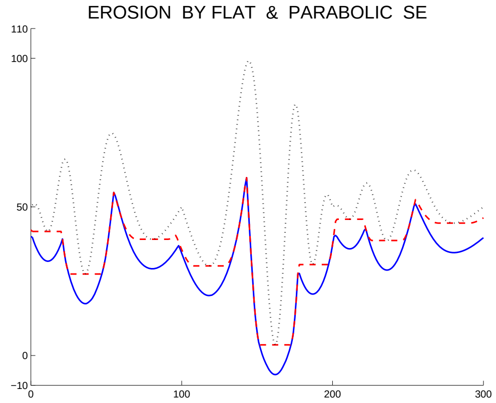
Dilation (Max-plus convolution): $(f \oplus g)(x) = \max_y f(y) + g(x-y)$

Erosion (Min-plus correlation): $(f \ominus g)(x) = \min_y f(y) - g(y-x)$







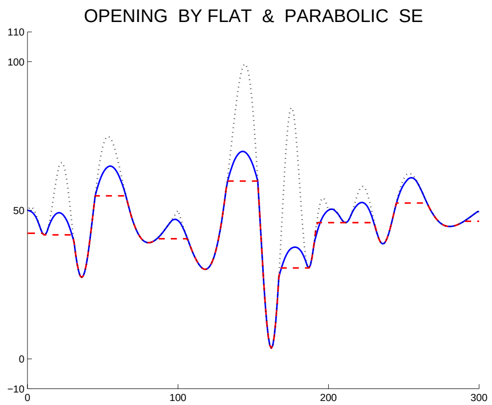


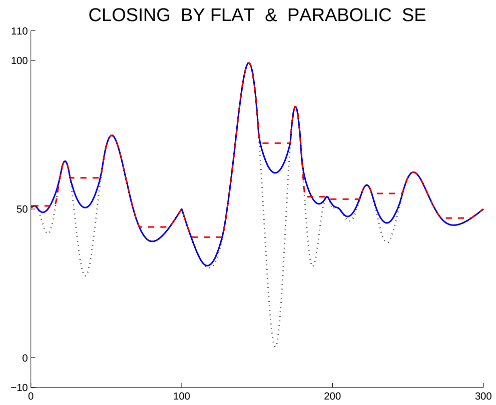
Opening:

$$f \circ g = (f \ominus g) \oplus g$$

Closing:

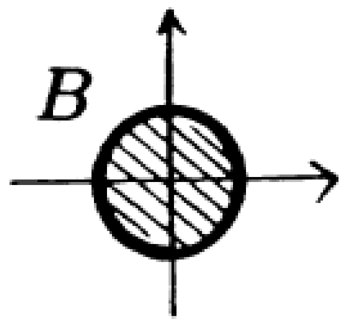
$$f \bullet g = (f \oplus g) \ominus g$$

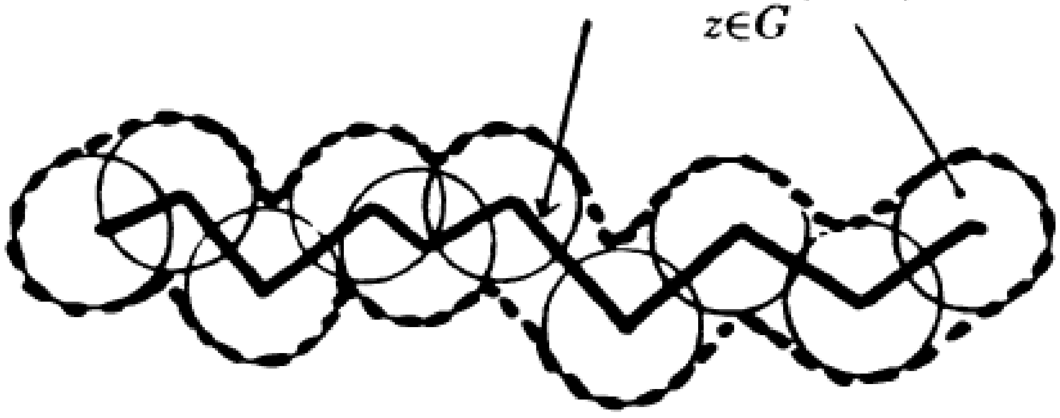




Minkowski Fractal Dimension of 1D Curve and Morphological Algorithm for 1D Signals

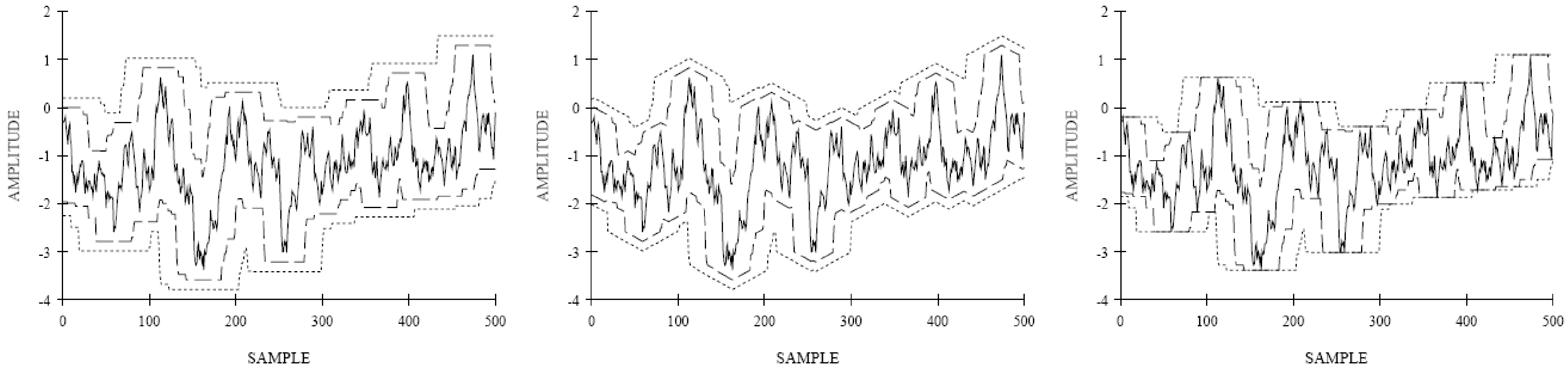
Minkowski cover of 2-dim curve $G = \bigcup_{z \in G} (rB)_{+z}$



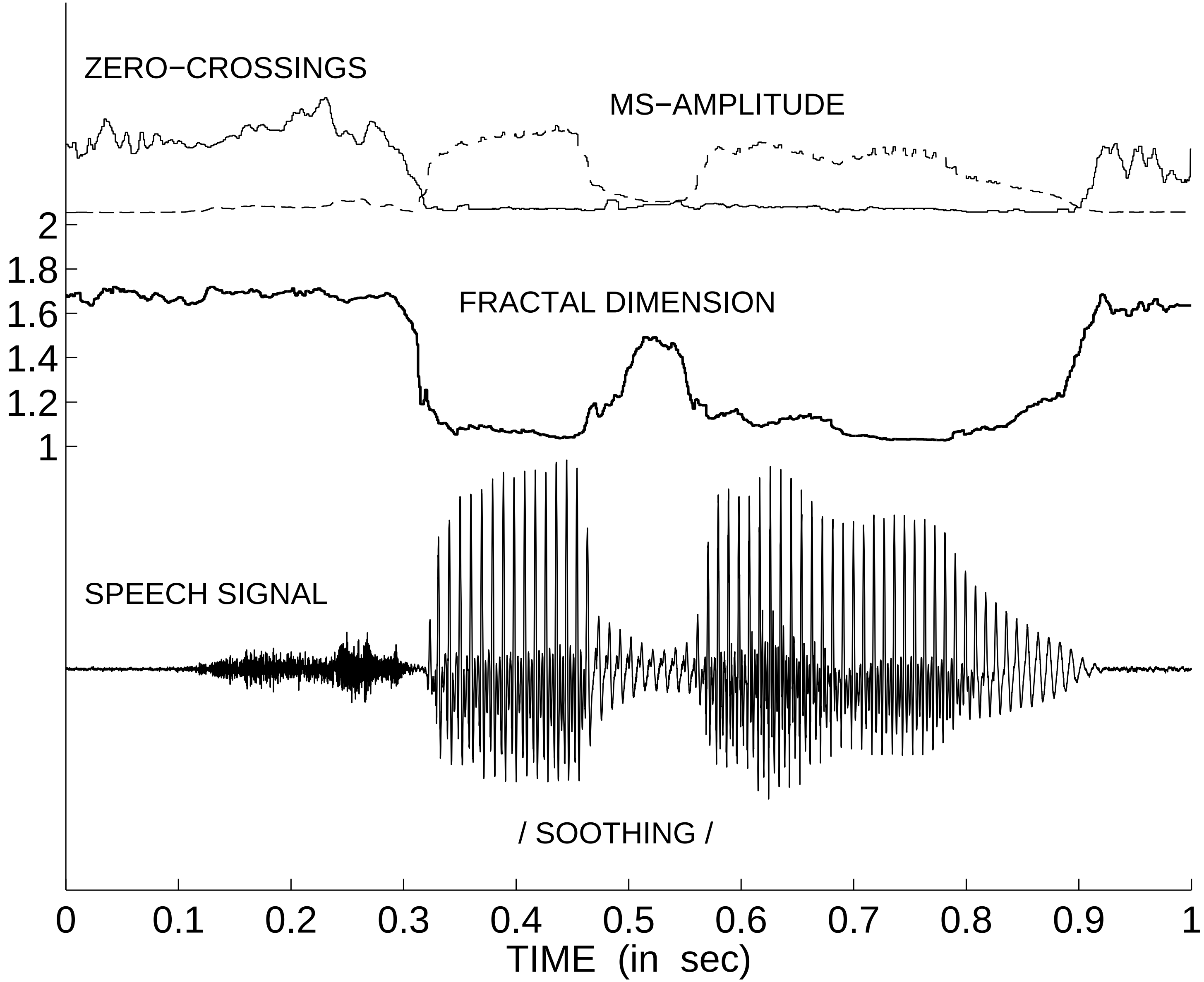


$$F_\epsilon = \bigcup_{z \in F} \{(\epsilon b + z) \in \mathbb{R}^2 : ||b|| \leq 1\}$$

$$D_M(F) = \lim_{\epsilon \to 0} \frac{\log[\mathrm{Area}(F_\epsilon)/\epsilon^2]}{\log(1/\epsilon)}$$



ST Speech & Fractal Dimension



Multiscale Speech Fractal Dimension

- **short-time speech signal**

  $$S(t), \quad 0 \le t \le T$$

- **signal graph**

$$G = \left\{ \left(t, S(t)\right) \in R^2 : 0 \le t \le T \right\}$$

- **fractal** $\Rightarrow$

  constant power law

$$area\left(G \oplus \varepsilon B\right) \approx C\varepsilon^{2-D}, \text{ as } \varepsilon \to 0$$

- **variable power law**

  $$area\left(G \oplus \varepsilon B\right) \approx C\varepsilon^{2-D(\varepsilon)}$$

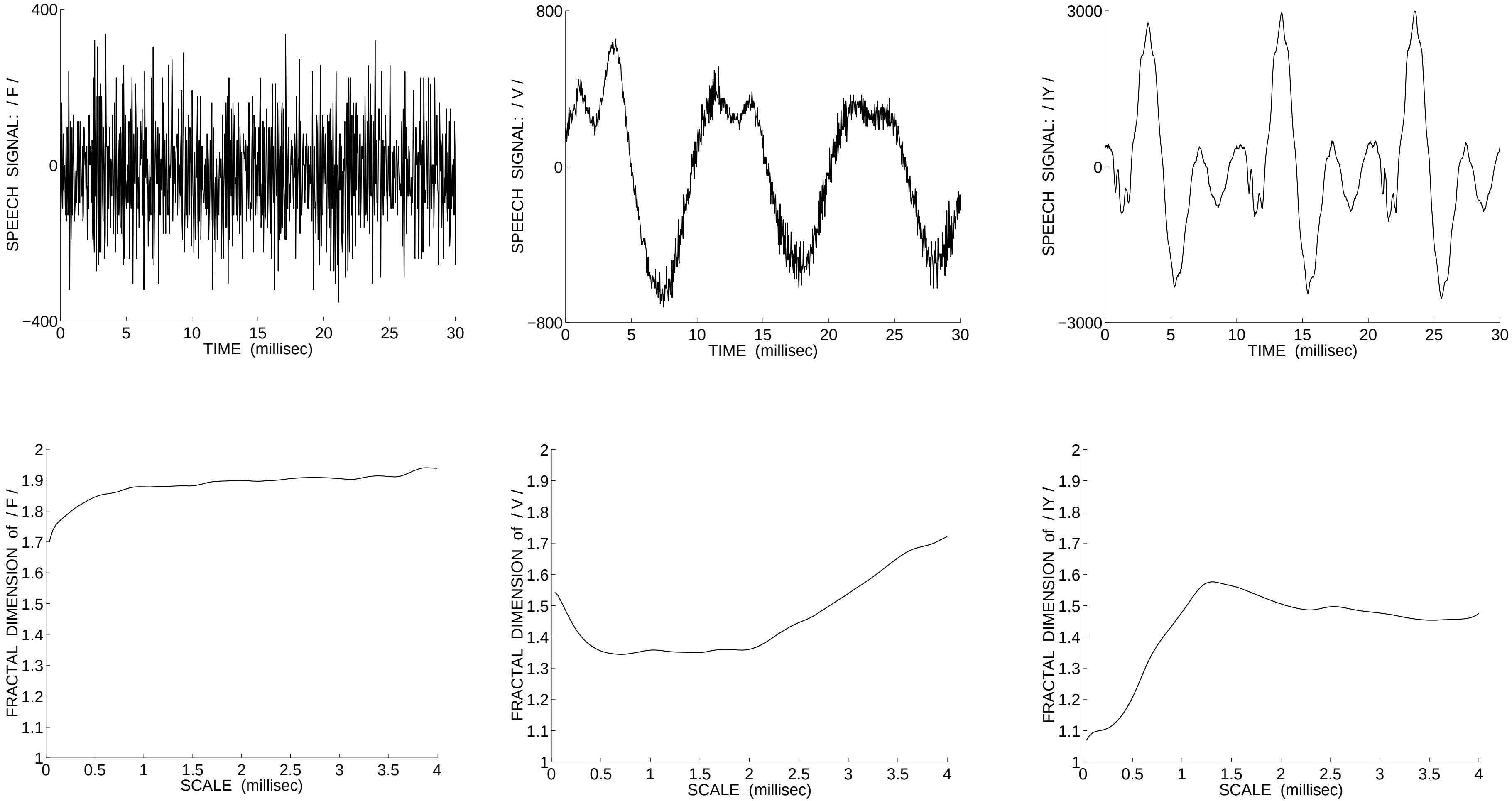
- **multiscale fractal "dimension"**

  (speech fractogram):

  $MFD(t, \varepsilon) = D(\varepsilon)$ **of**

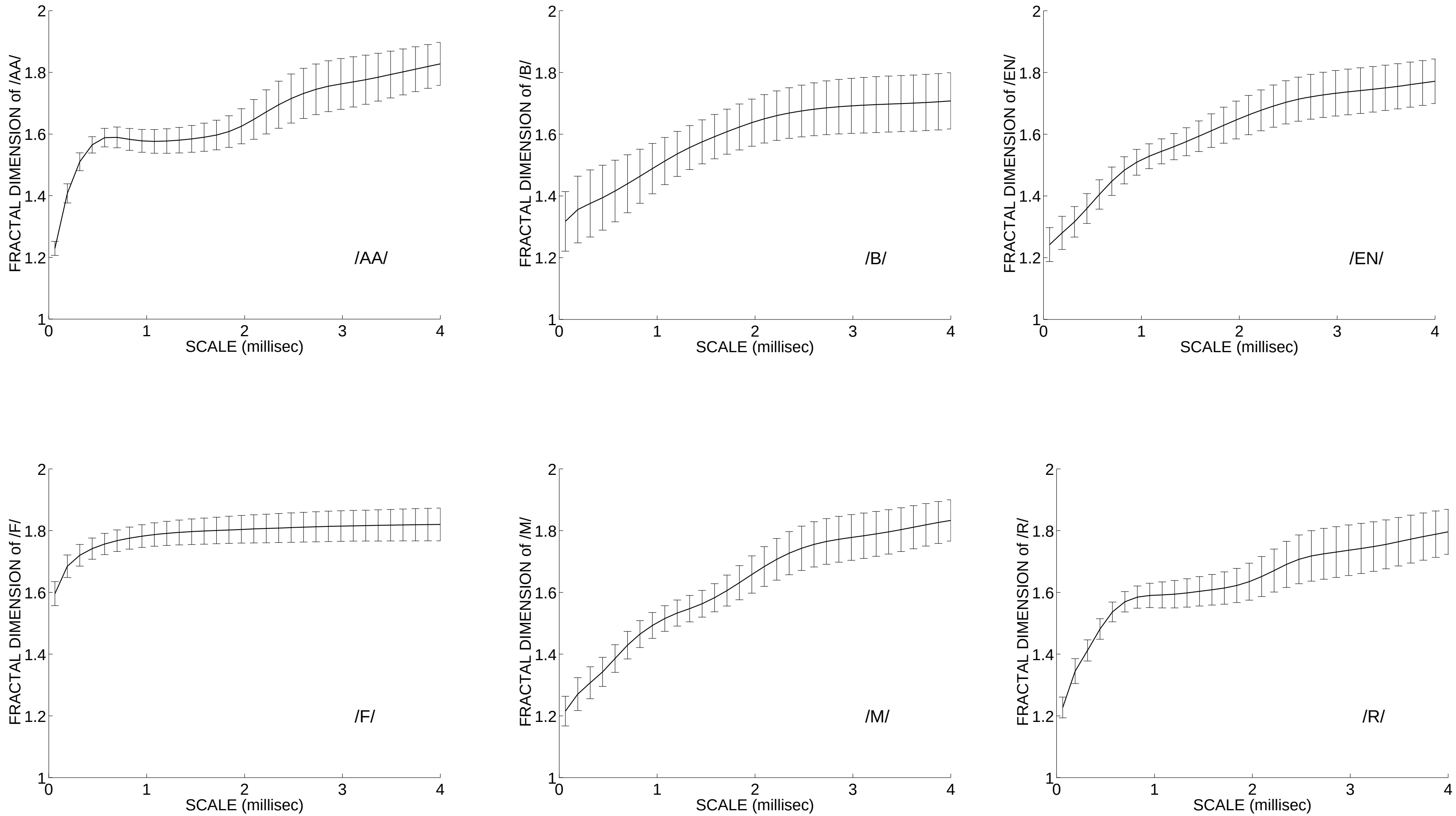
short-time speech segment around time t

Multiscale Fractal Dimension of Speech Spounds



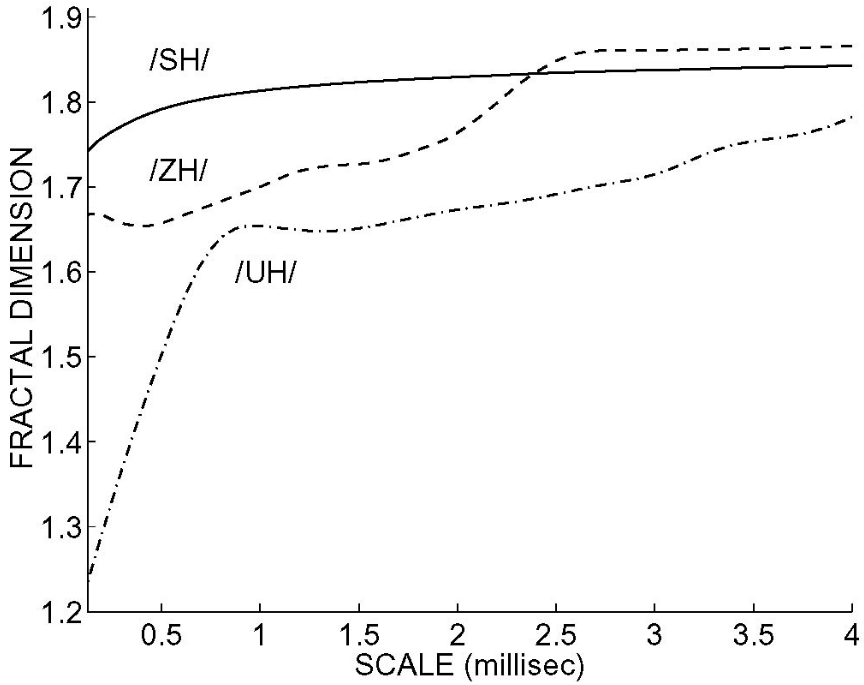
/f/ **/v/** **/iy/**

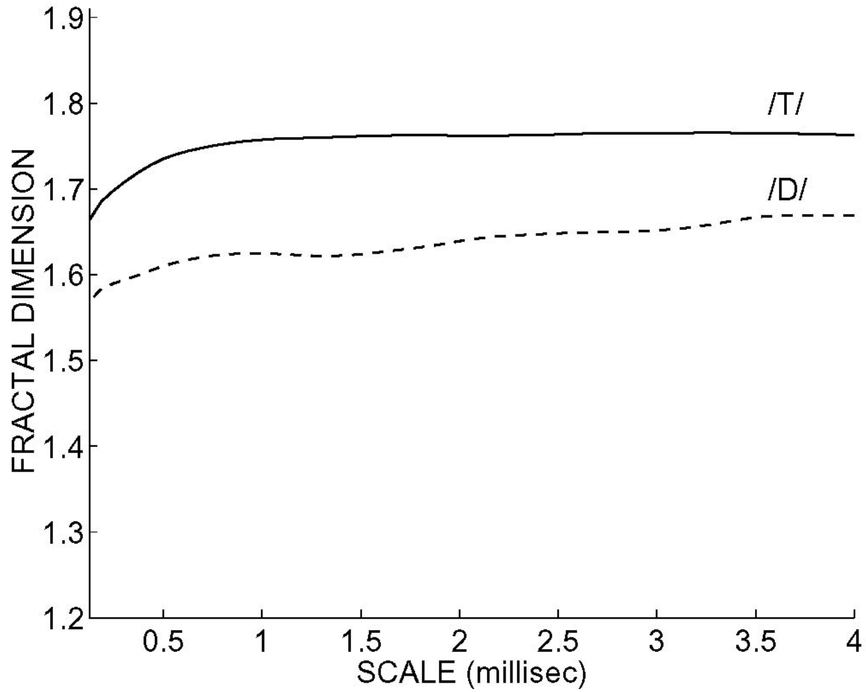
[P. Maragos & A. Potamianos, JASA 1999]



Mean and standard deviation (error bars) of the multiscale fractal dimension for the phonemes /aa/, /b/, /en/, /f/, /m/, /r/ from the TIMIT database (20 ms window, updated every 10 ms. Average over 200 phonemic instances.)

Mean MFD for /sh/, /zh/, /uh/, /t/, /d/





Word Percent Correct For the E-set Recognition Task (ISOLET Database, 5-Mixture Gaussians per HMM State)

$\{E, C, \Delta E, \Delta C\}$	$\{E, C, \Delta E, \Delta C\} + \{D_1, \Delta D_1\}$	$\{E, C, \Delta E, \Delta C\} + \{D_{1\cdots16}, \Delta D_{1\cdots16}\}$
81.2%	83.5%	84.5%

Word Percent Correct for the E-set Recognition Task

Features / Models	$\{E, C, \Delta E, \Delta C, \Delta\Delta E, \Delta\Delta C\}$	$\{E, C, \Delta E, \Delta C, \Delta\Delta E, \Delta\Delta C\} + \{D, \Delta D\}$
5-mixture Gaussians	85.6%	86.3%
10-mixture Gaussians	88.6%	88.9%

Maragos & Potamianos, JASA 1999

Fractal Modulations
for
Fricative Sounds

Ref:

- A. G. Dimakis and P. Maragos, "*Phase Modulated Resonances Modeled as Self-Similar Processes With Application to Turbulent Sounds*", IEEE Transactions on Signal Processing, Nov. 2005.

1/f Noises

- An important class of statistically self-similar random processes defined by their measured power spectra**:**

$$S(\omega) \propto \frac{\sigma^2}{|\omega|^\gamma}$$

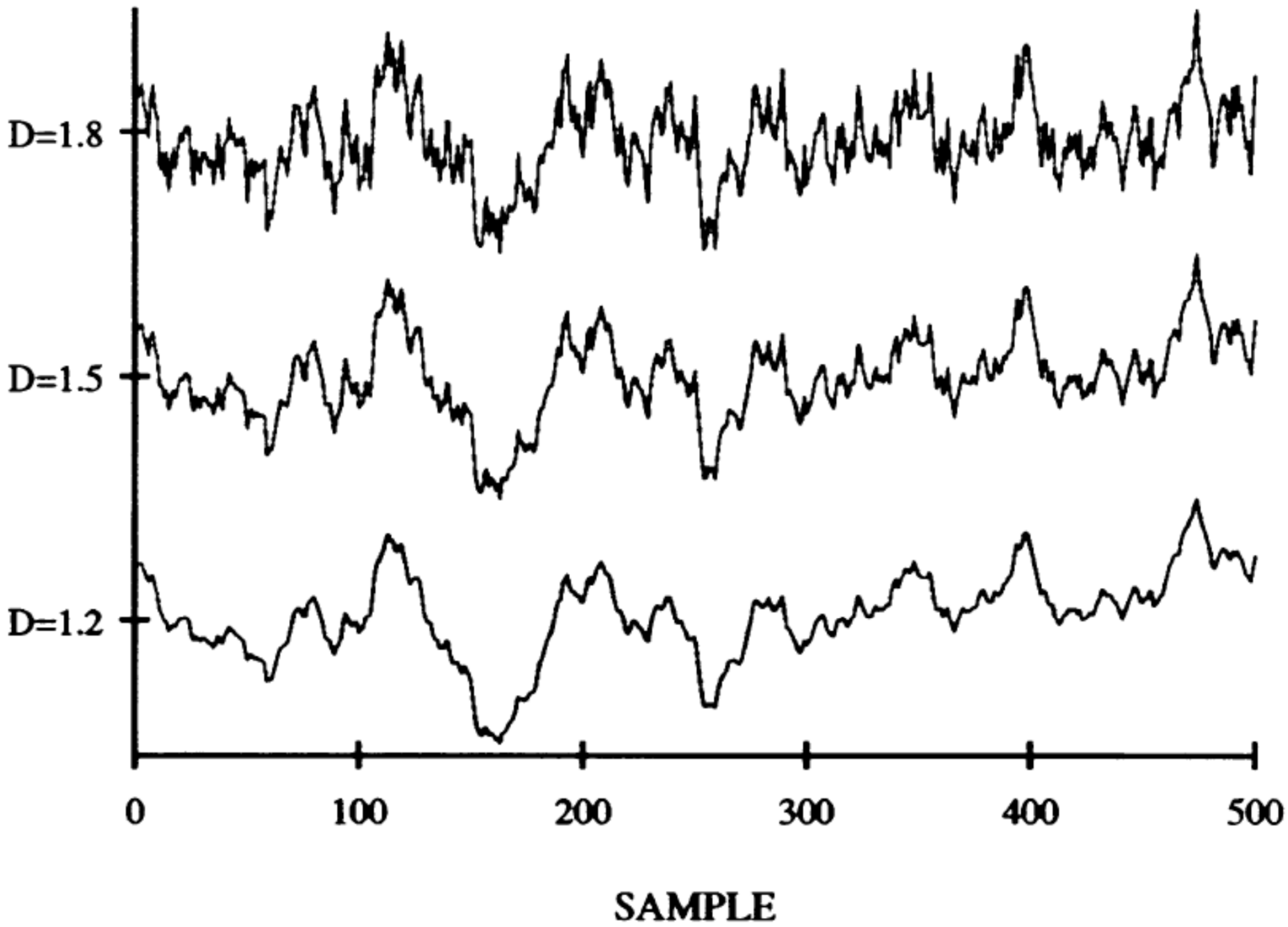
A truly enormous collection of natural phenomena exhibit 1/f-type spectral behavior over a wide frequency range: (frequency variations in quartz crystal oscillators, geophysical variations, heart rate variations, electronic device noises, network traffic flow and economic time series.)

- Most popular mathematical model for Gaussian 1/f processes: Fractional Brownian Motion (FBM)

$1/f^{\beta}$ Noises

- Stochastic processes with power spectrum $\propto 1/f^{\beta}$
- Filtering white noise with convolution kernel $\propto t^{\beta/2-1}$ (Fractional Integration)
- Non – exponential autocorrelation $\propto |t|^{\beta-1}$
- $\beta = 0 \Rightarrow$ White noise
- $\beta = 1 \Rightarrow$ Pink noise
- $1 < \beta < 3 \Rightarrow$ Fractal Brownian Motion
- $\beta = 2 \Rightarrow$ Brown noise
- $\beta > 2 \Rightarrow$ Black noise
- Applications: electronics, geophysics, astronomy, music, acoustics, optics, economics, traffic flows, communications, geometry of nature

Examples of FFT-based Synthesis of 1D FBM



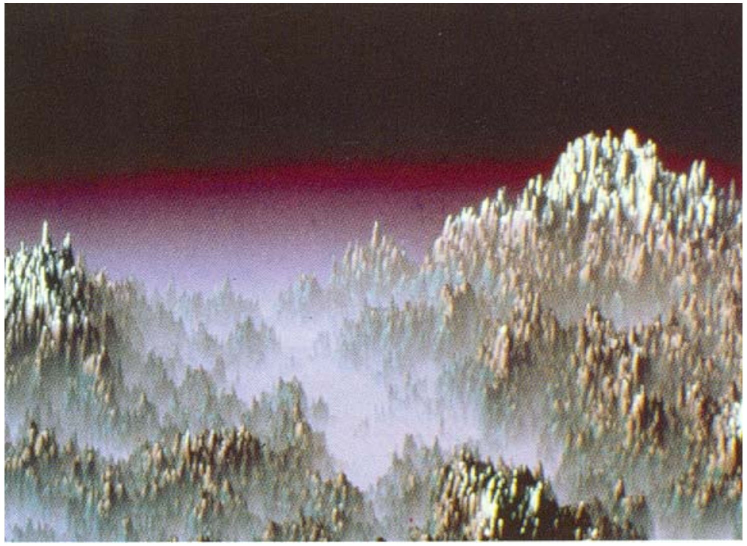
FBM Synthesis of Fractal Landscapes

R. Voss, 1988

D = 2.5

D = 2.15

D = 2.8



1/f Speech Modulation Model

- Model a resonance of a random speech phoneme as a phase-modulated 1/f signal:

$$S(t) = A\cos\underbrace{\left(\omega_c t + P(t)\right)}_{\phi(t)}$$

- **Nonlinear phase signal P(t) modeled as 1/f random process.**
- Useful model for broad resonances often observed in fricative voiced or unvoiced sounds and probably caused by nonlinear phenomena during speech production.

Parameter Estimation in 1/f-PM

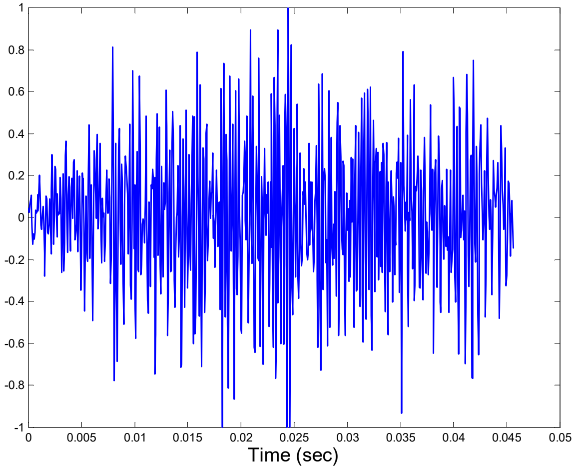
- Isolate resonance: Bandpass filter the speech signal.
- Demodulate filtered signal using ESA, obtain instant frequency F(t), and median filter to reduce spikes.
- Estimate phase modulation signal P(t) by integrating IF:

$$P(t) = 2\pi \int_0^t (F(\tau) - \bar{F}) d\tau$$

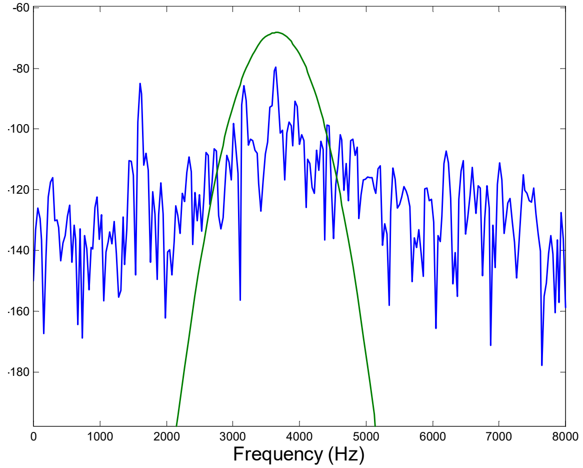
- **Fit $1/f^{\gamma}$ model to P(t).** Methods tested include:
  - Linear regression on Periodogram
  - Estimation using variance of wavelet coefficients
  - Maximum Likelihood estimation

/S/ phoneme experiment

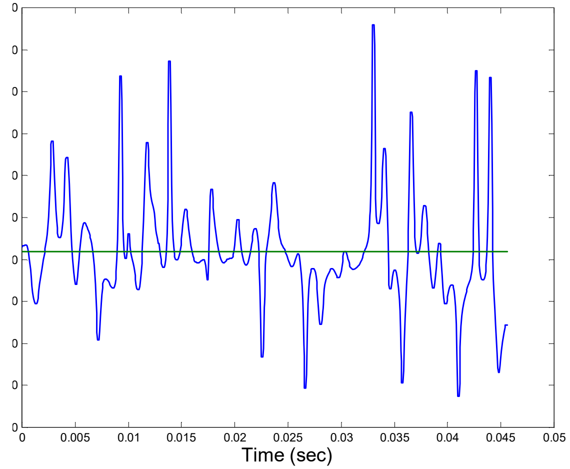
Speech signal



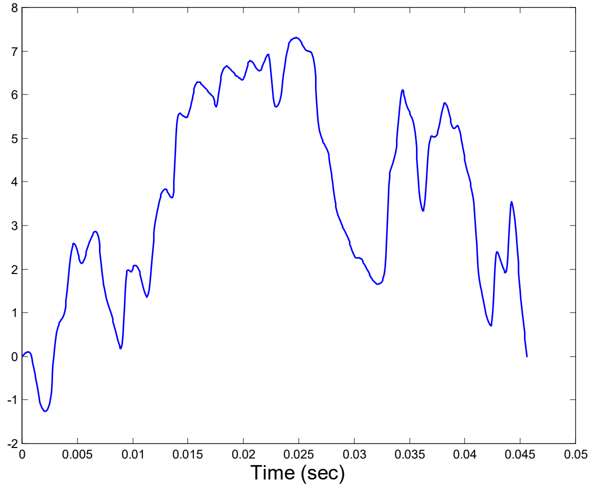
Power Spectrum



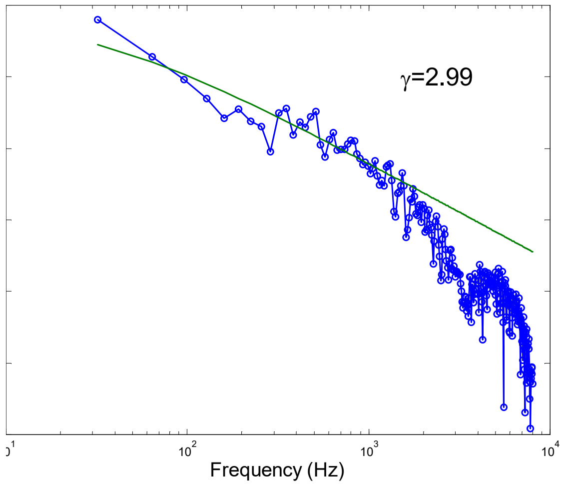
Instant Frequency



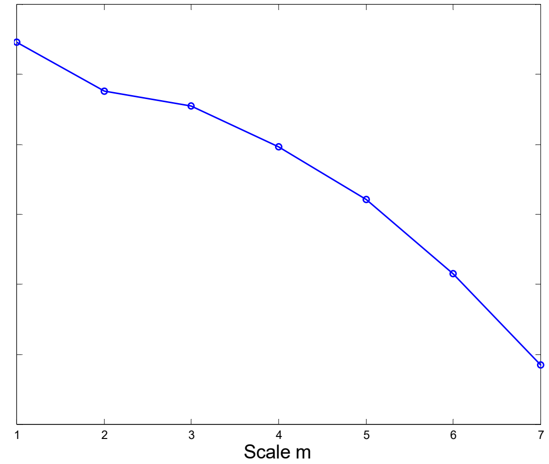
Phase modulation P(t)



PSD of P(t)

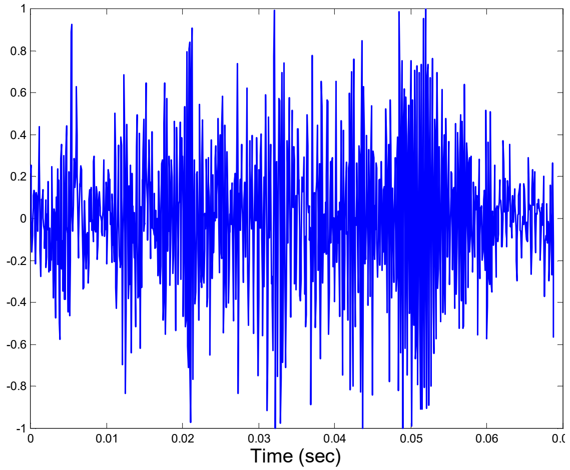


Var. of wavelet coefficients

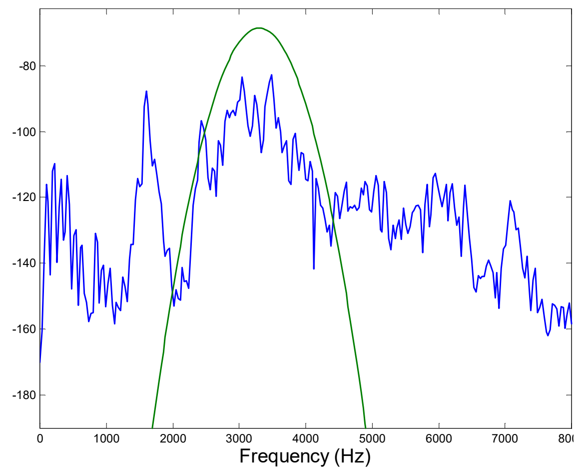


/Z/ phoneme experiment

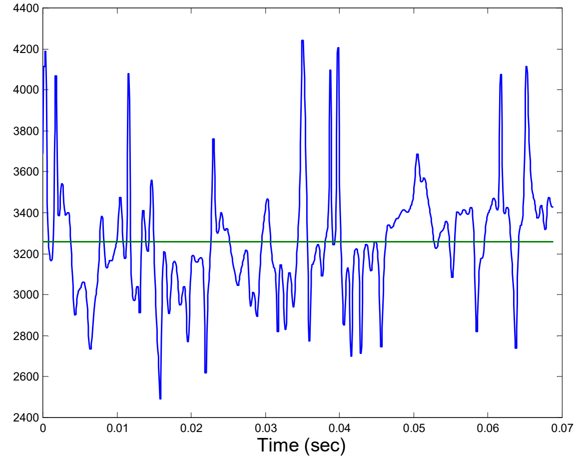
Speech signal



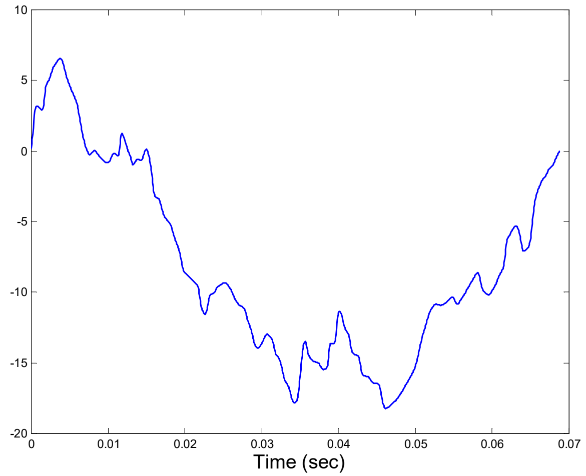
Power Spectrum



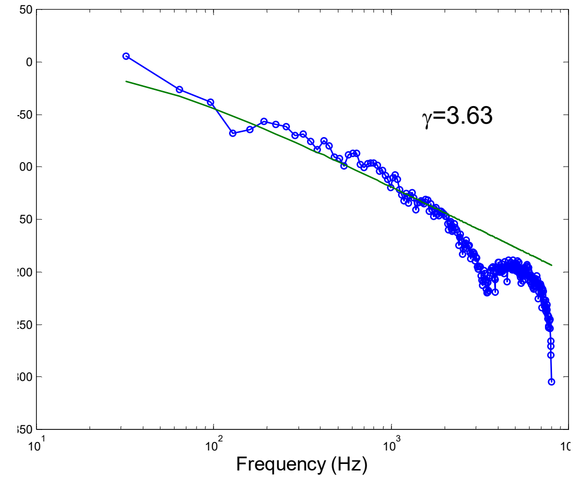
Instant Frequency



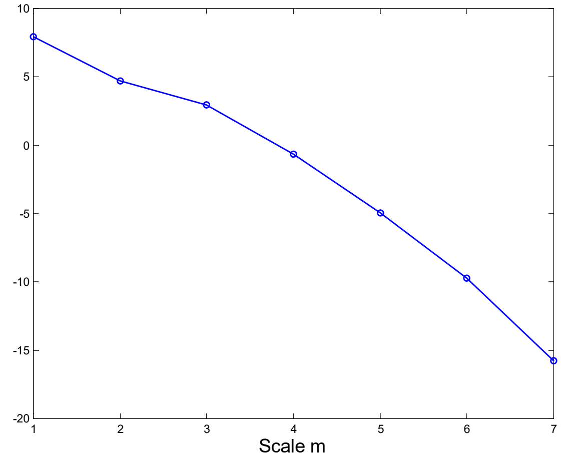
Phase modulation P(t)



PSD of P(t)



Var. of wavelet coefficients

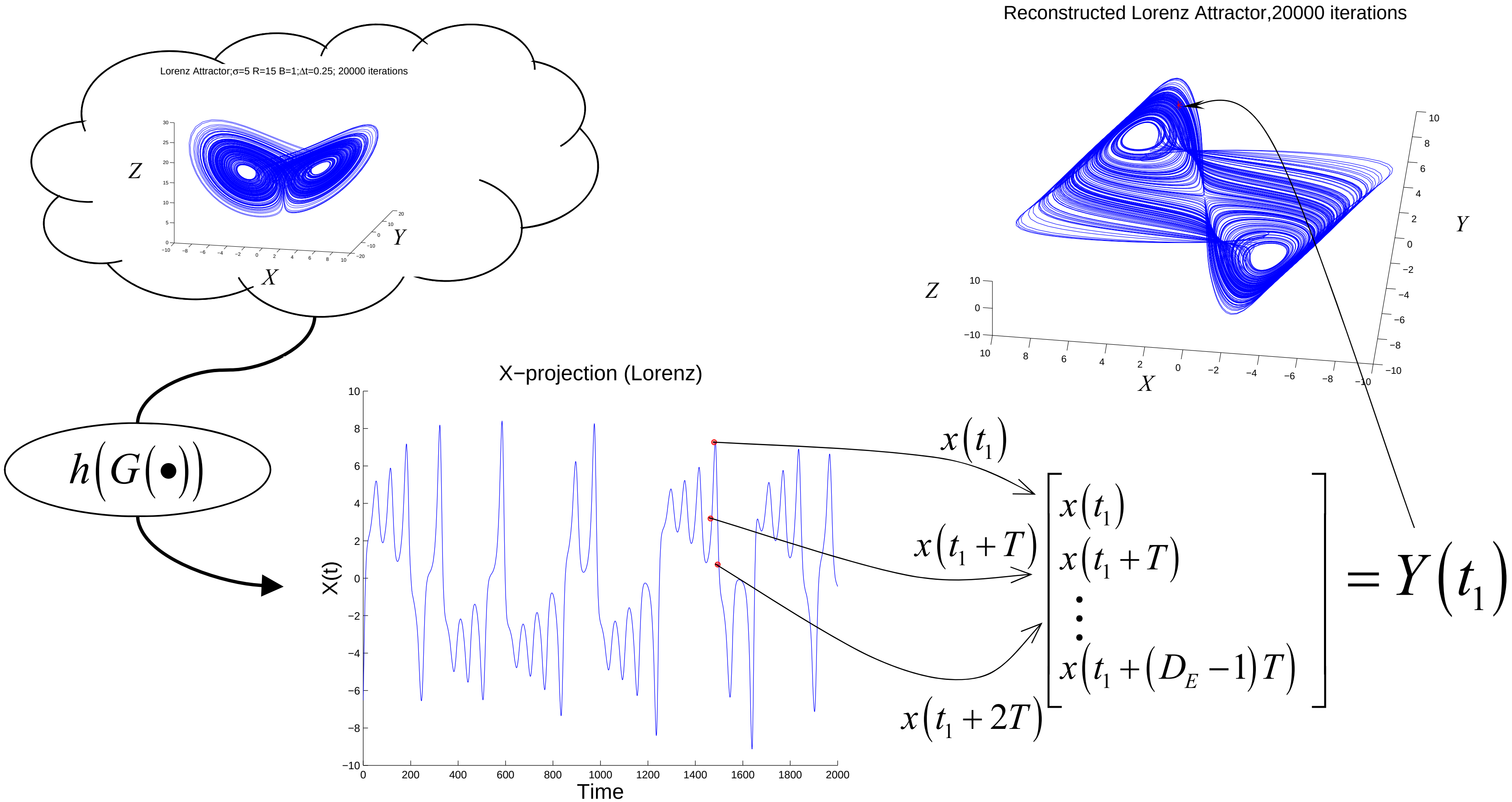


Chaotic Dynamics of Speech Sounds

Refs:

- V. Pitsikalis and P. Maragos, "*Filtered Dynamics and Fractal Dimensions for Noisy Speech Recognition*", IEEE Signal Processing Letters, Nov. 2006.
- V. Pitsikalis and P. Maragos, "*Analysis and Classification of Speech Signals by Generalized Fractal Dimension Features*", Speech Communication, Dec. 2009.
- I. Kokkinos and P. Maragos, "*Nonlinear Speech Analysis Using Models for Chaotic Systems*", IEEE Transactions Speech and Audio Processing, Nov. 2005.

Embedding-Attractor Reconstruction



•Parameters to specify: T, D_E

- **Nonlinear Dynamic System (Lorenz)**

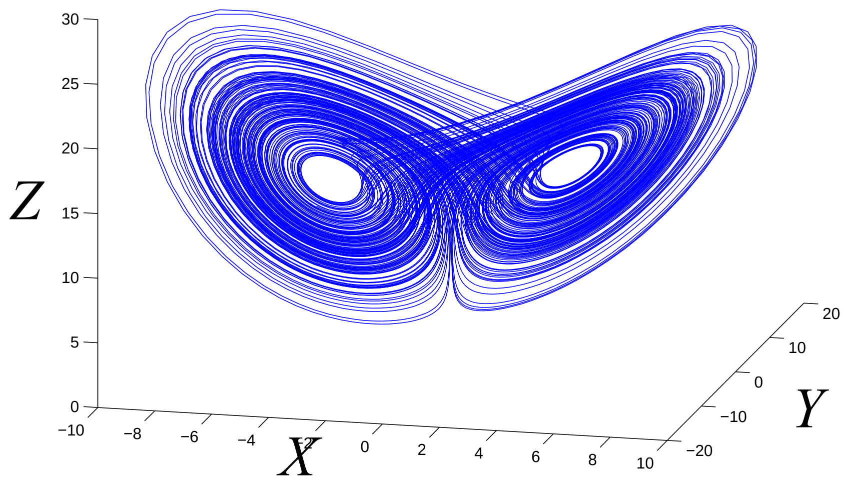
$$\frac{dx}{dt} = -\sigma \cdot x + \sigma \cdot y$$

$$\frac{dy}{dt} = R \cdot x - y - x \cdot z$$

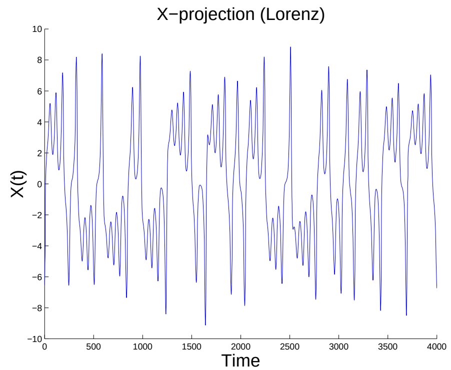
$$\frac{dz}{dt} = -B \cdot z + x \cdot y$$

- **Attractor**

Lorenz Attractor;σ=5 R=15 B=1;Δt=0.25; 20000 iterations



- **1D Projection**



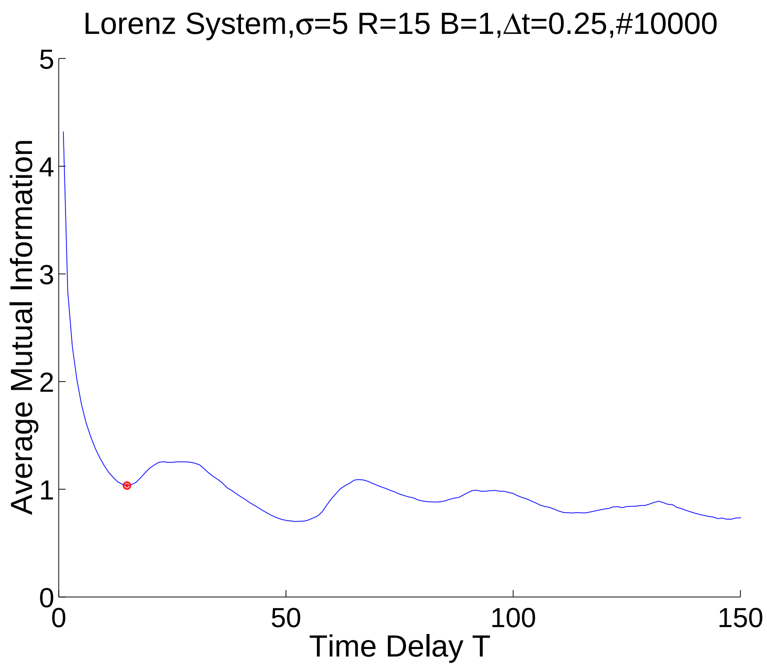
Time Delay

- **Average Mutual Information** between $x(t), x(t+T)$

$$I(T) = \sum \Pr\left(x(t), x(t+T)\right) \cdot \log\left[\frac{\Pr\left(x(t), x(t+T)\right)}{\Pr\left(x(t)\right) \cdot \Pr\left(x(t+T)\right)}\right]$$

- **“Optimum” Time Delay**

$$T_{opt} = \min\left\{\arg\min_{T} I(T)\right\}$$



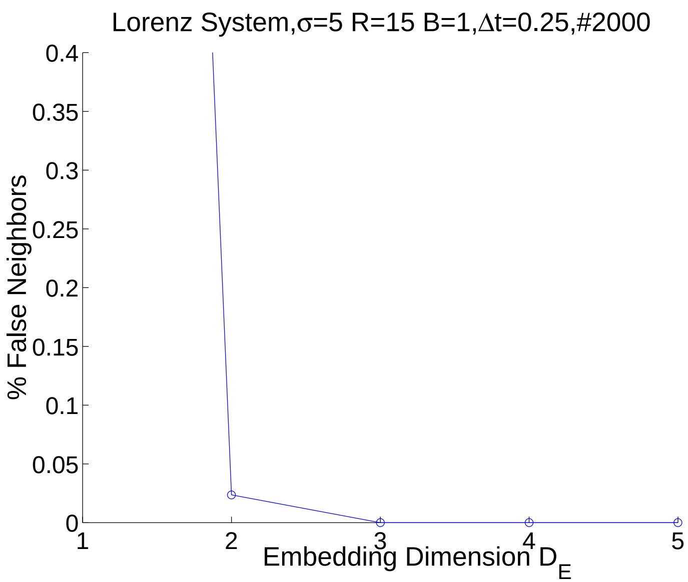
Embedding Dimension

- Sufficient: $D_E > 2 \cdot D_{Attractor}$
- False Neighbors: from projection
- True Neighbors: from dynamics
- False Neighbors Criterion

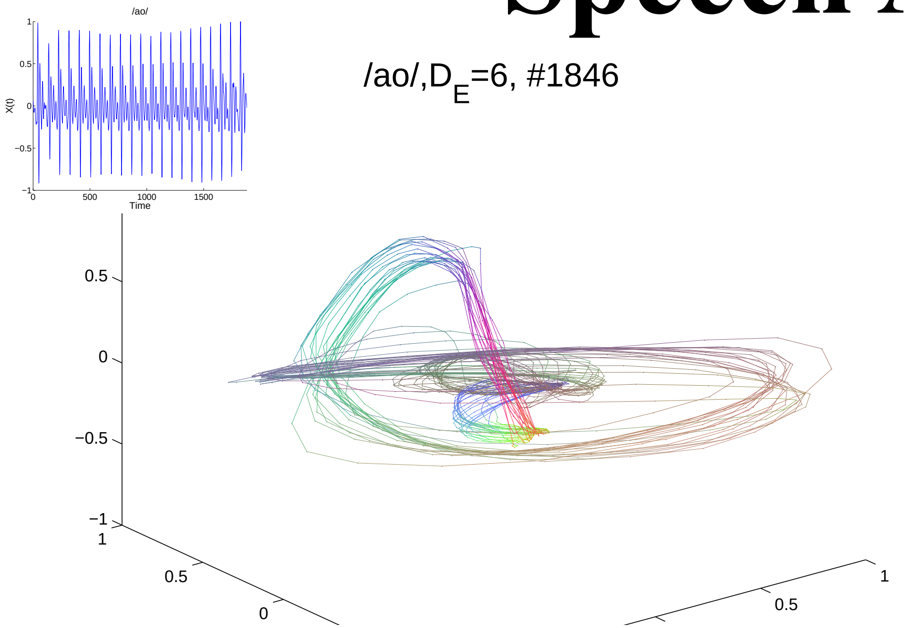
$$R_{i,j} = \frac{\left\| y_{d+1}(i) - y_{d+1}(j) \right\| - \left\| y_d(i) - y_d(j) \right\|}{\left\| y_d(i) - y_d(j) \right\|} > Threshold$$

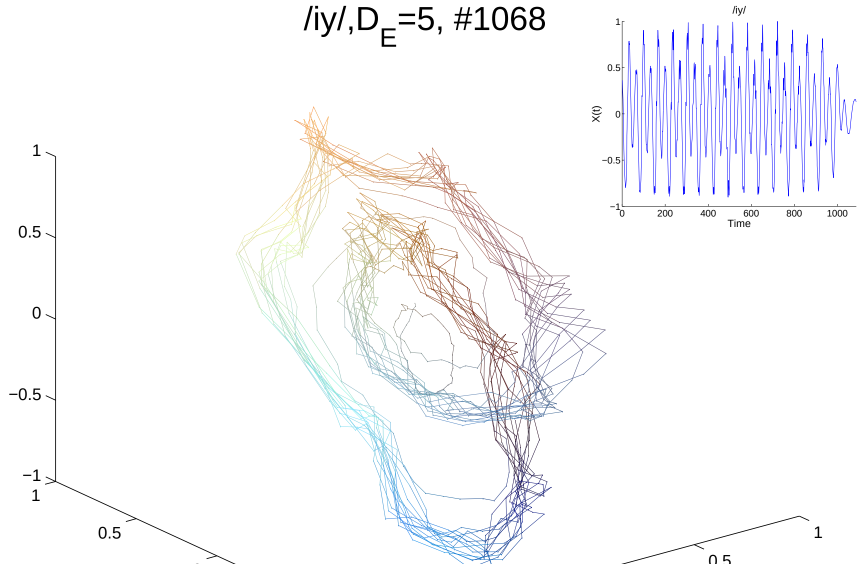
- When % false neighbors =0,

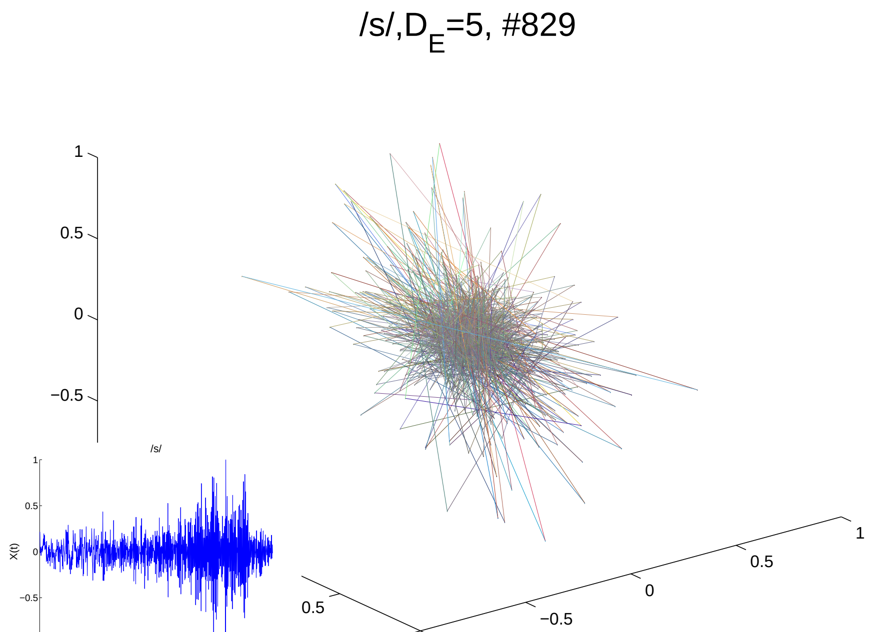
  Attractor is unfolded

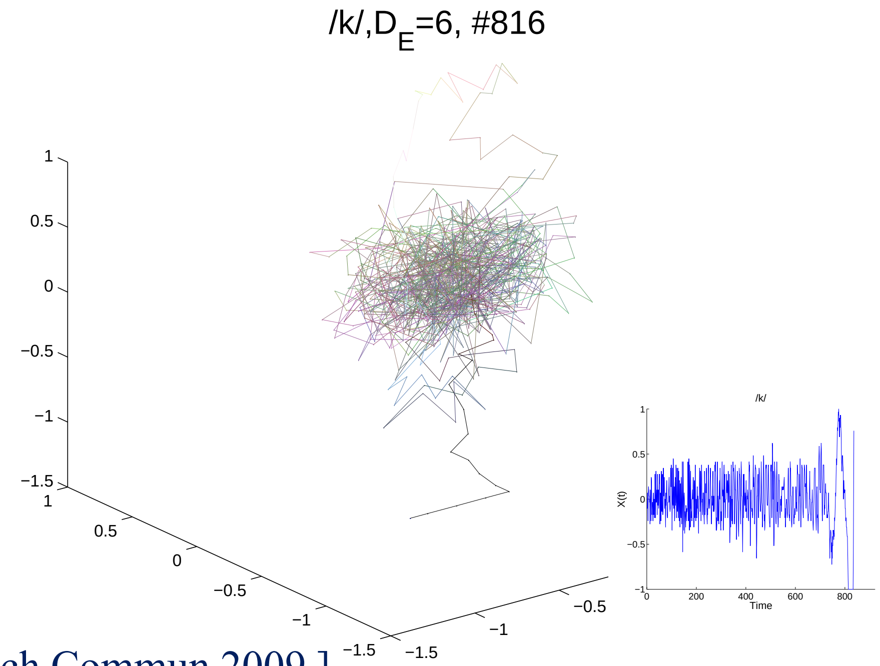


Speech Attractors





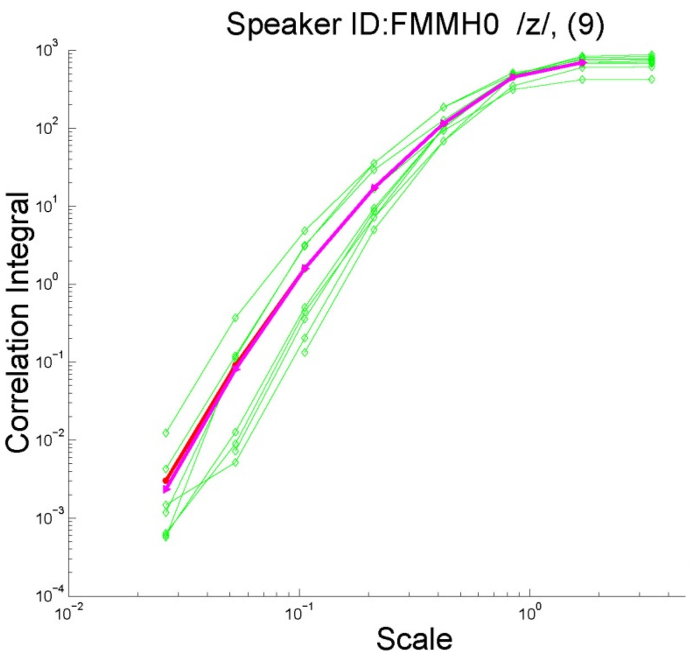


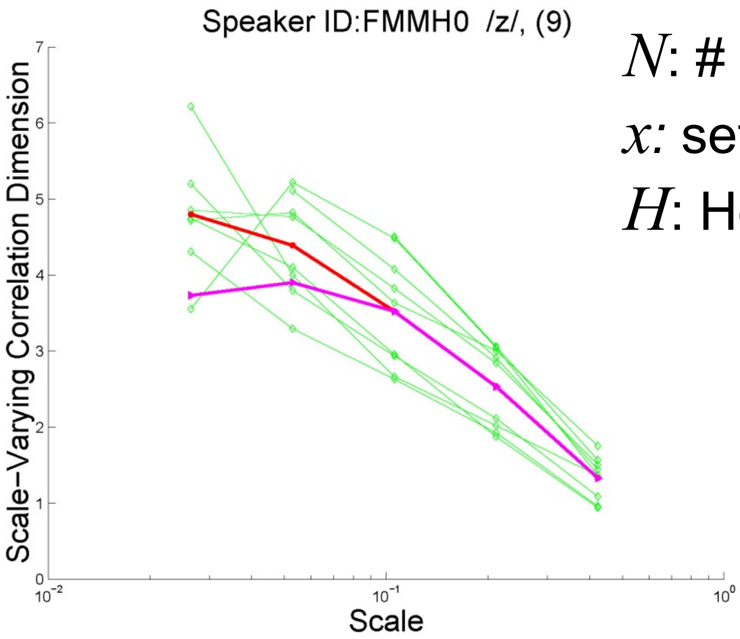


[Pitsikalis & Maragos, Speech Commun 2009]

Correlation Dimension (Speech)

- **Correlation Dimension**: $$D_C = \lim_{r\to 0}\lim_{N\to\infty}\frac{\log C(N,r)}{\log r}$$
- Correlation integral: $$C(N,r)=\frac{1}{N\cdot(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}H\left(r-\left\|x_i-x_j\right\|\right)$$



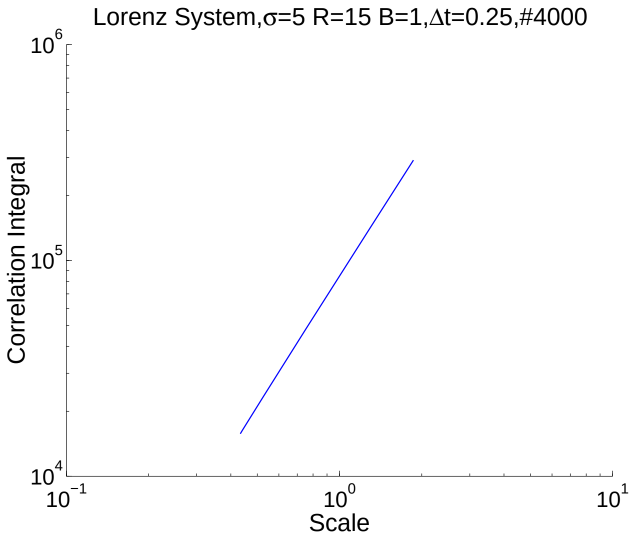


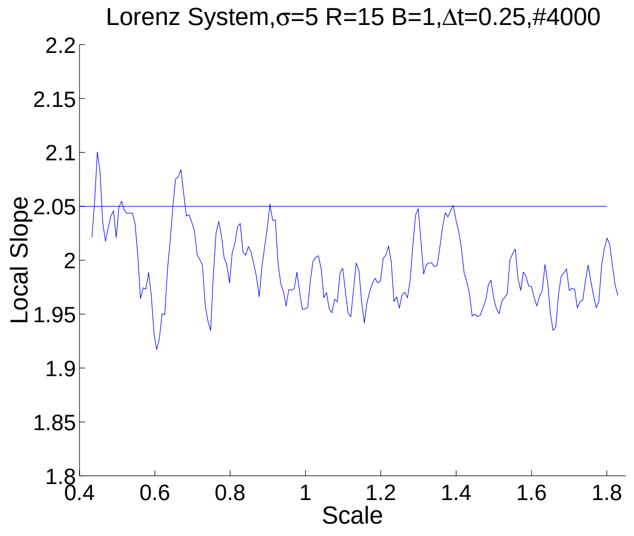
N: # of points, r: scale,
x: set points,
H: Heavyside function

Correlation Dimension (Lorenz)

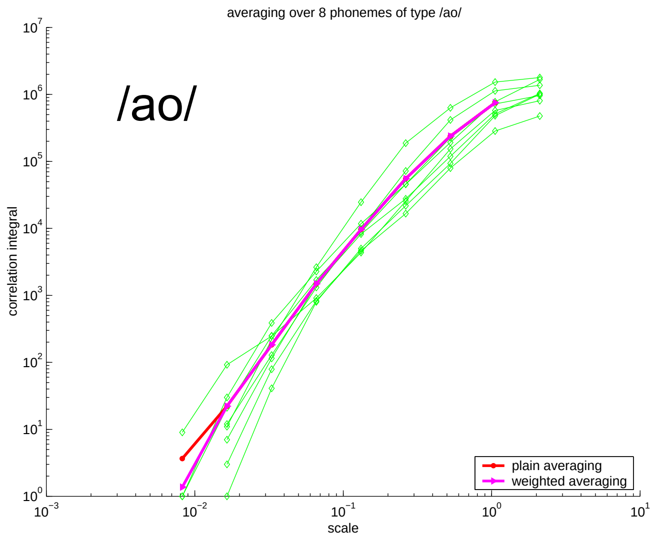
- $$C(N,r) = \frac{1}{N \cdot (N-1)} \sum_{i=1}^{N} \sum_{j \neq i} H\left(r - \left\| x_i - x_j \right\|\right)$$

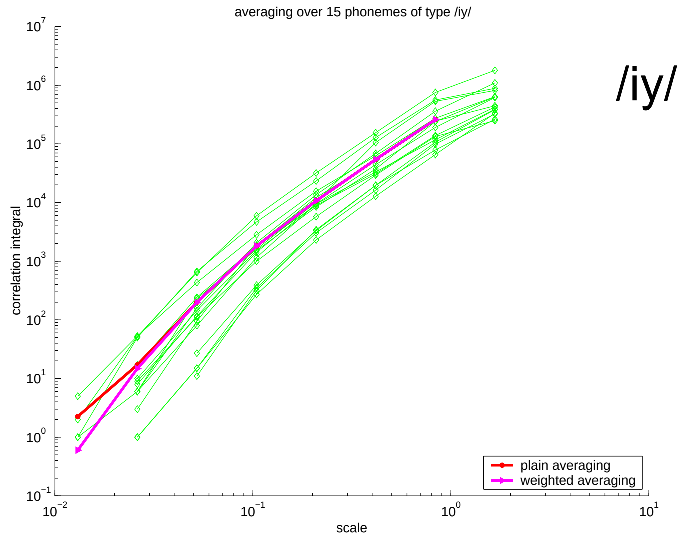
- $$D_C = \lim_{r \to 0} \lim_{N \to \infty} \frac{\log C(N,r)}{\log r}$$

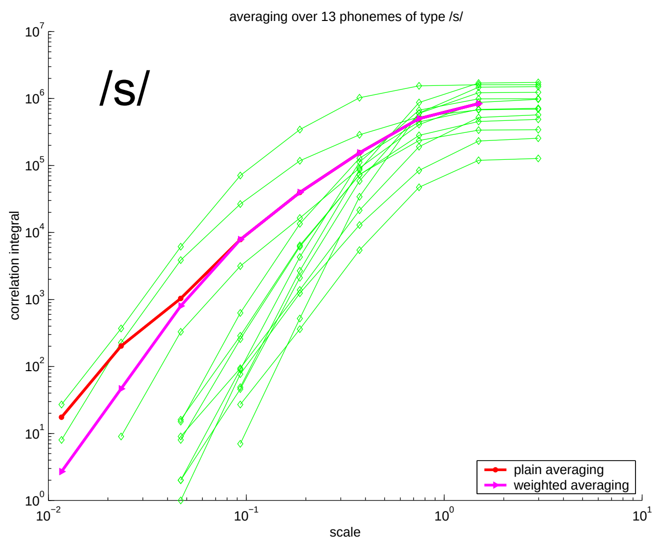


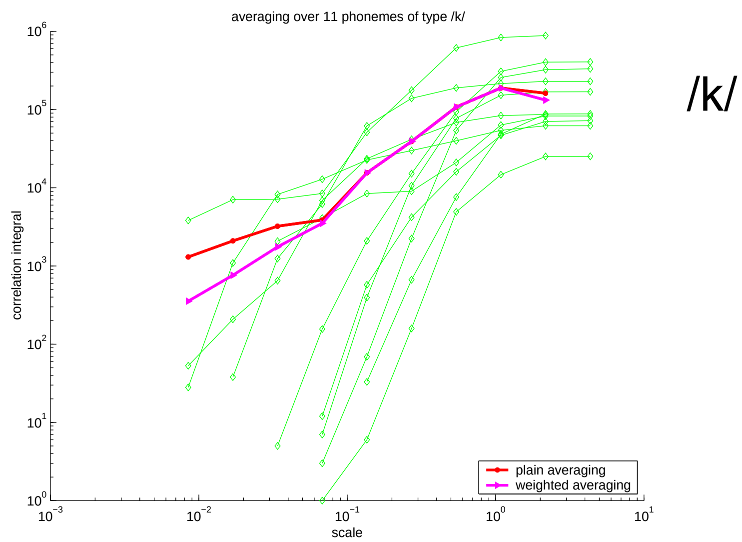


Correlation Integrals of Speech Sounds









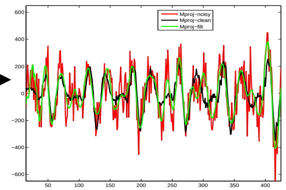
Fractal Features

speech signal → Embedding → N-d Signal → Local SVD → N-d Cleaned

FDCD

Filtered Dynamics - Correlation Dimension (8)

Cleaned → Projection →

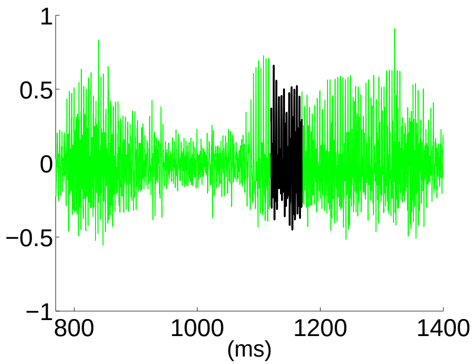


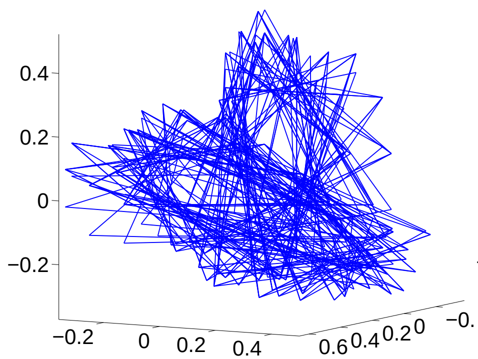
Enhanced Speech

speech signal → Geometrical Filtering →

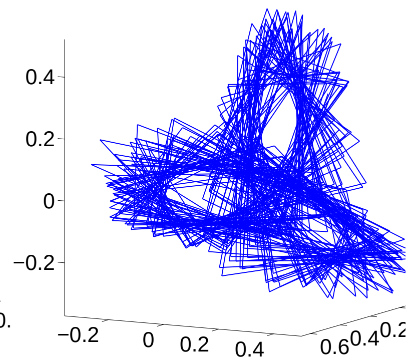
MFD

Multiscale Fractal Dimension (6)



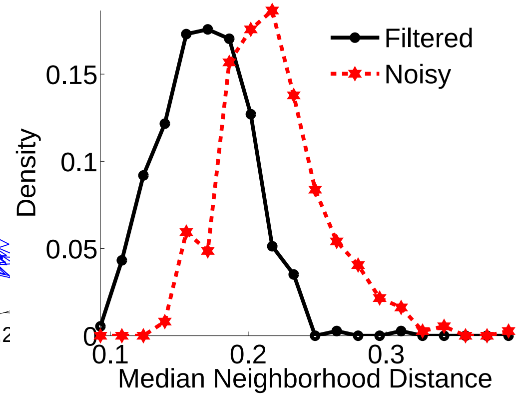


Noisy Embedding



Filtered Embedding

Neighborhood Distance Reduction

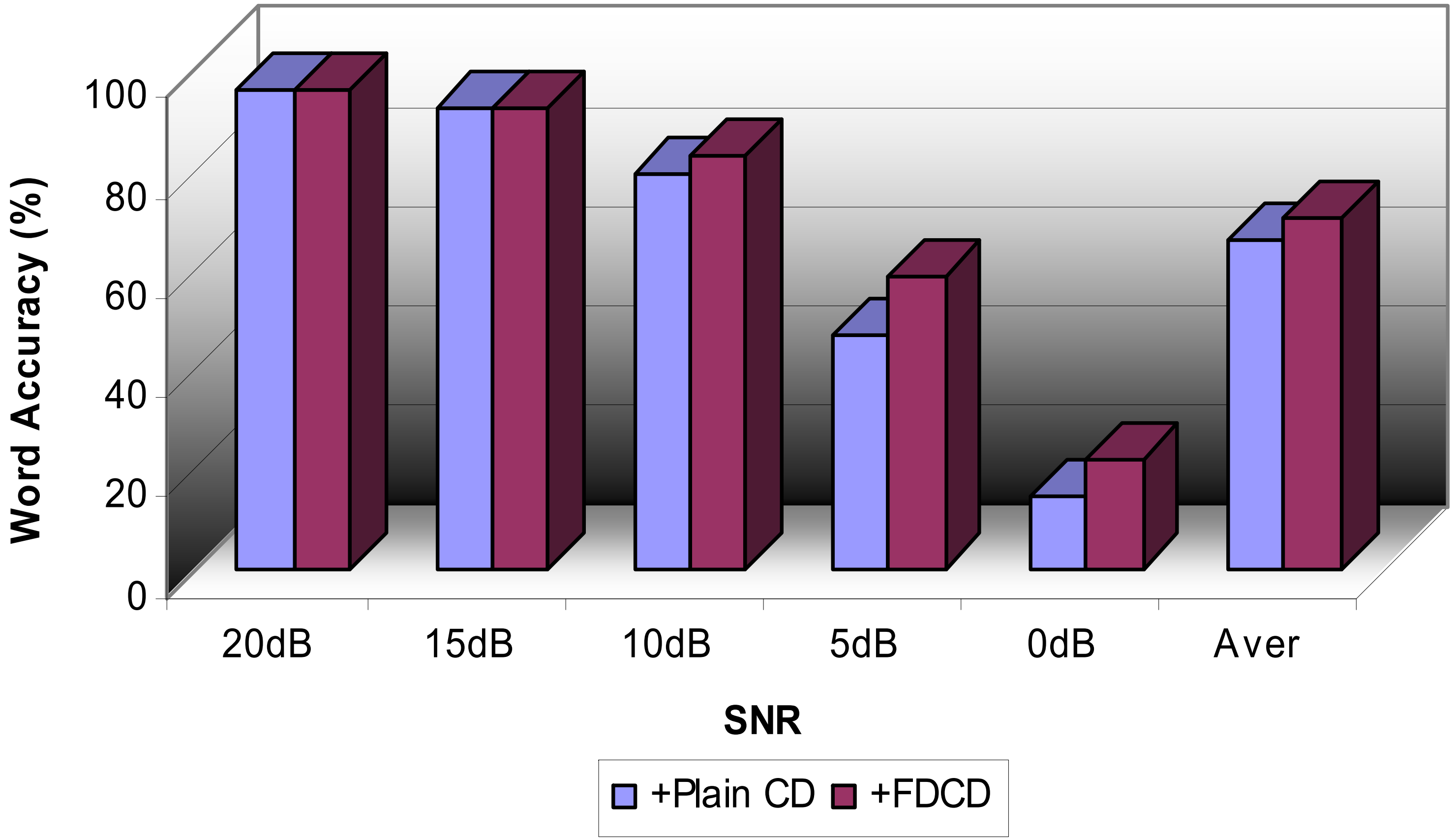


[Pitsikalis & Maragos, IEEE SPL 2006]

Noisy Speech Database: Aurora 2

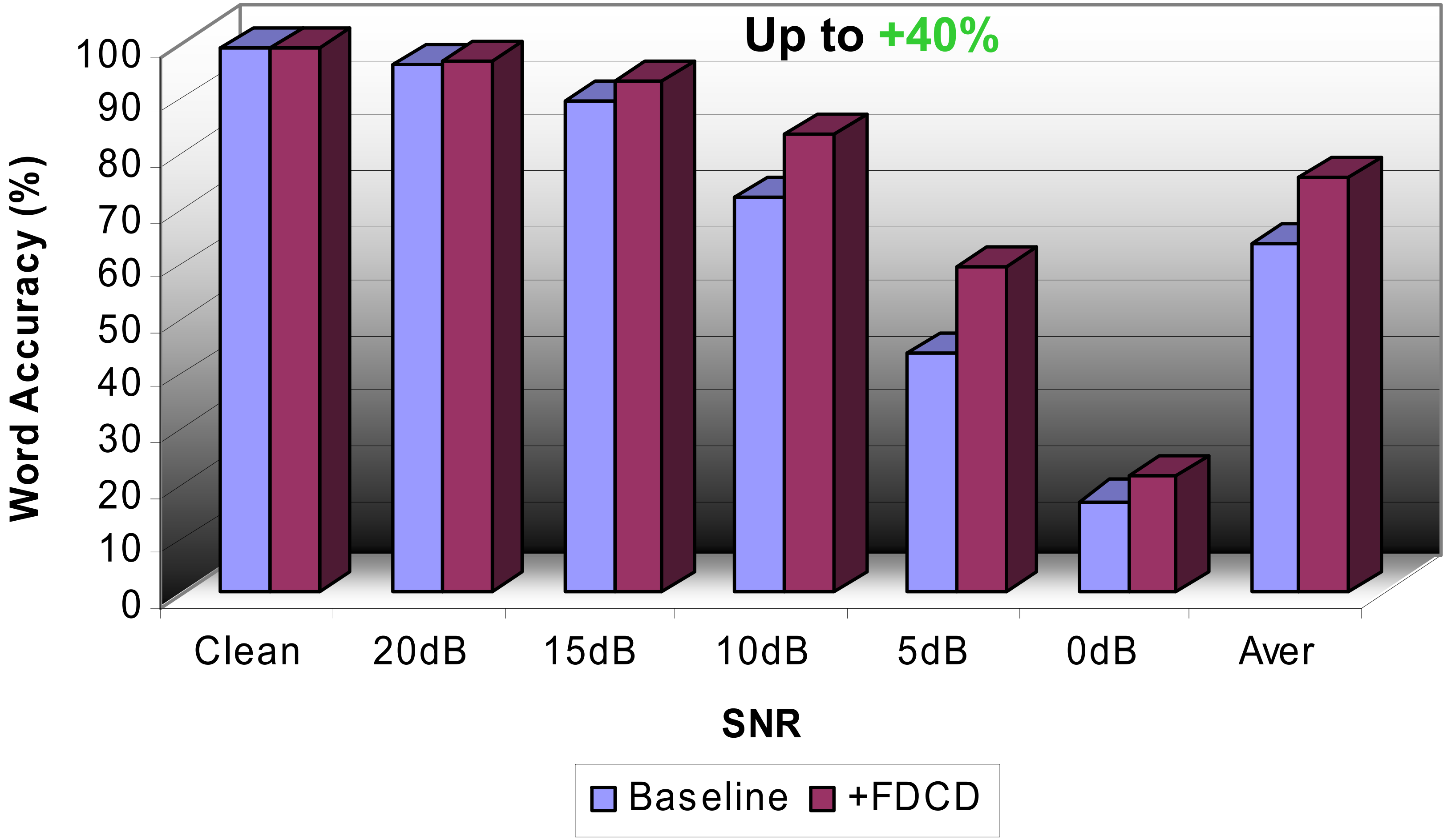
- Task: Speaker Independent Recognition of Digit Sequences
- TI - Digits at 8kHz
- Training (8440 Utterances per scenario, 55M/55F)
  - Clean (8kHz, G712)
  - Multi-Condition (8kHz, G712)
    - 4 Noises (artificial): subway, babble, car, exhibition
    - 5 SNRs : 5, 10, 15, 20dB , clean
- Testing, artificially added noise
  - **7 SNRs**: [-5, 0, 5, 10, 15, 20dB , clean]
  - **A**: noises as in multi-cond train., G712 (28028 Utters)
  - **B**: restaurant, street, airport, train station, G712 (28028 Utters)
  - **C**: subway, street (MIRS) (14014 Utters)

Average Recognition Results on Aurora 2: plain CD vs FDCD

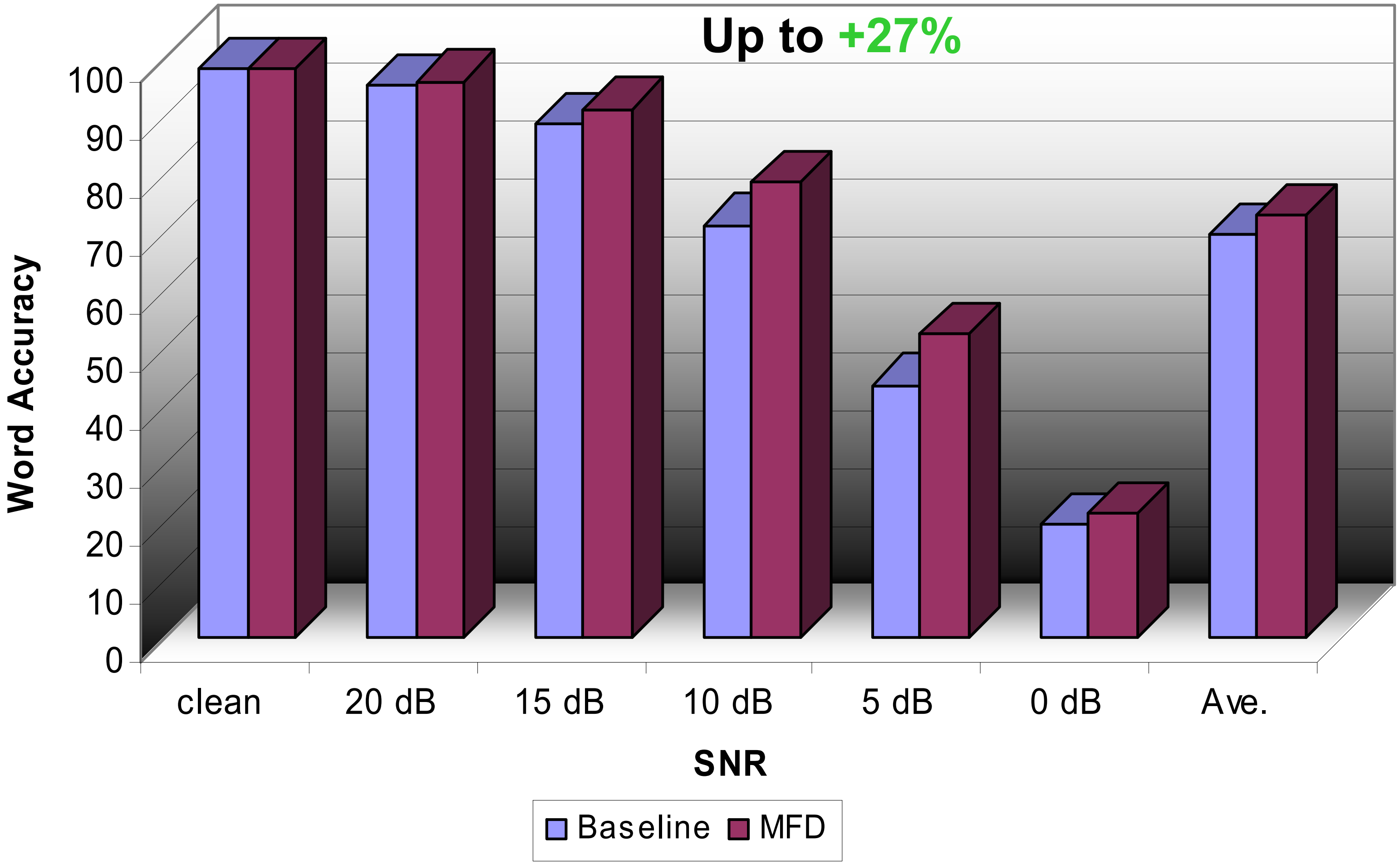


Plain CD: Correlation Dimension without Dynamical Filtering

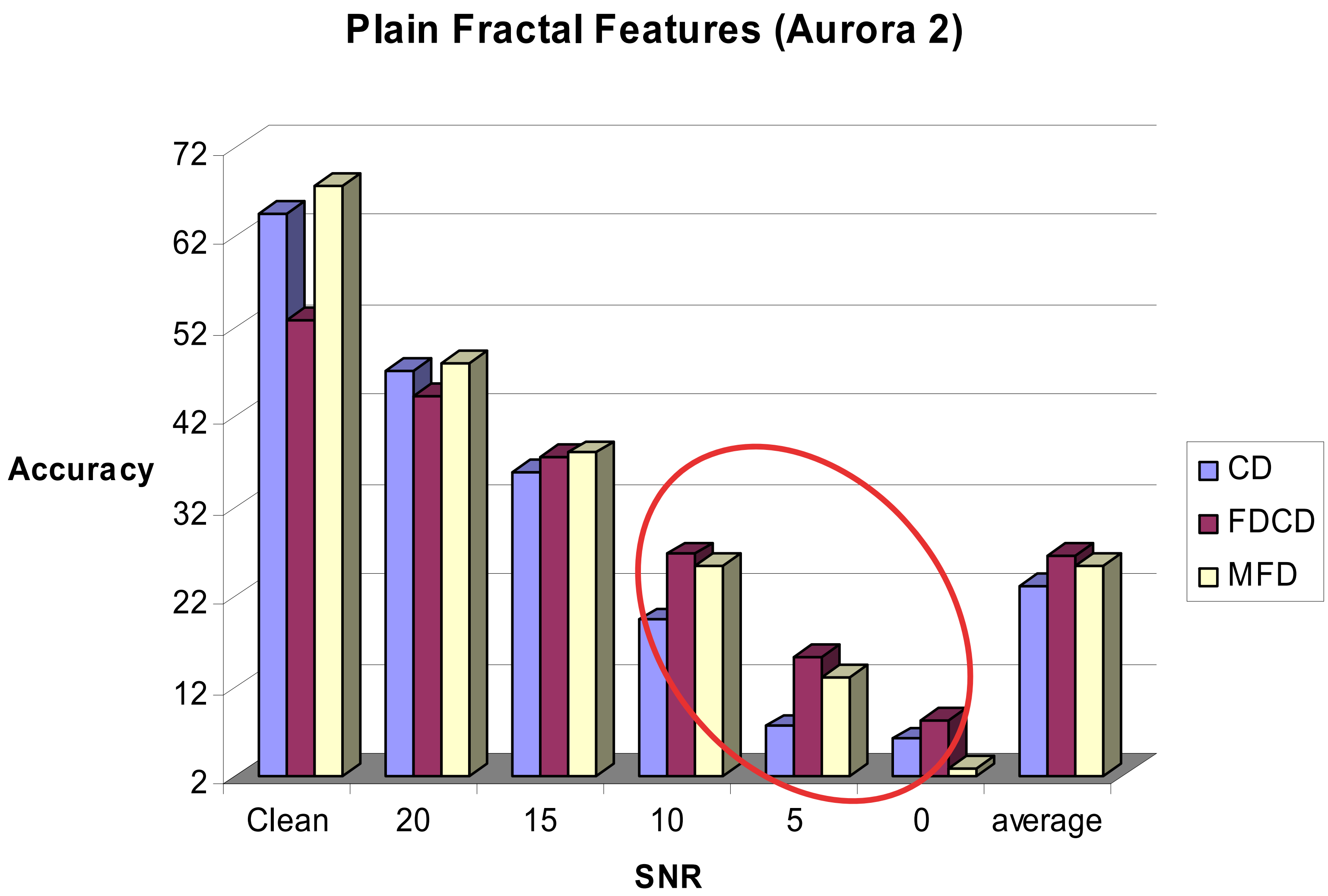
Average Recognition Results on Aurora 2: FDCD



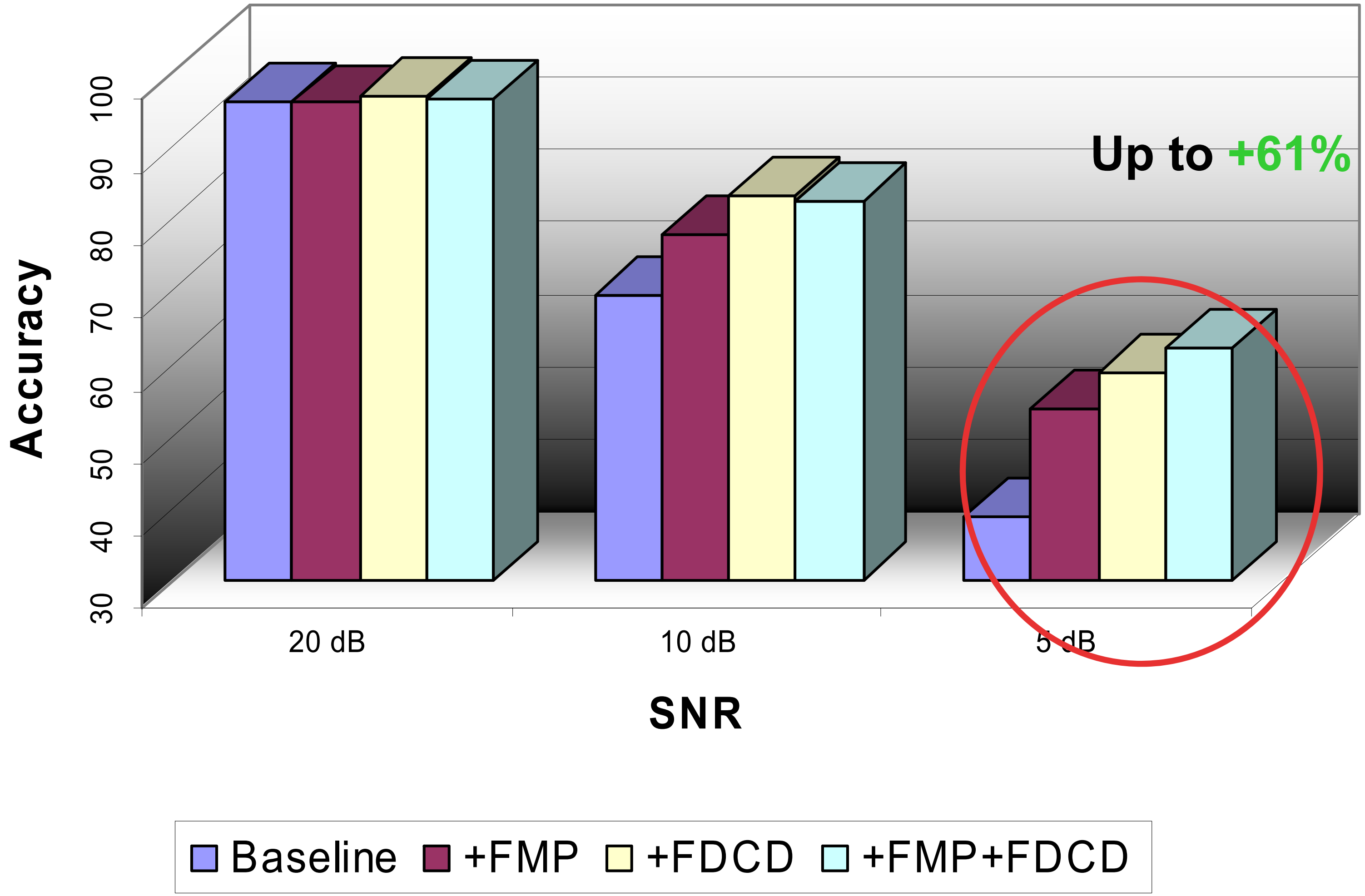
Average Recognition Results on Aurora 2: MFD



Average Recognition Results on Aurora 2

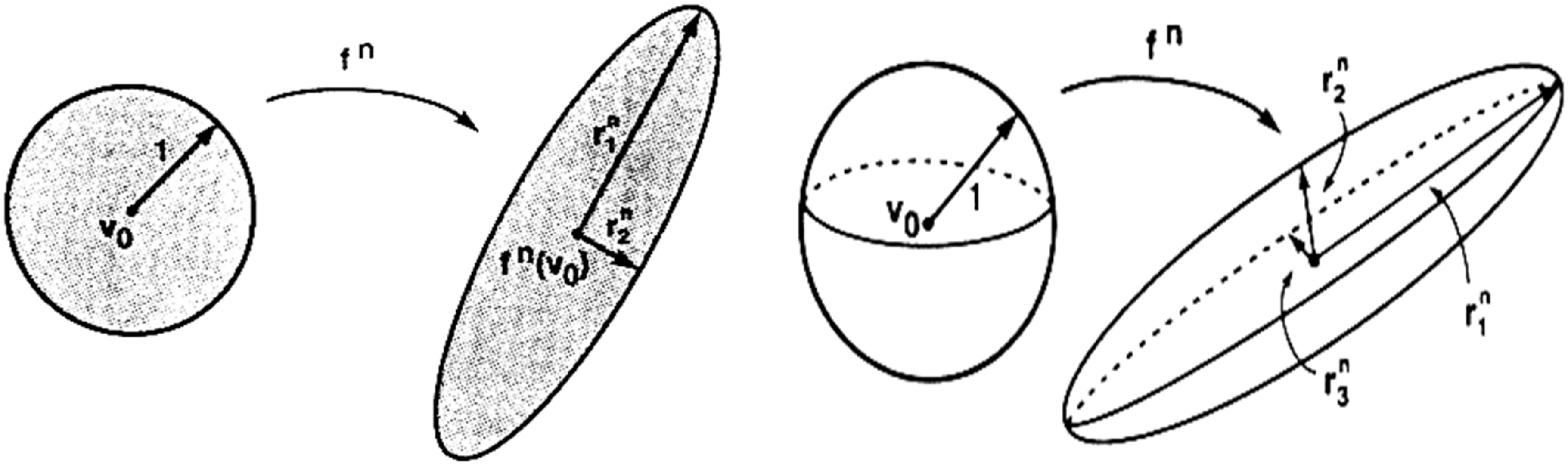


Average Recognition Results on Aurora 2: Hybrid Features: Fractals and Modulations



[Pitsikalis & Maragos, IEEE SPL 2006; Speech Commun 2009]

Lyapunov Exponents (L.E.s)



$$L_k = \lim_{n\to\infty}(r_k^n)^{1/n}$$

$$\lambda_k = \ln(L_k)$$

$$\lambda_1 > \lambda_2 > \ldots \lambda_k > \lambda_{k+1} > \ldots \lambda_{D_e}$$

Lyapunov Exponents (II)

- Quantify signal predictability
  (orbits convergence-divergence rates in phase space)
- Positive L.E. → exponential divergence
  Negative L.E. → exponential convergence
- Dissipative system → sum of L.Es <0
  Chaotic system → at least one L.E >0
- Invariants of system dynamics → useful for characterization /recognition purposes
- Determine prediction horizon
  (upper bound of system predictability)

Prediction on Reconstructed Attractor

(Kokkinos & Maragos, T-SAP 2005)

Goal: capture dynamics of MIMO system
from input-output pairs

$$X_{n+1} = \mathbf{f}(X_n)$$

Models tested: $X_{n+1} = F(X_n)$

- Local Polynomials
- Global Polynomials
- Radial Basis Function networks
- Takagi-Sugeno-Kang models
- Support Vector Machines

Computation of Lyapunov Exponents

- Consider an **orbit** $X_{n+1} = \mathbf{f}(X_n), \quad n = 1, 2, \dots, N$
- Oseledec matrix:

$$\mathbf{OSL} = \lim_{N \to \infty} \left[\mathbf{J}_F^T(X_N) \bullet \cdots \bullet \mathbf{J}_F^T(X_1) \bullet \mathbf{J}_F(X_1) \bullet \cdots \bullet \mathbf{J}_F^T(X_N) \right]$$

- **i-th L.E.** $\lambda_i = \log(s_i)$, s_i is **i-th eigenvalue** of **OSL**
- Limitations:
- Only approximation of Jacobian **J** of **f** is available

  (F is an approximation to **f**)
- Ill-conditioned nature of **OSL** →

  recursive QR decomposition technique
- Limited data set → **local L.E.s**

Validation of Lyapunov Exponents

- Inverse time sequencing of data
  - True exponents flip sign (divergence of nearby orbits becomes convergence & vice versa)
  - False exponents remain negative

    (artifact of embedding process ⇨

    no dependence on system dynamics)
- Models that learn the data (and not the system dynamics) fail to give such results.
- RBF nets, TSK-0, Global Polynomials ... *failed*
- SVM, TSK-1 *succeeded*

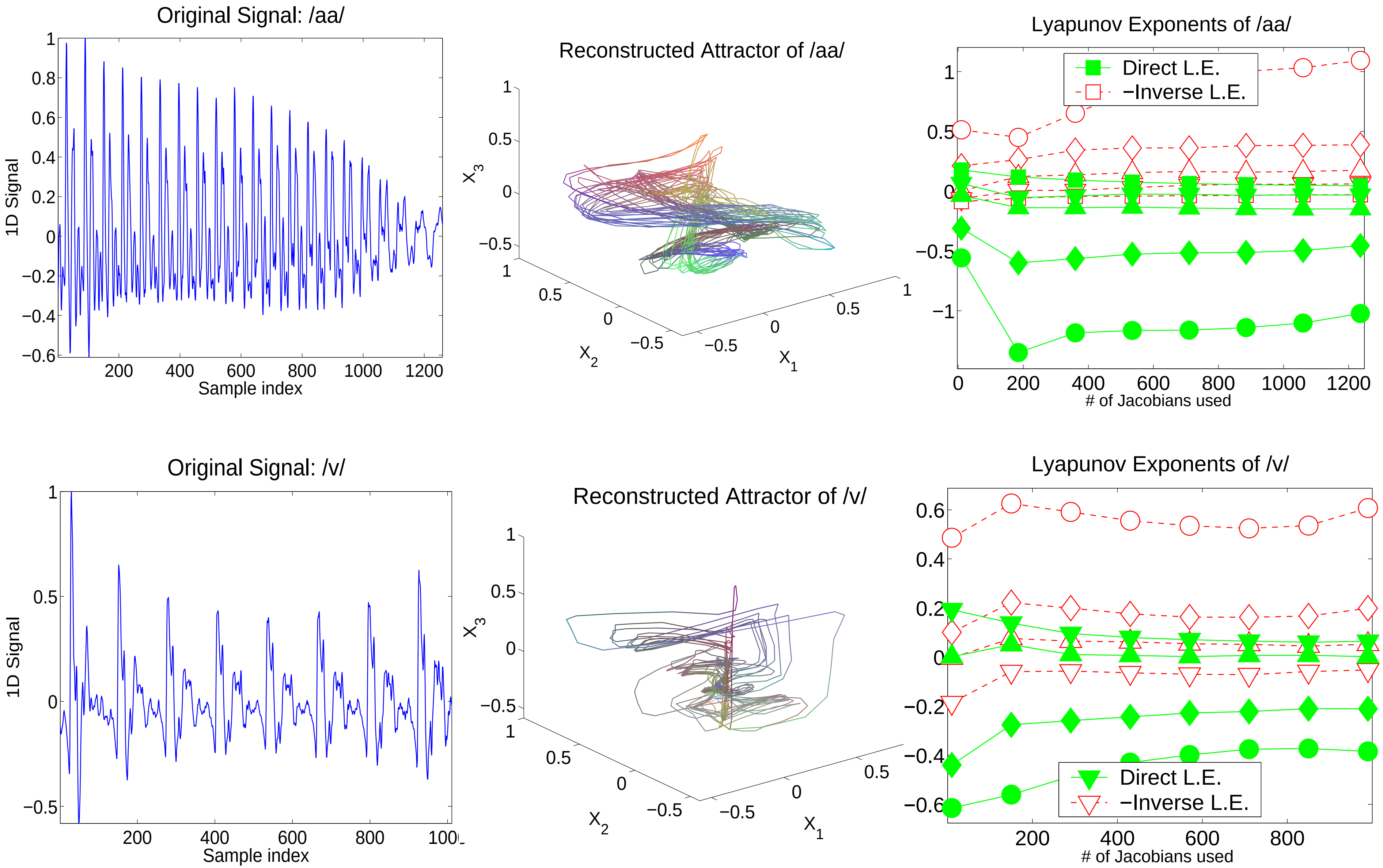
Applications to Speech Signals

(Kokkinos & Maragos 2005)

- Prediction – coding with global polynomials (smaller MSE than LPC with same # of params)
- Speech analysis using Lyapunov exponents
  - Vowels have small positive L.E.s
  - Voiced fricatives have bigger positive L.E.s
  - Unvoiced fricatives have no validated L.E.s (too noisy)
  - Stop sounds have no validated L.E.s (non-stationary)
  - Non-validated L.E.s are still useful

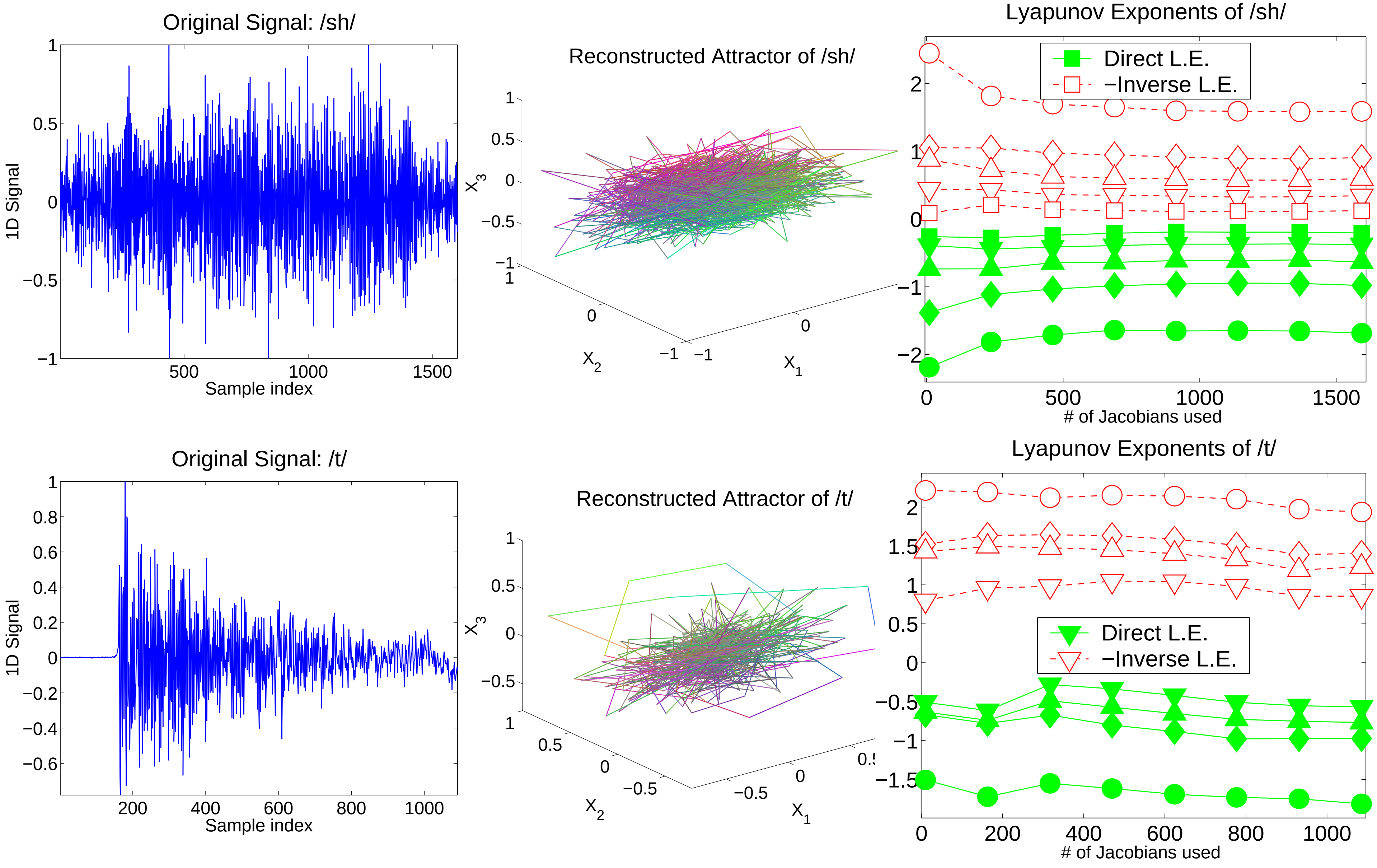
Speech Results: validated L.E.s

- Phoneme: /aa/



Speech Results: Non-validated L.E.s

- Phoneme: /sh/



Speech Lyapunov Exponents

THREE FIRST VALIDATED LYAPUNOV EXPONENTS FOR SPEECH PHONEMES

Phon./LEs	/aa/(70)	/eh/ (64)	/ih/ (59)	/ow/ (56)	/w/ (39)	/m/ (36)
λ_1	0.047±0.028	0.093±0.040	0.084±0.045	0.069±0.042	0.036±0.024	0.029±0.034
λ_2	−0.004±0.018	−0.014±0.027	−0.001±0.041	0.052±0.025	−0.009±0.015	−0.096±0.068
λ_3	−0.078±0.038	−0.139±0.048	−0.156±0.079	−0.083±0.052	−0.096±0.042	−0.289±0.142

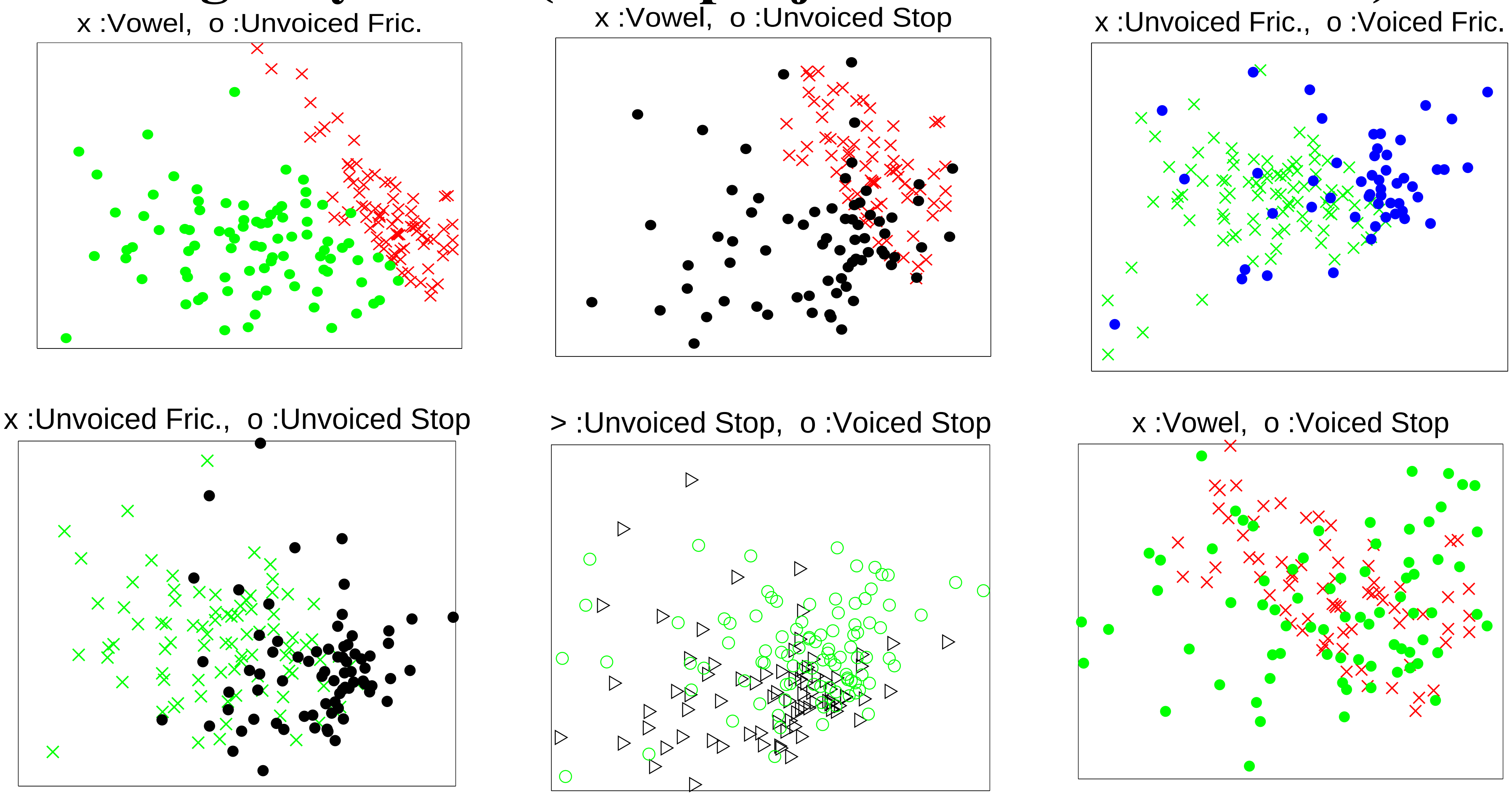
Phon./LEs	/r/ (52)	/l/ (39)	/f/ (50)	/s/ (102)	/b/ (37)	/t/ (35)
λ_1	0.074±0.038	0.048±0.035	−0.561±0.249	−0.312±0.157	−0.012±0.152	−0.296±0.254
λ_2	−0.012±0.030	−0.013±0.022	−0.772±0.260	−0.504±0.172	−0.047±0.277	−0.492±0.293
λ_3	−0.118±0.096	−0.099±0.069	−0.997±0.274	−0.725±0.217	−0.361±0.303	−0.710±0.323

Next to each phoneme is given the number of time series from which the statistics have been calculated; for robustness the median and the mean absolute deviation from the median are used instead of the mean and the standard deviation. The phonemes have been uttered by 11 speakers. For all phonemes, approximately the same number of pronunciations is used from every speaker. For all the vowels/semivowels in this table the exponents have been validated using the LEs of the inverse time series. For fricatives and unvoiced stops these are not validated, but used merely as features for classification; no conclusions should be drawn from these. One should note the increase in the variation of the LEs for the latter classes.

[Kokkinos & Maragos, IEEE T-SAP 2005]

Speech Sound Classification

➢ Using only L.E.s (PCA projection of 3 first L.E.s):



➢When combined with MFCC: (4 classes)

~12% smaller error using K-NN classifier

Other Works on Speech Fractals or Chaotic Dynamics

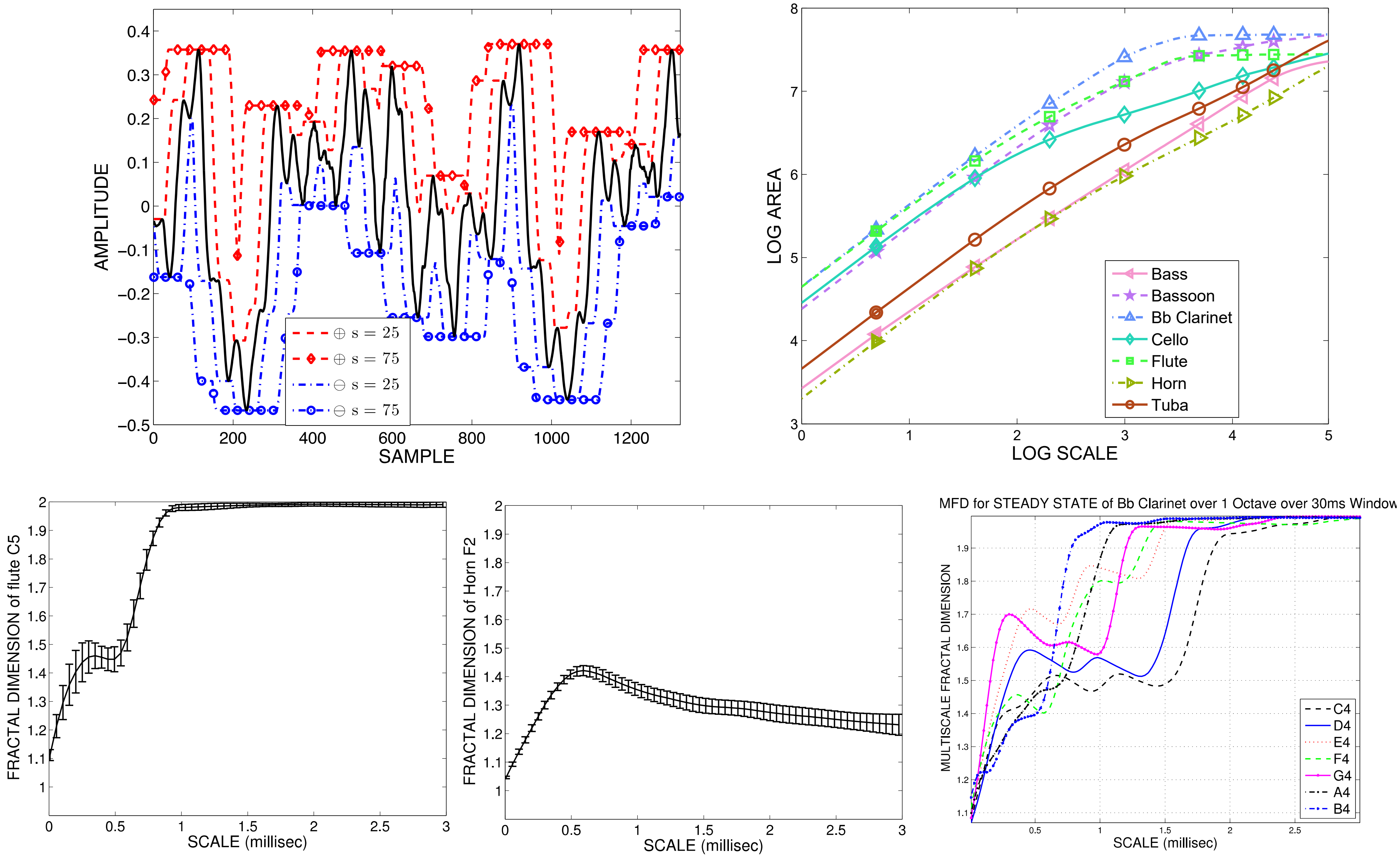
- C. A. Pickover and A. Khorasani, "*Fractal Characterization of Speech Waveform Graphs*," Computer Graphics 1986.
- P. J. B. Jackson and C. H. Shadle, "*Frication noise modulated by voicing, as revealed by pitch-scaled decomposition*", J. Acoust. Soc. Amer. 2000.
- S. McLaughlin and P. Maragos, "*Nonlinear Methods for Speech Analysis and Synthesis*", in Advances in Nonlinear Signal and Image Processing, edited by S. Marshall and G. L. Sicuranza, EURASIP Book Series on Signal Processing and Communications, Hindawi Publ. Corp., 2006, pp.103-140.
- M. Zaki, J. N. Shah and H. A. Patil, "*Effectiveness of Multiscale Fractal Dimension-based Phonetic Segmentation in Speech Synthesis for Low Resource Language*", in Proc. Int'l Conf. on Asian Language Processing (IALP) 2014.
- K. López-de-Ipina, J. Solé-Casals, H. Eguiraun, J.B. Alonso, C.M. Travieso, A.Ezeiza, N Barroso, M. Ecay-Torres, P. Martinez-Lage, Blanca Beitia, "*Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach*", Computer Speech & Language 2015.
- E. Tzinis, G. Paraskevopoulos, C. Baziotis, A. Potamianos, "*Integrating Recurrence Dynamics for Speech Emotion Recognition*", in Proc. Interspeech 2018.

Fractals and Music

Ref:

- A. Zlatintsi and P. Maragos, "*Multiscale Fractal Analysis of Musical Instrument Signals with Application to Recognition*", IEEE Transactions on Audio, Speech and Language Processing, Apr. 2013.

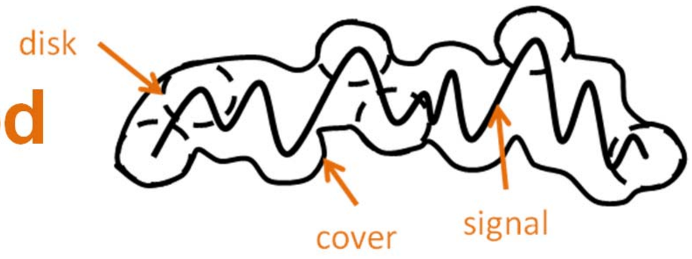
Multiscale Fractal Dimension of Music Sounds



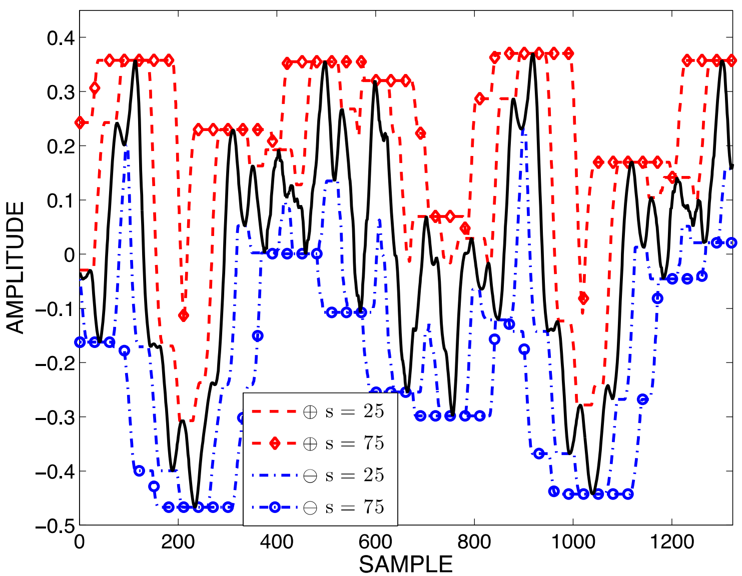
[Zlatintsi & Maragos, T-ASLP 2013]

Multiscale Fractal Analysis of Musical Instrument Signals

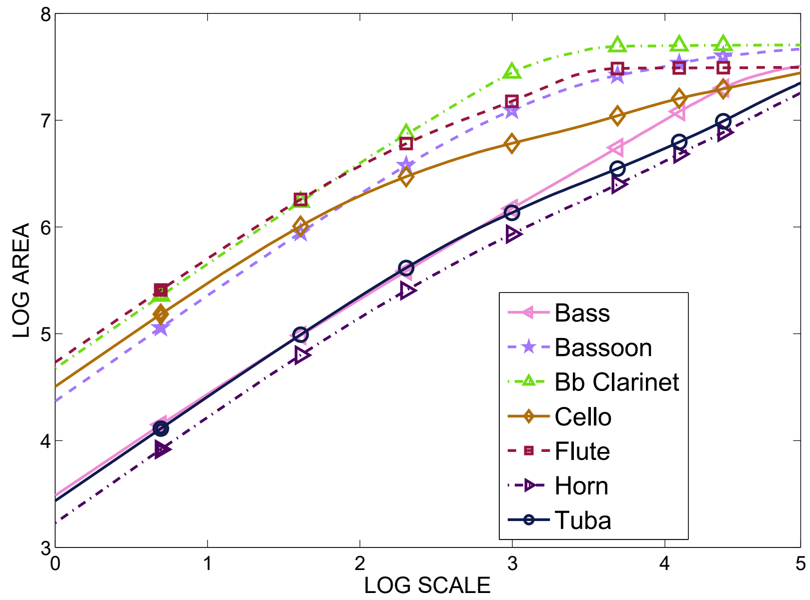
Morphological Covering Method



$$D = 2 - \lim \frac{\log[A_B(s)/s^2]}{\log(1/s)}$$



Double Bass steady state (solid line), its multiscale flat dilations and erosions at scales s=25,75, where B is a 3-sample symmetric horizontal segment with zero height.

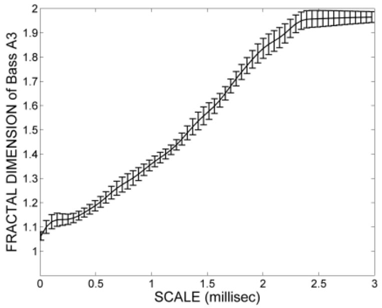


log[AB(s)] vs log(s) for the seven analyzed instruments for the note C3 except for Bb Clarinet and Flute shown for C5 instead. Note the difference in the slope for larger scales . (for 30ms analysis window).

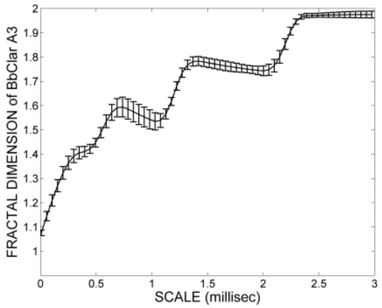
P. Maragos and A. Potamianos. *J. Acoust. Soc. Amer.*, 1999.

MFD Analysis for Steady State of the Note

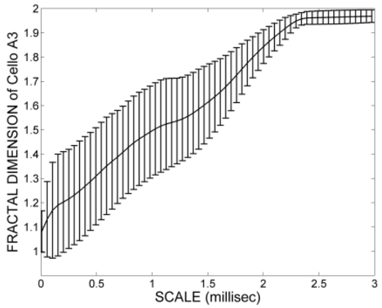
A3



Upright Bass

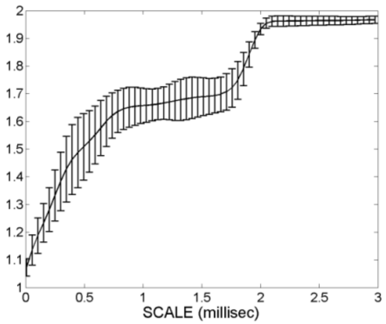


Clarinet

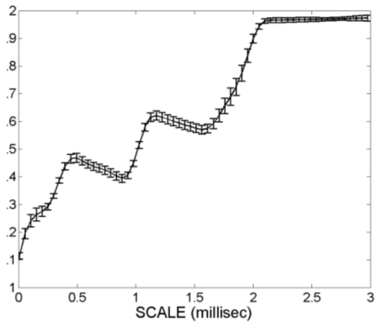


Cello

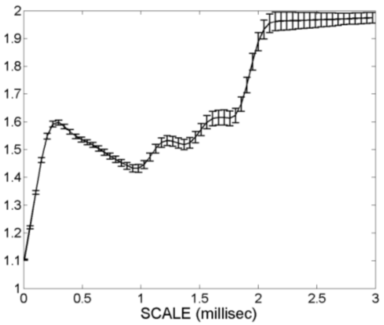
B3



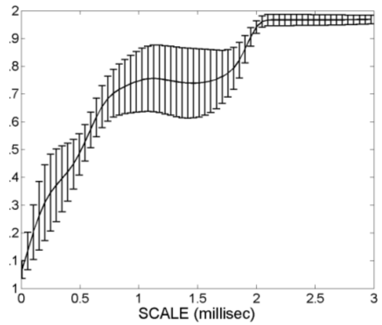
Flute



Clarinet



Oboe

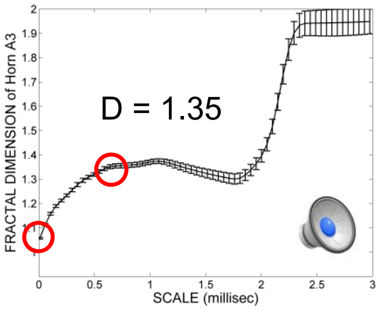


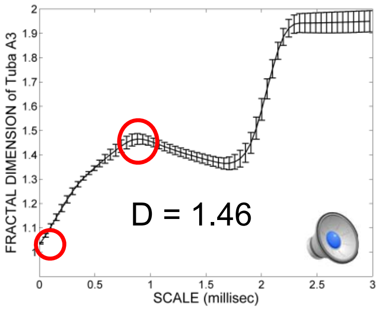
Piano

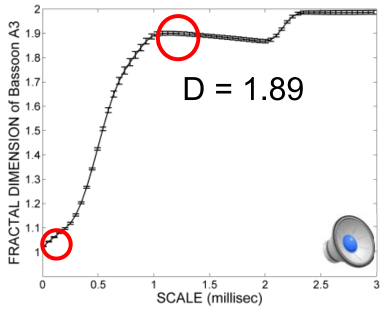
Mean MFD (middle line) and standard deviation (error bars)
(for 30 ms analysis window, updated every 15 ms).

MFD Analysis for Steady State of the Note

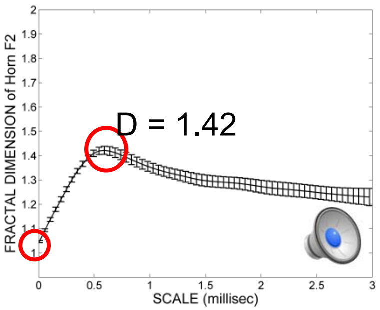
A3

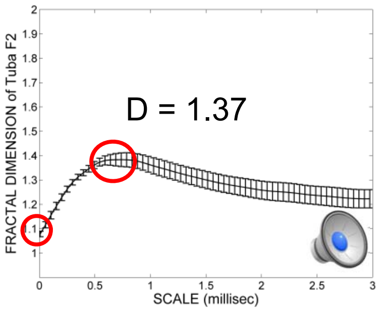


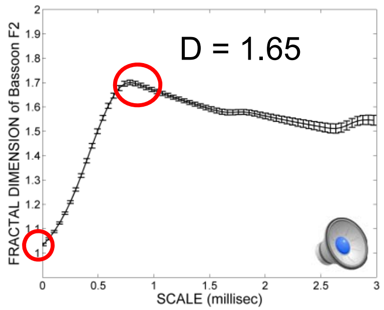




F2

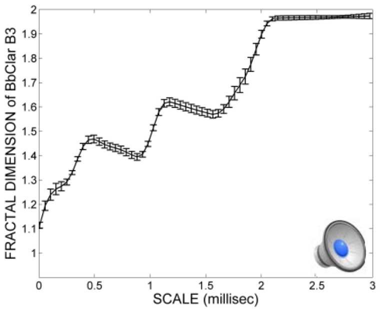


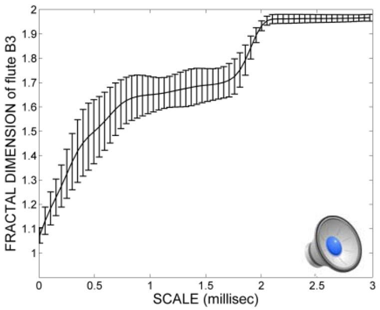


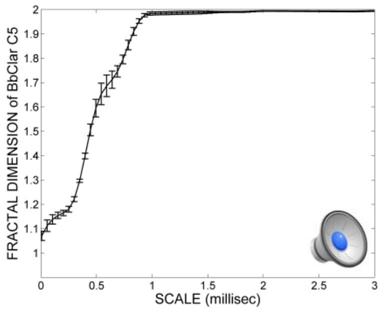


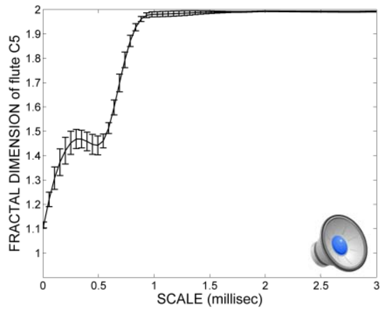
Horn Tuba Bassoon

B3 C5





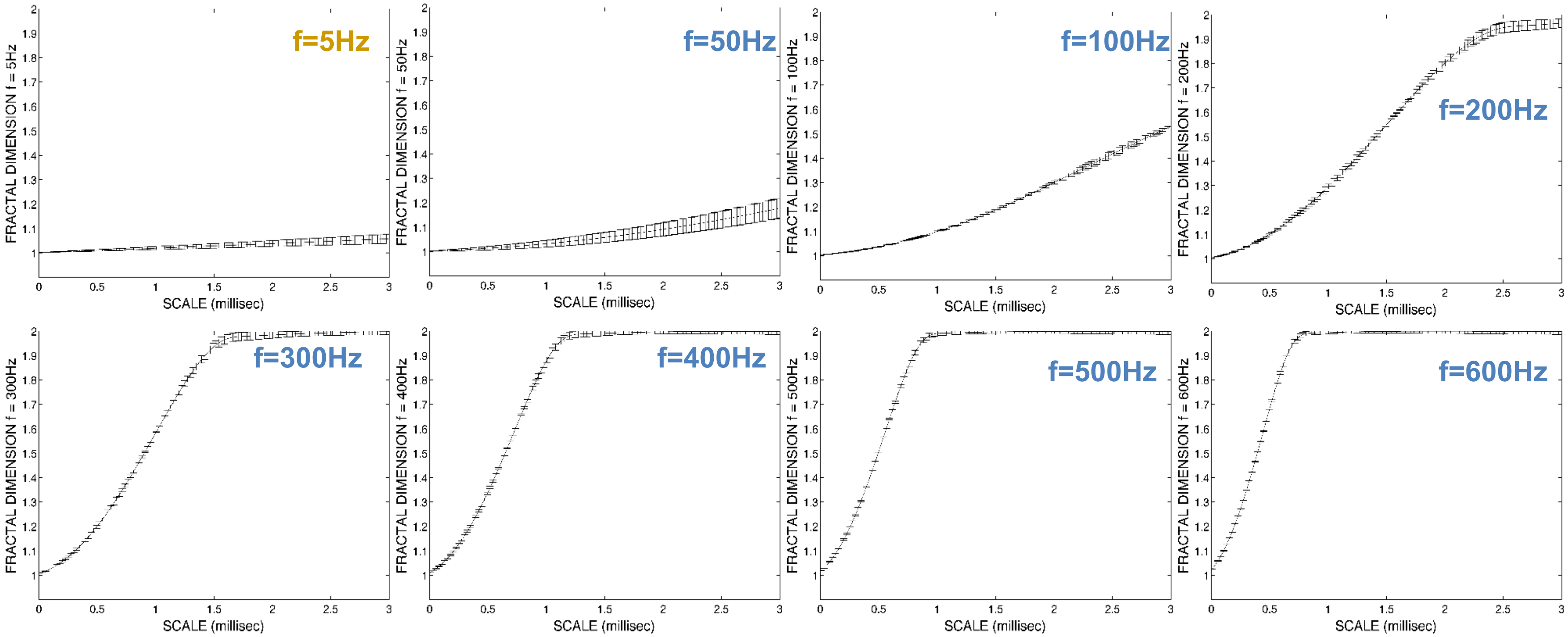




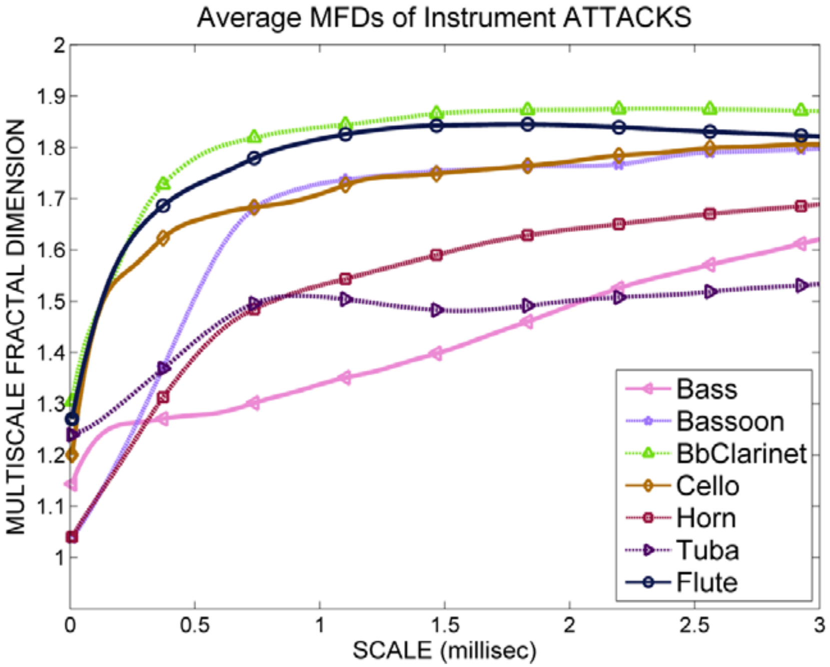
Bb Clarinet Flute Bb Clarinet Flute

MFD Analysis on Synthesized Signals

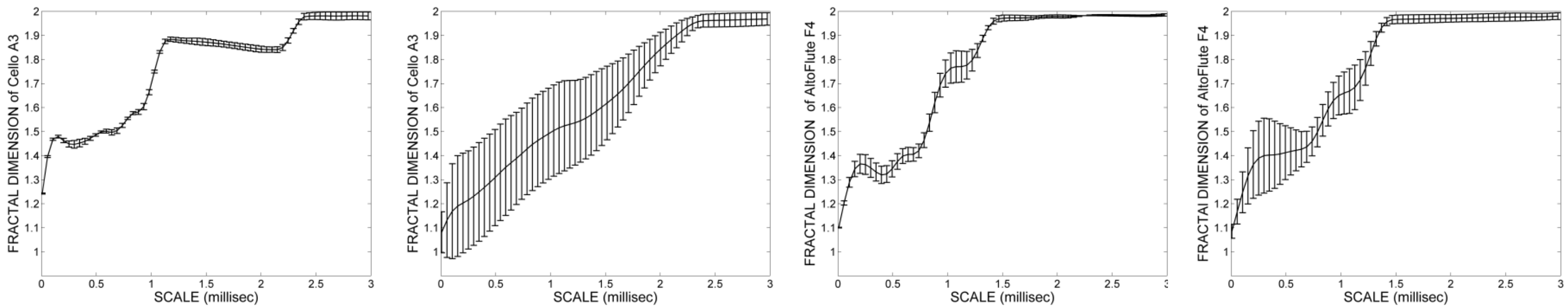
- Strong dependence on the frequency



Analysis of MFD during the Attack

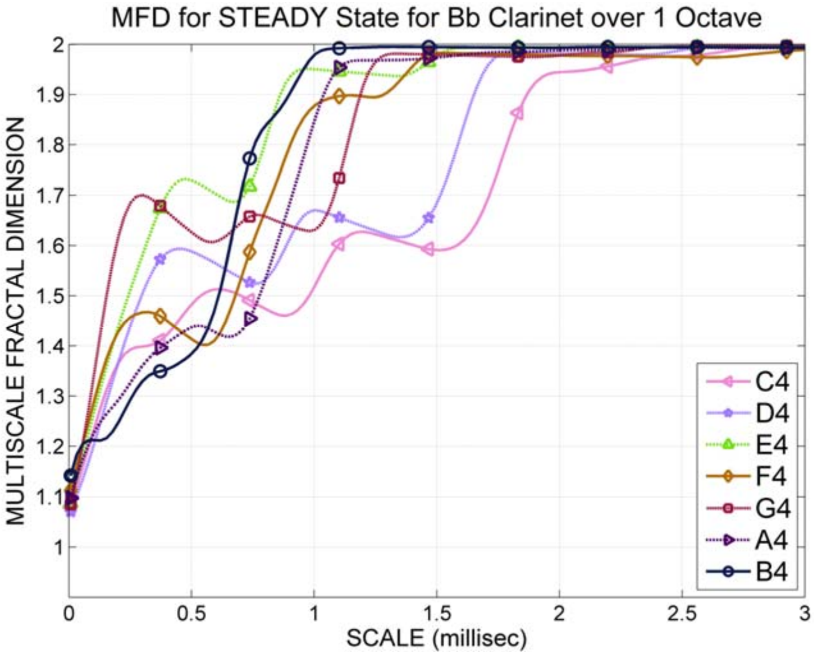


- MFDs estimated for the 7 analyzed instruments attacks, averaged over the whole range (using 30 ms analysis windows).
  - Similar as for the steady state
  - Higher D for small scales s_t and more fragmentation.
  - Increased value of D(s = 1).
  - Clear distinction of D among some of the analyzed instruments.



Mean MFD and standard deviation of the attack and steady state of A3 for Cello (left images) and F4 for Flute (right images).

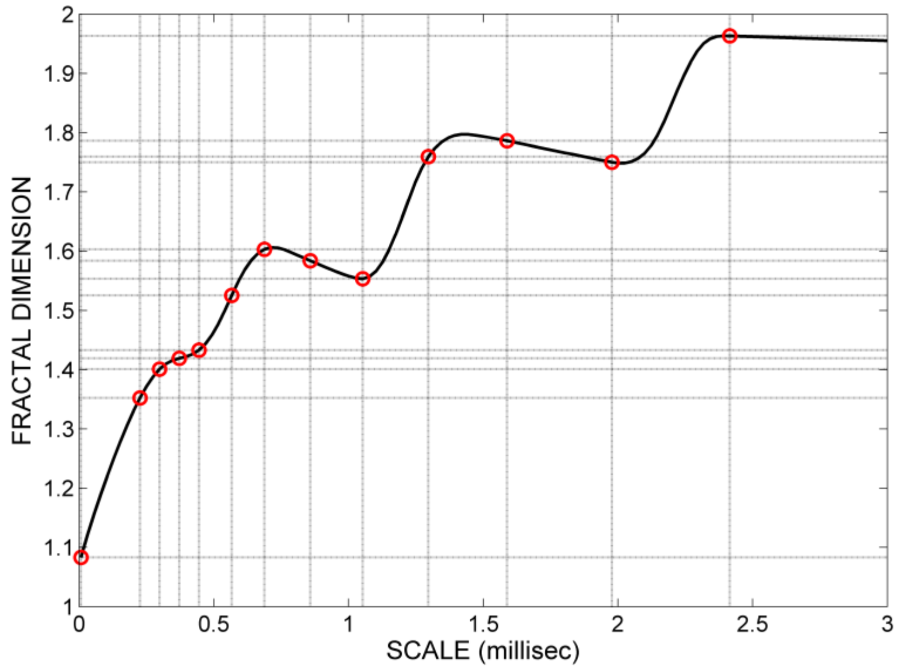
MFD Variability of the Steady State for the Same Instrument over One Octave



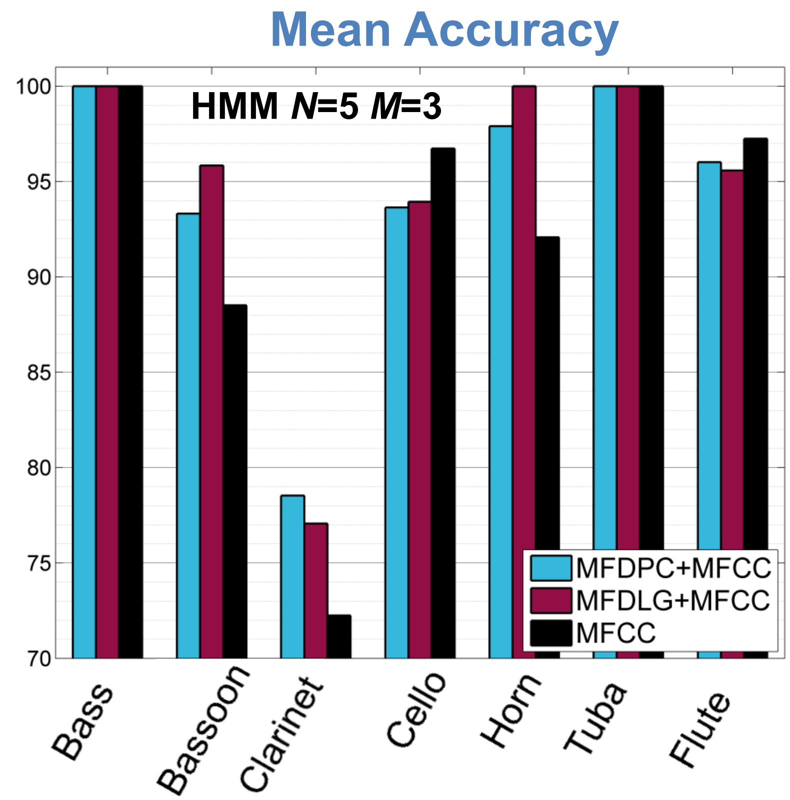
MFD of Bb Clarinet steady state notes, over one octave for one 30 ms analysis window.

- Dependence on the acoustical frequency and the MFD profile that increases rapidly for higher frequency sounds.
- The instruments' specific MFDs beholds the shape observed for the specific octave.

Experimental Evaluation



Example of the 13 logarithmically sampled points of the MFD, for Bb Clarinet (A3), forming the MFDLG feature vector.



- Double Bass, Bassoon & Tuba best recognized
- Low discriminability between Bb Clarinet & Flute

- Enhanced discriminability for Bassoon, Bb Clarinet and Horn
- Decreased for Cello & Flute
- On average MFD+MFCC features improve the recognition over the baseline

Conclusions

- **Existence-Importance of nonlinear speech structure of turbulence type (fractals, chaotic dynamics)**
- **Speech technology systems can benefit from including such nonlinear features**
- **Find/extract robust nonlinear features of turbulence type**
- **Improve computational algorithms**
- **Fuse nonlinear with linear features**
- **Applications also to other sound signals, e.g. music**

For more information, demos, and current results:
http://cvsp.cs.ntua.gr and http://robotics.ntua.gr