Towards Automatic Speech Recognition In Adverse Environments

D. Dimitriadis, A. Katsamanis, P. Maragos, G. Papandreou and V. Pitsikalis

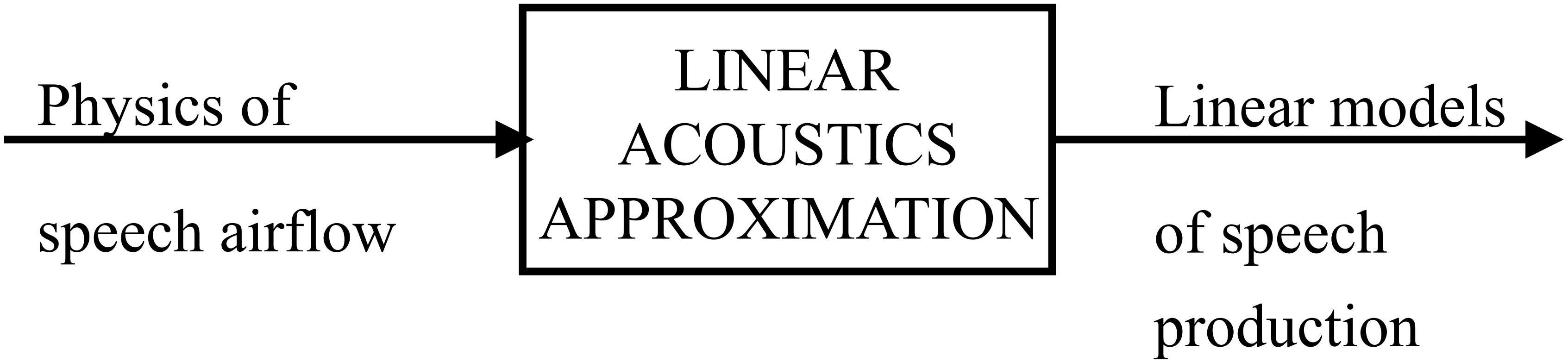
NTUA, School of ECE, http://cvsp.cs.ntua.gr

Computer Vision, Speech Communication and Signal Processing Research Group

Outline

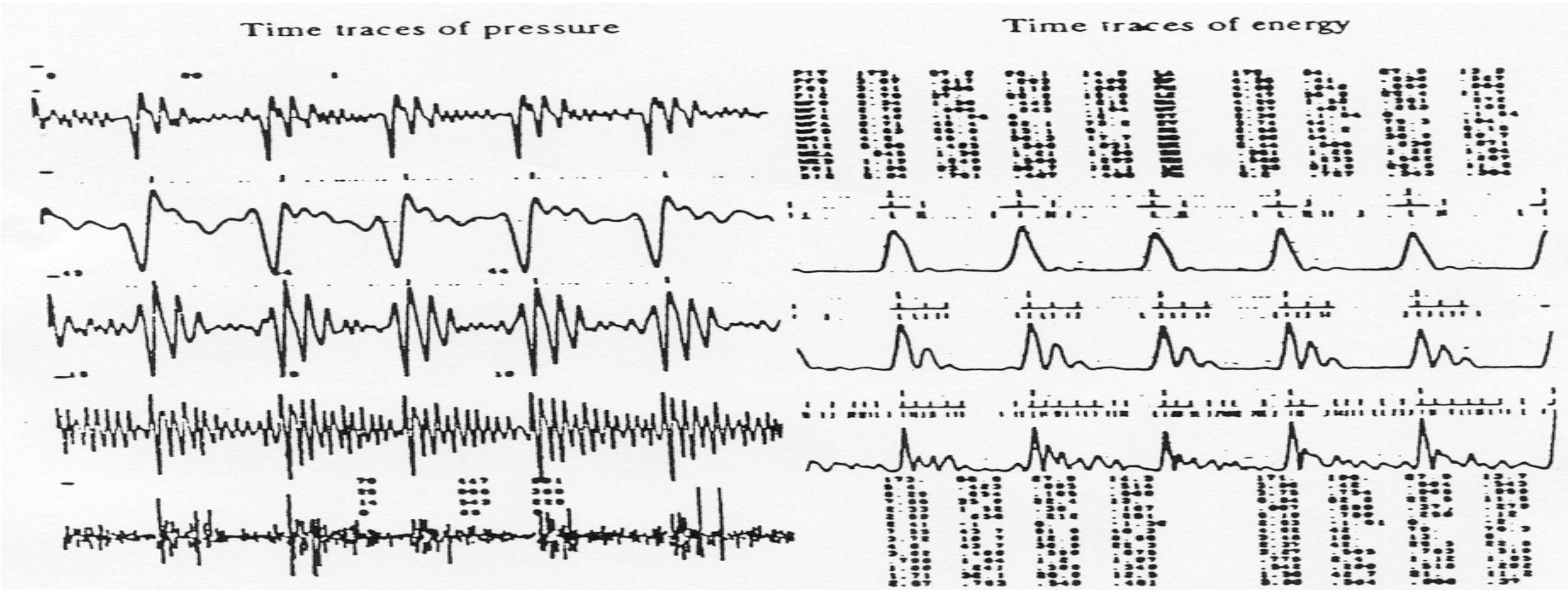
- Nonlinear Speech Processing: Background and Recent work
  - Modulations
  - Fractals
- Audio-Visual Processing
- Adaptation
- Applications to Robust ASR

Linear Model



Evidence for Speech Modulations

- separated & unstable airflow
- vortices
- oscillators with time-varying elements
- energy pulses (Teager)



Speech Modulation Model

(Maragos, Kaiser & Quatieri 1991)

- **One Single Resonance as damped AM–FM:**

$$S(t) = \underbrace{A(t)e^{-\sigma t}}_{a(t)} \cos\underbrace{\left(\omega_c t + \int_0^t q(\tau)d\tau + \theta \right)}_{\phi(t)}$$

Inst.Frequency: $\omega(t) = 2\pi f(t) = \frac{d}{dt}\phi(t) = \omega_c + q(t)$

- **If due to 2nd-order LTI system**

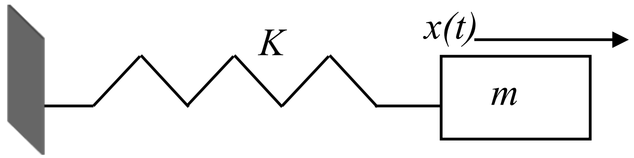
$$A(t) = \text{constant}, \qquad \omega(t) = \omega_c$$

- **Speech Signal as multi-component AM-FM:**

$$\text{Speech}(t) \approx \sum_k a_k(t)\cos(\phi_k(t))$$

Energy Tracking in Oscillators

- harmonic oscillator



- motion equation

$$\boxed{m\ddot{x} + kx = 0}$$

- response

$$\boxed{\begin{aligned} &x(t) = A\cos(\omega t + \theta), \\ &\omega^2 = k/m \end{aligned}}$$

- energy

$$\boxed{E = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}kx^2 = \frac{m}{2}(A^2\omega^2) = \text{constant}}$$

- energy tracking

$$\boxed{\Psi(x) = (\dot{x})^2 - x\ddot{x} = A^2\omega^2 = \frac{E}{(m/2)}}$$

Energy Operators

(Teager, Kaiser 1990)

- Continuous-time signals *x*(*t*):

property:

$$\boxed{\Psi_c\left[x(t)\right] \equiv [\dot{x}(t)]^2 - x(t)\ddot{x}(t)}$$

- Discrete-time signals *x(n)* :

$$\Psi_c\left[A e^{rt} \cos(\omega_c t + \theta)\right] = A^2 e^{2rt} \omega_c^2$$

property:

$$\boxed{\Psi_d\left[x(n)\right] \equiv x^2(n) - x(n+1)x(n-1)}$$

$$- + = \Psi_{dc} \otimes A r^{rn} \Omega \cos \sin(\Omega) \quad N \quad ^{222}$$

Energy Separation Algorithm (ESA)

(Maragos, Kaiser & Quatieri 1991)

- **Cosine:** $$x(t) = A\cos(\omega_c(t) + \theta)$$

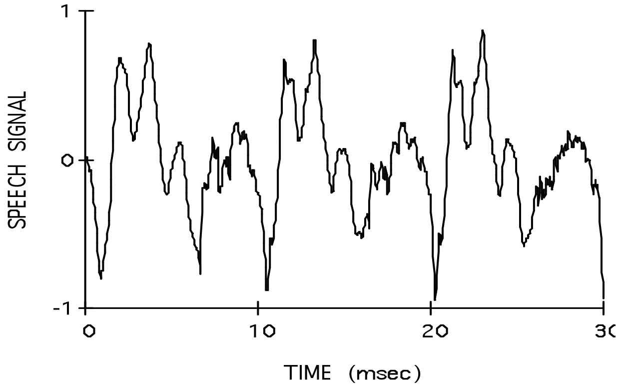
$$\Psi[x(t)] = A^2\omega_c^2 \qquad \Psi[\dot{x}(t)] = A^2\omega_c^4$$

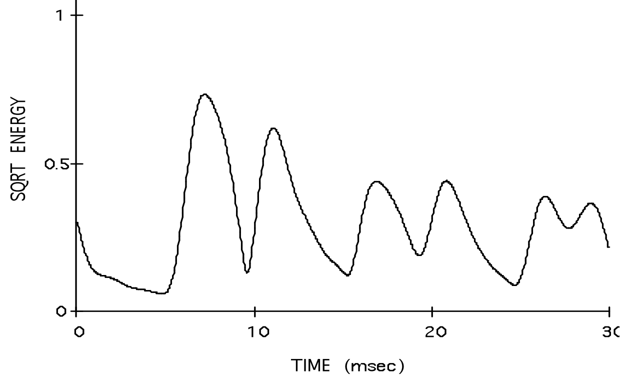
- **AM-FM signal:** $$x(t) = a(t)\cos\left(\int_0^t \omega(\tau)d\tau\right)$$

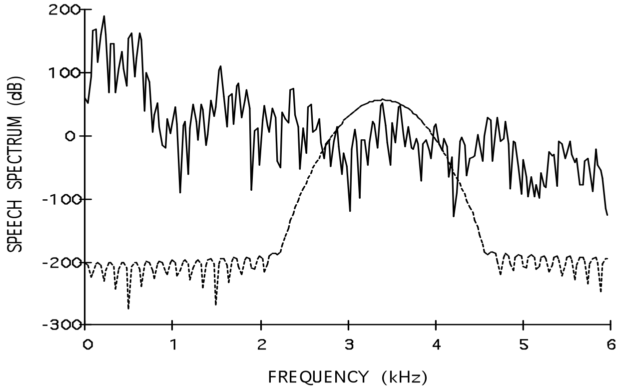
$a(t)$, *ω(t)* do not vary too fast or too much w.r.t. ω_c

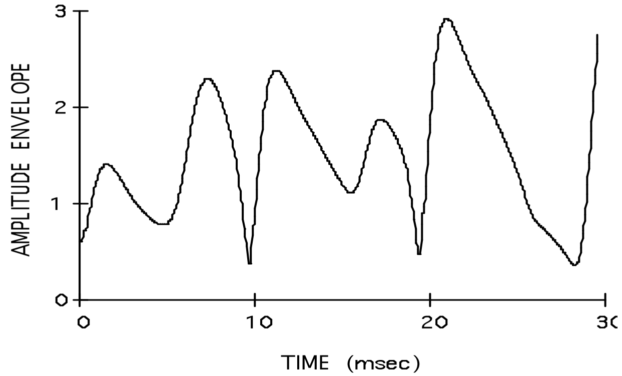
$$\frac{-[x(t)]}{\sqrt{-[\dot{x}(t)]}} \} \left|a(t)\right| \qquad \sqrt{\frac{-[\dot{x}(t)]}{-[x(t)]}} \} T(t)$$

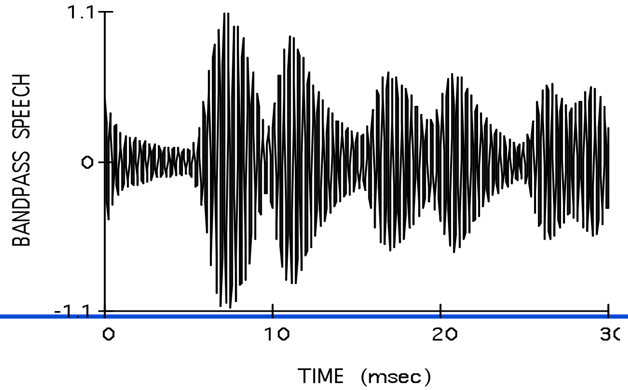
ESA Applied to Speech Resonance

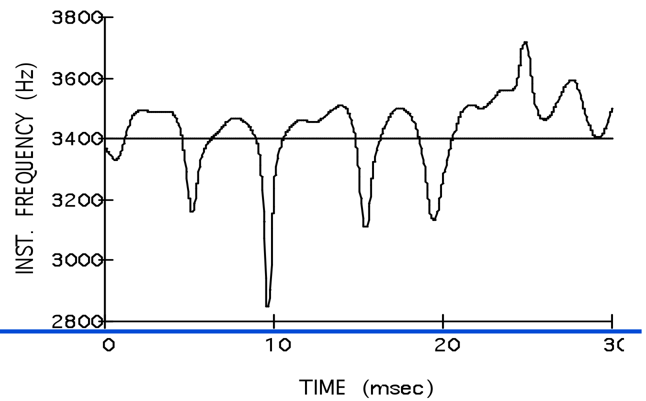






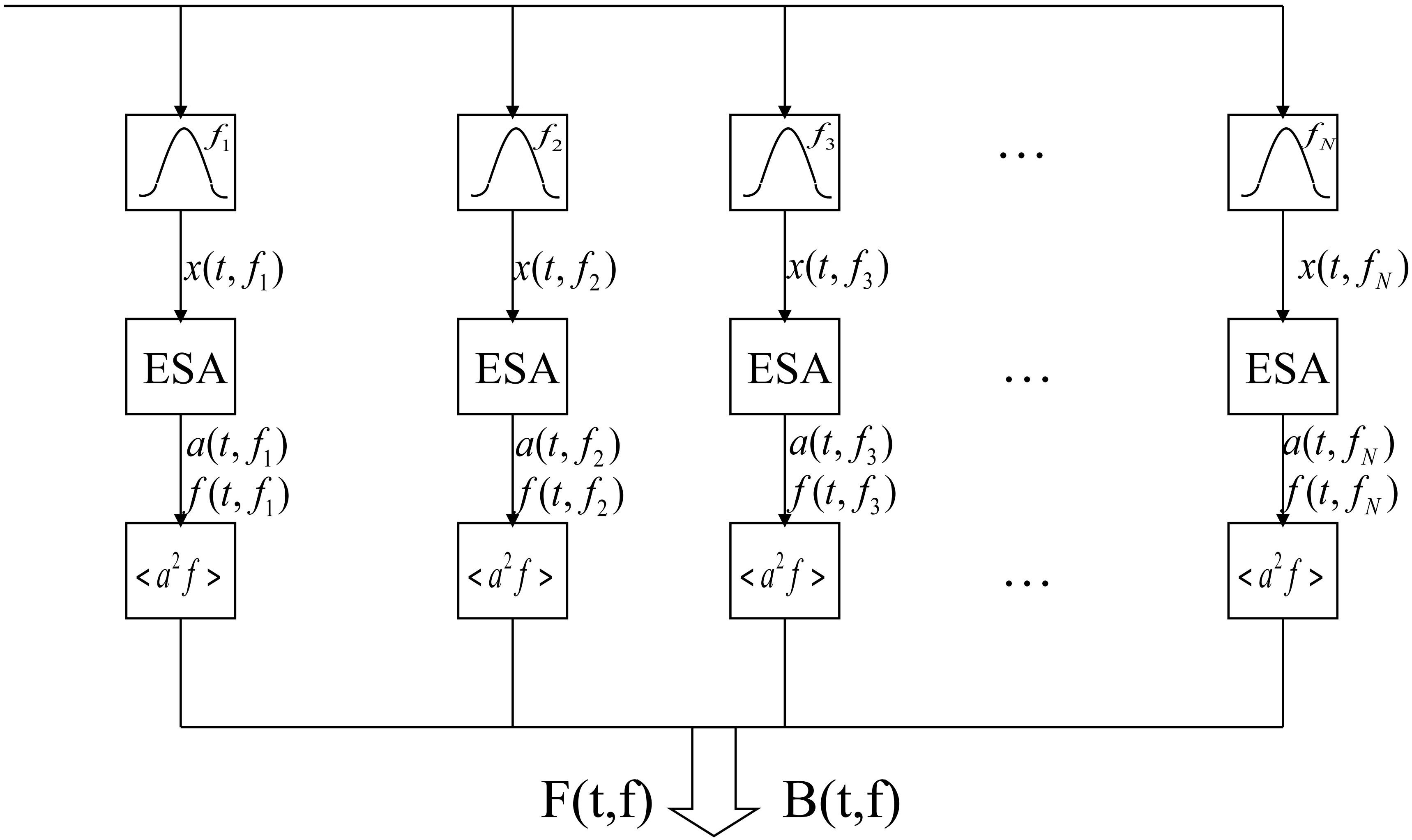






Multiband Demodulation and F/B Tracking

(Potamianos & Maragos 1996)



Frequency and Bandwidth Estimates

- **Center Frequency Estimates:**

$$F_u = \frac{1}{T}\int_0^T f(t)\,dt$$

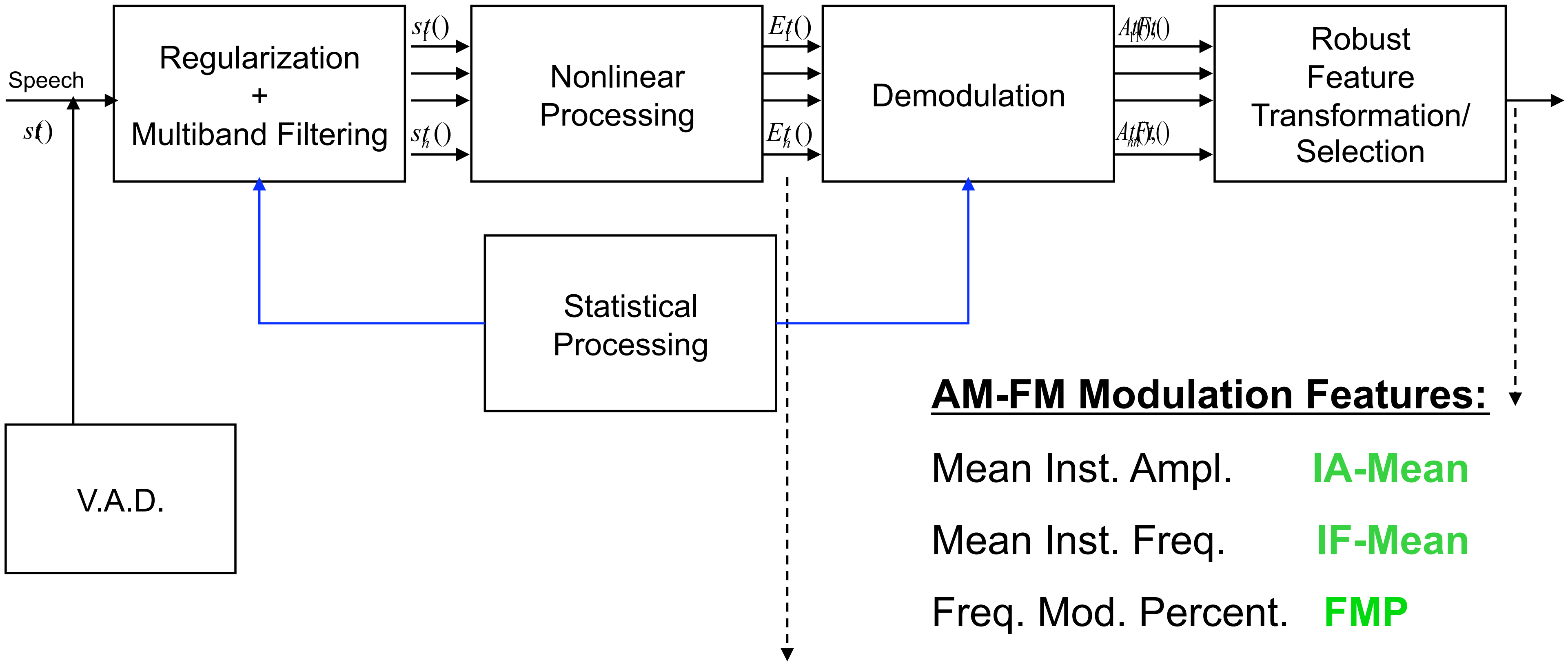
$$F_w = \frac{\int_0^T f(t)\,a^2(t)\,dt}{\int_0^T a^2(t)\,dt}$$

- **Bandwidth Estimates:**

$$B_u^2 = \frac{1}{T}\int_0^T \left(f(t) - F_u\right)^2 dt$$

$$B_w^2 = \frac{\int_0^T \left[\left(\frac{\dot{a}(t)}{2\pi}\right)^2 + \left(f(t) - F_w\right)^2 a^2(t)\right] dt}{\int_0^T a^2(t)\,dt}$$

Modulation Acoustic Features



AM-FM Modulation Features:

Mean Inst. Ampl. **IA-Mean**

Mean Inst. Freq. **IF-Mean**

Freq. Mod. Percent. **FMP**

Energy Features:

Teager Energy Cepstrum Coeff. TECC

Auditory Filter-banks

- Dense asymmetrical filters estimate activity in each frequency band
- Equivalent Rectangular Bandwidth (ERB) used to quantify bandwidth of filters
- Gammatone filters used to approximate auditory filters
- Bi-parametric filter-bank design
  - Parameter 1: Number of filters
  - Parameter 2: Filter bandwidth as percent of ERB

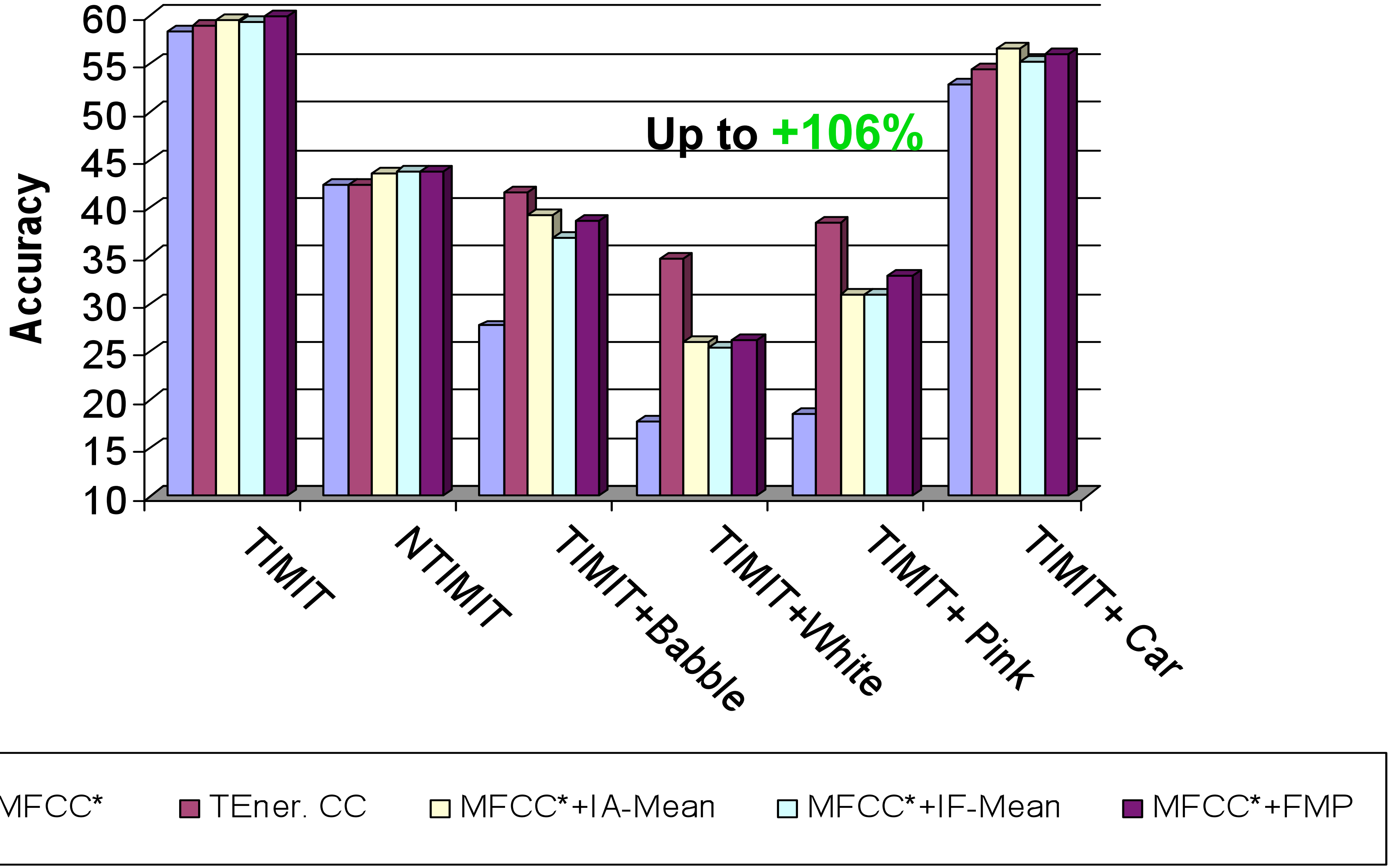
Teager Energy Cepstrum Coefficients (TECCs)

- Use a Gammatone filterbank (with 20-40 filters) to bandpass the speech signal. The filter spacing is linear in bark scale.
- Estimate the logarithm of short-time average of the Teager-Kaiser energy operator for each band-passed signal. The short-time averaging window duration and window shift are the same as for the standard MFCC front-end.
- Estimate the cepstrum coefficients of the short-time average Teager Energy using the discrete cosine transform (DCT).
- Truncate the TE cepstrum coefficients to keep the first 13 coefficients (including the zeroth coefficient C0), similarly to the standard MFCC front-end.

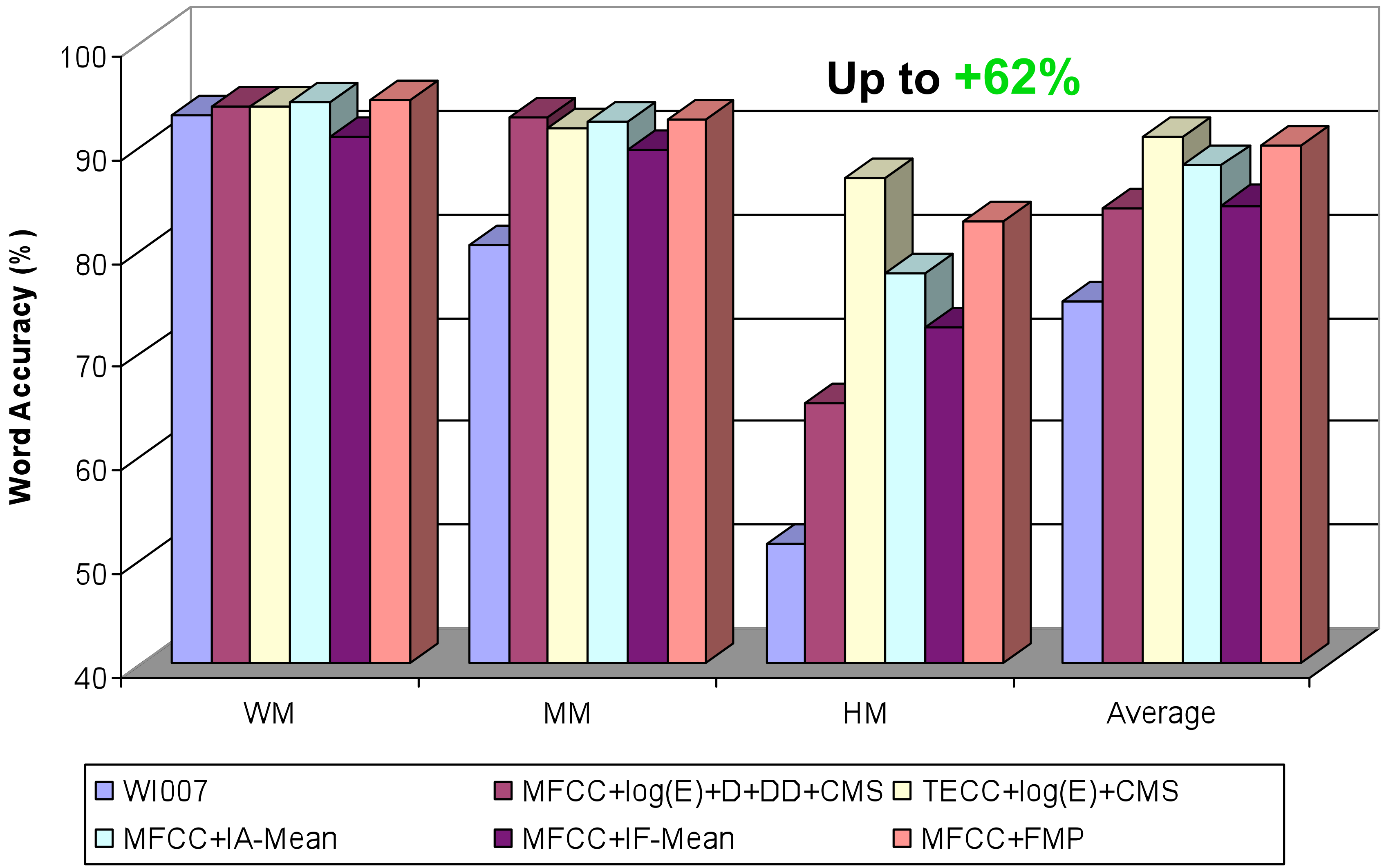
Speech Recognition Experiments

- **TIMIT**
  - Phone recognition task
  - Mono-phone ASR models—no grammar
- **TIMIT + Noise**
  - Phone recognition task
  - Noise added artificially from NOISEX dB
- **AURORA 3 Spanish Task**
  - Connected digit recognition task
  - Whole digit ASR models

Results: TIMIT+Noise



Results: Aurora 3 (HTK)



ASR Results

TIMIT-Based Speech Databases (Correct Phone Accuracies (%))							
Features	TIMIT	NTIMIT	TIMIT+ Babble	TIMIT+ White	TIMIT+ Pink	TIMIT+ Car	Av. Rel. Improv.
MFCC*	58.40	42.42	27.71	17.72	18.60	52.75	-
TEner. CC	58.89	42.40	41.61	34.74	38.40	54.35	24.26
MFCC*+ IA-Mean	59.61	43.53	39.25	26.03	31.05	56.50	17.62
MFCC*+ IF-Mean	59.34	43.70	36.87	25.38	30.92	55.30	15.58
MFCC*+ FMP	59.92	43.69	38.60	26.15	32.84	55.97	18.17

* MFCC+C_0+D+DD, # states=3, # mixtures=16

Experimental Results IIa (HTK)

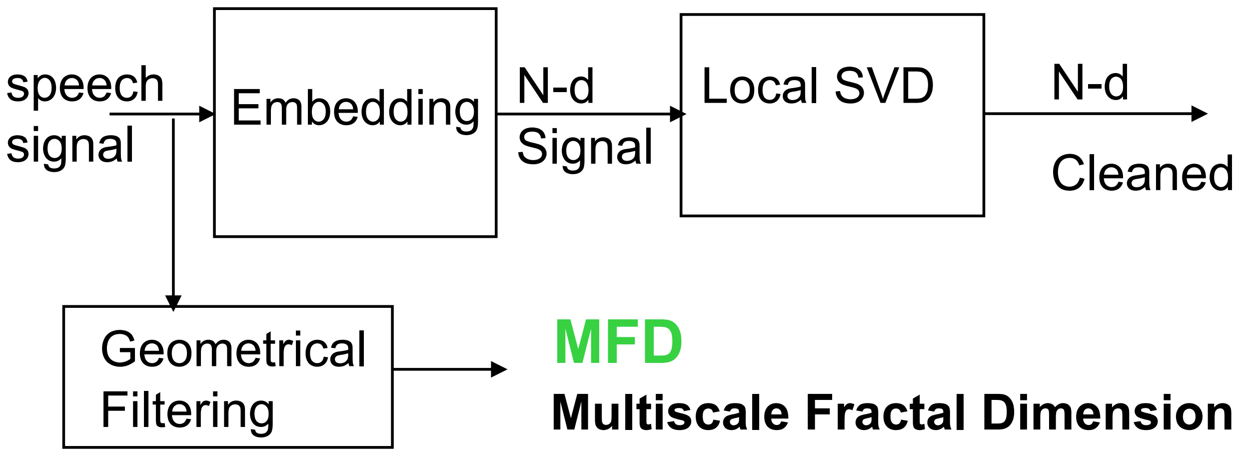
Aurora3 (Spanish Task) (Correct Word Accuracies (%))					
Scenario / Features	WM	MM	HM	Average	Av. Rel. Improv.
Aurora Front-End (WI007)	92.94	80.31	51.55	74.93	-
MFCC*	93.68	92.73	65.18	83.86	+35.62 %
TEnerCC+log (Ener) +CMS	93.64	91.61	86.85	90.70	62.90 %
MFCC*+IA-Mean	94.05	92.22	77.70	87.99	52.09 %
MFCC*+IF-Mean	90.71	89.52	72.36	84.20	36.98 %
MFCC*+FMP	94.41	92.46	82.73	89.87	59.59 %

* MFCC+log(Ener)+D+DD+CMS, # states=14, # mixtures=14

Future Work on Modulation Features

- Refinements w.r.t. AM-FM Features
- Fusion w. Other Features

Fractal Features



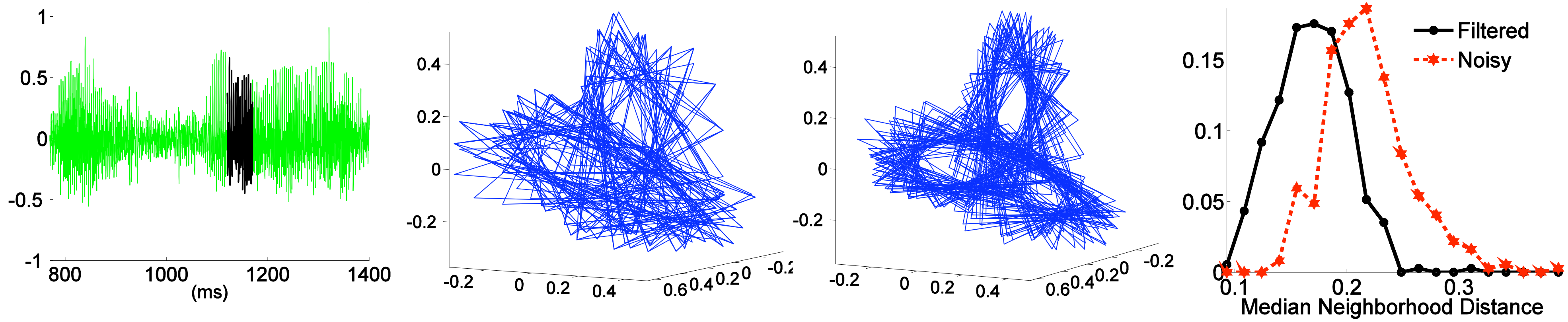
FDCD

Filtered Dynamics - Correlation Dimension (8)

MFD

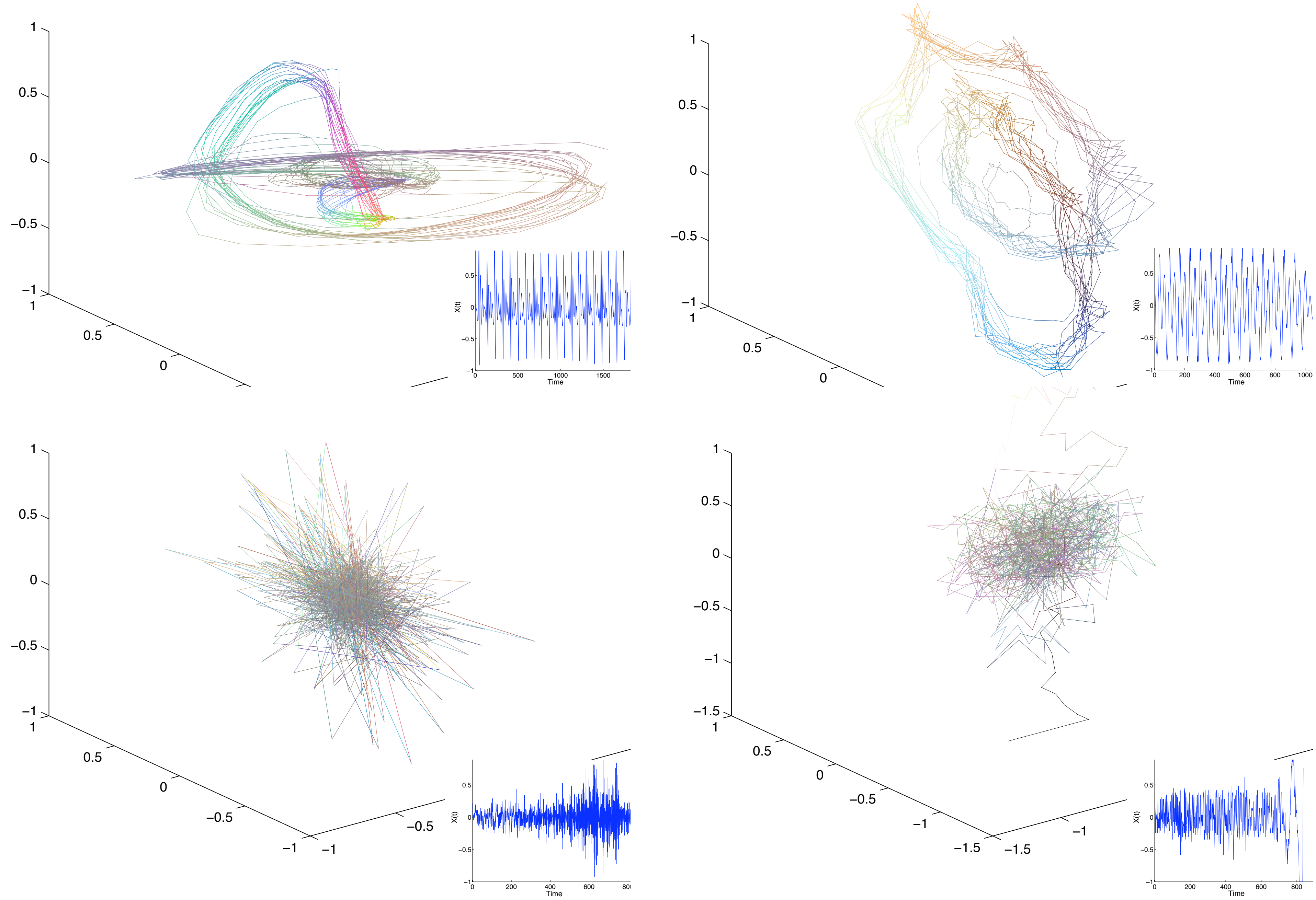
Multiscale Fractal Dimension (6)

Neighborhood Distance Reduction



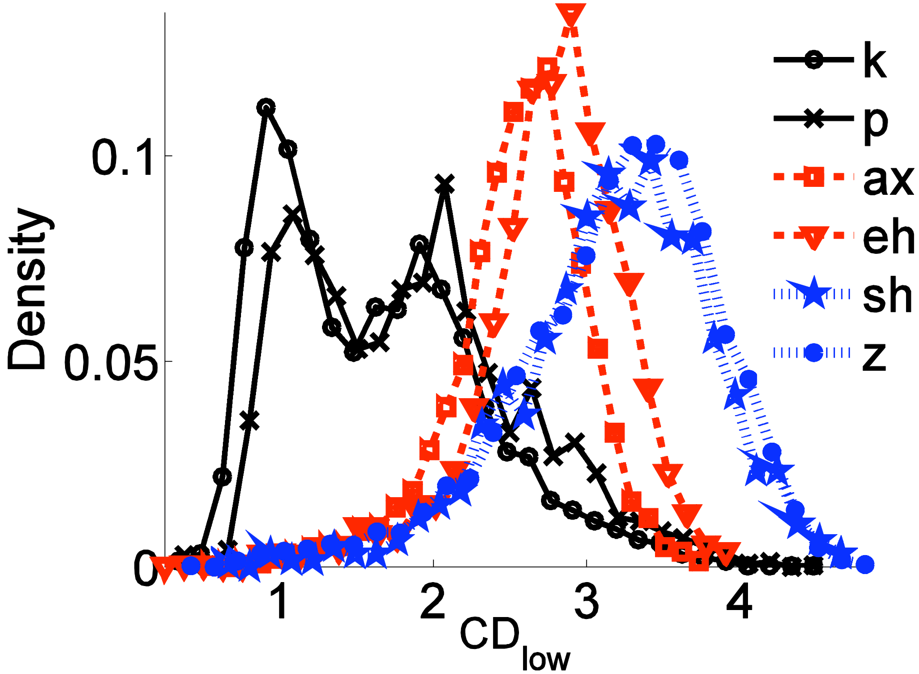
Noisy Embedding **Filtered Embedding**

Reconstructed Attractors



Fractal Features: Correlation Dimension

- Correlation Dimension : $$D_C = \lim_{r \to 0} \lim_{N \to \infty} \frac{\log C(N, r)}{\log r}$$
- Correlation sum $$C(N, r) = \frac{2}{N(N-1)} \sum_{i \neq j}^{N} f\left(r - \| x_i - x_j \|\right)$$

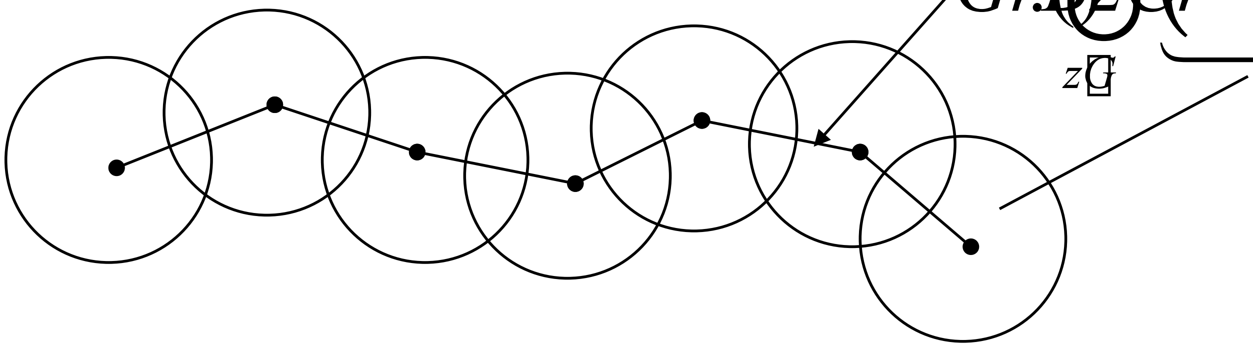


N: # of points, *r*: scale,
x: set points,
f : heavyside function

Histogram of D_C component for selected phoneme classes
(stops, vowels, fricatives)

Morphological Measurement of Fractal Dimension

- **Minkowski cover of curve** $G \oplus rB = \bigcup_{z \in G} \underbrace{(z + rB)}_{B}$



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

$$A_B(r) = \text{area}(G \oplus rB); \qquad \text{length}(r) = \frac{A_B(r)}{2r} \propto r^{1-D}$$

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

$$\left(\log(1/r),\ \log\left(A_B(r)/r^2\right)\right) \rightarrow D$$

Multiscale Speech Fractal Dimension

- **short-time speech signal**

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

- **signal graph**

$$G = \{(t, S(t)) : t \in [0, T]\} \subset \mathbb{R}^2$$

$$\Rightarrow$$

- **fractal**

constant power law

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

- **variable power law**

$$area(G \oplus \varepsilon B) \propto \varepsilon^{2-D(\varepsilon)}$$

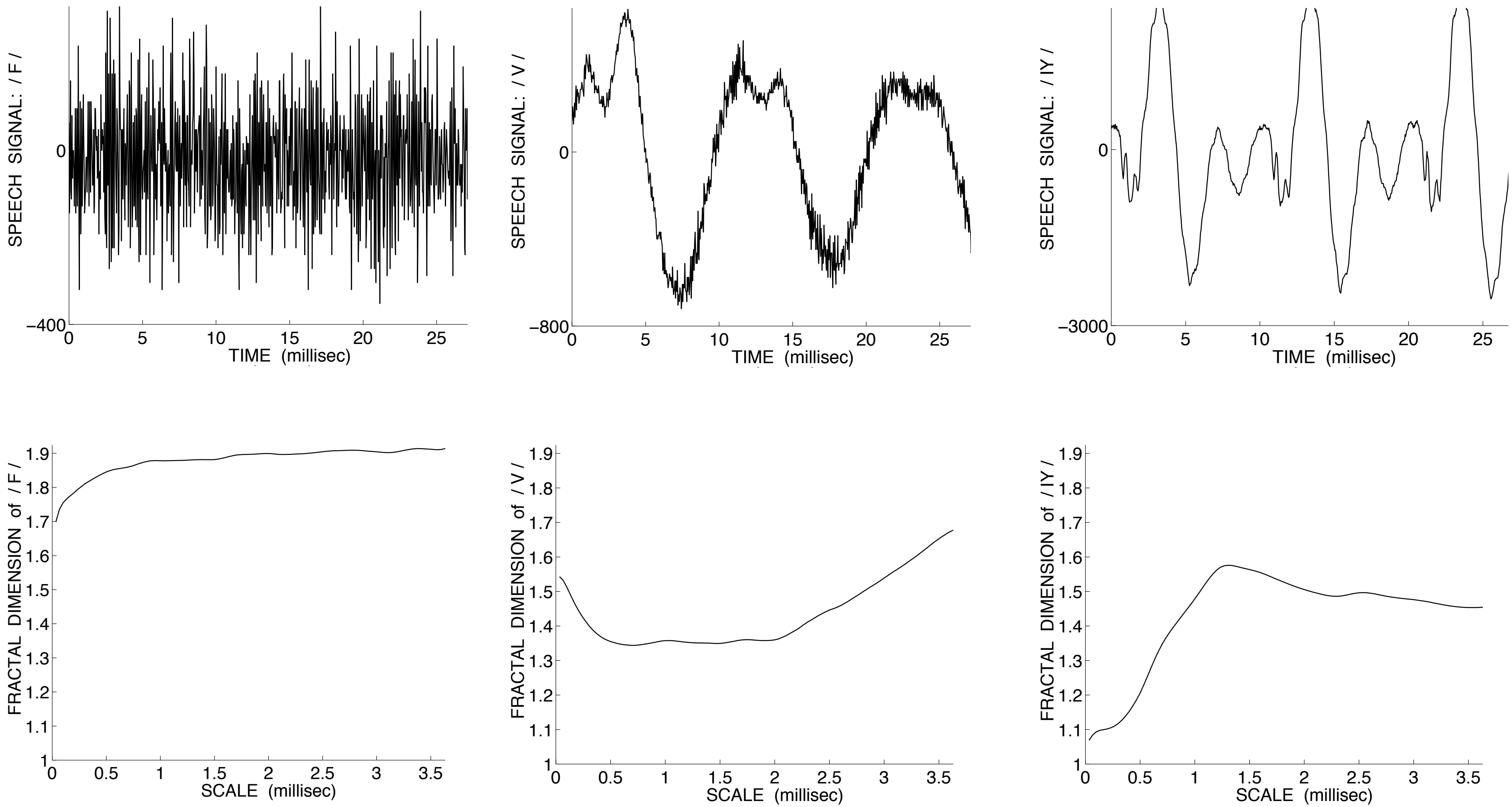
- **multiscale fractal "dimension"**

(speech fractogram):

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

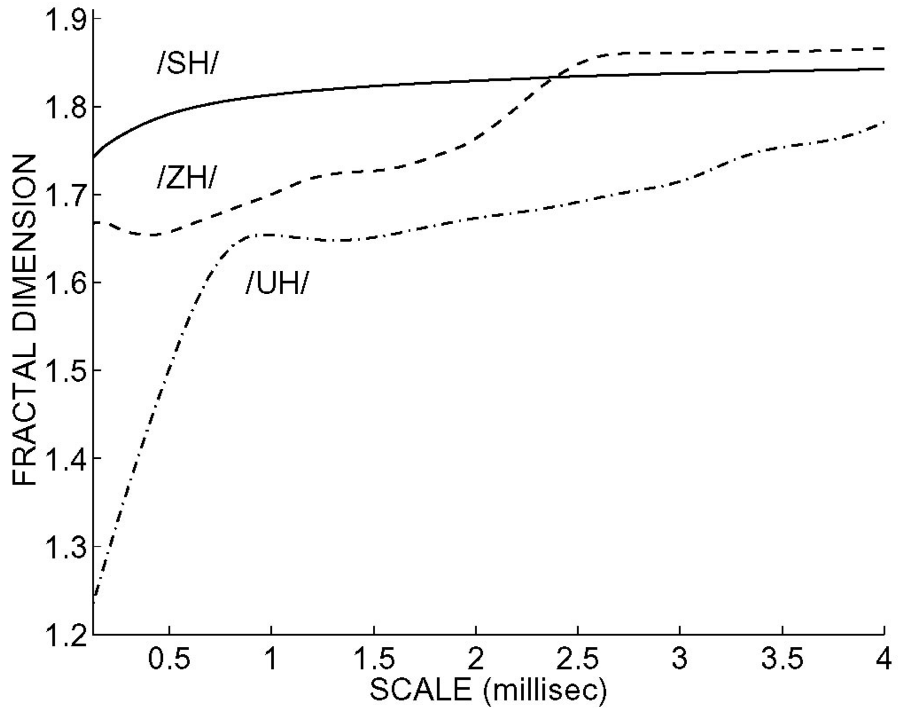
of short-time speech segment around time t

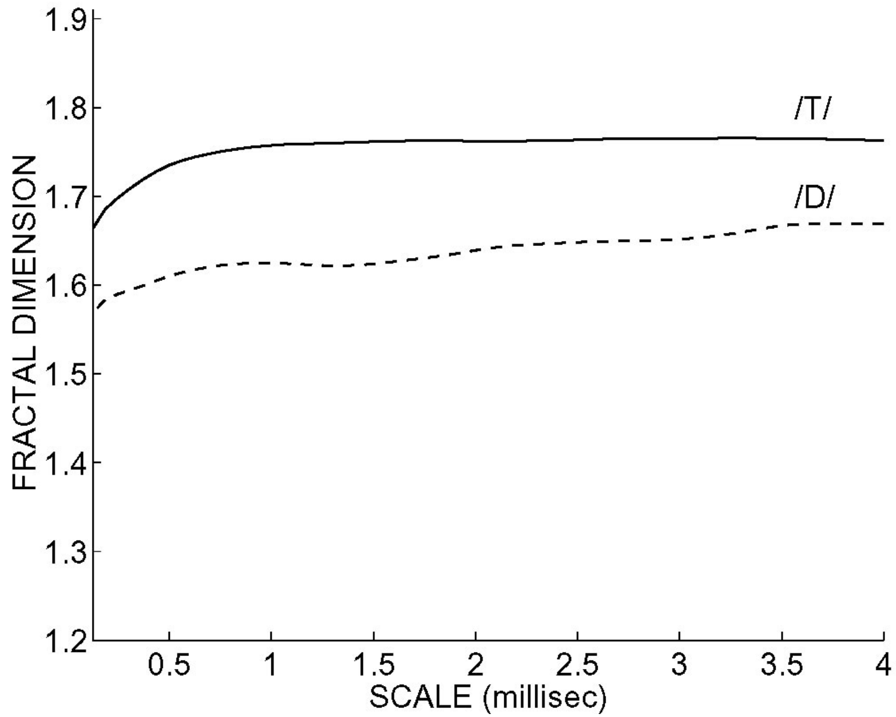
Multiscale Fractal Dimension



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

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)

$\{E, C, E \infty \infty\}$	$\{E, C, \Delta E, \Delta C\} + \{D_1, \Delta D_1\}$	$\{E, C, E \infty \infty\} + \infty\{D D_{116116}\}$
81.2%	83.5%	84.5%

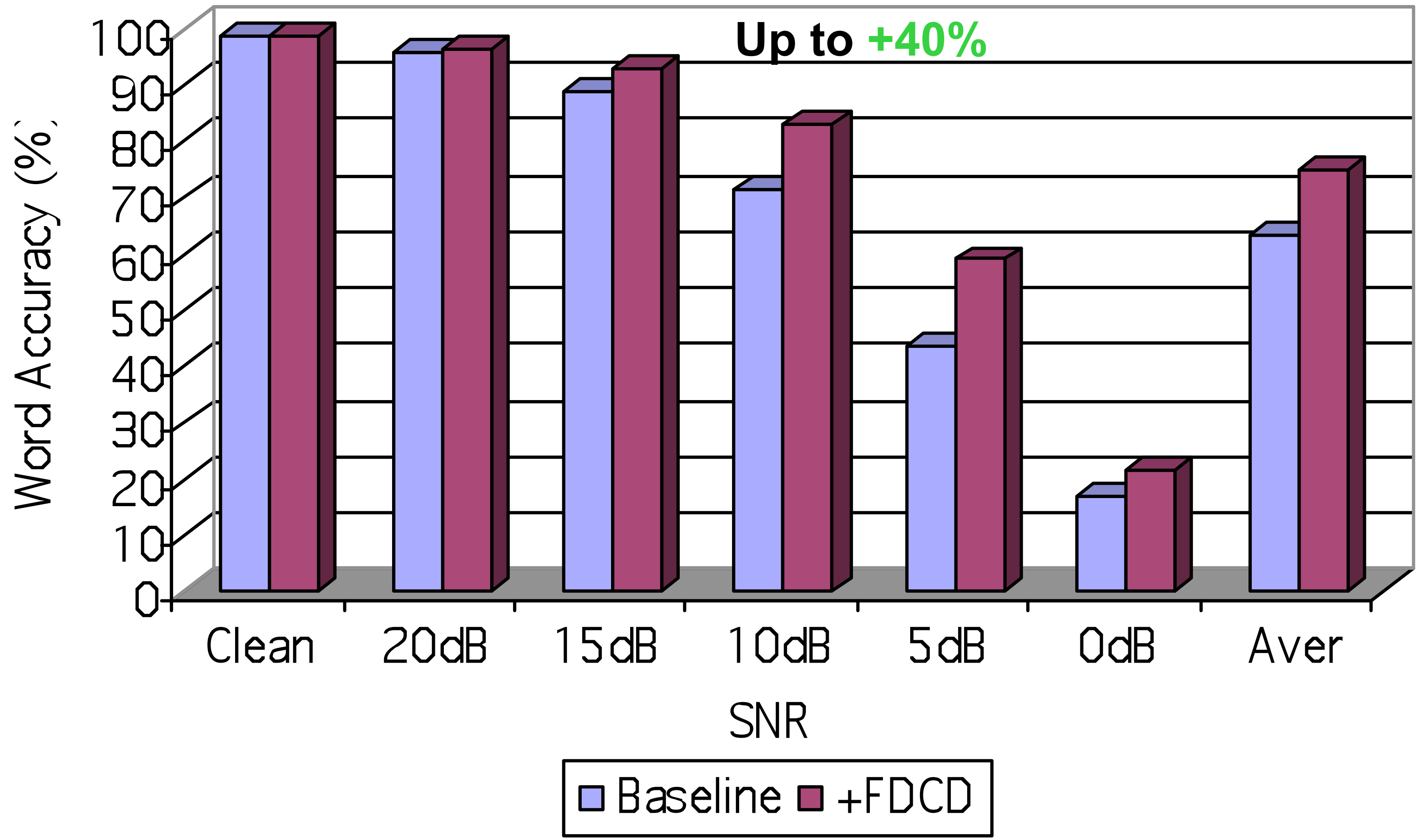
Word Percent Correct for the E-set Recognition Task

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%

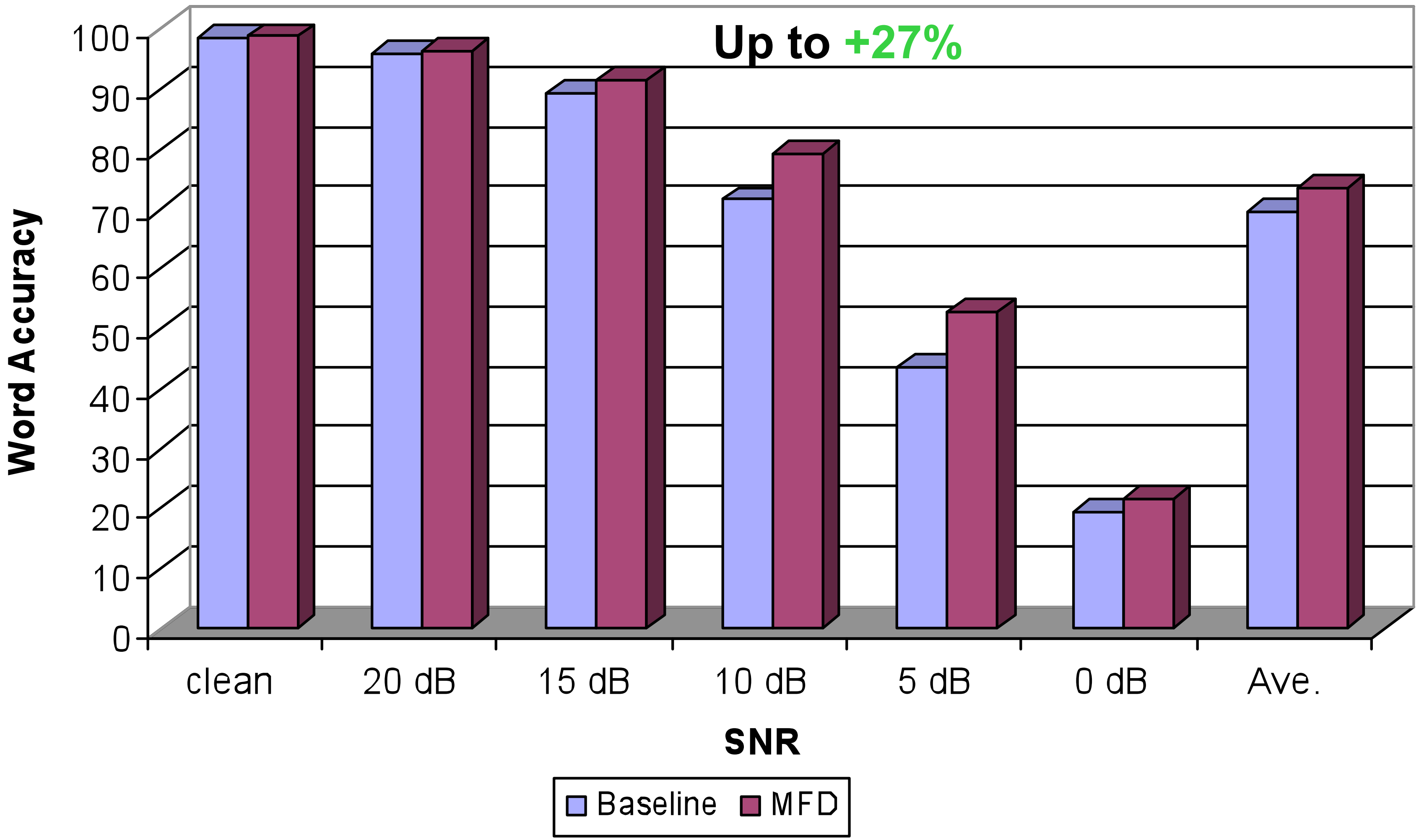
Databases: Aurora 2

- 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)

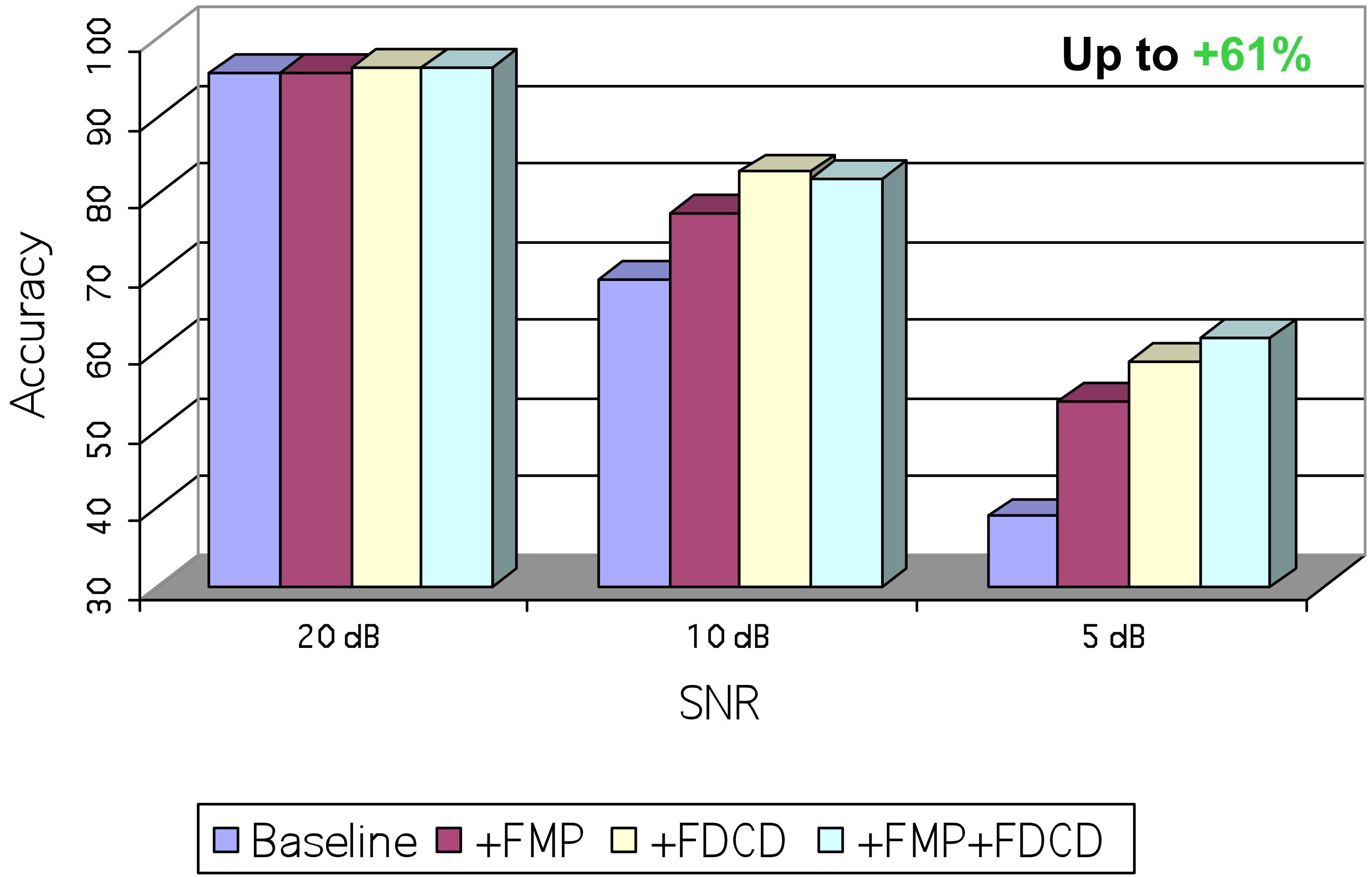
Results: Aurora 2



Results: Aurora 2



Results: Aurora 2



Future Directions on Fractal Features

- Refine Fractal Feature Extraction.
- Application to Aurora 3.
- Fusion with other features.

Visual Front-End

- Aim:
  - Extract low-dimensional visual speech feature vector from video
- Visual front-end modules:
  - Speaker's face detection
  - ROI tracking
  - Facial Model Fitting
  - Visual feature extraction
- Challenges:
  - Very high dimensional signal - which features are proper?
  - Robustness
  - Computational Efficiency

Face Modeling

- A well studied problem in Computer Vision:
  - Active Appearance Models, Morphable Models, Active Blobs
- Both *Shape* & *Appearance* can enhance lipreading
- The shape and appearance of human faces “live” in low dimensional manifolds

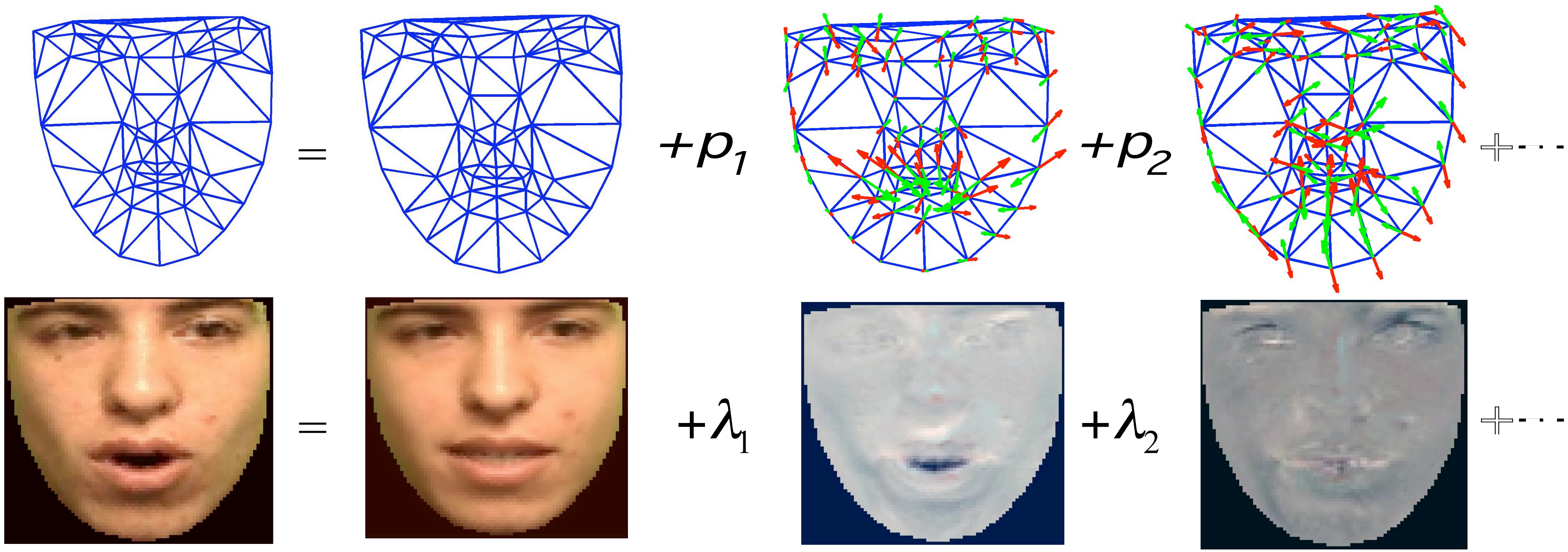
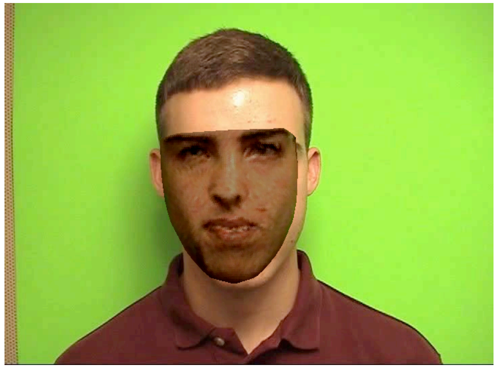
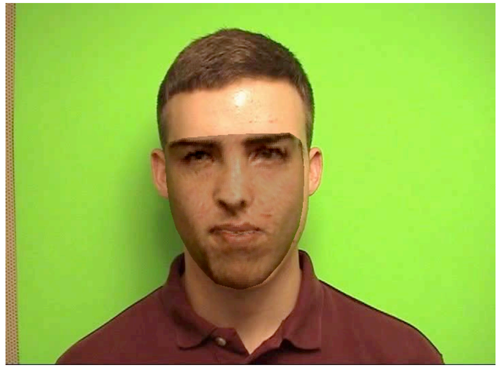


Image Fitting Example

step 2



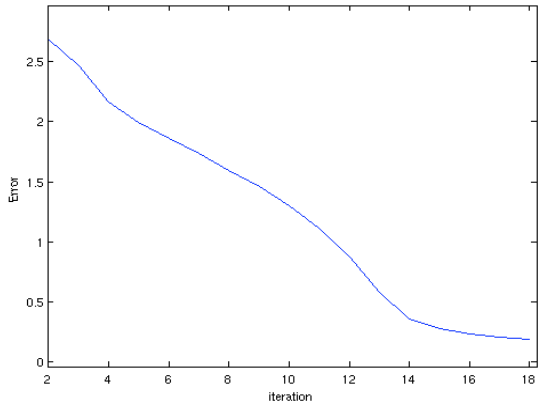
step 6



step 10

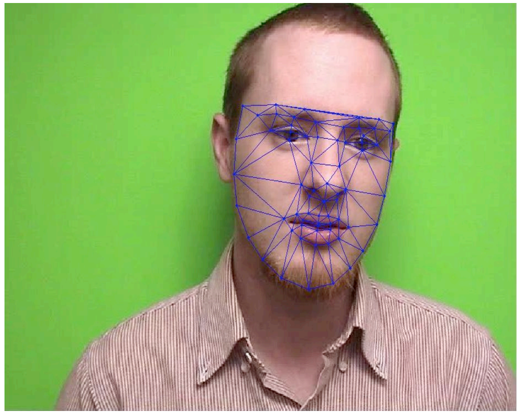
step 14

step 18

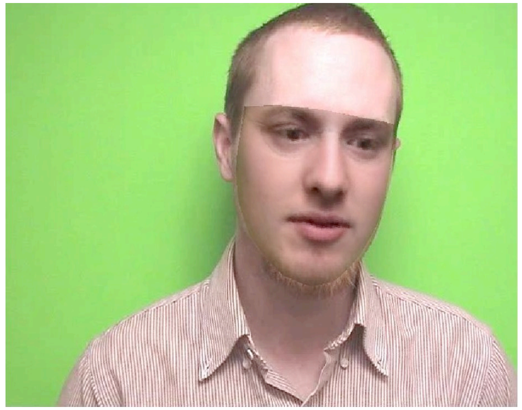


Example: Face Interpretation Using AAM

original video



shape track superimposed on original video



reconstructed face

This is what the visual-only speech recognizer "sees"!

- **Generative models like AAM allow us to evaluate the output of the visual front-end**

Evaluation on the CUAVE Database

Audio-Visual ASR: Database

- Subset of CUAVE database used:
  - 36 speakers (30 training, 6 testing)
  - 5 sequences of 10 connected digits per speaker
  - Training set: 1500 digits (30x5x10)
  - Test set: 300 digits (6x5x10)

- CUAVE database also contains more complex data sets:
  speaker moving around, speaker shows profile, continuous digits, two speakers (to be used in future evaluations)

- CUAVE was kindly provided by the Clemson University

Recognition Results (Word Accuracy)

- Data
  - Training: ~500 digits (29 speakers)
  - Testing: ~100 digits (4 speakers)

	Audio	Visual	Audiovisual
Classification	99%	46%	85%
Recognition	98%	26%	78%

Future Work

- Visual Front-end
  - Better trained AAM
  - Temporal tracking
- Feature fusion
  - Experimentation with alternative DBN architectures
  - Automatic stream weight determination
- Integration with non-linear acoustic features
- Experiments on other audio-visual databases
- Systematic evaluation of visual features

User Robustness, Speaker Adaptation

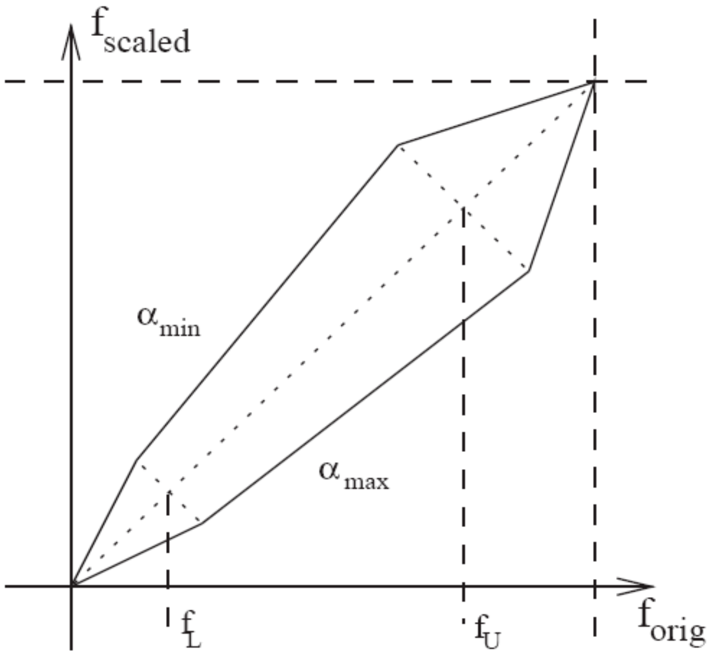
- VTLN **Baseline**
  - Platform: HTK
  - Database: AURORA 4
  - Fs = 8 kHz
  - Scenarios: Training, Testing
  - Comparison with MLLR
- Collection of non-Native Speech Data **Completed**
  - 10 Speakers
  - 100 Utterances/Speaker

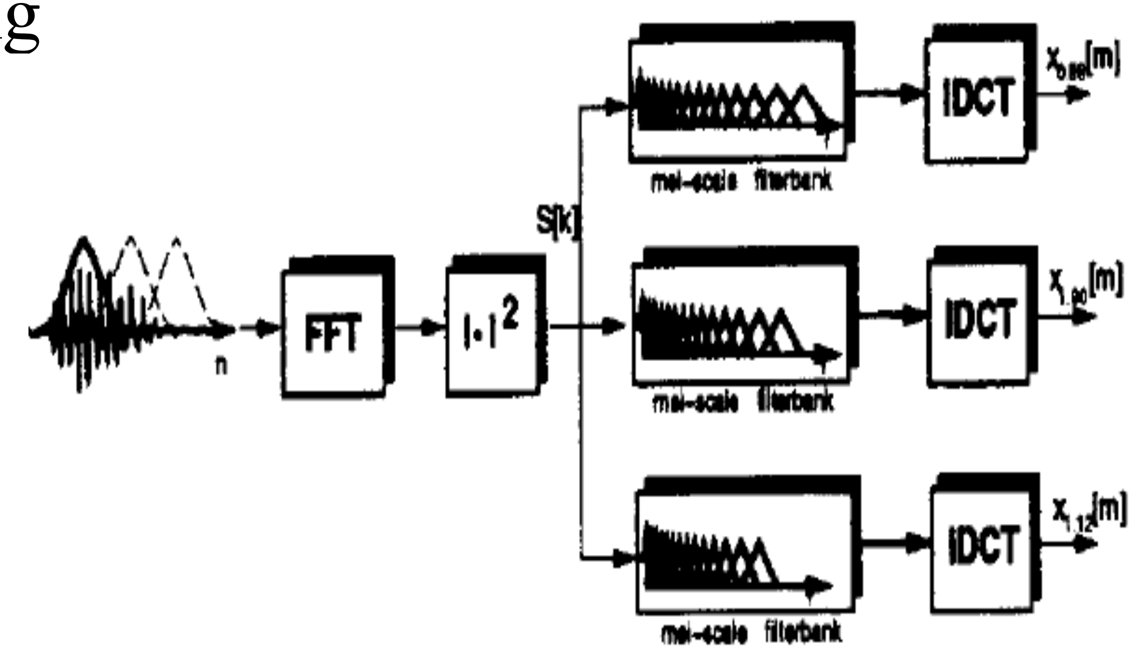
Vocal Tract Length Normalization

Implementation: HTK

- **Warping Factor Estimation**
  - Maximum Likelihood (ML) criterion

- **Frequency Warping**



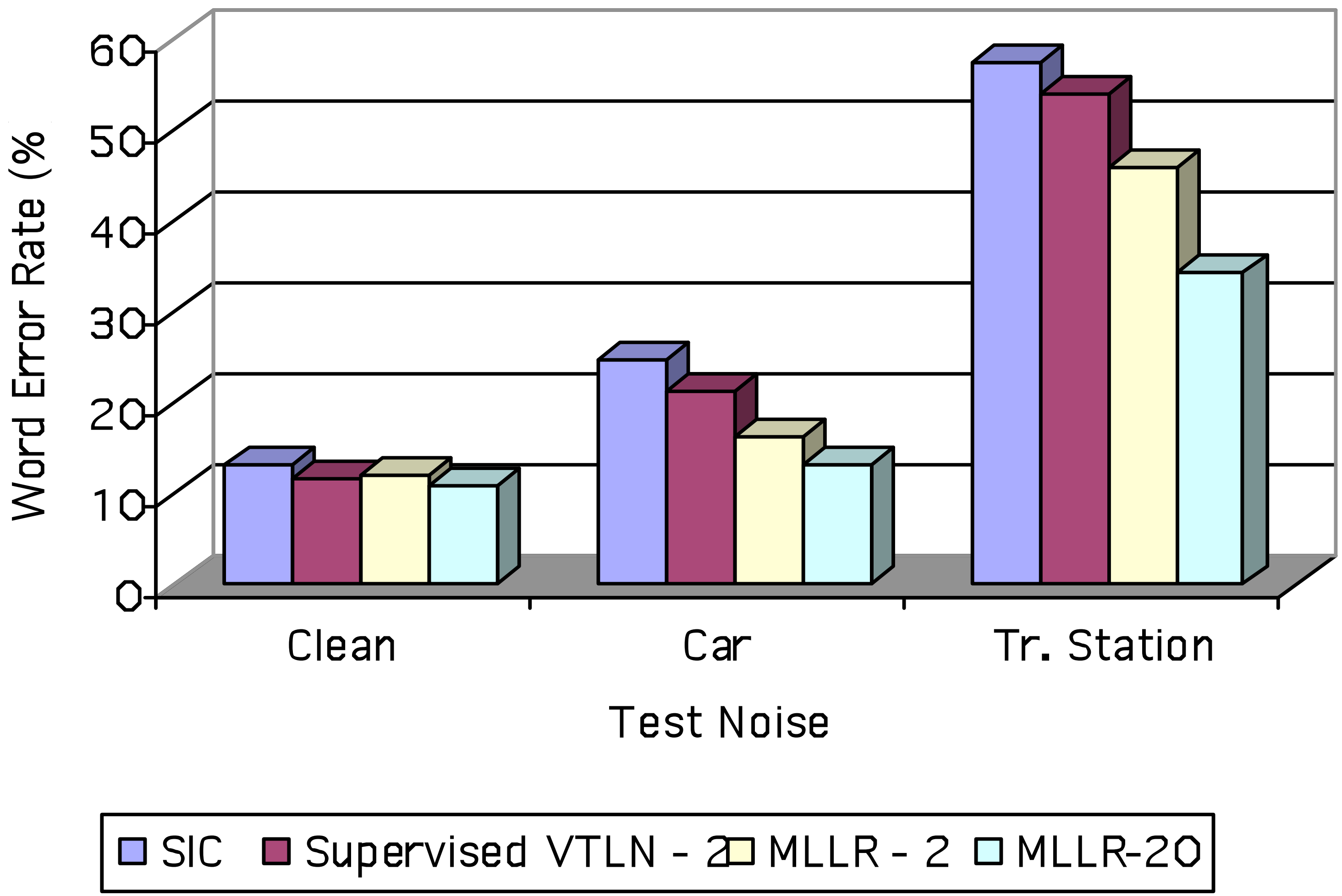


Figures from Hain99, Lee96

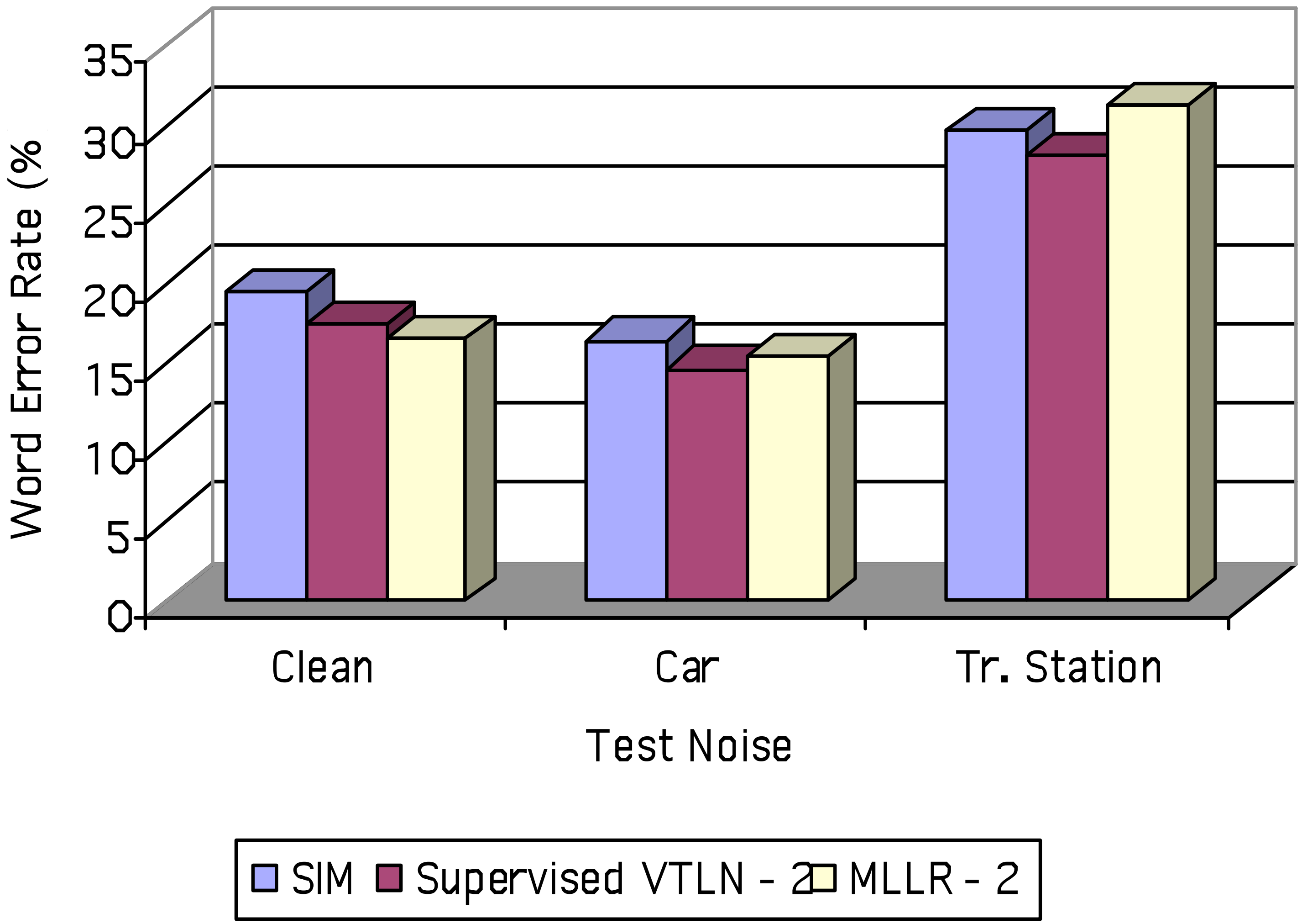
VTLN

- Training
  - AURORA 4 Baseline Setup
  - Clean (**SIC**), Multi-Condition (**SIM**), Noisy (**SIN**)
- Testing
  - Estimate warping factor using adaptation utterances (Supervised VTLN)
    - Per speaker warping factor (1, 2, 10, 20 Utterances)
  - 2-pass Decoding
    - 1st pass
      - Get a hypothetical transcription
    - Alignment and ML to estimate per utterance warping factor
    - 2nd pass
      - Decode properly normalized utterance

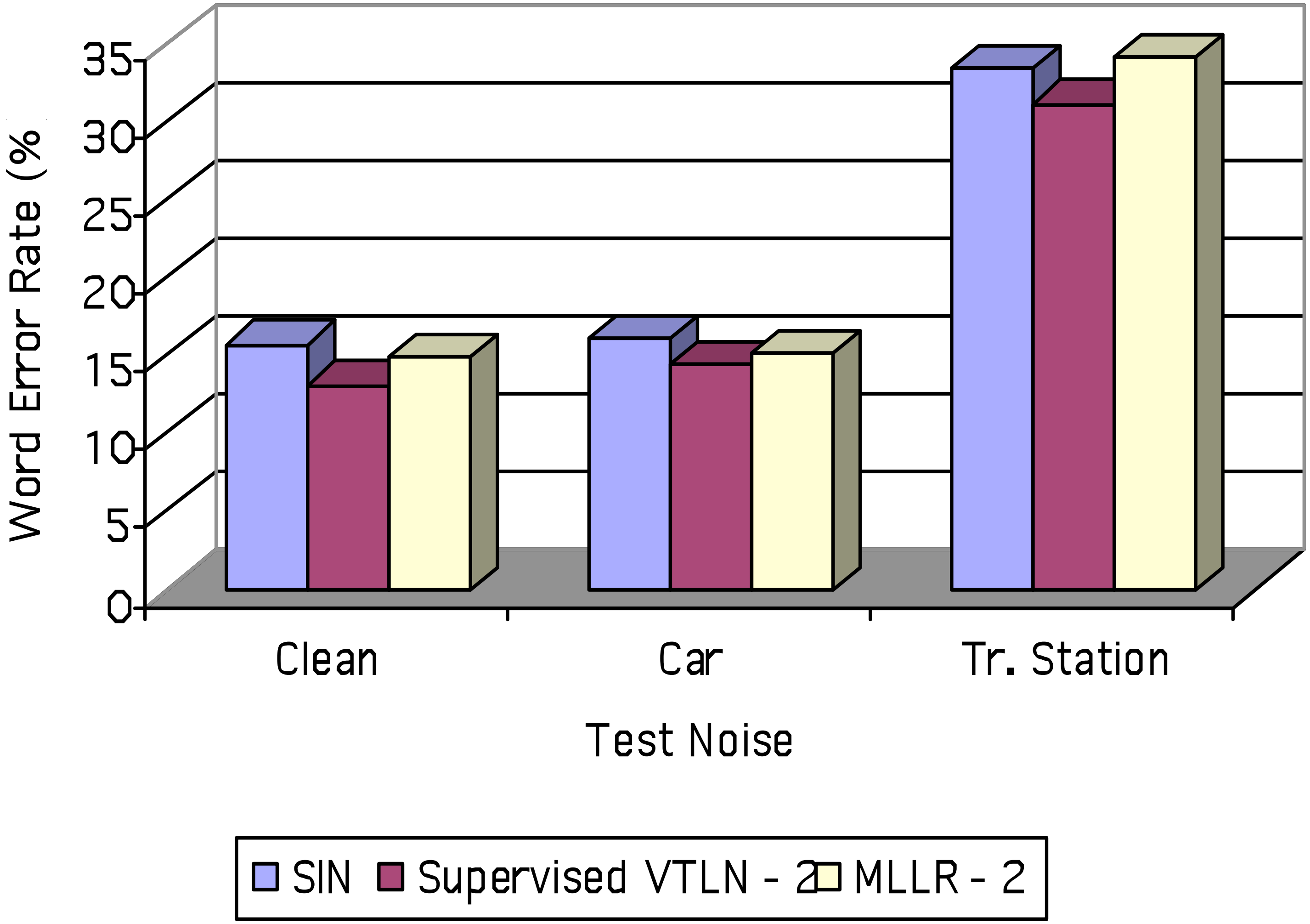
VTLN Results, Clean Training



VTLN Results, Multi-Condition Training



VTLN Results, Noisy Training



Future Directions for Speaker Normalization

- Estimate warping transforms at signal level
  - Exploit instantaneous amplitude or frequency signals to estimate the warping parameters, Normalize the signal

- Effective integration with model-based adaptation techniques (*collaboration with TSI*)

Conclusions

- Results on Modulation (AM-FM and TECC) Features
- Results on Fractal (MFD and FDCD) Features
- Results on AV-ASR
- Results on Adaptation

Our Publications in **http://cvsp.cs.ntua.gr**