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)

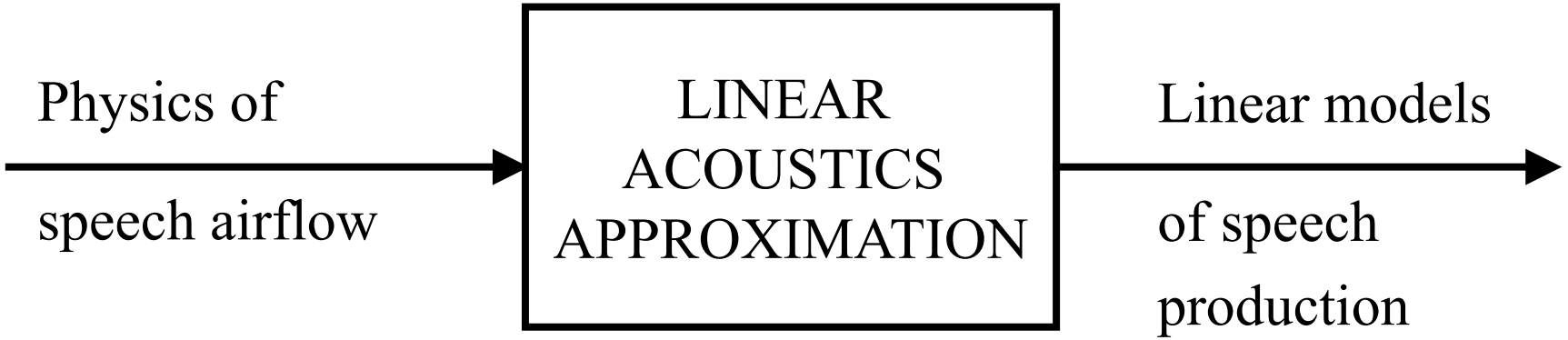
Nonlinear Aspects of Speech Production: Modulations and Energy Operators

Petros Maragos

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

Outline

- Nonlinear Speech Processing → Modulations
- Energy Operators
- AM-FM Speech Model, Demodulation Algorithms
- Applications to Speech Recognition
- Applications to Music Recognition
- Application to Audio Summarization
- Application to Distant Speech Recognition
- Applications of Spatio-Temporal Modulations to Image and Video Processing


Physics of
speech airflow
LINEAR
ACOUSTICS
APPROXIMATION
Linear models
of speech
production

Physics of Speech Airflow

- **airflow variables:** ρ **= air density;** p **= pressure**

  $\vec{u}$ **= 3D air particle velocity**

- **governing equations:**

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

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

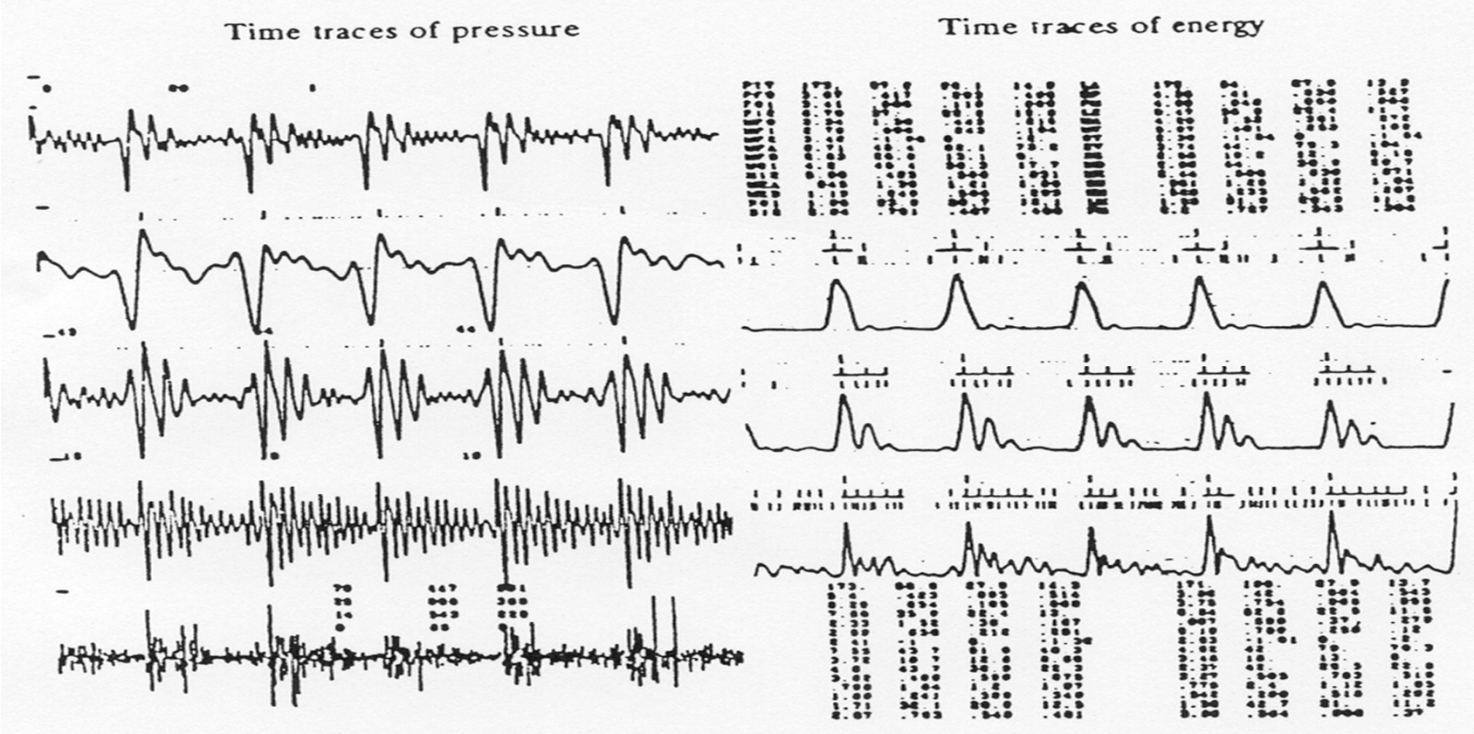
- **time-varying boundary conditions**

Nonlinear Speech Processing

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

Evidence for Speech Modulations

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



Time-varying Oscillators → AM-FM

Simple second-order oscillators with time-varying elements produce modulations:

- If mass or compliance are time-varying → FM
  [Van der Pol, Proc. IRE 1930]

- If damping is time-varying → AM
  [Van der Pol, IEE J. London 1946]

AM-FM Speech Model, Energy Demodulation Algorithms

AM-FM Speech Modulation Model

[Maragos, Kaiser & Quatieri, IEEE T-SP Oct.1993]

- **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: $\omega(t) = 2\pi \cdot f(t) = \dfrac{d}{dt}\varphi(t) = \omega_c(t) + q(t)$

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

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

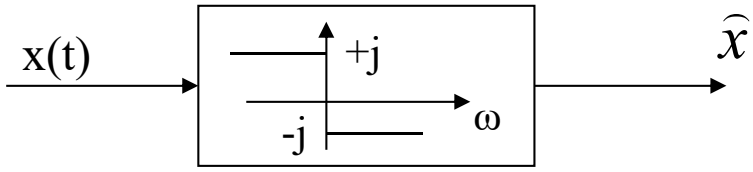
- **Speech Signal as multi-component AM-FM:**

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

AM-FM Demodulation Problem

Given $x(t) = a(t) \cdot \cos(\phi(t))$, estimate $a(t), \dot{\phi}(t)$

- **Variational approach**
- **Hilbert Transform:**



$$\hat{x}(t) = x(t) * \frac{1}{\pi t}$$

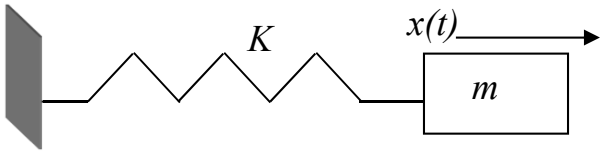
$$\sqrt{x^2 + \hat{x}^2} \approx a$$

$$\frac{d}{dt} \arctan\left(\frac{\hat{x}}{x}\right) \approx \dot{\phi}$$

- **Energy Operators**

Energy Tracking in Oscillators

- **harmonic oscillator**



- **motion equation**

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

- **response**

$$x(t) = A\cos(\omega t + \theta),$$
$$\omega^2 = k/m$$

- **energy**

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

- **energy tracking**

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

1D Energy Operators

(Teager, Kaiser ICASSP 1990)

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

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

property:

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

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

$$\Psi_d[x(n)] \equiv x^2(n) - x(n+1)x(n-1)$$

-Discretize Derivatives [Maragos, Kaiser & Quatieri, T-SP Apr.1993]

-Special case of Quadratic opers [Atlas & Fang, T-SP 1995]

property:

$$\Psi_d\left[Ar^n\cos(\Omega_c n + \theta)\right] = A^2 r^{2n}\sin^2(\Omega_c)$$

Energy Separation Algorithm (ESA)

(Maragos, Kaiser & Quatieri, IEEE T-SP Oct. 1993)

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

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

$$\frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \cong |a(t)| \qquad \sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \cong \omega(t)$$

Discrete ESA (DESA-2)

- **AM-FM Signal:** $x[n] = a[n]\cos(\int_0^n \Omega(m)dm)$
- **Energy Tracking:**

$$\Psi\left[x[n]\right] \cong a^2[n]\sin^2\left(\Omega[n]\right)$$

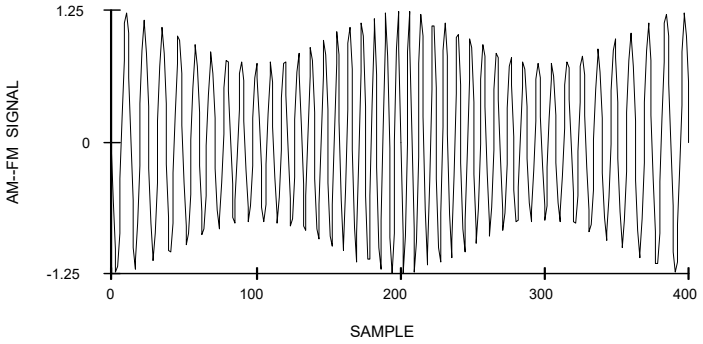
$$\Psi\left[x[n+1] - x[n-1]\right] \cong 4a^4[n]\sin^4\left(\Omega[n]\right)$$

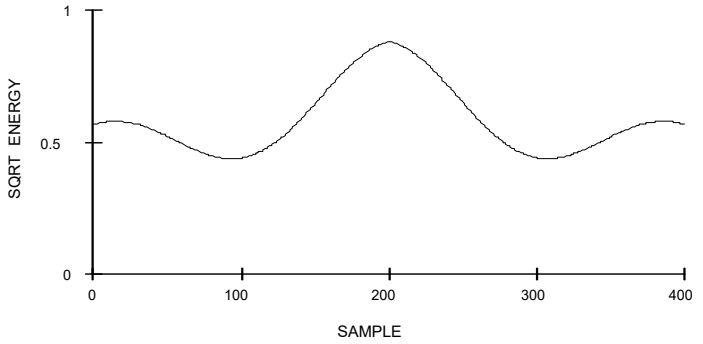
- **DESA-2:**

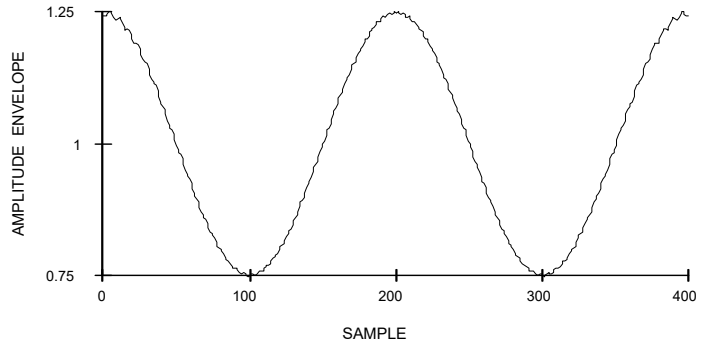
$$\frac{2\Psi\left[x[n]\right]}{\sqrt{\Psi\left[x[n+1] - x[n-1]\right]}} \cong \left|a[n]\right|$$

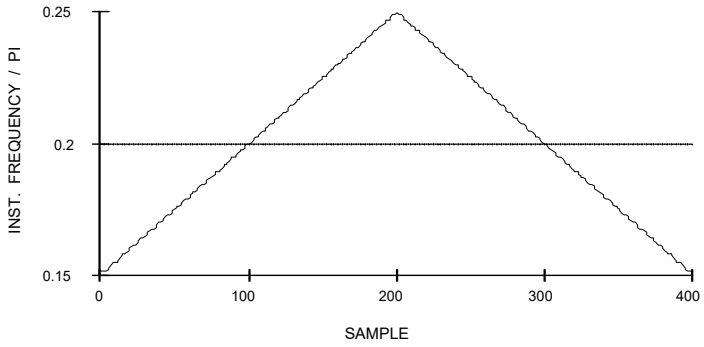
$$\arcsin\sqrt{\frac{\Psi\left[x[n+1] - x[n-1]\right]}{4\Psi\left[x[n]\right]}} \cong \Omega[n]$$

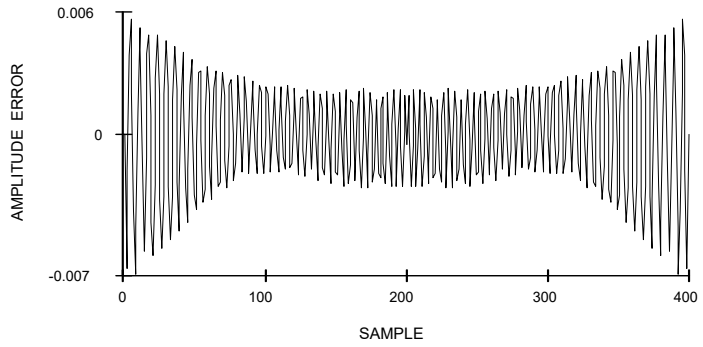
ESA Applied to Synthetic AM-FM

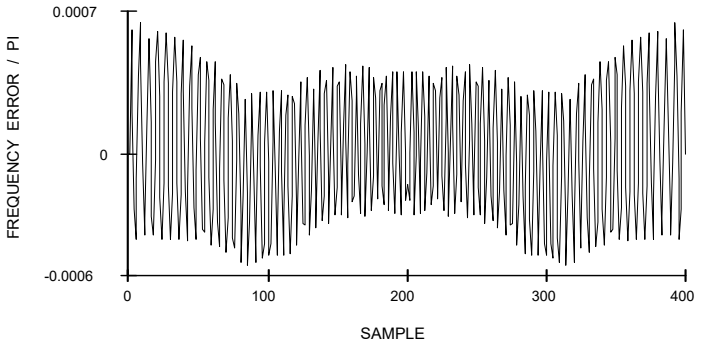




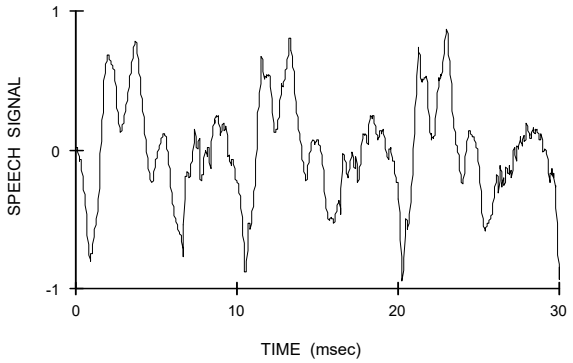


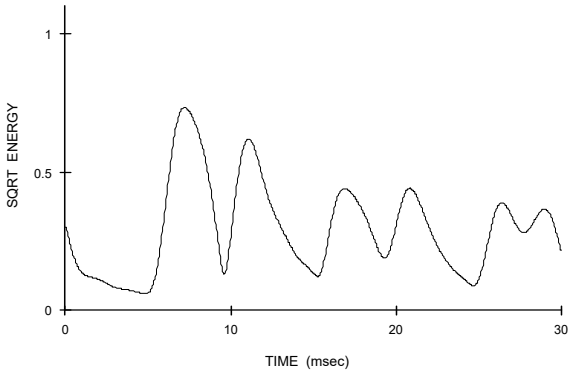


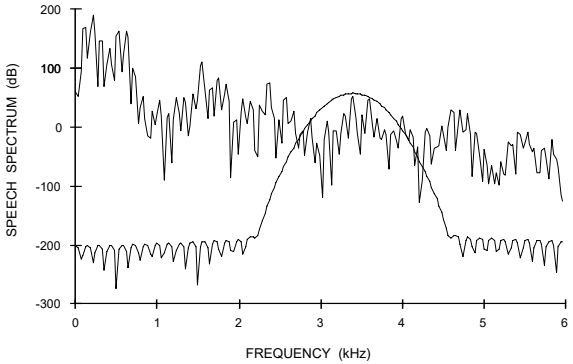


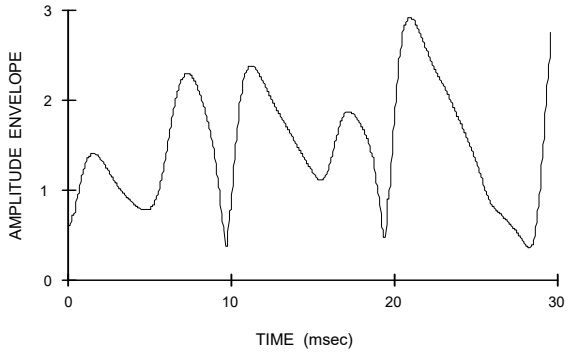


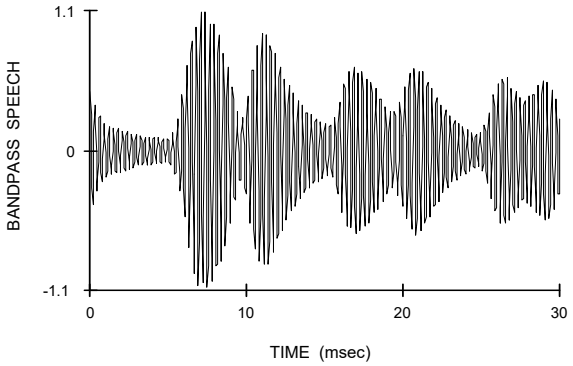
ESA Applied to Speech Resonance

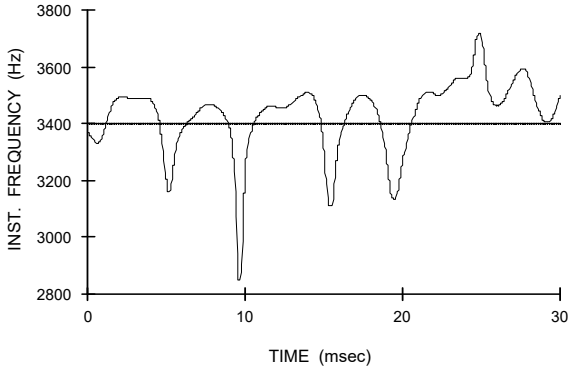












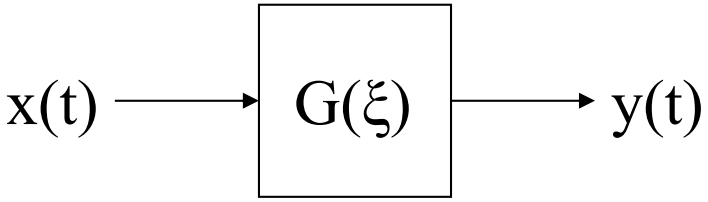
ESA in Noise and BP Filtering

(Bovik, Maragos & Quatieri, IEEE T-SP Dec. 1993)

- **AM-FM signal:**

$$x(t) = \underbrace{a(t)\cos\left(\int_0^t \omega(\tau)d\tau\right)}_{signal} + n(t)$$

- **Noise: wss Gaussian zero-mean, p.spectrum $N(\xi)$**
- **Bandpass Filter:**



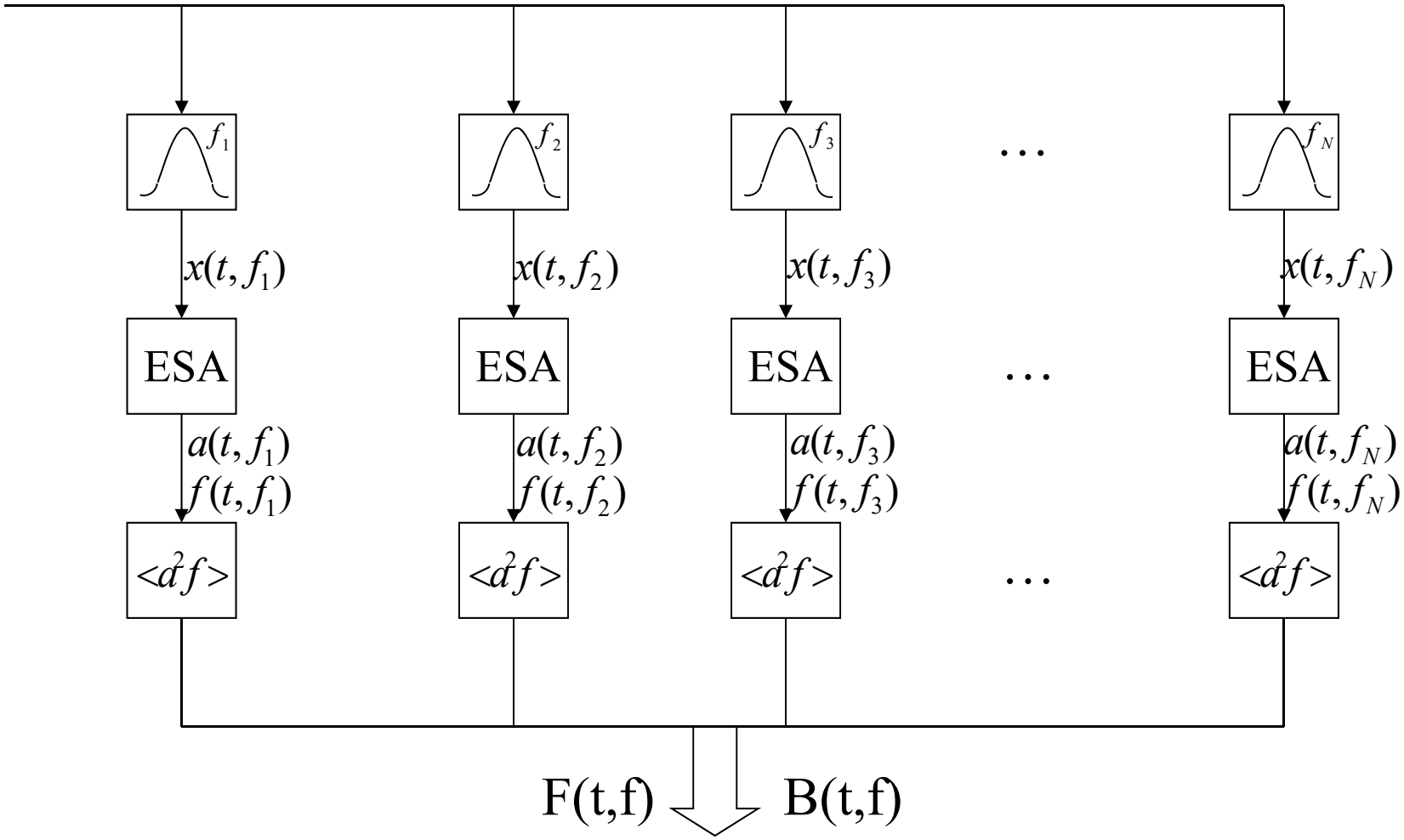
$$SNR(t) = \frac{a^2(t)}{\int_{passband} N(\xi)d\xi}$$

- **ESA Ampl./Freq. Estimates:** $\hat{a}(t), \hat{\omega}(t)$

$$E[\hat{\omega}^2(t)] \approx \omega^2(t)\left(1 + \frac{4SNR(t)}{[SNR(t)+2]^2}\right)$$

$$E[\hat{a}^2(t)] \approx a^2(t)\left(1 + \frac{10SNR(t)+4}{SNR(t)[SNR(t)+2]}\right)\left|G(\omega(t))\right|^2$$

Multiband Demodulation and F/B Tracking



[A. Potamianos & P. Maragos, JASA 1996]

Frequency and Bandwidth Estimates

- **Center Frequency Estimates:**

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

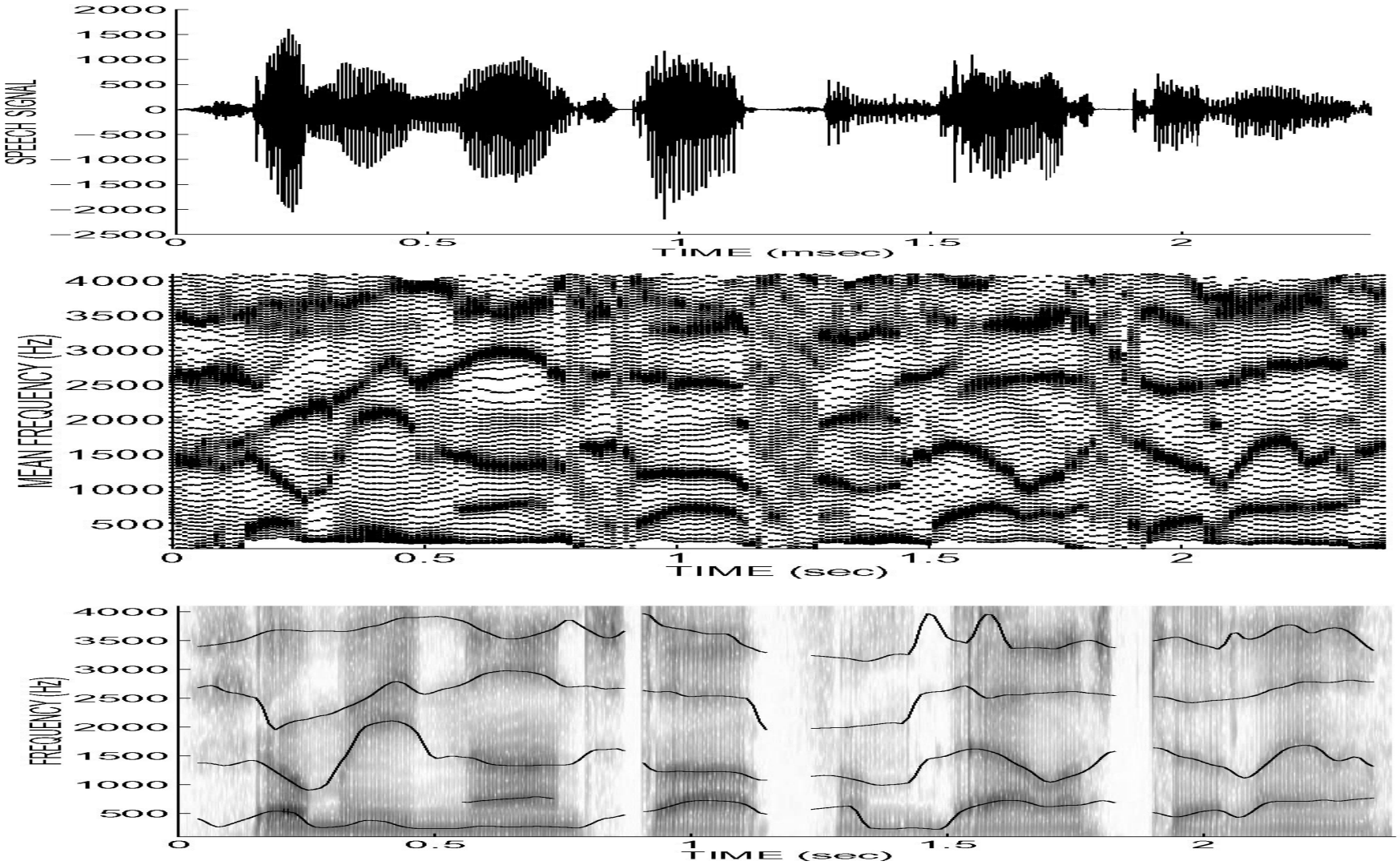
$$F_w = \frac{\int_o^T f(t)a^2(t)dt}{\int_o^T a^2(t)dt}$$

- **Bandwidth Estimates:**

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

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

Speech Pyknogram



[A. Potamianos & P. Maragos, JASA 1996]

Smooth Energy Operators and tracking

- Teager-Kaiser Energy Operator (TKEO): $\Psi[s(t)] \triangleq [s'(t)]^2 - s(t)s''(t)$
- AM-FM signals $s(t) = \alpha(t)\cos(\phi(t))$: $\Psi[s(t)] \approx \underbrace{[\alpha(t)\phi'(t)]^2}_{\text{Energy of the oscillatory source}}$
- **Regularized or Gabor TKEO**:

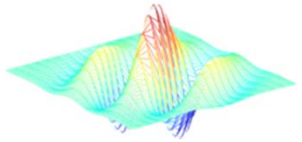
$$\Psi_h\left[s(t) * h(t)\right] = \left[s(t) * h'(t)\right]^2 - \left[s(t) * h(t)\right]\left[s(t) * h''(t)\right]$$

where $h(t)$ the Gabor filter's impulse response

  - Wideband signals (sum of non-stationary sinusoids)
  - Simultaneous narrowband component separation, energy tracking and denoising

- **2D Gabor TKEO**:

$$\Phi_g\left(I * g\right) = \left\|I * \nabla g\right\|^2 - \left(I * g\right)\left(I * \nabla^2 g\right)$$



Refs: Dimitriadis & Maragos, Speech Com 2006. Kokkinos, Evangelopoulos & Maragos, T-PAMI 2009

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.

[Dimakis & Maragos, IEEE T-SP 2005]

Other Works in AM-FM and/or Energy Operators

- Higher-Order EO [PM & A.Potamianos, IEEE SPL 1995], Iterative ESA [H. Hanson, PM & A.Potamianos, T-SAP 1994], Speech Emotion Classific. [Chaspari et al, EUSIPCO 2014]
- Energy Demodulation of Multi-component AM-FM [B. Santhanam & PM, IEEE SPL 1996; T-COM 2000], ED for Large Freq. Deviations & Wideband Sig [Santhanam SPL 2004]
- Kumaresan & Rao [JASA 1999]: Envelope and Positive IF estimation (pole-zero modeling of analytic signal), speech applications.
- P. Doerschuk, S.Lu, W.C.Pai, [T-SP 1996, T-SP 2000]: AM-FM model, Kalman filtering
- T.Quatieri et al: [T-SAP'97] AM-FM Auditory Separation, FM-AM Transduction [T-SP'99]
- J. Hansen et al: Vocal Fold Pathology [T-BE 1998], Nonlinear Features, Speech Classification Under Stress [T-SAP 2001], [Springer 2007 LNAI 4343]
- H. Patil et al: TEO-MFCC, Voice Biometrics [ICONIP'04, PReMI'07, ICASSP10], Spoofed Speech Detection [Interspeech 2018]
- A. Boudraa et al [JOSA'07, JASA'08]: Cross TEO, Generalized HOEO
- Y. Stylianou et al [T-ASLP 2011]: AM-FM decomposition, Sinusoidal model
- L. Atlas et al: Quad. Energy Oper., Modulation Spectrum (JASP 2003)
- N. Huang et al [Proc.R.Soc.Lond.A 1998]: EMD - Hilbert Spectrum
- Monogenic Signal (2D general Anal.Sig., Riesz Transf.) [Felsberg & Sommer, T-SP'01]

Applications of AM-FM Modulations and Energy Operators in Speech Recognition

Properties related to Time Duration of Energy Averaging Window

D. Dimitriadis, A. Potamianos, and P. Maragos, "*A Comparison of the Squared Energy and Teager-Kaiser Operators for Short-Time Energy Estimation in Noise* ," IEEE Transactions on Signal Processing, July 2009.

Signal and Noise Models

- Clean AM-FM Signal: $x(t) = a(t) \cdot \cos\left(\phi_x(t)\right)$

- Noise: Sinusoidal Approximation: $n(t) = \sum_{i=1}^{K} b_i \cdot \cos\left(\omega_i t + \vartheta_i\right)$

  (Refs: Deng, Droppo & Acero, IEEE T-SAP 2004. Seltzer, Droppo & Acero, Eurospeech 2003)

- Teager-Kaiser Energy of Noise:

$$\Psi[n(t)] = \sum_{i=1}^{K} \left(b_i \cdot \omega_i\right)^2 + \frac{1}{2}\sum_{i=1}^{K}\sum_{j\neq i} b_i b_j \omega_i \left(\omega_i + \omega_j\right)\cos\left(\phi_i(t) - \phi_j(t)\right)$$

$$+ \frac{1}{2}\sum_{i=1}^{K}\sum_{j\neq i} b_i b_j \omega_i \left(\omega_i - \omega_j\right)\cos\left(\phi_i(t) + \phi_j(t)\right)$$

Noisy Signal Energy Estimation

- Teager-Kaiser Energy:

$$\Psi\left[x(t)+n(t)\right]=\Psi\left[x(t)\right]+\Psi\left[n(t)\right]+\underbrace{2\dot{x}(t)\dot{n}(t)-\ddot{x}(t)n(t)-x(t)\ddot{n}(t)}_{Cross-Terms}$$

- Squared-Amplitude Energy:

$$S\left[x(t)+n(t)\right]=x^2(t)+n^2(t)+\underbrace{2x(t)n(t)}_{Cross-Terms}$$

Normalized Energy Deviations in Steady-state

- Teager-Kaiser Energy Normalized Deviation (**p=1**)

- Squared-Amplitude Energy Normalized Deviation (**p=0**)

$$D = \frac{\sum_{i=1}^{K} b_i^2 \cdot \omega_i^{2p}}{\left\langle a^2(t) \cdot \omega_x^{2p}(t) \right\rangle_T}$$

T: is the length of the averaging time window
~ steady state (long-term) when ***T*** > 50-100 msec

Energy Deviations terms

Deviation = Steady state + Lowpass Transient + Highpass Transient

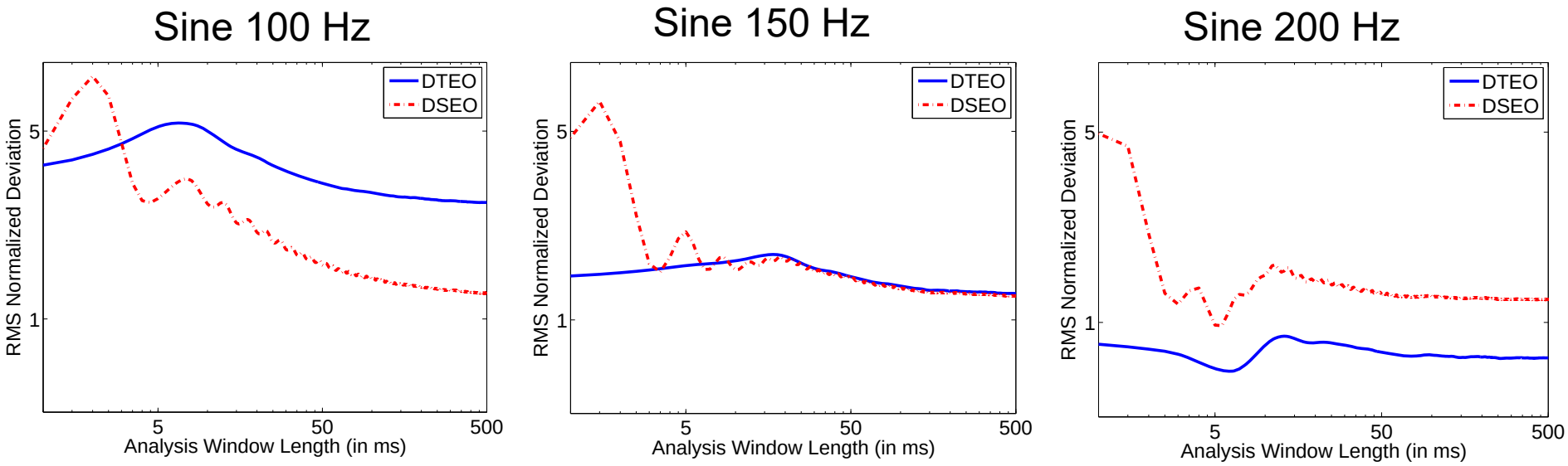
(Long-term) (Medium-term) (Short-term)

$$\mathcal{D}_T = \frac{\sum_i {b_i}^2 {\omega_i}^2}{\langle a^2 \omega_x^2 \rangle} + \mathcal{D}_T{}^- + \mathcal{D}_T{}^+$$

$$\mathcal{D}_S = \frac{\sum_i {b_i}^2}{\langle a^2 \rangle} + \mathcal{D}_S{}^- + \mathcal{D}_S{}^+.$$

Experiments with Sinusoids

- Signal is a sinusoid at 100, 150, 200 Hz (constant amplitude, random phase)
- Noise is white Gaussian band-passed in [100--200] Hz
- Log RMS normalized energy deviation shown for **SEO** (red) and **TEO** (blue)
- x-axis is duration of averaging window (short-, medium-, long-term)



Main Results

- TEO always better than SEO for Short averaging windows. This is even more important for the low-freq filters.

- For long- and mid-term averaging, TEO better than SEO when spectral content of noise is in lower frequencies than the signal's

Applying Energy Operators to Signal Derivatives

- ℓ^{th} - Order Signal Derivatives

$$x^{(\ell)}(t) \cong a(t) \cdot \omega_x^{\ell}(t) \cdot \cos\left(\phi_x(t) + \ell\frac{\pi}{2}\right)$$

- Teager-Kaiser Energy Deviation

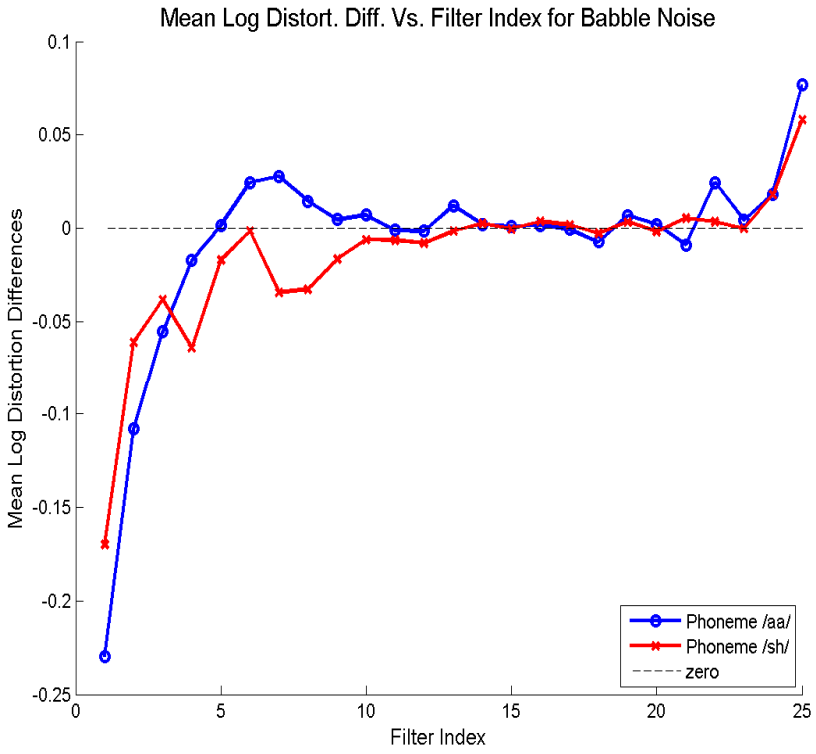
$$D_T \cong \frac{\sum_{i=1}^{K} b_i^2 \cdot \omega_i^{2(\ell+1)}}{\left\langle a^2(t) \cdot \omega_x^{2(\ell+1)}(t) \right\rangle_T}$$

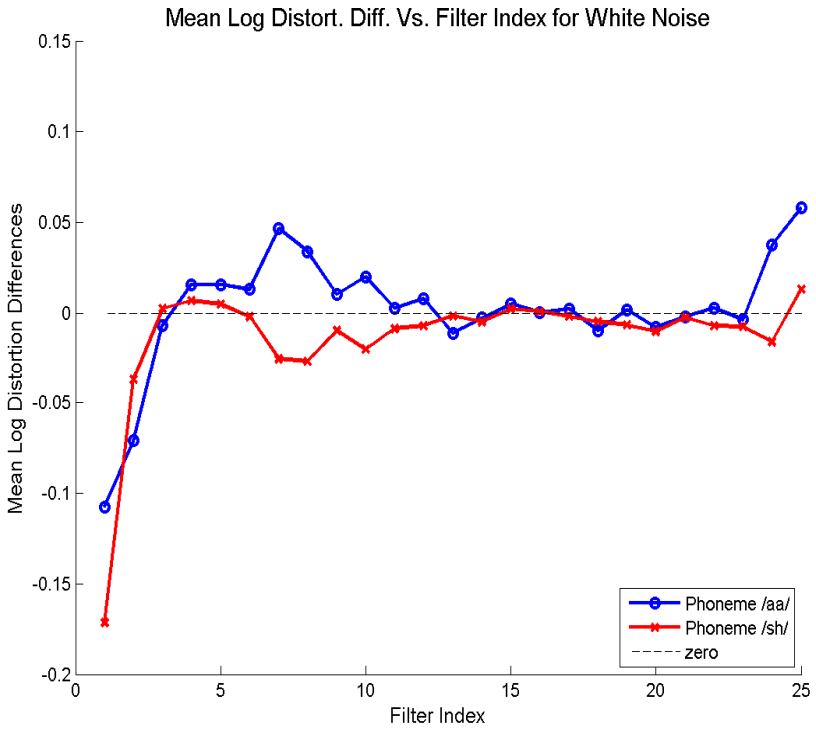
- Squared-Amplitude Energy Deviation

$$D_S \cong \frac{\sum_{i=1}^{K} b_i^2 \cdot \omega_i^{2\ell}}{\left\langle a^2(t) \cdot \omega_x^{2\ell}(t) \right\rangle_T}$$

Results on Noisy Speech Signals (Short/Med-term, T=30ms)

- Signal are 1000 instances of **/aa/** and **/sh/** from TIMIT database + noise
- Noise is Babble (left) or White (right): average global SNR = 5 dB
- Mean log distortion diff as a function of frequency: when < 0 TEO is better





Main Results

- In general (for discrete TEO):
  - TEO better than SEO for first few filters (short/mid-term averaging)
  - TEO better than SEO for fricative sounds
  - TEO better than SEO for low pass noise
  - SEO better than TEO for last few filters (for the discrete approximation of TEO)

Feature Extraction

- D. Dimitriadis, J. Segura, L. Garcia, A. Potamianos, P. Maragos and V. Pitsikalis, "*Advanced front-end for robust speech recognition in extremely adverse environments*", Proc. Interspeech 2007.
- D. Dimitriadis, P. Maragos and A. Potamianos, "*Robust AM-FM Features for Speech Recognition*", IEEE Signal Processing Letters, 2005.
- D. Dimitriadis, P. Maragos and A. Potamianos, "*Auditory Teager Energy Cepstrum Coefficients for Robust Speech Recognition*", Proc. Interspeech 2005.

Energy and Modulation Features

- Energy-Related Features – **TECC**

- Inst. Frequency-Related Feature Sets
  - **IF-Mean, IF-Var**
  - **FMP**

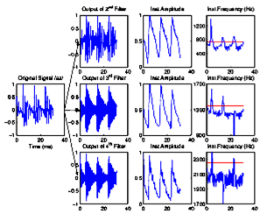
- Inst. Amplitude-Related Features
  - **IA-Mean, IA-Var**
  - **BandW-Mean, BandW-Var**
  - **ΔBandW-Mean, ΔBandW-Var**

Advanced Front-End

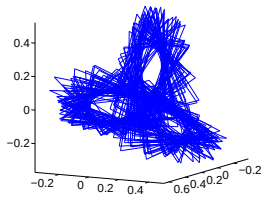
Speech

$s_i(t)$

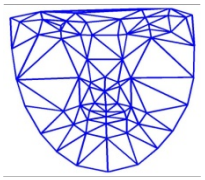
M-Array Processing



Modulations – Energy
- Multiband Filtering
- Nonlinear Processing
- Demodulation



Dynamics - Fractals
- Embedding
- Geometrical Filtering
- Fractal Dimensions



Visual
- Active Appearance Model
- Face Detection/Tracking
- Mouth R.O.I. Features

Feature Transform./ Selection

Fusion

MFCC

Feature Stream

IA-Mean

IF-Mean

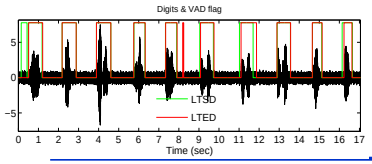
FMP

TECC

FDCD

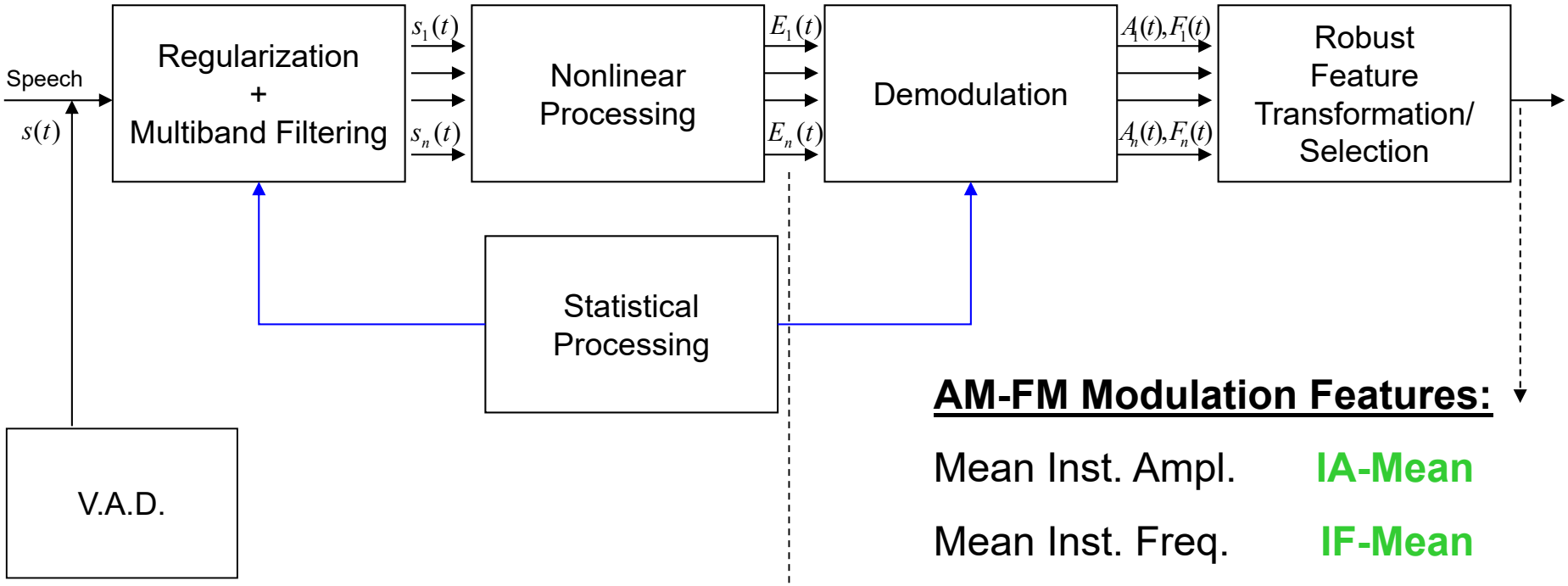
MFD

VAD



Speaker Normalization

Modulation – Teager-Energy Acoustic Features (Overview)



$$\Psi[x(t)] \equiv [\dot{x}(t)]^2 - x(t)\ddot{x}(t)$$

AM-FM Modulation Features:

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

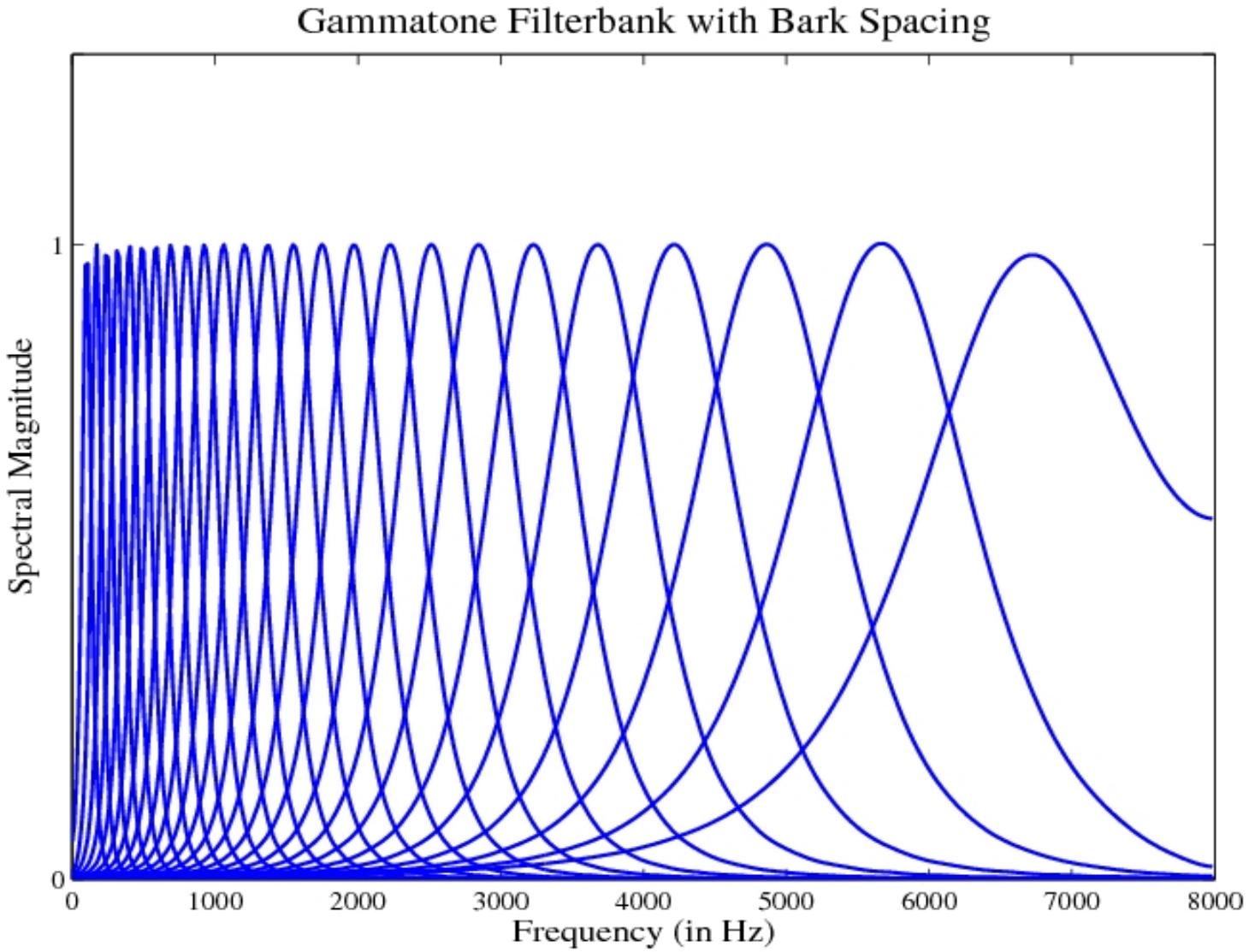
Mean Inst. Freq. **IF-Mean**

Freq. Mod. Percent. **FMP**

Energy Features:

Teager Energy Cepstrum Coeff. **TECC**

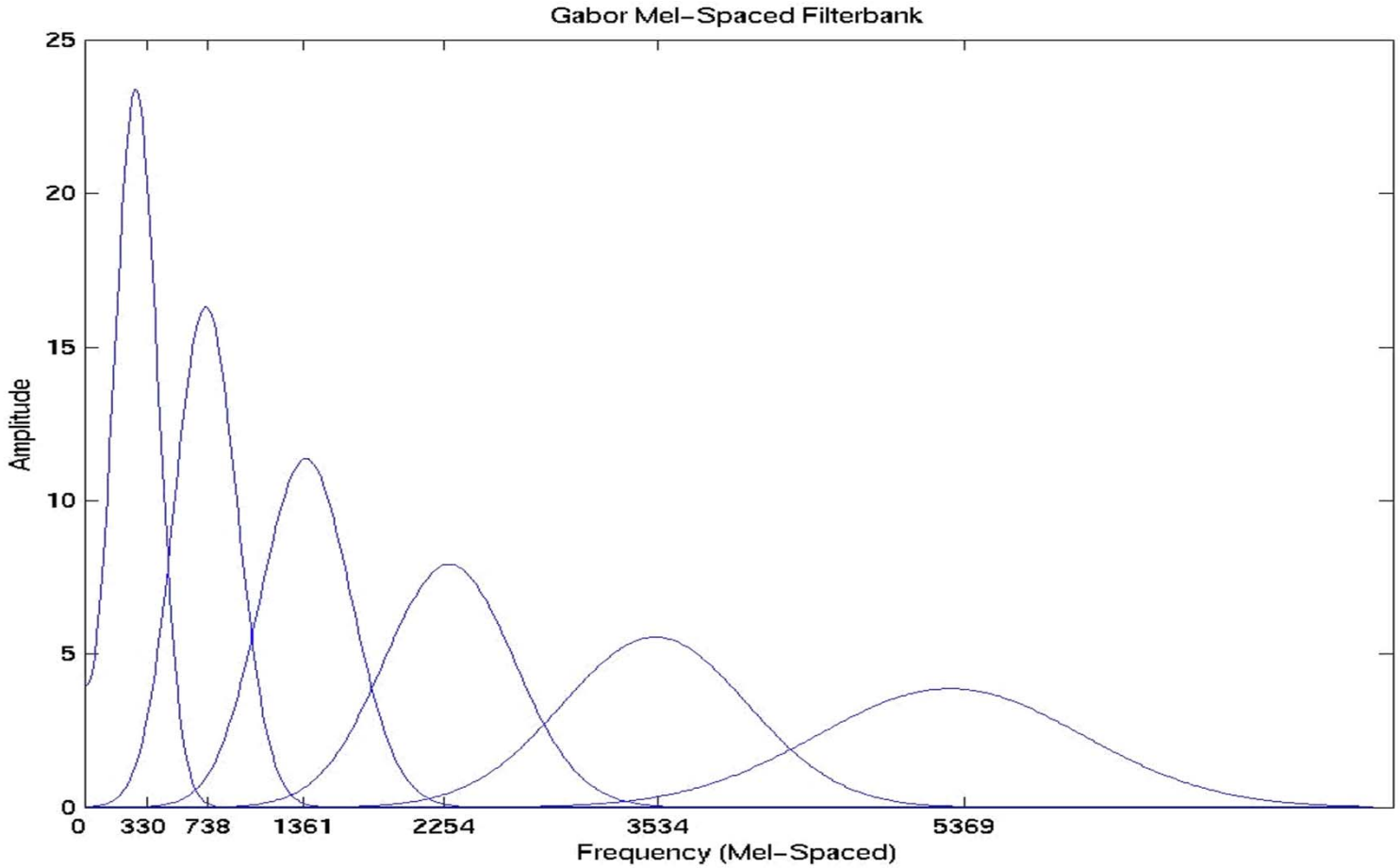
Filterbank Design (I)



Filterbank Design (II)

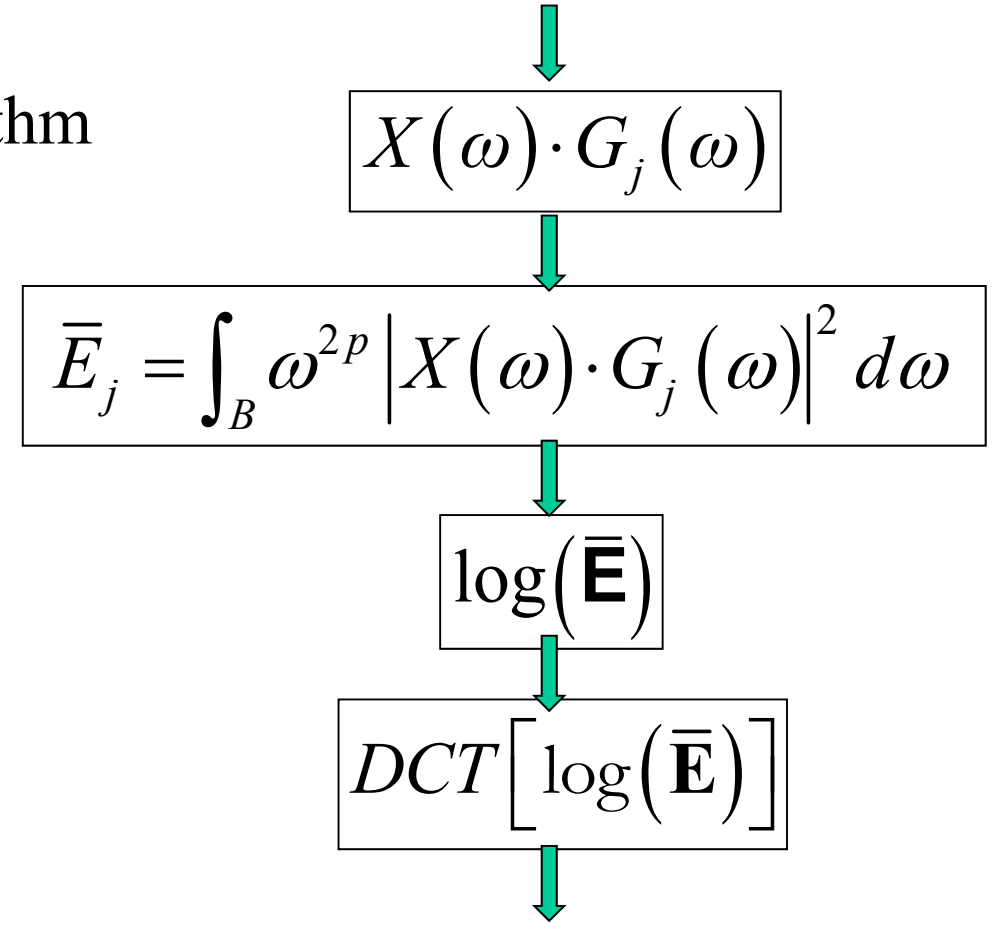
Mel-spaced Gabor Filterbank for Feature Extraction

(filters are normalized to have constant energy)



Teager-Energy Cepstral Coefficients (TECC)

- TECC Extraction Algorithm
  - Filter Speech
  - Estimate Mean Energy
  - Log mean Energy
  - Truncate of Cepstrum



FM-Based Feature Extraction

- Weighted Mean Instantaneous Frequency and Bandwidth Estimates

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

- Un-weighted Mean Instantaneous Frequency and Bandwidth Estimates

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

FM-Based Feature Extraction: FMPs and IFMs

- *FMP* Features: Frequency Modulation Percentages.

$$Coeff_i = \frac{B_{w,i}}{F_{w,i}}$$

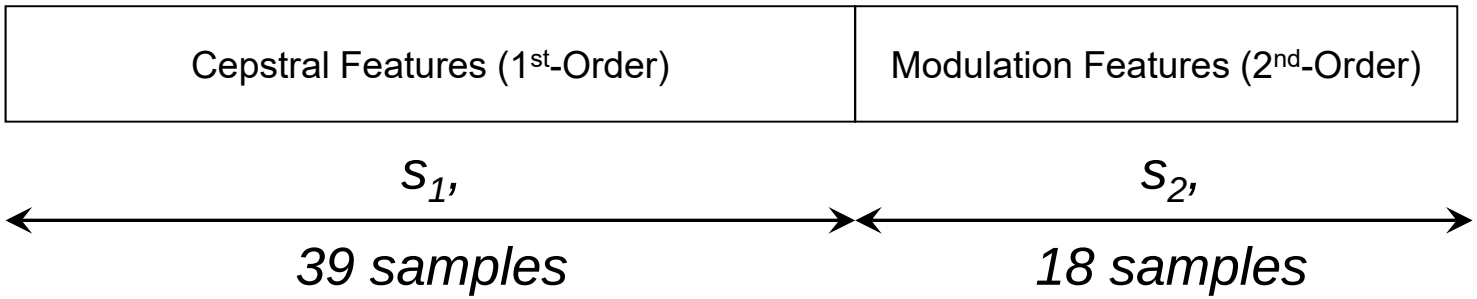
- *IFM* Features: Instantaneous Frequency Mean Values.

$$Coeff_i = F_{w,i}$$

- Concatenated as 2nd Data Stream to *MFCCs* or *TECCs*.

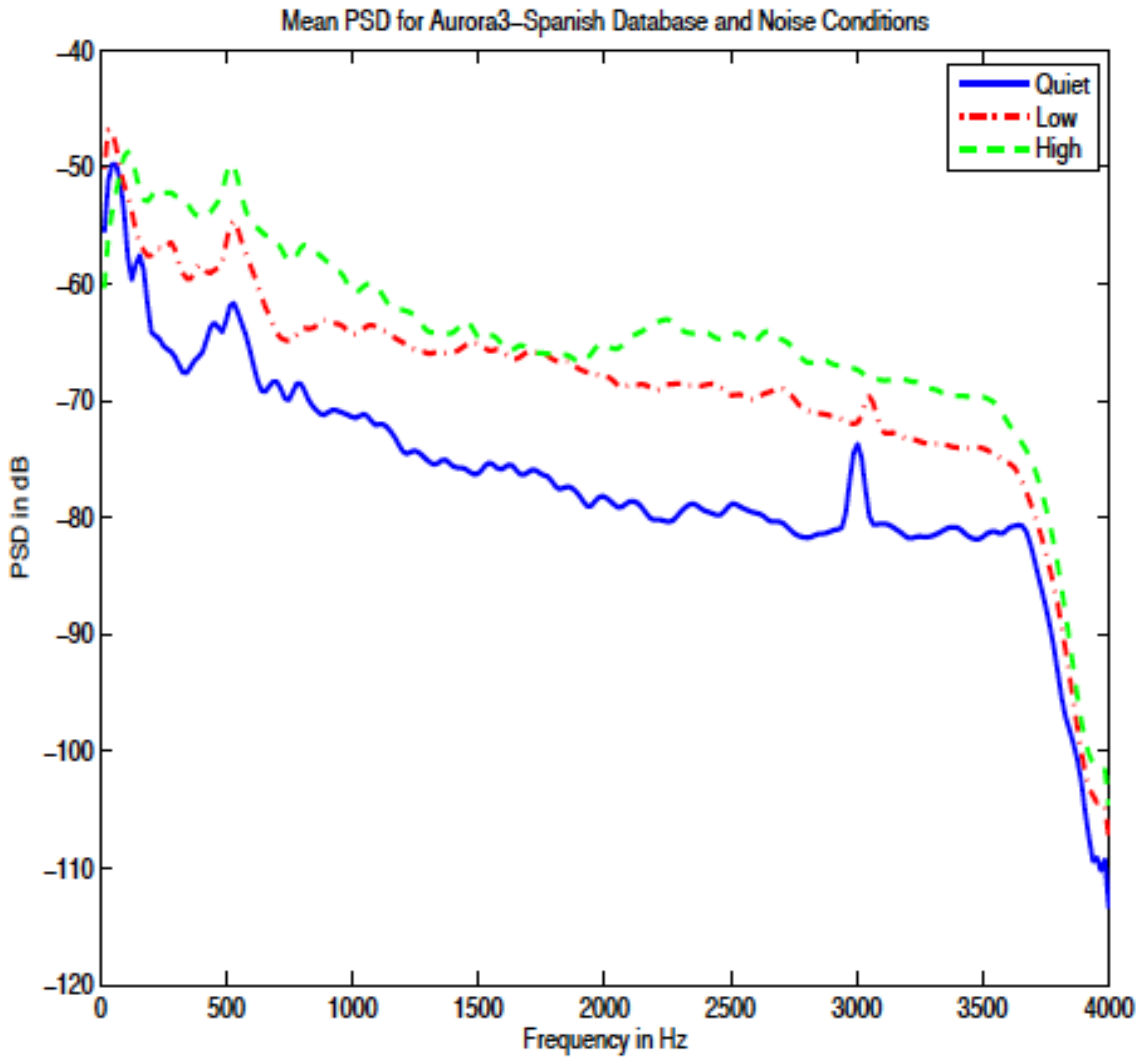
Feature Combination

- Hybrid Feature Vector



- Separate Streams for Cepstral and Modulation features: MFCCs, PLPs, TECCs, …
- Modulation features can be any of: IA-Mean, IA-Var, IF-Mean, IF-Var, FMP,…
- First and Second time derivatives also included

Aurora-3 Spanish: Spectral "Fingerprint"



Quiet: 12 dB
Low noise: 9 dB
High noise: 5 dB

HAFE and ETSI AFE Comparison in Noisy Conditions (Additive Noise)

HAFE = TECC & FMP & CMS & Wiener & FD & PEQ

	Word Accuracy for the AURORA 3 Spanish Task				
	WM	HM	Average	% error reduct. over baseline	**% error reduct. over ETSI AFE**
BASELINE	93.7%	65.2%	79.5%	0.0%	
ETSI AFE standard	96.6%	90.8%	93.7%	69.3%	**0.0%**
HAFE	97.4%	92.7%	95.1%	75.9%	**21.4%**

	Word Accuracy for the HIWIRE DB					
	Clean	10dB SNR	5dB SNR	Average	% error reduct. over baseline	**% error reduct. over ETSI AFE**
BASELINE	91.4%	46.5%	24.7%	54.2%	0.0%	
ETSI AFE standard	89.0%	71.1%	58.0%	72.7%	40.4%	**0.0%**
HAFE	93.9%	81.1%	61.8%	78.9%	54.0%	**22.8%**

D. Dimitriadis, J. C. Segura, L. Garcia, A. Potamianos, P. Maragos, and V. Pitsikalis, “Advanced front-end for robust speech recognition in extremely adverse environments,” Proc. Interspeech 2007.

Aurora 3 - Spanish Task

- Connected-Digits, Fs: 8 kHz
- 2 Feature Vectors:
  - ***MFCC*** or ***TECC + C0***
  - ***FMP (Modulation Features) or MFD (Fractal Features)***
    + Wiener Filtering (WF) + Cepstral Mean Subtraction (CMS) + Parameter Equalization (PEQ) + Regr. Coefficients + Frame Dropping (FD)
- ***PEQ***-Statistics Calculation held only on High-Noise Data
- All-Pair, Unweighted Grammar (or Word-Pair Grammar)
- Performance Criterion: Word (digit) Accuracy Rates

Aurora-3, Spanish Task

Correct Word Accuracies (%)

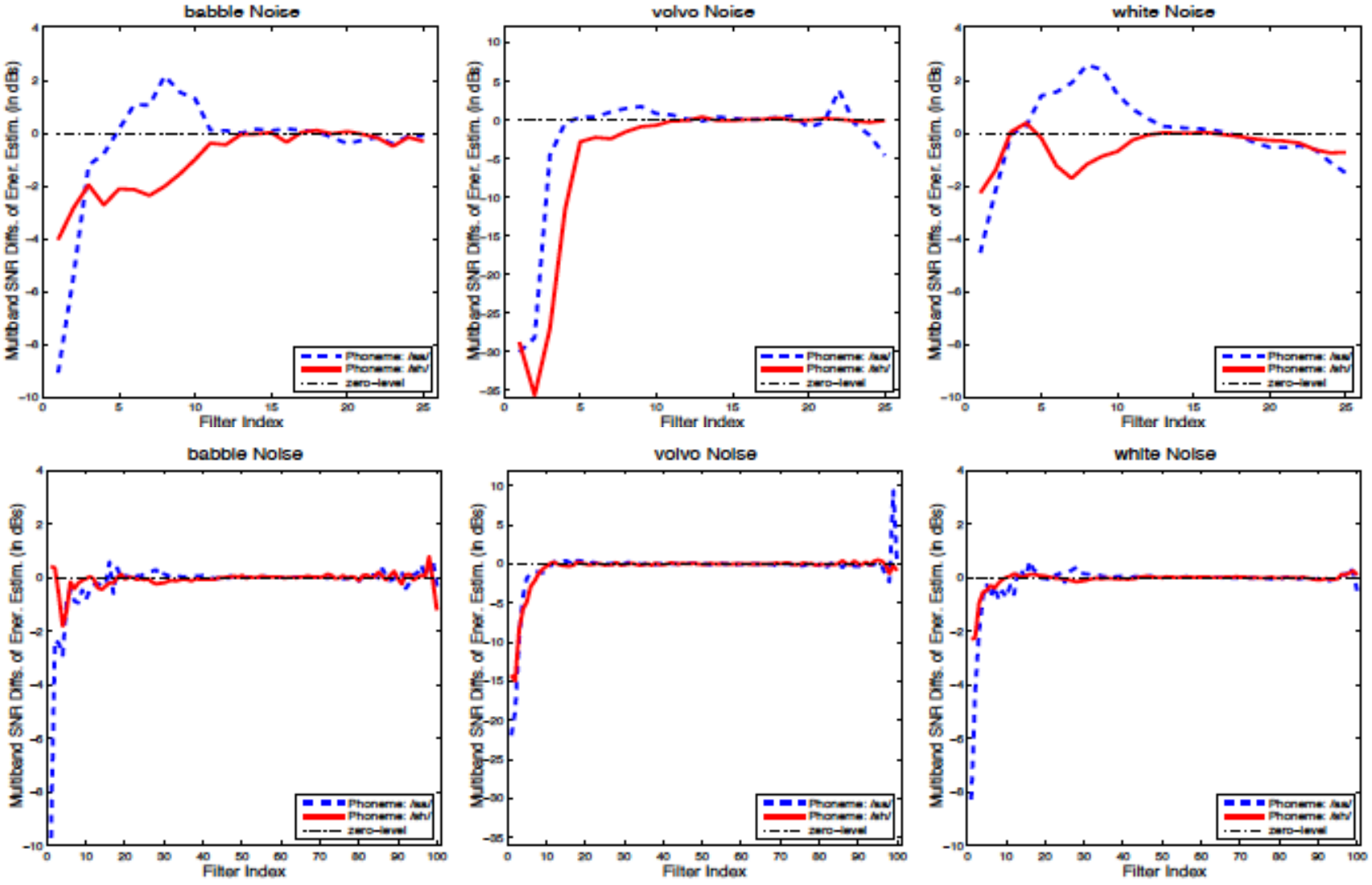
	WM	MM	HM
MFCC+c0+D+DD+CMS (Baseline - HTK)	93.68	92.73	65.18
MFCC (HAFE)	96.93	92.98	91.46
TECC (HAFE)	96.90	92.56	91.82
TECC+FMP (HAFE)	**97.39**	**93.64**	**92.72**
MFCC+MFD (HAFE)	96.96	92.67	92.42

Investigating Filterbank Configurations and Energy Computations

D. Dimitriadis, P. Maragos, and A. Potamianos, "*On the Effects of Filterbank Design and Energy Computation on Robust Speech Recognition*",
IEEE Transactions on Audio, Speech and Language Processing, Aug. 2011.

Energy Deviation (on TIMIT + Noise)

Mel-spaced Gammatone filterbanks with 50% overlap (Top: 25, Bottom: 100 filters)



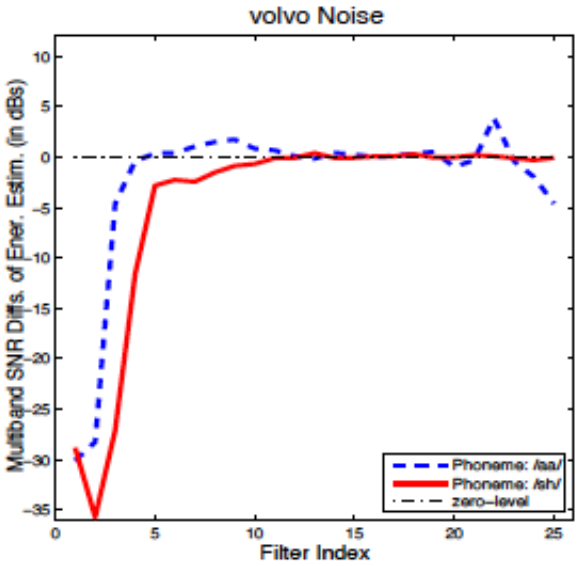
Cepstral Coefficient Deviations

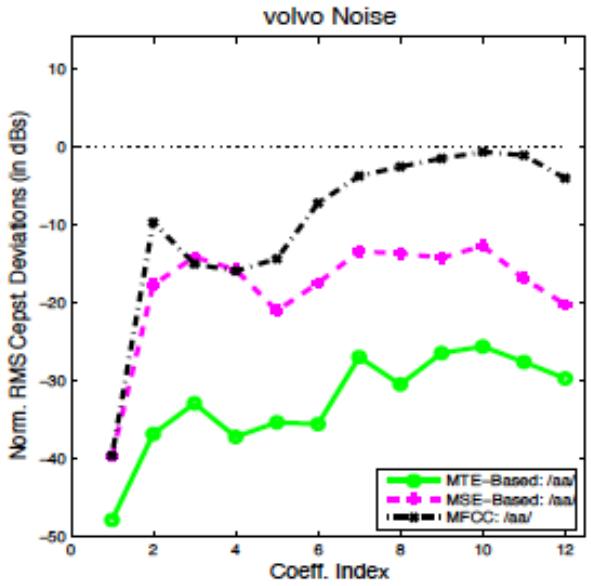
- Energy Estimation Deviations Propagate to Cepstral Coefficient Deviations

  (Refs: Deng et.al, T-SAP 2004. Moreno, CMU 1996. Raj, Gouvea, Moreno & Stern, ICSLP 1996)

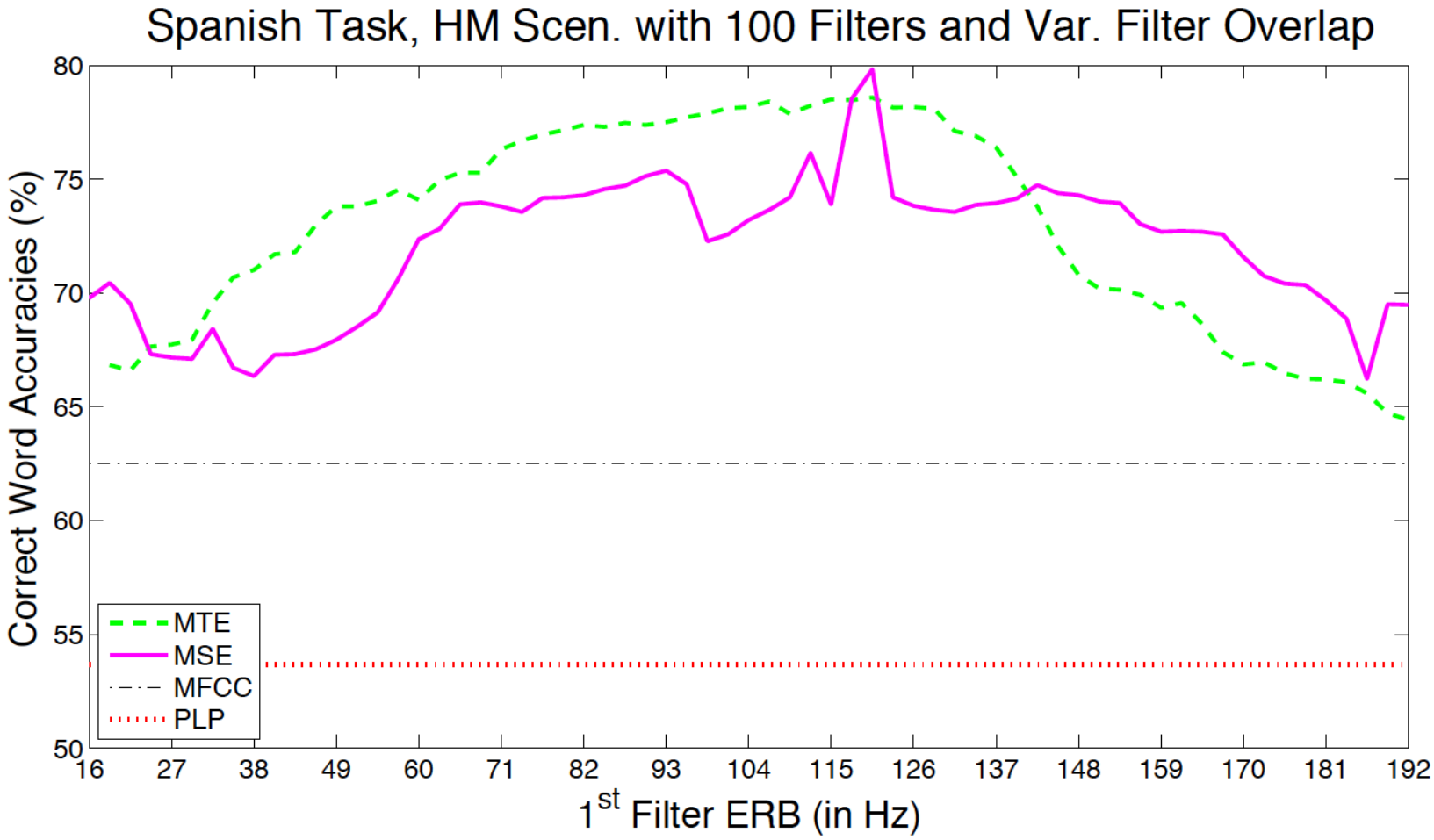
$$\Delta C[i] = \sum_{j=1}^{J} W_{ij} \log\left(1 + D_j\right)$$

where i: the cepstral coeff index, j the energy coeff index, W_{ij} the DCT coeff and D_j the energy deviation at the filter j .





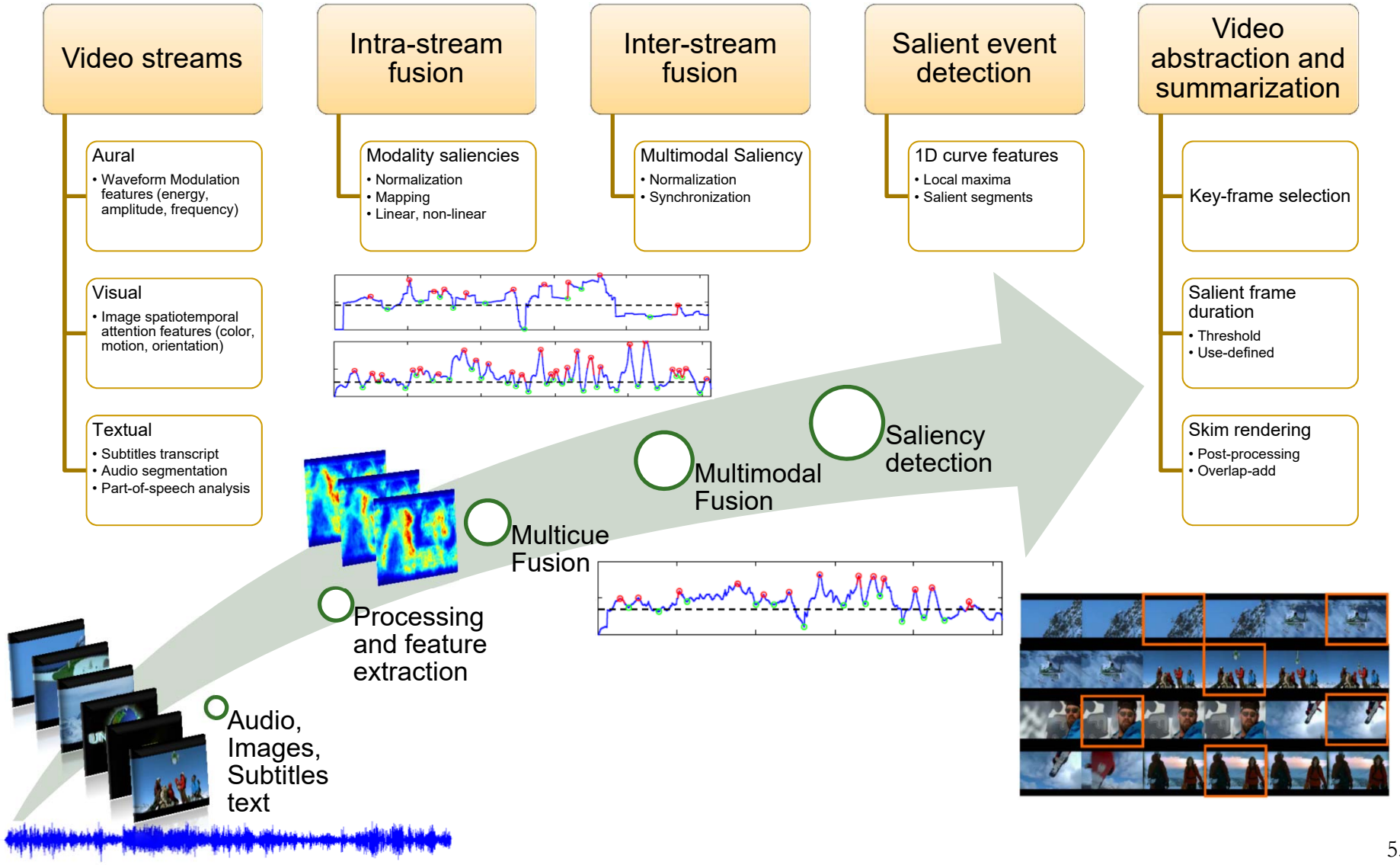
Word Accuracies for Aurora-3 Spanish Task (High-Mismatch): MTE vs MSE-based Features



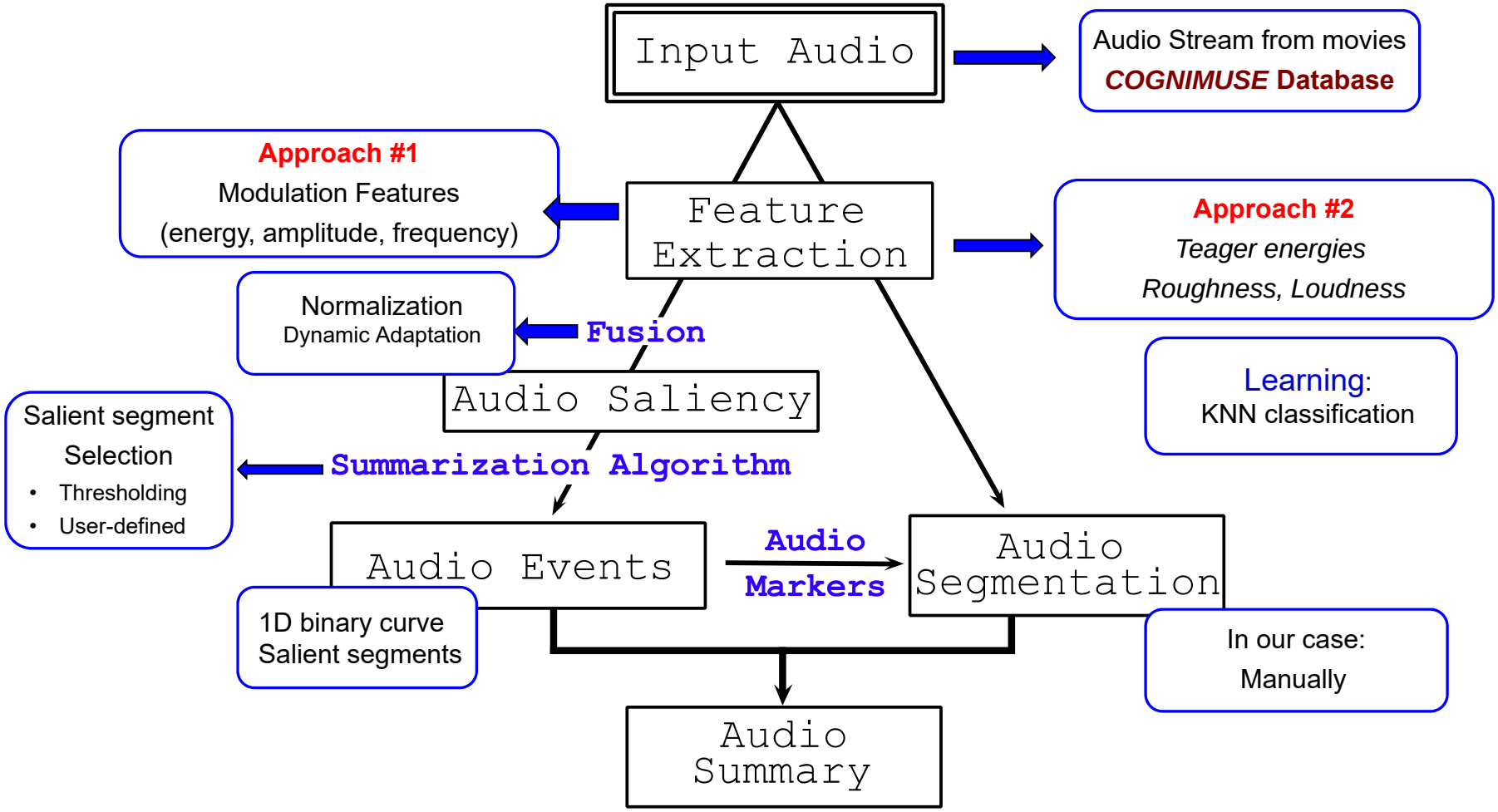
Dominant Speech Modulations and **Audio Summarization**

A. Zlatintsi, P. Maragos, A. Potamianos and G. Evangelopoulos, "*A Saliency Based Approach to Audio Event Detection and Summarization*", Proc. EUSIPCO 2012.

Movie Video Event Detection and Summarization



Audio Summarization System Overview



[A. Zlatintsi, E. Iosif, P. Maragos and A. Potamianos, *Audio Salient Event Detection And Summarization Using Audio And Text Modalities*, EUSIPCO 2015]

[P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos and A. Potamianos, *Predicting Audio-visual Salient Events based on A-V-T Modalities For Movie Summarization*, ICIP 2015.]

Audio Analysis I (Feature extraction)

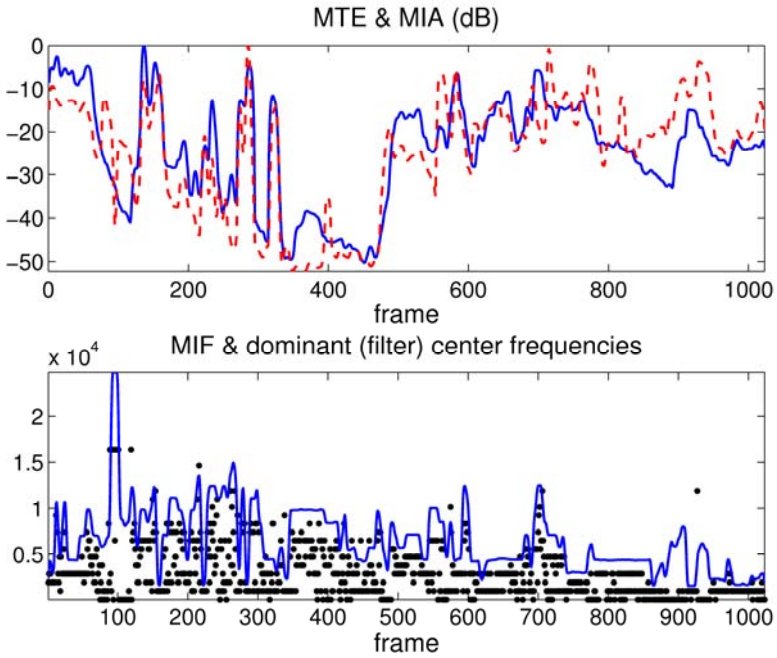
- Audio AM-FM model: $s[n]=\sum_{k=1}^{K} A_k[n]\cos\left(\int_0^n \Omega_\kappa[m]\,dm\right)$

- Modulation bands
  - K Gabor filters h_k , narrowband components
- Nonlinear energy tracking
  - Teager-Kaiser energy operator
  - ESA demodulation
- Dominant modulation features

$$\mathrm{MTE}[m]=\max_{1\le k\le K}\frac{1}{N}\sum_{n=1}^{N}\Psi\left(s*h_k[n]\right)$$

$$i=\underset{k}{\mathrm{argmax}}\left\{\mathrm{MTE}[m;k]\right\}$$

$$\mathrm{MIA}[m]=\frac{1}{N}\sum_{n=1}^{N}\left|A_i[n]\right| \qquad \mathrm{MIF}[m]=\frac{1}{N}\sum_{n=1}^{N}\left|\Omega_i[n]\right|$$



G. Evangelopoulos and P. Maragos, “Multiband modulation energy tracking for noisy speech detection,” *IEEE Trans. Audio Speech Language Processing,* 2006.

Audio Analysis II (Fusion and Saliency)

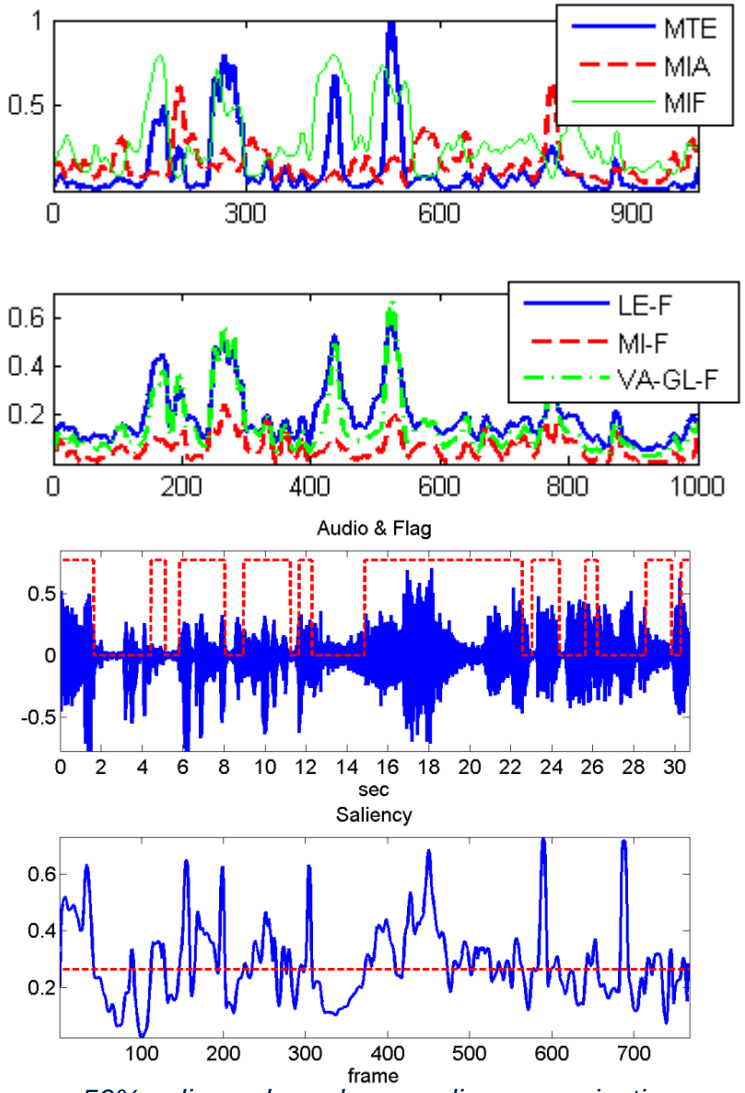
- Audio saliency cues
  - extracted through nonlinear operators

  Convey information on:
  - excitation level
  - frequency content
  - source energy tracking

- 3D Feature vector formation

$$\vec{F}_{\mathrm{a}}[m]=(\mathrm{MTE},\ \mathrm{MIA},\ \mathrm{MIF})[m]$$

- Monomodal saliency curve
  - Continuous-valued indicator of salient events, in [0,1]



50% saliency-based raw audio summarization

G. Evangelopoulos, K. Rapantzikos, P. Maragos, Y. Avrithis, and A. Potamianos, "Audiovisual attention modeling and salient event detection," in *Multimodal Processing and Interaction: Audio, Video, Text,* P. Maragos, A. Potamianos, and P. Gros, Eds. Springer, 2008.

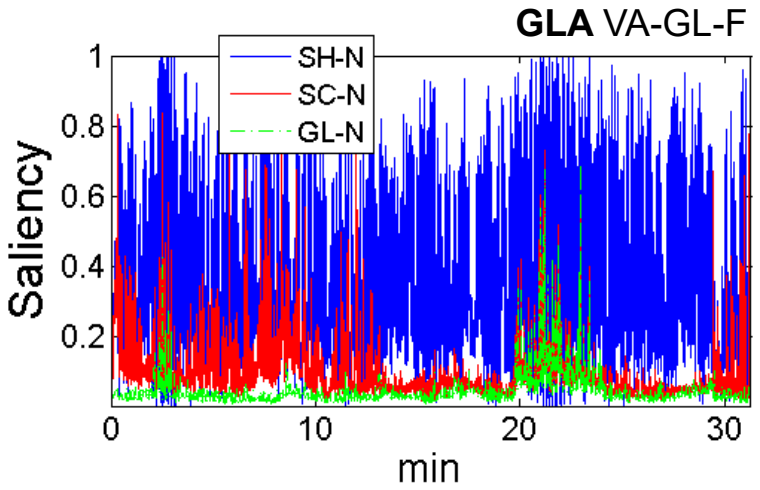
Monomodal Fusion I (Event detection)

- Nine Fusion schemes $S_A = \text{fusion}(S_1, S_2, S_3)$
  - **Linear** (equal weights) (Low-level, memoryless) $S_{\text{LIN}} = w_1 S_1 + w_2 S_2 + w_3 S_3$
  - **Variance-based** (adaptive weights) $S_{VAR} = \sum_i \left(\frac{S_i}{\text{var}(S_i)} \right) / \sum_i \left(\frac{1}{\text{var}(S_i)} \right)$
  - **Nonlinear**
    - MIN $S_{MIN} = \min\{S_1, S_2, S_3\}$
    - MAX $S_{MAX} = \max\{S_1, S_2, S_3\}$
    - Weighted MIN $S_{MIVA} = \min(S_1 - w_1, S_2 - w_2, S_3 - w_3\} + \max(w_1, w_2, w_3)$

$$\text{where } w_i = \log\left(\frac{1}{\text{var}(S_i)} \right)$$

Monomodal Fusion II - Normalization

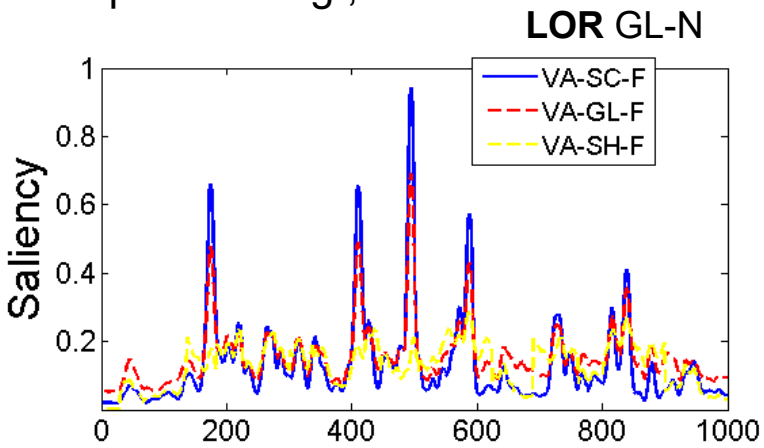
- Normalization intervals
  - Global linear normalization (GL)
  - Scene-based linear normalization (SC)
  - Shot-based linear normalization (SH)



- Dynamic Adaptation levels

  i.e., weight updating with respect to Global or Local windows

  Inverse Variance & Weighted Min fusion can be computed at e.g.,
  - Global level (VA-GL)
  - Scene level (VA-SC)
  - Shot level (VA-SH)



Summarization Algorithm

Median-filtered Saliency
(length $2M + 1$)

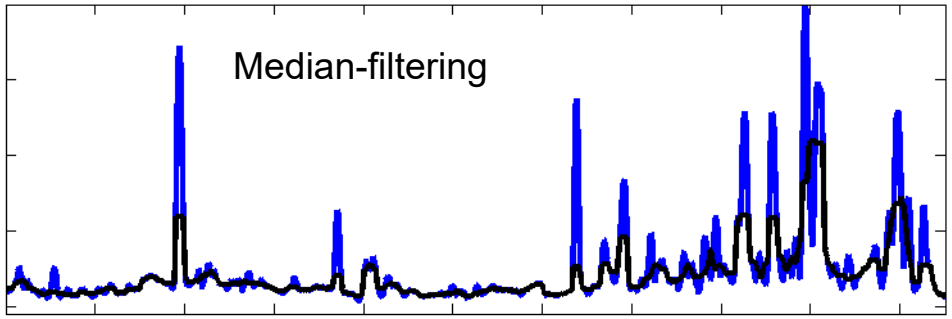
Threshold selection

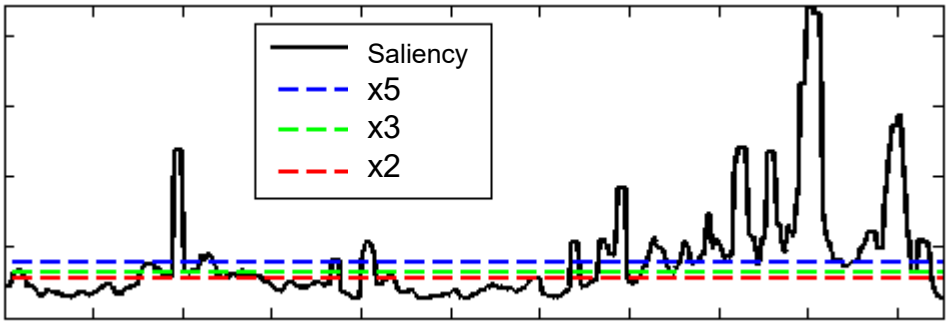
Segment selection

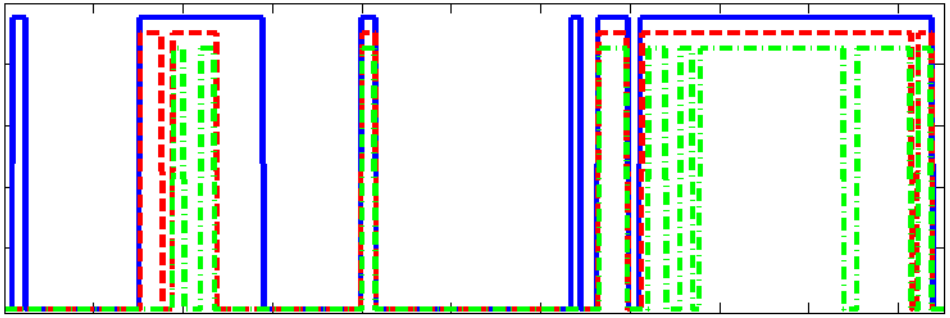
Reject: segments shorter than
N frames
(Morphological opening)

Join: segments less than
K frames apart
(*Morphological closing*)

Render:
Linear overlap-add







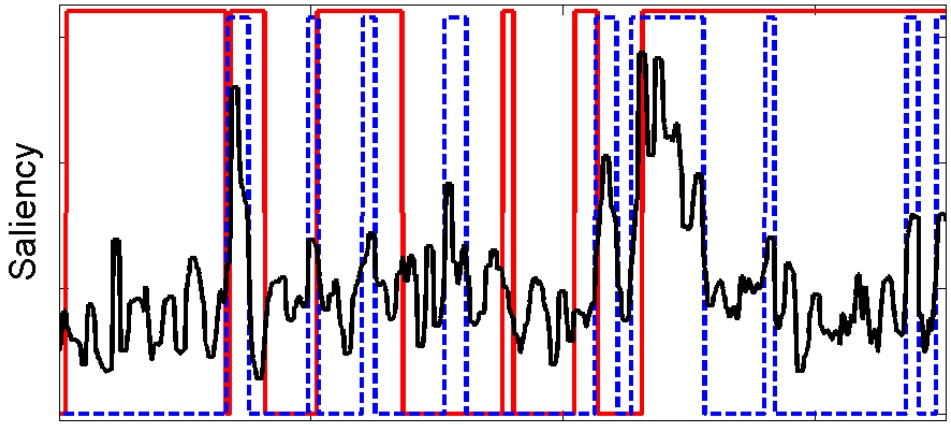
Selected Segments

Objective Evaluation

- Audio from Academy awarded movies (MovSum database)

  ca. 30 min. duration segments (on average 13 scenes/movie, 560 shots/movie)
  - **GLA**, *"Gladiator"*: DreamWorks SKG, 2000
  - **CHI,** *"Chicago"*: Miramax Films, 2002
  - **LOR**, *"Lord Of the Rings III: The Return of the King"*: New Line Cinema, 2003
  - **CRA,** *"Crash"*: Bob Yari Productions, 2005
  - **DEP,** *"Departed"*: Warner Bros. Pictures, 2006
  - **FNE**, *"Finding Nemo"*: Walt Disney Pictures, 2003

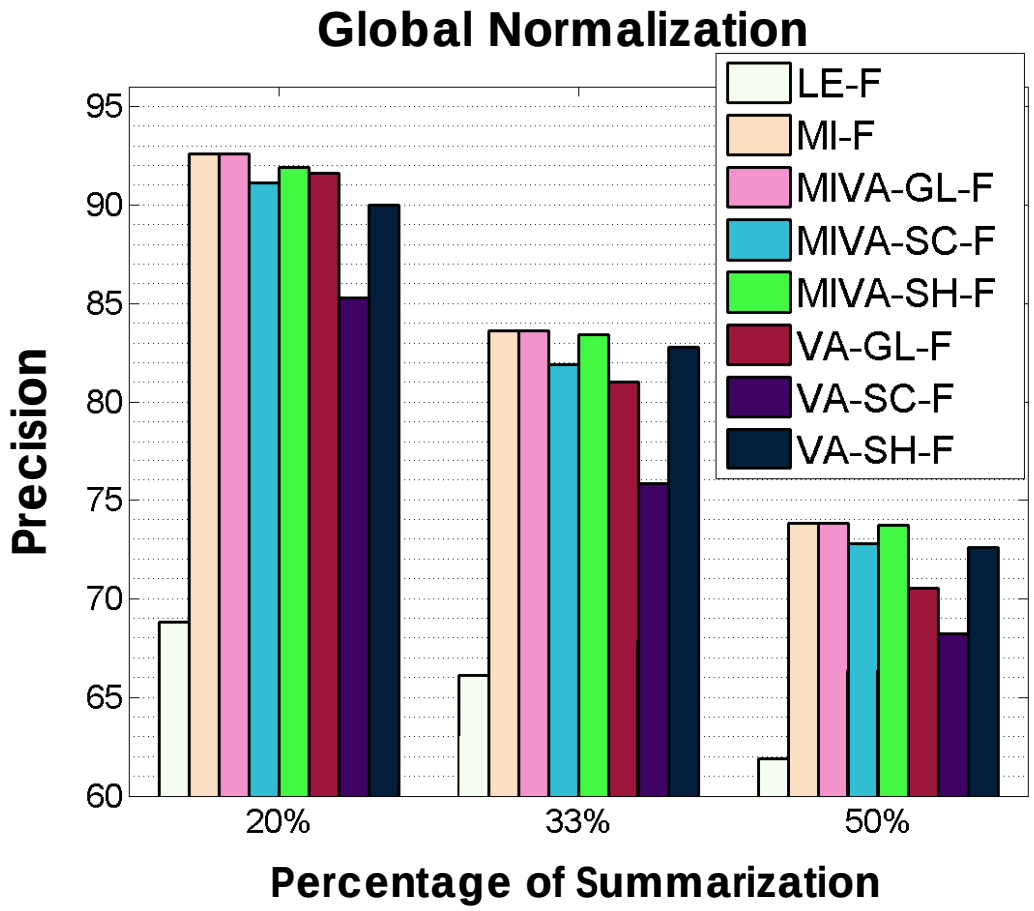
- Skimming rates: c = 20%, 33%, 50% (x5, x3, x2 real time summaries)
- Correspondence with manually labelled saliency



Labeled Saliency
SUM x2
'MIVA-SH-F'

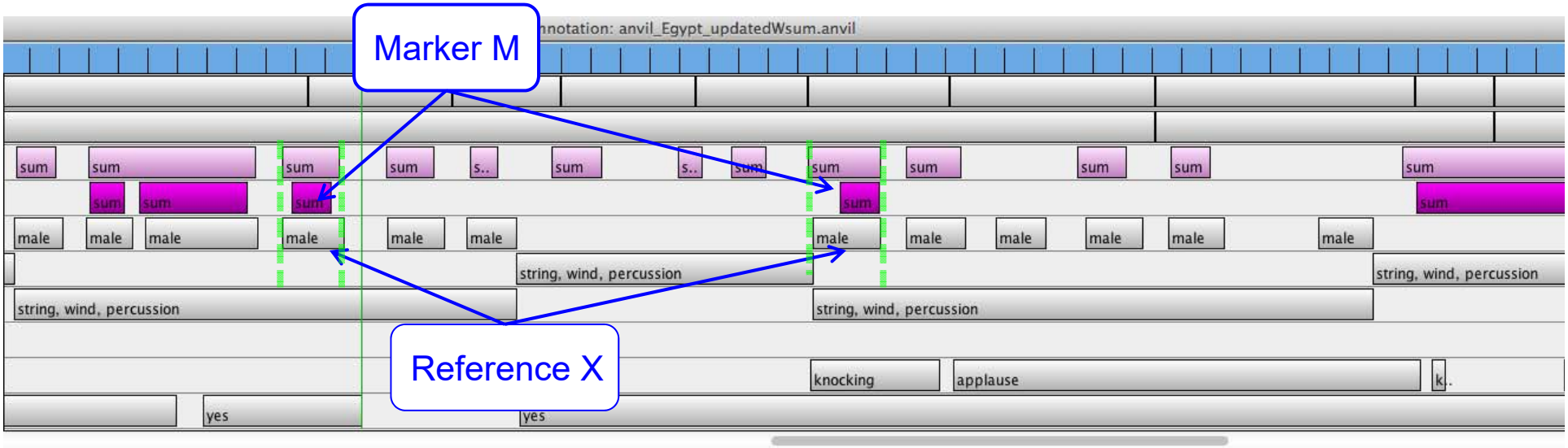
Results for System #1

- Results in terms of frame-level precision



Audio summarizer

- Choose segments salient and meaningful: perform boundary correction
- Reconstruction Opening ➡ connected components of X intersecting M
- VAD-like algorithms could provide automatic segmentation



[P. Maragos, The Image and Video Processing Handbook, chapter Morphological Filtering for Image Enhancement and Feature Detection, Elsevier Acad. Press, 2005]

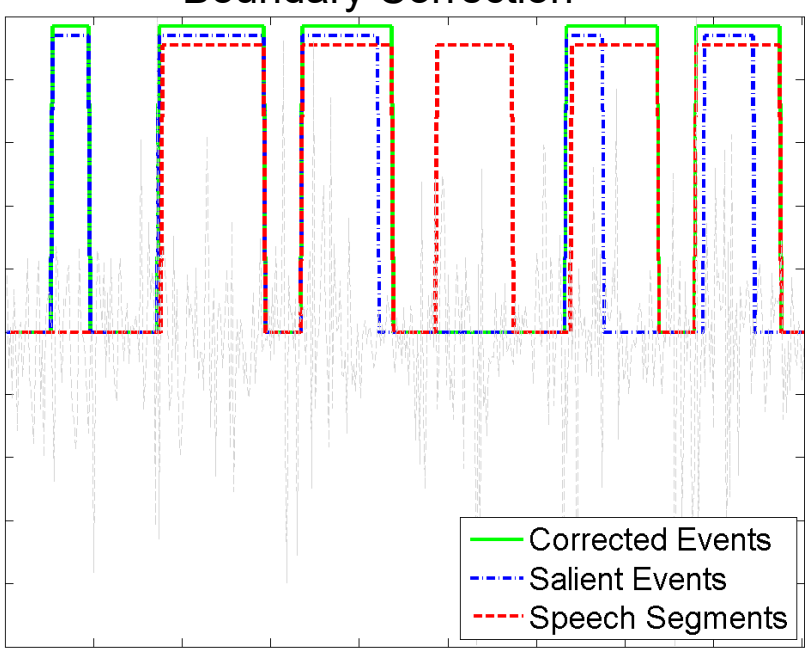
Demo

Audio Summary Example

- Audio extracted from documentary

- Duration of original segment 3 min
  - Including: speech (narration), music, diverse “bang”-sounds

- Summary x3 : duration 1.02 min
  - Corrected boundaries regarding speech



Demo I: Movie Summarization (System #1)

LOR VA-SH-F, rate: x5 (6:50 min from 37:33 min)
Inform: 78.7 %
Enjoy: 80.9 %

AM-FM Modulation Features for Music Analysis & Classification

Refs:

- A. Zlatintsi and P. Maragos, "*Comparison of Different Representations Based on Nonlinear Features for Music Genre Classification*", Proc. EUSIPCO 2014.
- A. Zlatintsi and P. Maragos, "*AM-FM Modulation Features for Music Instrument Signal Analysis and Recognition*", Proc. EUSIPCO 2012.

Motivation

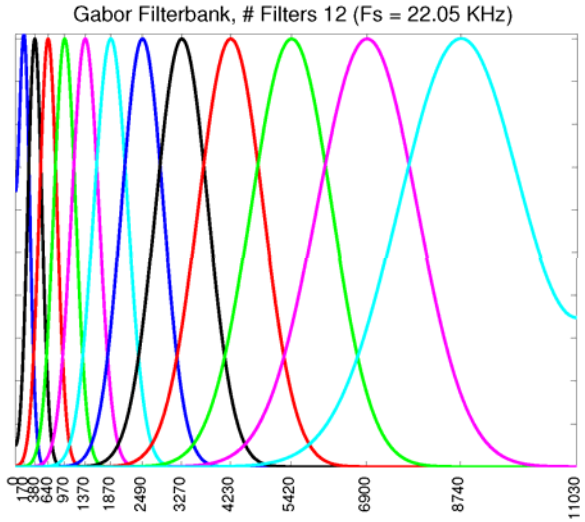
Methodology

- Existence of modulations in music (e.g., vibrato, tremolo)
- *Claims* that music is mimetic reg. nature, human emotions, properties of certain objects;
  and that nature contains structures (e.g., mountains, coastlines, the structures of plants), which could be described by fractals
- Methodology success in **speech recognition, musical instrument classification** and **audio saliency & event detection**
- Parallel evolution of speech and music

Experimental Evaluation

Gabor Filterbanks

- Baseline Gabor filterbank
  - 12 bandpass mel-spaced filters
  - bandwidth overlap equal to 50%

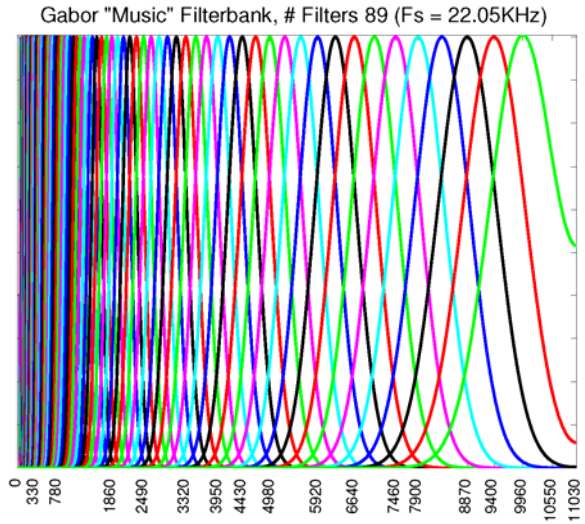


- *"Music"* filterbank

  Center Frequencies f_c: of each filter determined by the frequency of the music tones

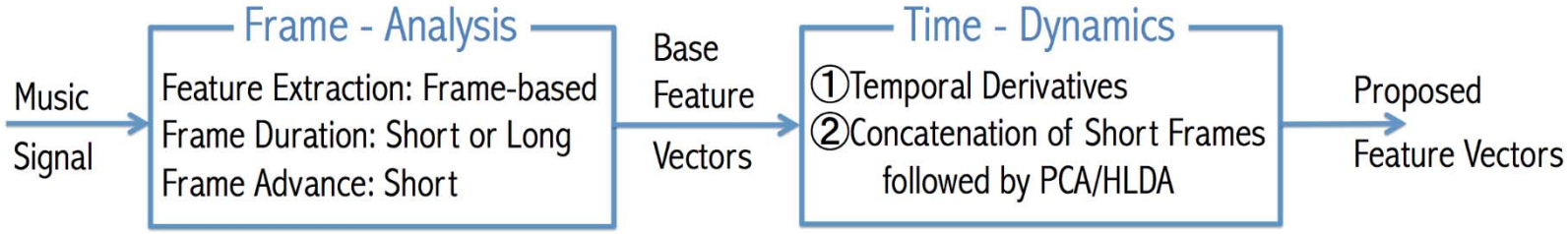
  1) 89 filters starting at C2=65.4 Hz
  2) 101 filters starting at C1=32.7Hz

  - bandwidth: $b_{1i} = [f_{i-1}, f_{i+1}]$ for center frequency f_i



Experimental Evaluation

Proposed Features and Feature Representations (FR)



Feature Sets

- **FR1:** Baseline Gabor filterbank

  Short-time analysis, 30 ms frames with 50% overlap (+ Δs)

- **FR2: "*Music*" Gabor filterbank**

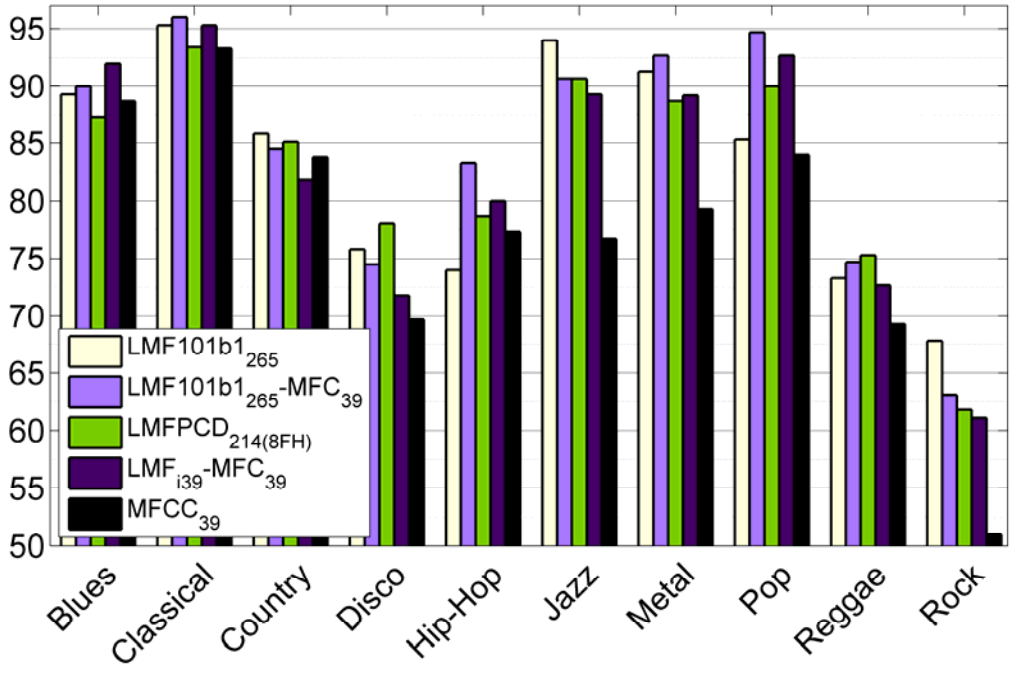
  Short-time analysis, 30 ms frames with 50% overlap (+ Δs)

  followed by PCA analysis for dimensionality reduction

Database for Experimentation:

- GTZAN Database incl. 10 musical genres
  blues, classical, country, disco, hiphop, jazz, metal, pop, reggae και rock
  1000 excerpts, 100 excerpts/genre, 30 seconds each

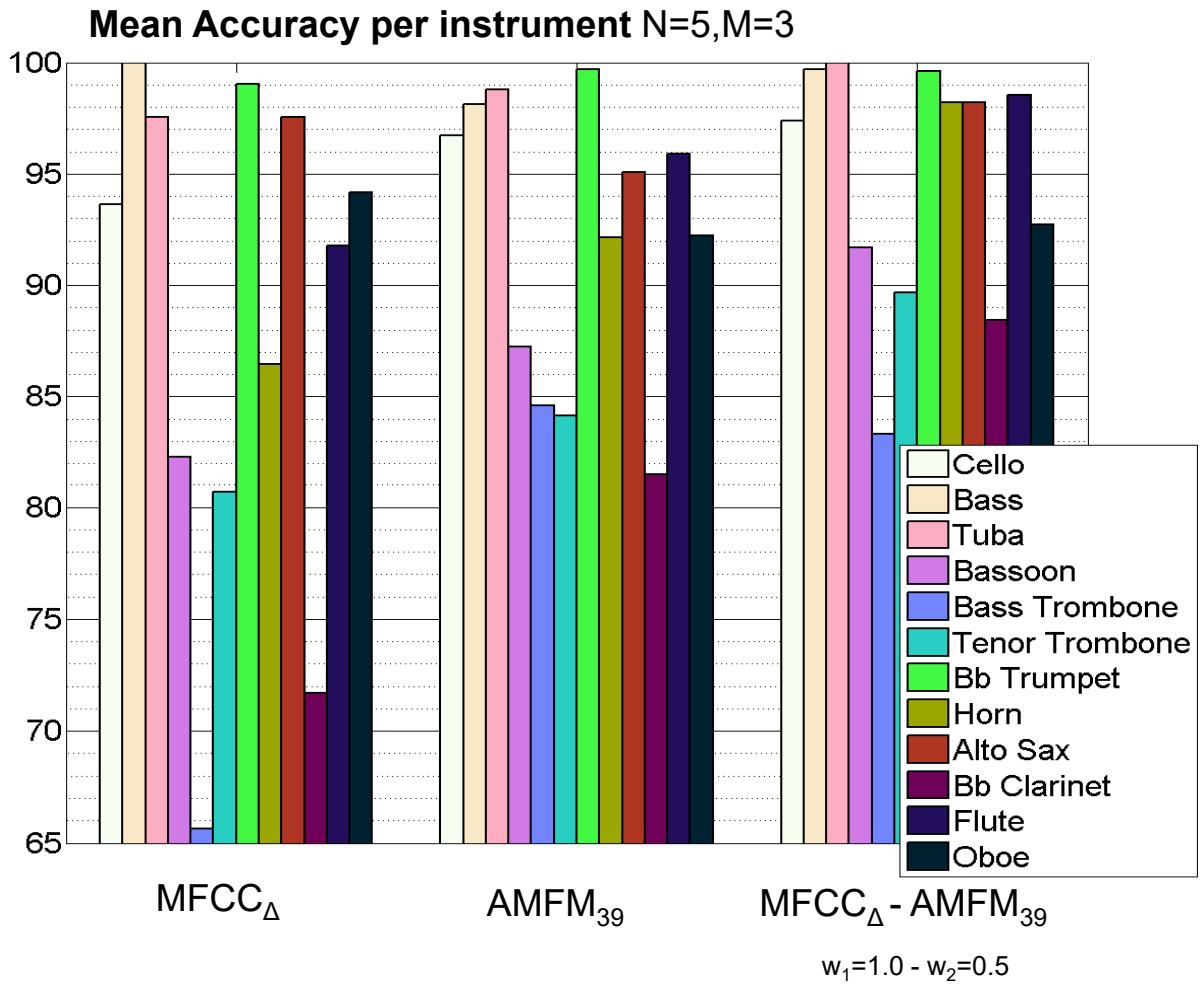
Experimental Evaluation: Different Genres



Conclusions:

- Best recognition:
  - classical – 96%
  - pop – 94.7%
  - jazz – 94%
  - metal – 92.7%
- Worst recognition: rock, reggae & disco
- Better classification for all genres and almost all proposed feature sets compared to MFCC.

Experimental Evaluation: Different Instruments



Conclusions:

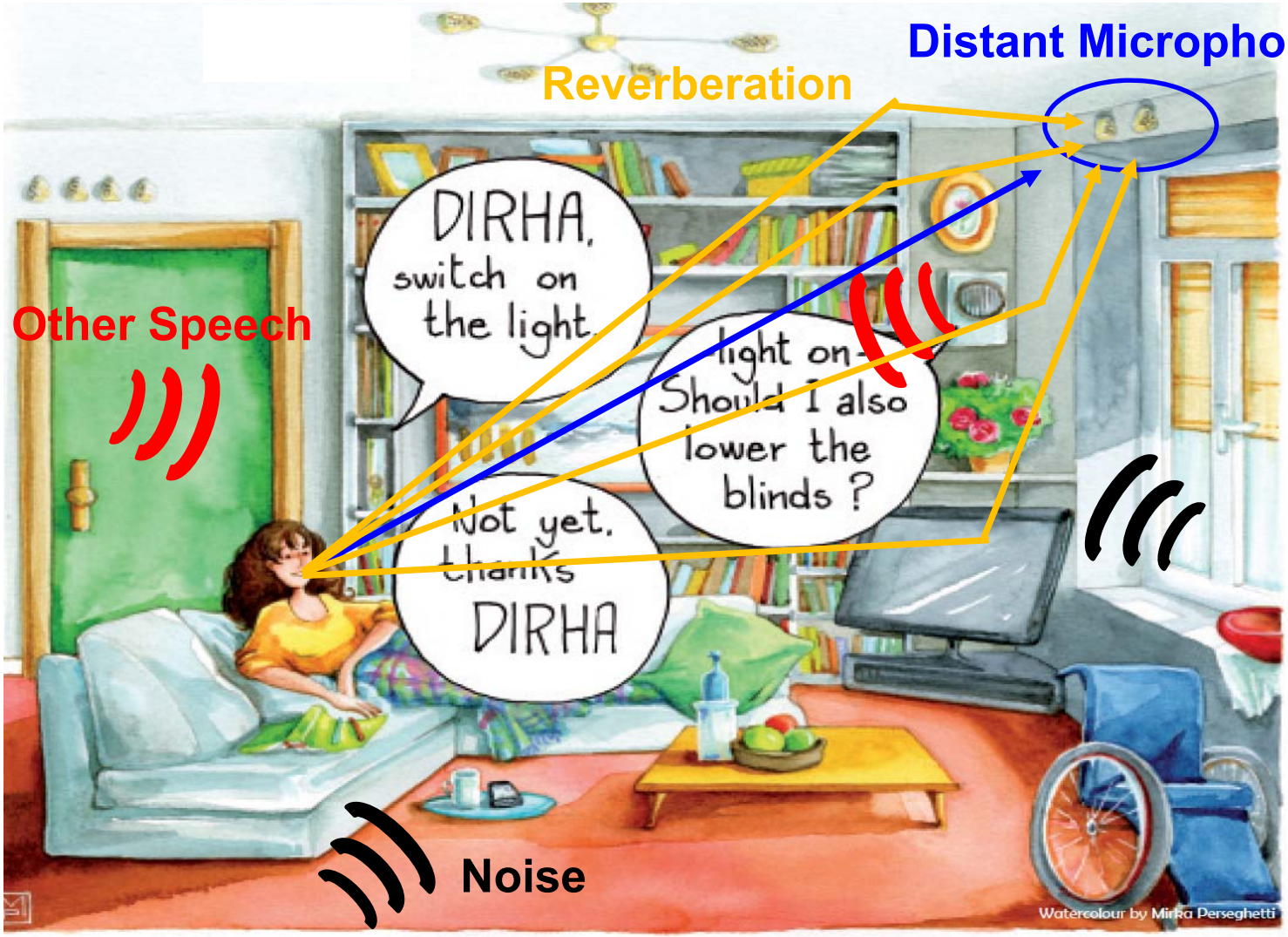
- Better recognition in all (#12) instruments except bass, saxophone and oboe
- Better discrimination between bass and tenor trombone as well as between bass and clarinet

Multi-Microphone Energy Tracking for Robust Distant Speech Recognition

References:

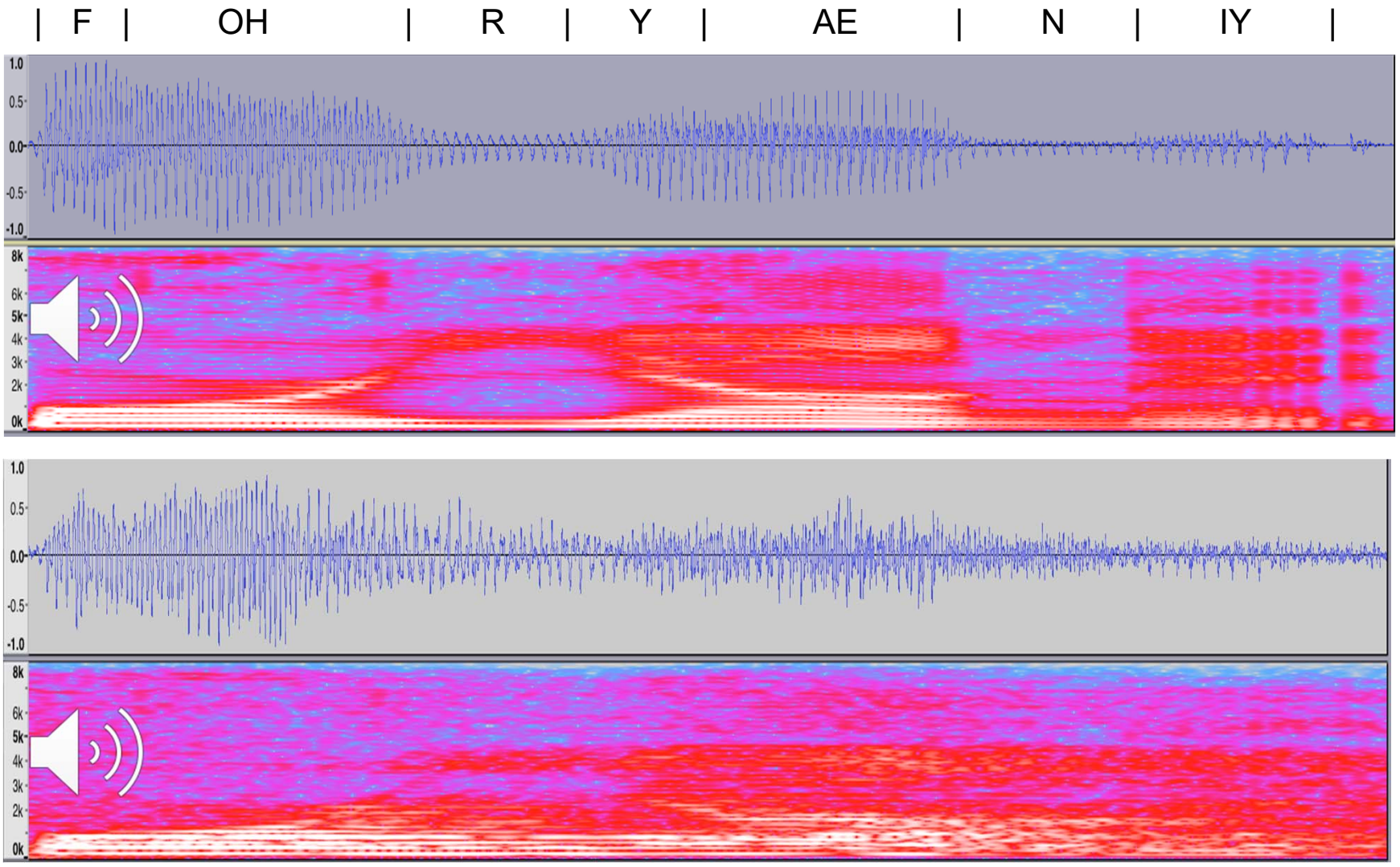
- I. Rodomagoulakis and P. Maragos, "*On the Improvement of Modulation Features Using Multi-Microphone Energy Tracking for Robust Distant Speech Recognition*", Proc. EUSIPCO 2017.
- I. Rodomagoulakis, G. Potamianos, and P. Maragos, "*Advances in Large Vocabulary Continuous Speech Recognition in Greek: Modeling and Nonlinear Features*", Proc. EUSIPCO 2013.

Distant Speech Recognition inVoice-enabled Interfaces

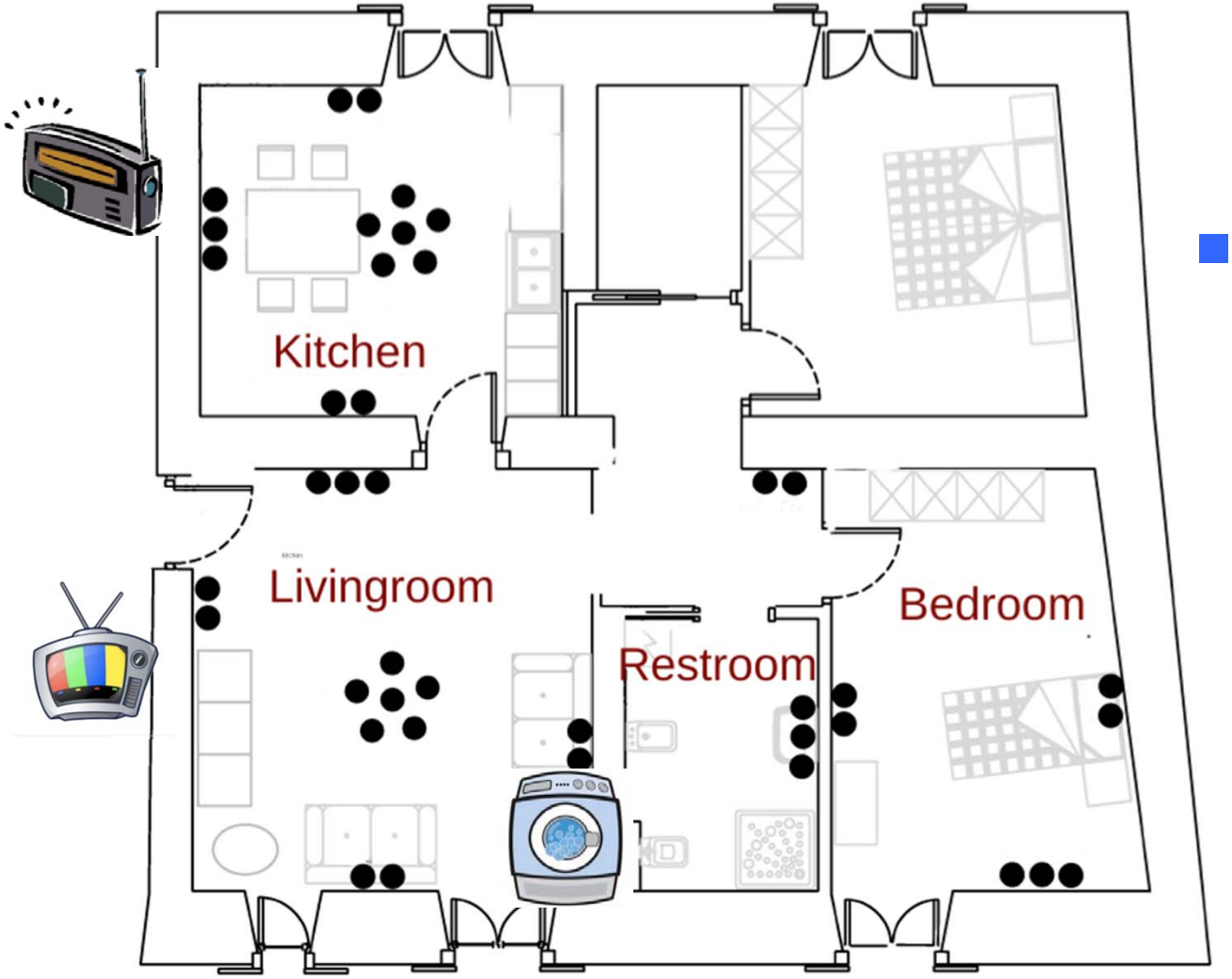


https://dirha.fbk.eu/

Near- v.s. Far-Field Speech



Smart Home Voice Interface



Sweet home listen!
Turn on the lights in the living room!

- Main technologies:
  - Voice Activity Detection
  - Acoustic Event Detection
  - Speaker Localization
  - Speech Enhancement
  - Keyword Spotting
  - Far-field command recognition

DIRHA demo (“spitaki mou”)

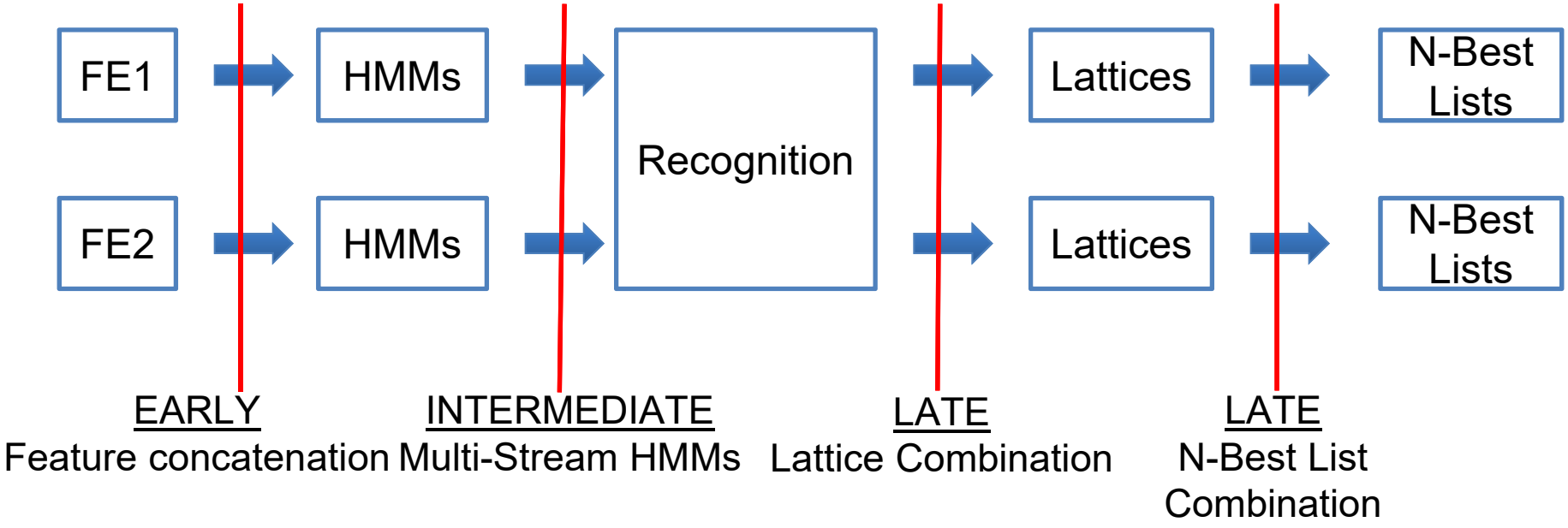
https://www.youtube.com/watch?v=zf5wSKv9wKs

- I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, P. Maragos, “Room-localized spoken command recognition in multi-room, multi-microphone environments”, *Computer Speech & Language*, 2017.
- A. Tsiami, I. Rodomagoulakis, P. Giannoulis, A. Katsamanis, G. Potamianos and P. Maragos, “ATHENA: A Greek Multi-Sensory Database for Home Automation Control”, Proc. Interspeech 2014.

AM-FM features for Distant Speech Recognition

- Features
  - Mean Instantaneous Amplitudes (MIA)
  - Mean Instantaneous Frequencies (MIF)
  - Frequency Modulation Percentages (FMP)
  - Mean Instantaneous Weighted Frequencies (Fw)
- Single-channel DSR
  - Fusion schemes with MFCC
- Multichannel Multiband Demodulation (MMD)
  - Improved estimations of instantaneous modulations of amplitudes and frequencies
- Multichannel DSR using MMD
- Experiments on challenging multichannel DSR databases
  - Baseline HMM-GMM recognizer
  - Ongoing work on DNN-based recognition

Fusion of MIA-MIF with MFCCs (Single-channel DSR)



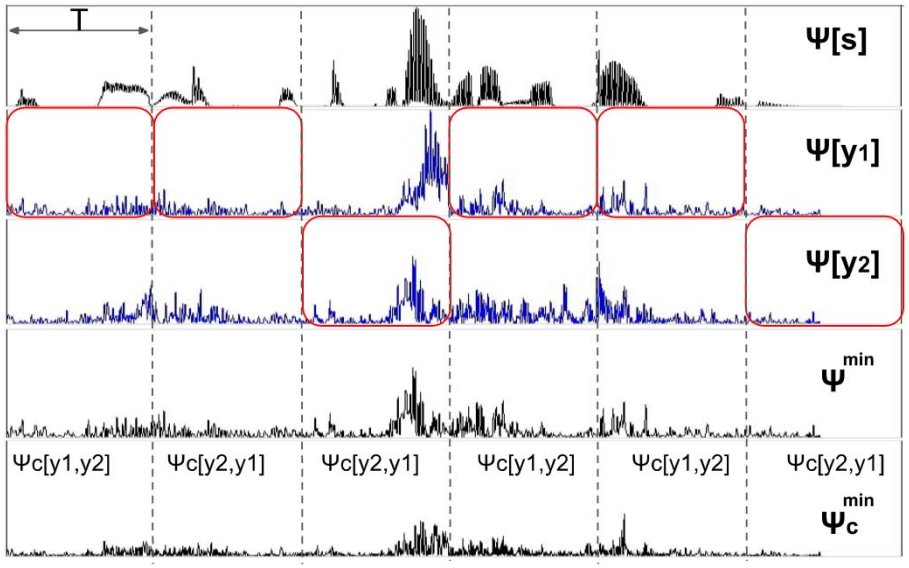
Word Error Rate

	fusion			**individually**	
conditions	early	intermediate	late	MFCCs	MIAs-MIFs
clean	22.27	**15.80**	15.85	16.56	21.18
reverb1	43.27	40.86	**40.08**	41.11	50.44
reverbR	45.44	43.58	**42.23**	44.52	55.12

[I. Rodomagoulakis, G. Potamianos and P. Maragos, EUSIPCO 2013]

Multichannel Estimation of Noisy Speech Energy

- Microphone array recordings: $y_m(t) = s(t) + u_m(t),\ m = 1, \dots, M$ mics
- Bandlimited components: $y_{mk}(t) = y_m(t) * g_k(t),\ k = 1, \dots, K$ freq. bands
- Correlation between recordings from adjacent microphones m, ℓ
  - Cross-Teager Energy [1]
    - $\Psi_c[y_{mk}, y_{\ell k}](t) = \dot{y}_{mk}(t)\dot{y}_{\ell k}(t) - y_{mk}(t)\ddot{y}_{\ell k}(t)$



- Noise is additive error on averaging:
  $\mathcal{E}\{\Psi_c[y_{mk}, y_{\ell k}]\} = \mathcal{E}\{\Psi[s_k]\} + error$ [2]
- **low cross energy → low error**
- Tracking minimum energy per band k
  - $(\hat{m}, \hat{\ell})$: $\Psi[\, y_{\hat{m}k}\,] < \Psi[\, y_{\hat{\ell}k}\,] < \cdots$
  - $\Psi_c^{min}(k) = \Psi_c[\, y_{\hat{m}k}, y_{\hat{\ell}k}]$
  - $\Psi^{min}(k) = \Psi[\, y_{\hat{m}k}\,]$

[1] P. Maragos & A. Potamianos, IEEE SPL 1995.
[2] S. Lefkimmiatis, P. Maragos & A. Katsamanis,, ICASSP 2008.

Multichannel, Multiband Demodulation (MMD)

- $\Psi[\, y_{mk}] \rightarrow \Psi_c^{min}(k) = \Psi_c[\, y_{\hat{m}k}, y_{\hat{\ell}k}]$

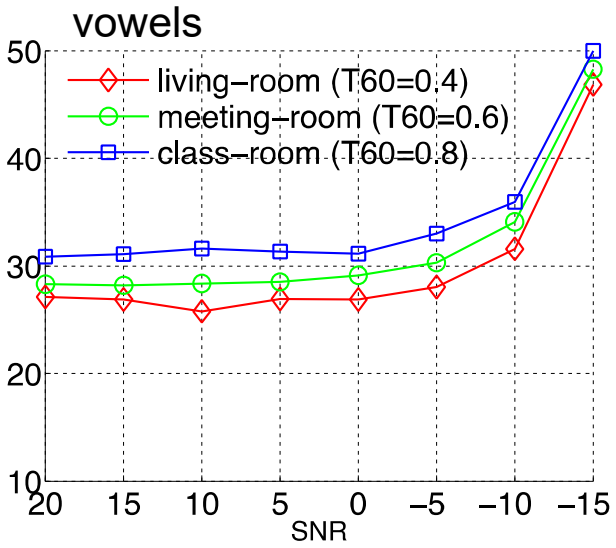
- $\omega_k(t) \approx \sqrt{\frac{\Psi_c[\, \dot{y}_{\hat{m}k}, \dot{y}_{\hat{\ell}k}]}{\Psi_c[\, y_{\hat{m}k}, y_{\hat{\ell}k}]}}$, $a_k(t) \approx \frac{\Psi_c[\, y_{\hat{m}k}, y_{\hat{\ell}k}]}{\sqrt{\Psi_c[\, \dot{y}_{\hat{m}k}, \dot{y}_{\hat{\ell}k}]}}$

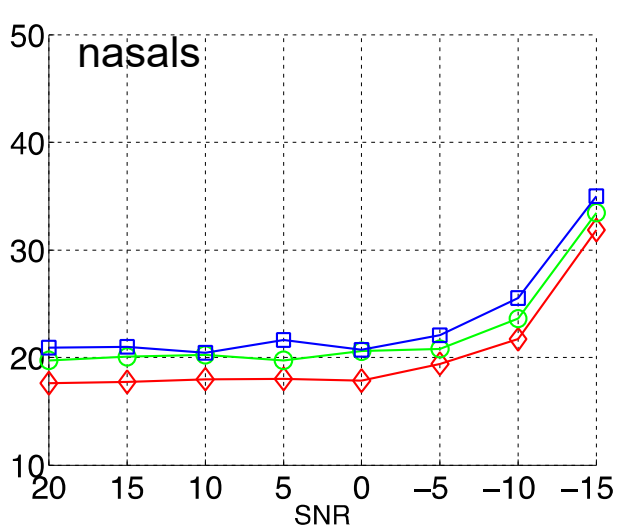
- Gabor-ESA:
  - $\Psi_c[\, y_{\hat{m}k}, y_{\hat{\ell}k}] = (\, y_{\hat{m}} * \dot{g}_k)(\, y_{\hat{\ell}} * \dot{g}_k) - (\, y_m * g_k)(\, y_{\hat{\ell}} * \ddot{g}_k)$
  - $\Psi_c[\dot{y}_{\hat{m}k}, \dot{y}_{\hat{\ell}k}] = (\, y_{\hat{m}} * \ddot{g}_k)(\, y_{\hat{\ell}} * \ddot{g}_k) - (\, y_m * \dot{g}_k)(\, y_{\hat{\ell}} * \dddot{g}_k)$

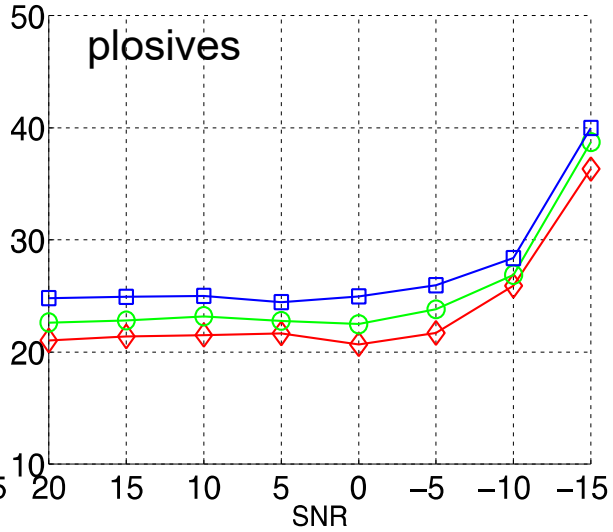
- **→ Improved estimations of $\omega_k(t)$, $a_k(t)$**

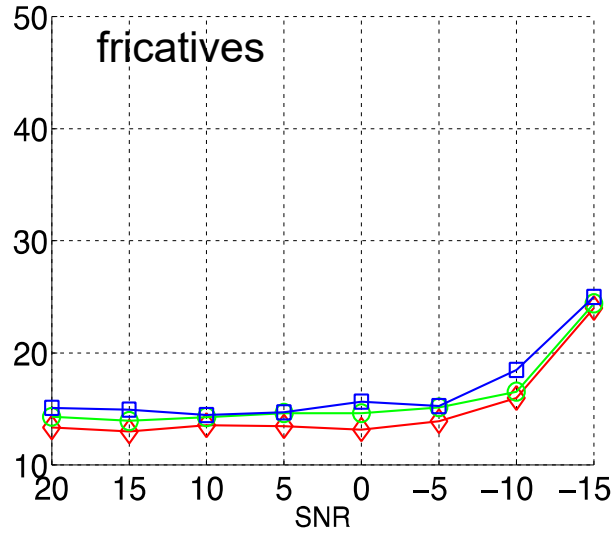
[I. Rodomagoulakis & P. Maragos, EUSIPCO 2017]

Single- vs Multi-channel Demodulation





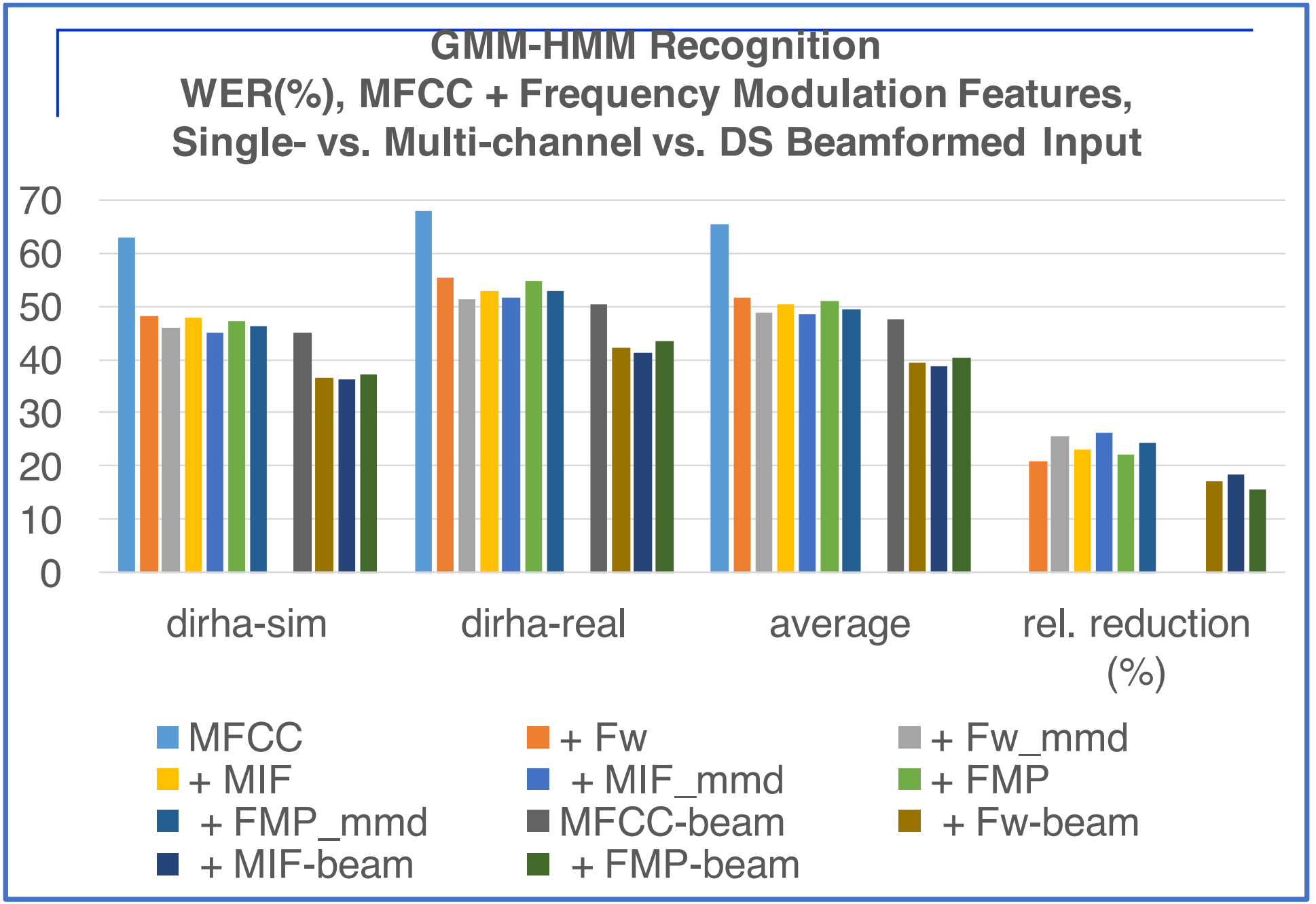




- TIMIT database (100 examples/phoneme)
- Simulations of small & medium room acoustics
  - Image Source Method + white noise ([-15…20] dB)
  - Speaker's moving in spiral trajectory 3m away from 3-mic linear array
- Demodulation error in estimating $\omega_k(t)$
  - Ground-truth from $s(t)$
  - Single-channel estimation from $y_2(t)$
  - Multi-channel estimation from $y_m(t),\ m = 1,2,3$
  - Average RMS error across bands
- Comparison
  - **Relative reduction (%) of RMS error**

DSR Experiments on Simulated and Real Data

- DIRHA-English corpus
  - Simulations of real-life scenarios of speech-based domestic control
  - Kitchen-Livingroom space with 21 condenser microphones arranged in distributed arrays
  - 15 hours of simulated multichannel training material
    - convolution of studio recordings with the apartment's RIRs mixed with typical domestic background noise.
  - 1000 utterances simulated (dirha-sim) and real (dirha-real) speech
- Experimental Framework
  - BeamformIt tool for state-of-the-art delay-and-sum beamforming
  - MMD: 12 Gabor filters with 70% overlap, Ψ_c^{min} energy
  - Kaldi baseline HMM-GMM recognizer with LDA, MLLT and fMLLR transformations


GMM-HMM Recognition
WER(%), MFCC + Frequency Modulation Features,
Single- vs. Multi-channel vs. DS Beamformed Input
70
60
50
40
30
20
10
0
dirha-sim
dirha-real
average
rel. reduction (%)
MFCC
+ Fw
+ Fw_mmd
+ MIF
+ MIF_mmd
+ FMP
+ FMP_mmd
MFCC-beam
+ Fw-beam
+ MIF-beam
+ FMP-beam

MODULATIONS FOR IMAGE & VIDEO PROCESSING

AM-FM Image Modulations and Image Segmentation

Ref:

I. Kokkinos, G. Evangelopoulos & P. Maragos, "*Texture Analysis & Segmentation Using Modulation Features, Generative Models, and Weighted Curve Evolution*", IEEE T-PAMI Jan. 2009.

AM-FM Texture Model

- Locally narrowband image texture (Bovik etal 1992, Havlicek etal. 2000)

$$f(x,y) = a(x,y)\cdot\cos\left[\phi(x,y)\right], \qquad \nabla\phi(x,y) = \vec{\omega}(x,y)$$

  - analogies between AM-FM and Y.Meyer's oscillating functions for texture

- Inst. Amplitude & Frequency estimation (Maragos & Bovik , JOSA1995):
  - Multiband Gabor filtering
  - **2D Energy Operator** $\Psi(f) = \|\nabla f\|^2 - f\nabla^2 f$
  - Demodulation via the **Energy Separation Algorithm (ESA):**

$$\frac{\Psi(f)}{\sqrt{\Psi(\partial f/\partial x)+\Psi(\partial f/\partial y)}} \approx |a(x,y)|$$

$$\sqrt{\Psi(\partial f/\partial x)/\Psi(f)} \approx |\omega_1(x,y)|, \qquad \sqrt{\Psi(\partial f/\partial y)/\Psi(f)} \approx |\omega_2(x,y)|$$

Modulation Features for Texture Analysis

- *Dominant Components Analysis* (DCA) chooses at each pixel the most prominent channel, *j*

$$a(x,y) = \left|a_j(x,y)\right|, \qquad |\vec{\omega}(x,y)| = |\vec{\omega}_j(x,y)|$$

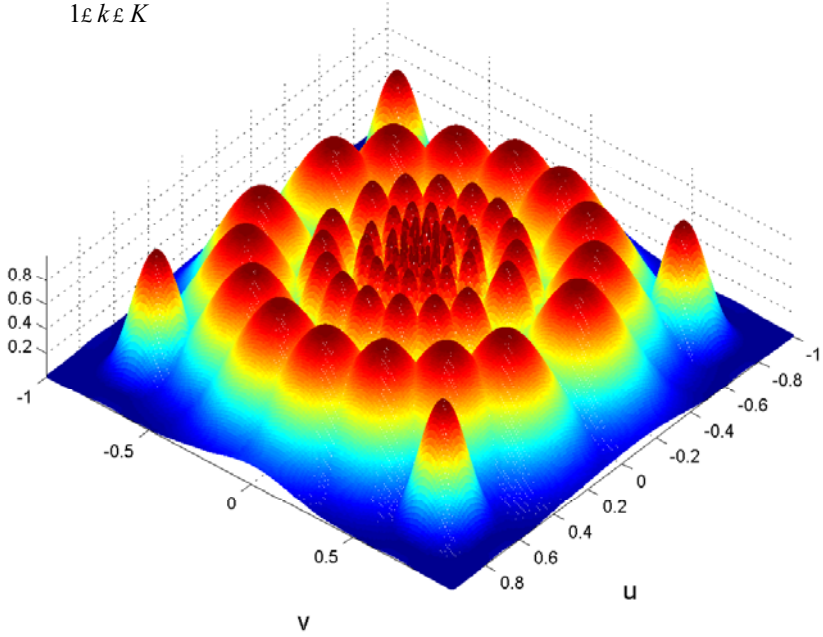
- Maximize criterion for choosing $j = \arg\max_{1 \le k \le K} \{G_k\}$, among K channels

Amplitude-DCA

$$\Gamma_k(x,y) = \frac{\left|a_k(x,y)\right|}{\max_{\vec{\omega}} |H_k(\vec{\omega})|}$$

Teager *Energy*-DCA

$$\Gamma_k(x,y) = \Psi\left[(I * h_k)(x,y)\right]$$



- Using a single channel amounts to locally modeling the texture with a Gabor-like '*texton*' whose characteristics are described by the DCA components.

Modulation Feature Extraction Examples

Real Modulation Parameters

Synthetic AM-FM

$a(x,y)$

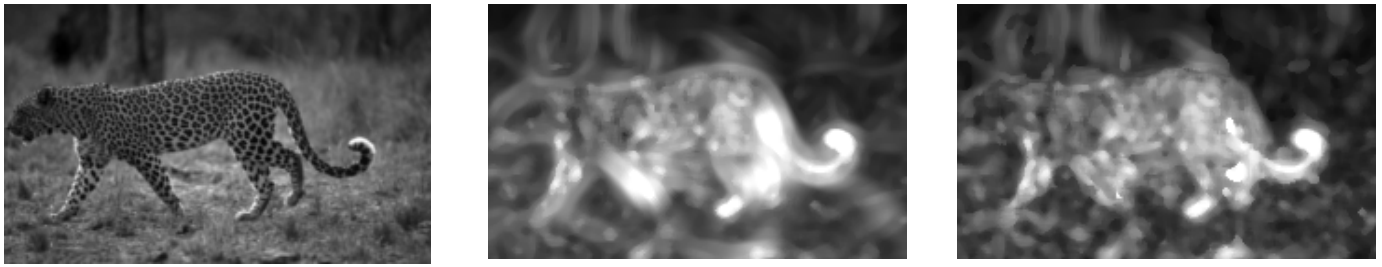
$\omega_1(x,y)$

$\omega_2(x,y)$

DCA Estimated

A-DCA

E-DCA



Amplitude

Frequency Magnitude

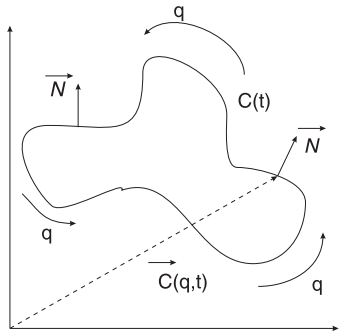
Unsupervised Variational Texture Segmentation

- Functional expressing segmentation cost (*Region Competition*):

$$J[C,\{\theta_i\}] = \sum_{i=1}^{M} \frac{\mu}{2} \int_{C_i} ds - \iint_{R_i} \log\left(P(I;\theta_i)\right) \qquad C = \{C_1,\ldots,C_M\}$$

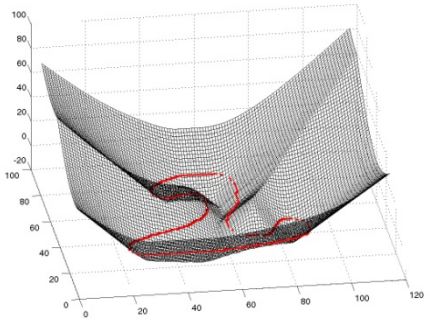
- *Euler-Lagrange* equations:

$$\frac{\partial C_i}{\partial t} = -\mu\kappa\vec{N} + \log\frac{P\left(I;\theta_i\right)}{P\left(I;\theta_j\right)}\vec{N}$$



- *Level Set* Implementation & Edge-based terms (*Geodesic Active Regions*):

$$\frac{\partial C_i}{\partial t} = \lambda \log\frac{P\left(I;\theta_i\right)}{P\left(I;\theta_j\right)}\vec{N} - (1-\lambda)\left[\, g(\mathrm{I})k\vec{N} + \left(\nabla g(\mathrm{I})\cdot\vec{N}\right)\vec{N}\right]$$



- *Active Contours without Edges, Statistical approach to Snakes*

Refs: Zhu & Yuille, T-*PAMI* 96, Paragios & Deriche *IJCV* 02, Vese & Chan T-*IP* 01, Yezzi, Chai & Willsky *ICCV* 99

2D Gabor ESA

- 2D energy operator with Gabor bandpass filtering

$$f(x,y) = I(x,y) * h(x,y)$$

- Gabor Energy Operator

$$\Psi(f) = \Psi(I*h) = \|I*\nabla h\|^2 - (I*h)(I*\nabla^2 h)$$

- Differential operators are replaced by derivatives of Gabor

- Estimation of inst. amplitude and frequency by ESA

$$\Psi(f_x = I*h_x) = \|I*\nabla h_x\|^2 - (I*h_x)(I*\nabla^2 h_x), \quad \Psi(f_y = I*h_y)$$

- 2D Gabor ESA: need seven Gabor differential formulae

$$(h_x, h_y, h_{xx}, h_{yy}, h_{xy}, \nabla^2 h_x, \nabla^2 h_y)$$

Regularized ESA

- Reduce complexity of applying Gabor ESA to all filters

- Bandpass Image $f_k(x,y) = I(x,y) * h_k(x,y)$

- Regularized Energy Operator

$$\Psi_\sigma(f_k) = \|f_k * \nabla G_\sigma\|^2 - f_k\left(f_k * \nabla^2 G_\sigma\right)$$

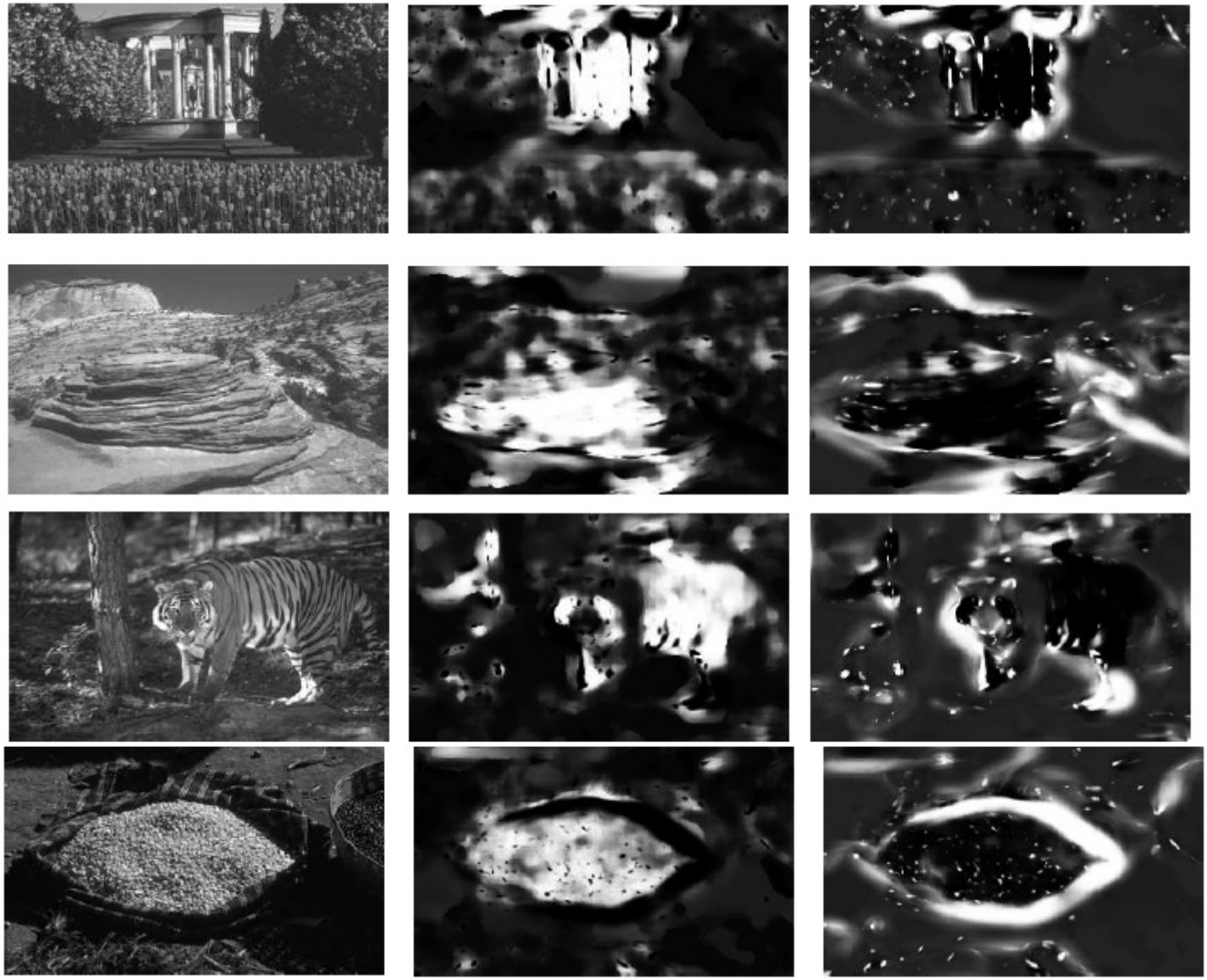
REO needs three convolutions of f_k with $\partial/\partial x, \partial/\partial y, \nabla^2$ of the Gaussian

- Apply Regularized ESA to each channel

$$\Psi_\sigma(\partial f_k/\partial x) = \|f_k * \nabla(\partial_x G_\sigma)\|^2 - f_k\left[f_k * \nabla^2(\partial_x G_\sigma)\right], \qquad \Psi_\sigma(\partial f_k/\partial y)$$

Model-based Cue Probabilities

Intensity | *P(texture)* | *P(edge)*



Cue Integration for Region Competition

- How can we introduce the confidence measures in the evolution?

- Modified RC with probability assignments to features

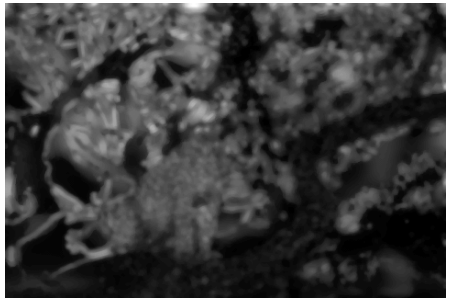
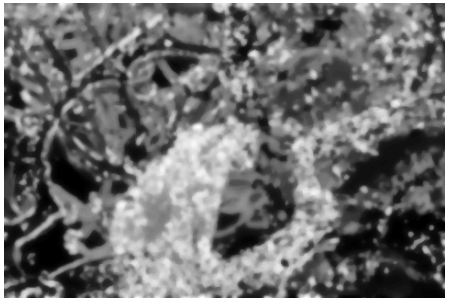
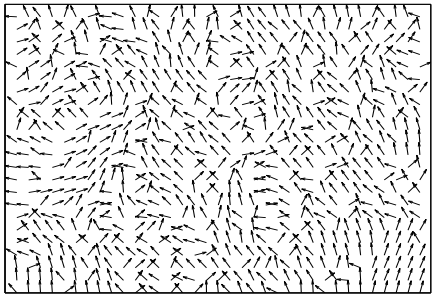
w_c : cue weight, w_e : edge.

$$\frac{\partial C_i}{\partial t} = \sum_{c \in T,S} w_c \log \frac{P^c\left(F_c;\theta_i\right)}{P^c\left(F_c;\theta_j\right)} \vec{N} - w_e\left(\nabla g \cdot \vec{N}\right)\vec{N} - gk\vec{N}$$

Features and Segmentation

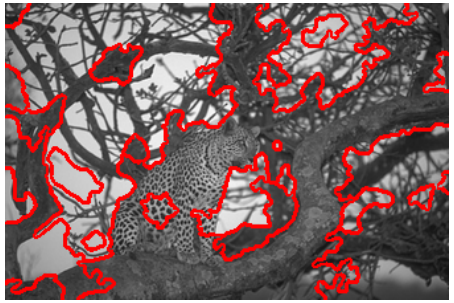
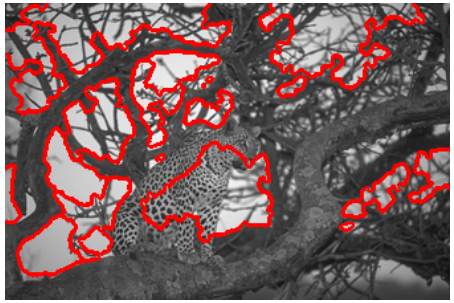
Features for Segmentation: Intensity, Amplitude, Freq. Magnitude, Freq. Orientation

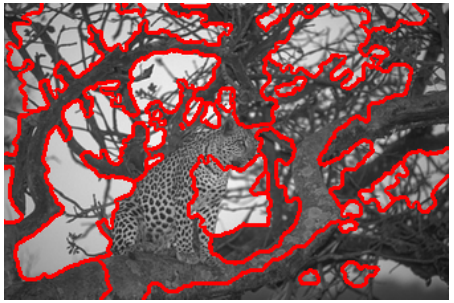
I a $|\vec{\omega}|$ $\angle\vec{\omega}$

Segmentation Results - Comparisons

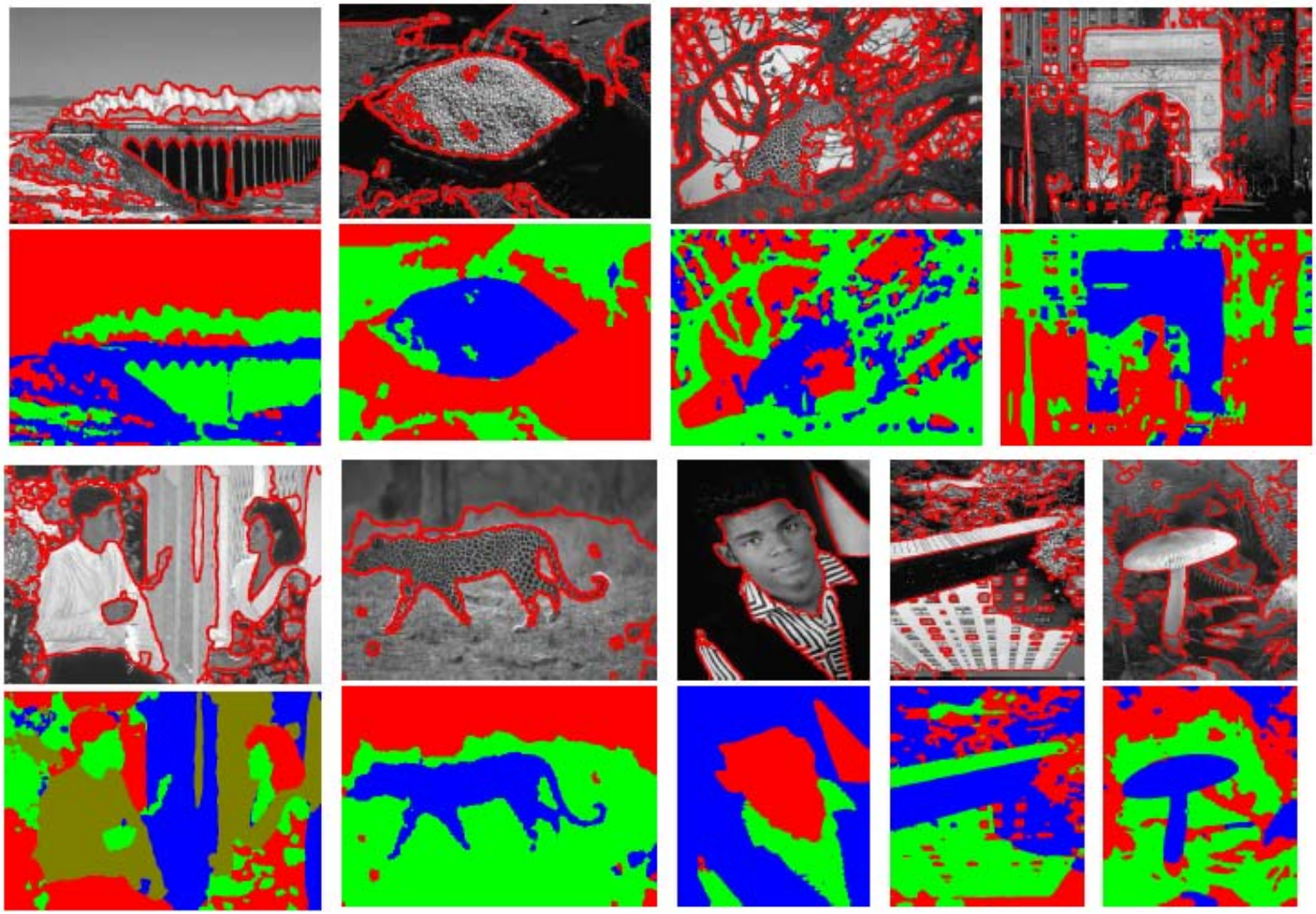
RC-GAR ***Weighted RC-GAR*** ***Weighted RC-GAR*** ***RC-GAR***



Feats: $[a,|\vec{\omega}|,\angle\vec{\omega},I]^{\mathrm{T}}$ $[a,|\vec{\omega}|,I]^{\mathrm{T}}$ $[a,|\vec{\omega}|,\angle\vec{\omega},I]^{\mathrm{T}}$ ***Baseline Diffusion features***

Unsupervis. Segmentation w. Weighted Curve Evolution



Spatio-Temporal Modulations and Video Action Recognition

C. Georgakis, P. Maragos, G. Evangelopoulos, and D. Dimitriadis, Proc. ICIP 2012.

Overview of the DCA3D Detector

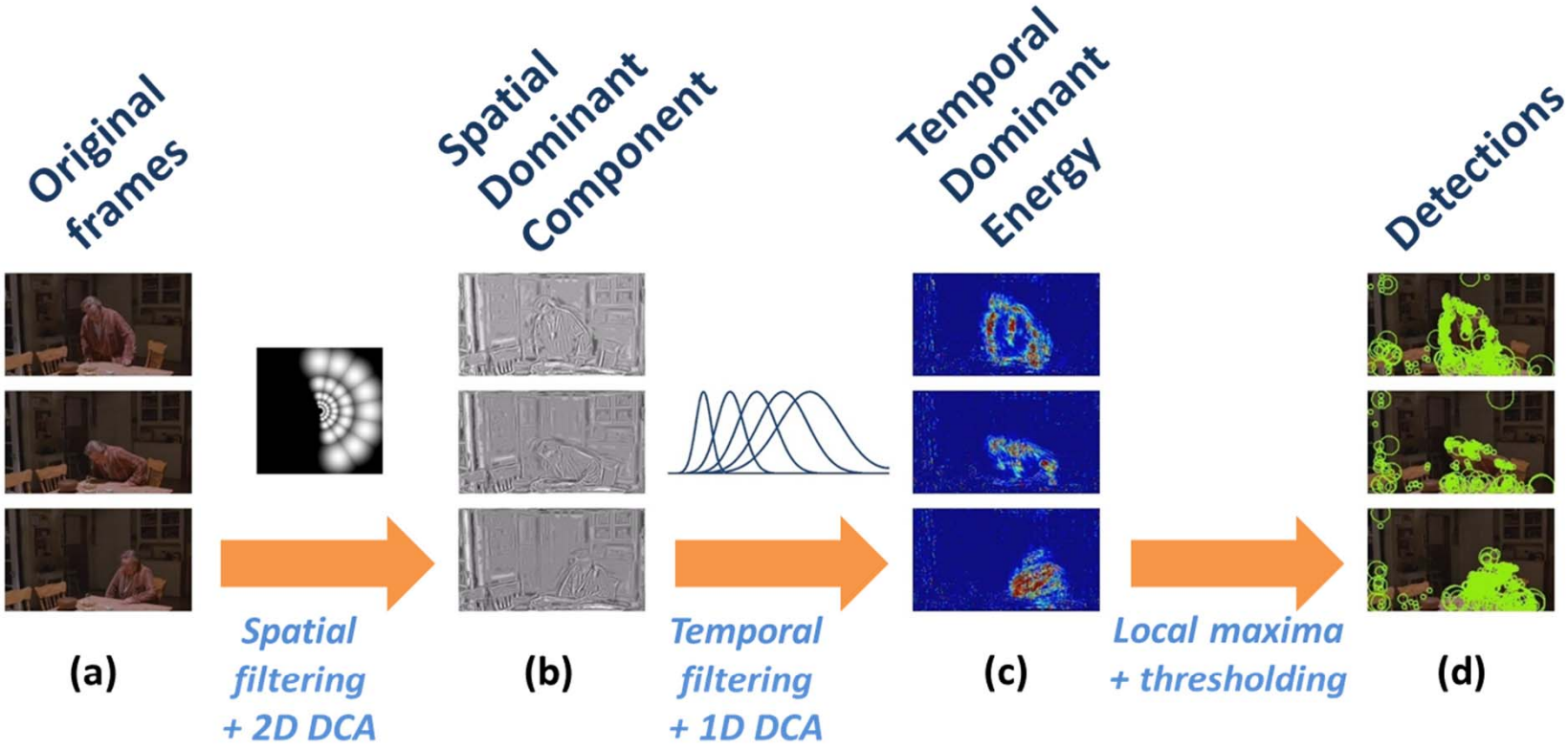


Figure 1. Discrete stages followed by DCA3D detector to yield final detections. **(a)** Original frames from a *SitDown* action clip, **(b)** DCA-synthesized images composed of locally dominant channel filtered outputs, **(c)** Energy values corresponding to 1D DCA dominant temporal channels, **(d)** Final detections as the prominent local maxima

Example

Spatial Energy-based DCA emphasizes the prominent texture variations and meaningful object boundary information

Original Color Image

Spatial Dominant Component

Spatial Dominant Energy

Figure 2. Spatial Energy-based DCA on a wideband image of complex structure. **(a)** Original color image, **(b)** Bandpass image values $(I * g_i)(x, y)$ from the dominant components, **(c)** Energy values corresponding to max-energy dominant channels $i(x, y)$

Example

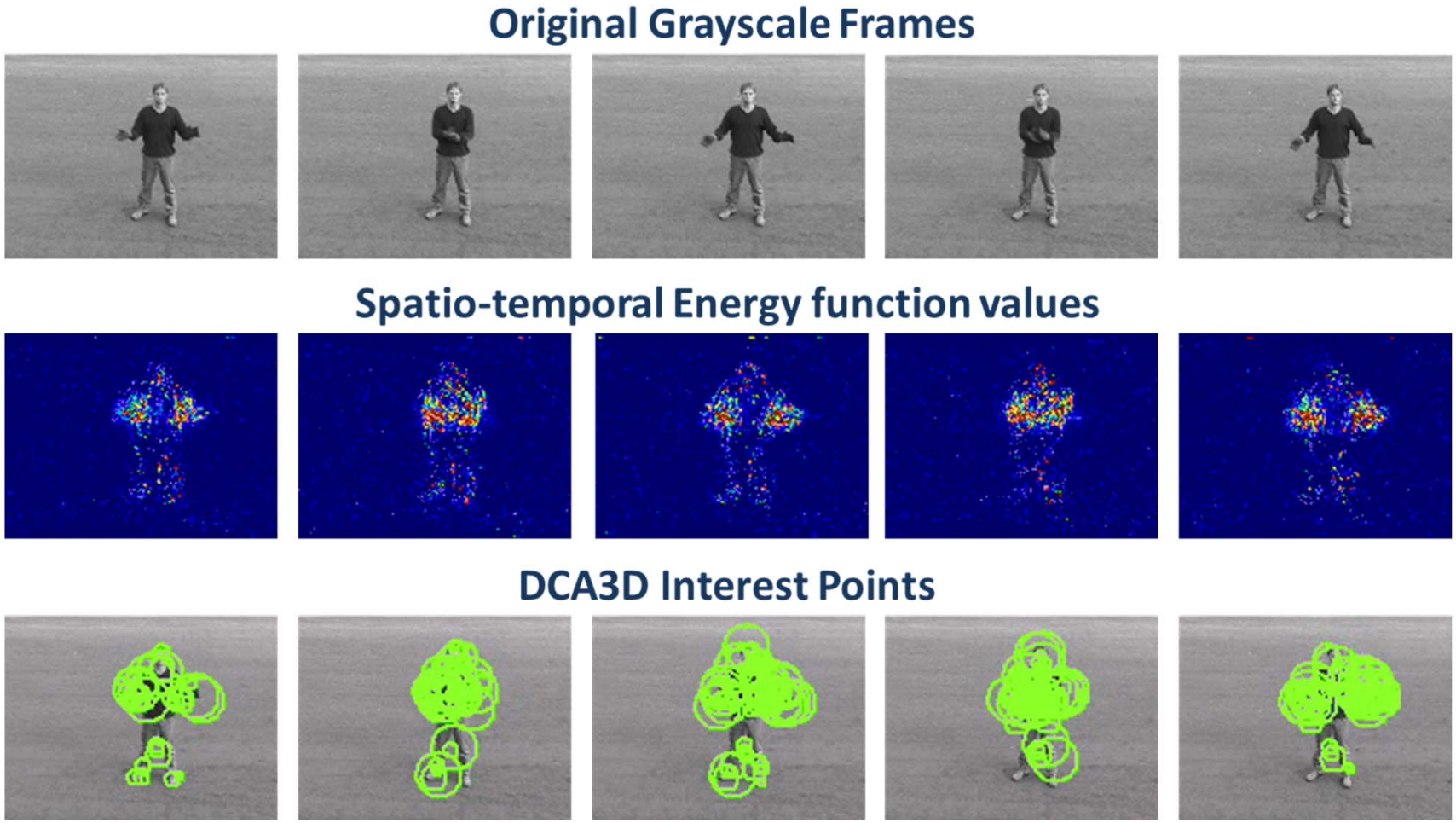


Figure 3. DCA3D detector outputs on five frames from a *handclapping* action clip. **Top**: Original grayscale action instances, **Middle**: DCA3D energy function values, **Bottom**: Interest points corresponding to globally thresholded local maxima

Conclusions

- AM and FM are fundamental phenomena in sound (speech, music, general audio) and other oscillatory signals (e.g. image textures, or space-time patterns in videos).
- Energy operators are related to physics, very simple/fast to compute, have excellent time-resolution and can demodulate efficiently AM-FM.
- Applications in speech, music, image/video processing & recognition.
- Open analytic problems: Optimality of EO, Variational approaches

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