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

Part 3
Audio Processing and Saliency

Athanasia Zlatintsi

Tutorial at IEEE International Conference on Acoustics, Speech and Signal Processing 2017,
New Orleans, USA, March 5, 2017

Part 3: Outline

- Auditory Saliency and Attention
- State-of-the-art on Auditory Saliency
- Audio Processing and Saliency Computation based on
  - AM-FM and ESA demodulation
  - Teager Energy Operator
- Application: Audio Summarization

Auditory Saliency and Attention

Auditory Information Processing

- Our brain is capable of parsing information in the environment by using various cognitive processes
  - regardless various prominent distractors (known as the ‘cocktail party problem’)

- Such processes allow us to navigate to the soundscape, focus on interesting conversations, enjoy the background music and be alert to any **salient** sound events, i.e. when someone is calling us

[T. Darrell, J. W. Fisher III, P. Viola and W. Freeman, *Audio-visual Segmentation and “The Cocktail Party Effect”*, ICMI 2000]
[C. Alain and L. Bernstein, *From sounds to meaning: the role of attention during auditory scene analysis*, Curr. Opin. Otolaryngol. Head Neck Surg. 16, 2008]

Auditory Attention

Adjusted by:

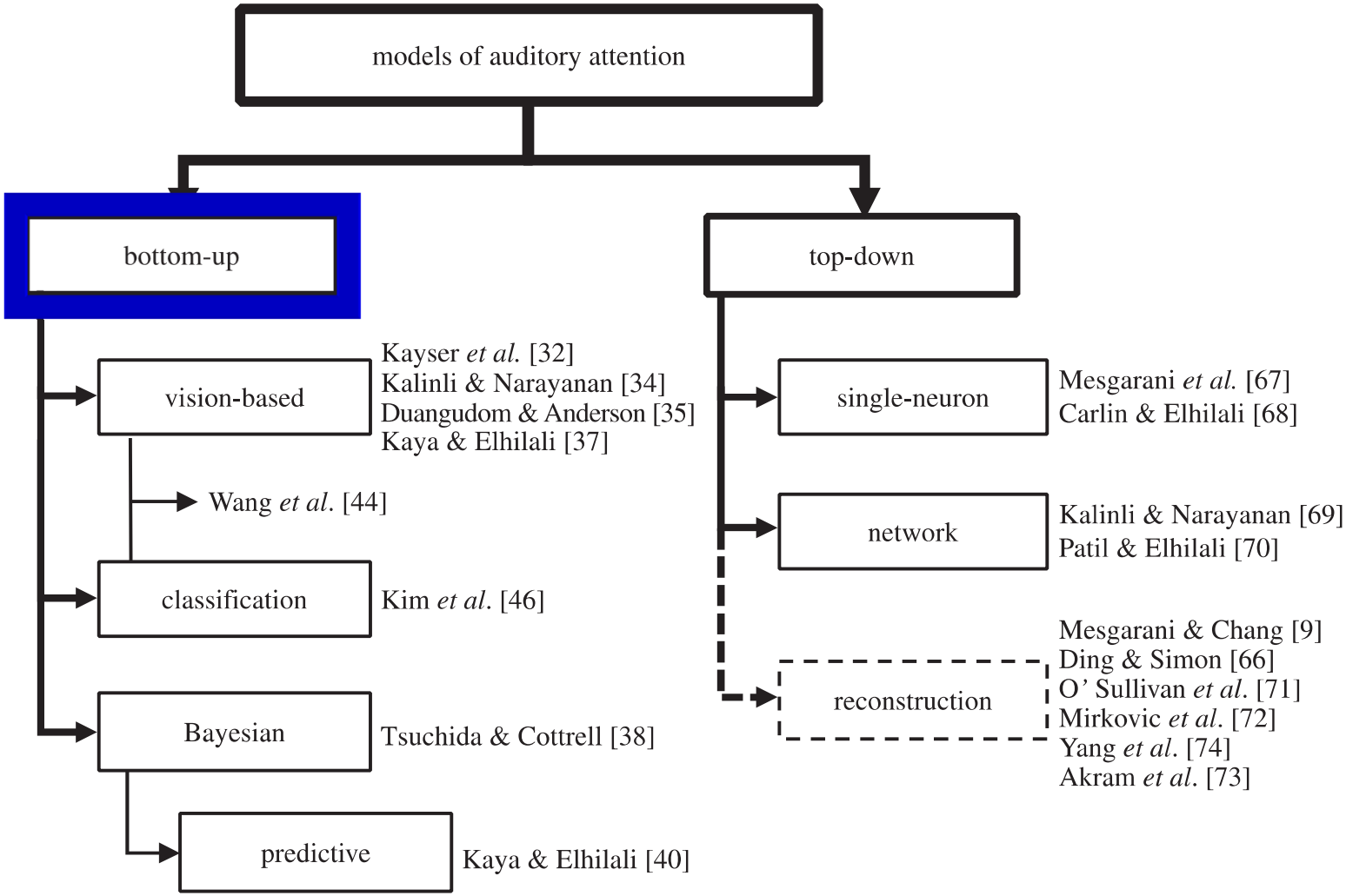
- 'Bottom-up' sensory-driven task-independent factors (automatic, 'unconscious', stimulus driven, little or no attention)
- 'Top-down' task-dependent goals, expectations and learned schemas ('conscious', effortful, selective, memory dependent)
- It acts as a selection process that focus both sensory and cognitive resources on the most relevant events in the soundscape, i.e.,
  - a sudden loud explosion
  - or a task at hand, e.g., listen to announcements in a busy environment

Auditory Saliency

- Quality to stand out relatively to the surrounding soundscape
  - Salient stimuli are able to attract our attention and are easier to detect
- Describes the potential influence of a stimulus on our perception and behavior
- Key attentional mechanism facilitating learning and survival
- Complements the frequently studied processes of attention and detection
- Introduces a qualitative description of those stimulus properties relevant for these processes

State-of-the-Art in Auditory Saliency

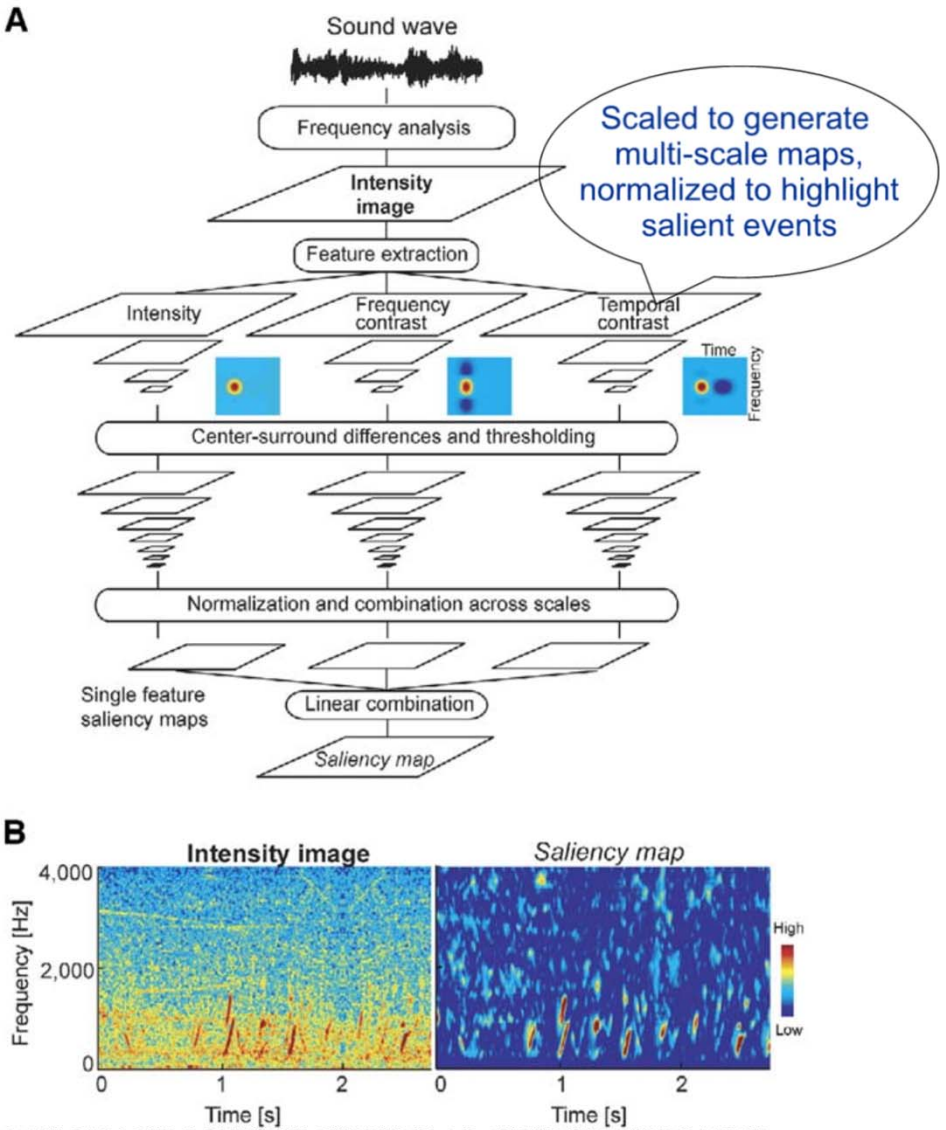
Classification of Models of Auditory Attention



[E.M. Kaya and M. Elhilali, *Modelling auditory attention*. Phil. Trans. R. Soc., 2016]

An Auditory Saliency Map

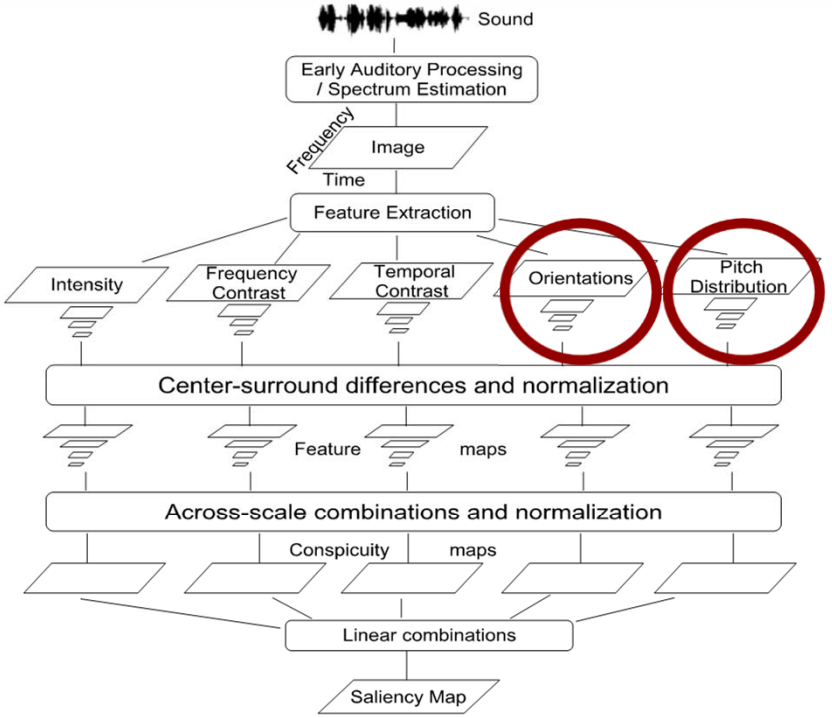
- Time - frequency representation of the sound waveform as an “auditory image”
- Spectro-temporal features such as intensity, frequency & temporal contrast
- The model was able to match both human and monkey behavioral responses for the detection of salient sounds in noisy scenes
- Demonstrated that saliency processing in the brain may share commonalities across sensory modalities
- Provided a guide for the design of psychoacoustical experiments to probe auditory bottom-up attention in humans



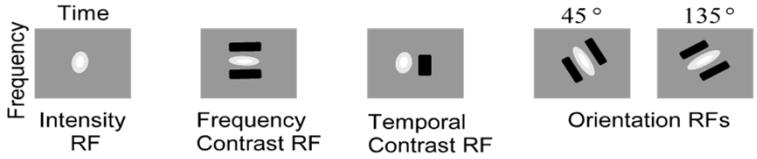
[C. Kayser, C.I. Petkov, M. Lippert and N.K. Logothetis, *Mechanisms for allocating auditory attention: an auditory saliency map*, Curr. Biol., 2005]

Saliency-based Auditory Attention for Prominent Syllable Detection

- Extension of Kayser's model, incorporating more relevant auditory cues
- Multi-scale auditory features, including also pitch and orientation along time and frequency
- Improved contrast computation in order to derive feature maps, making them more robust to noise



Adapted Auditory Saliency Map

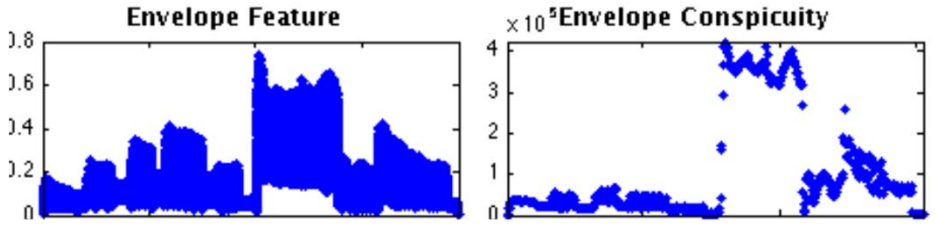


Receptive filters

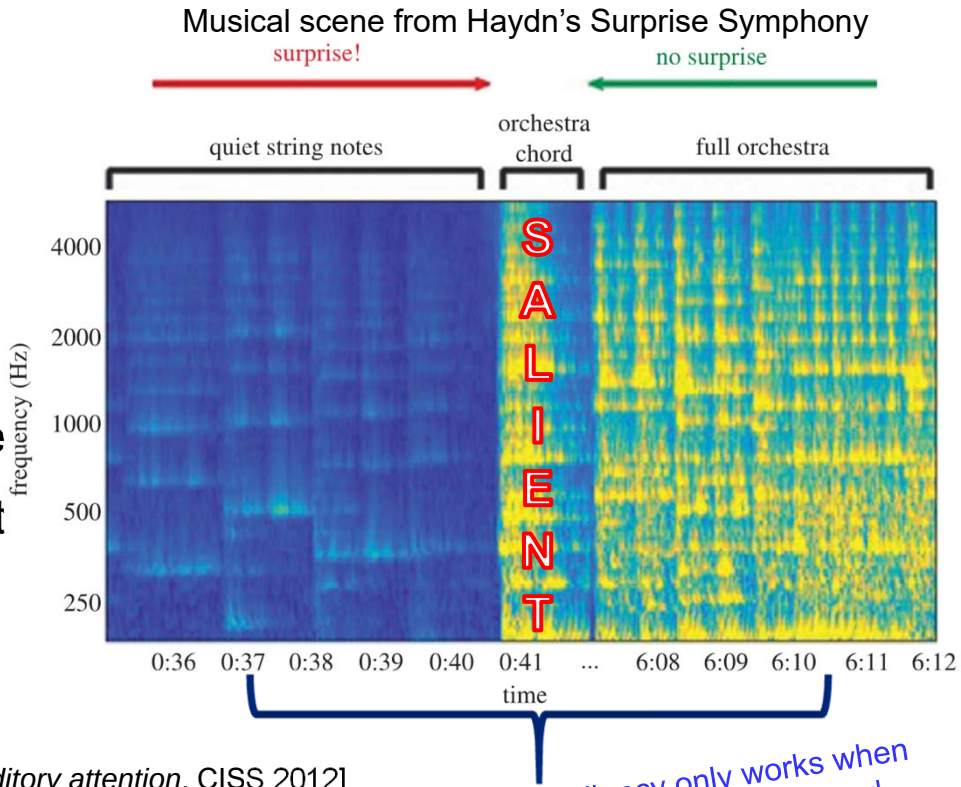
[O. Kalinli and S. Narayanan, *A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech*, Interspeech 2007]

Temporal Saliency Map

- Based on the fact that sound is actually a naturally temporally evolving entity
- Saliency is measured by how much an event differs from its surrounding, thus. sounds preceding in time
- Auditory scene is treated as single dimensional temporal input (at all times), rather than as an image
- Employing perceptual properties of sound, i.e., loudness, pitch and timbre
- Feature analysis over time to highlight their dynamic quality before normalizing and integrating across feature maps

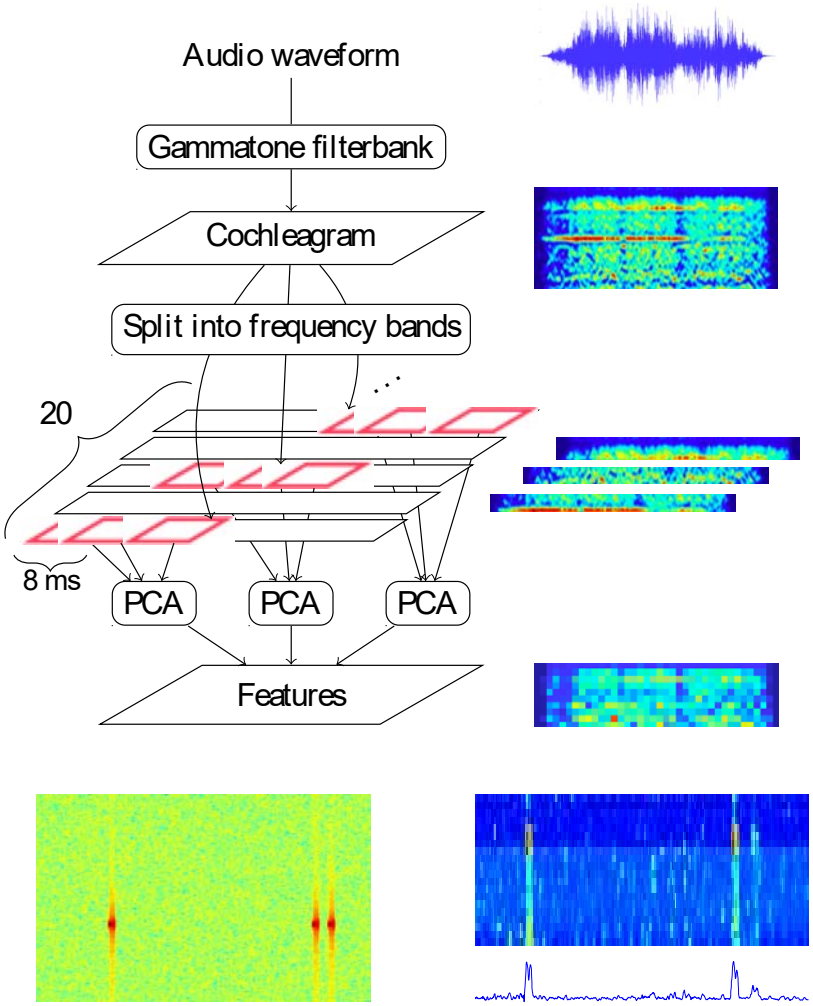


Differences in saliency selection between the temporal saliency model and Kayser's saliency model



[E.M. Kaya and M. Elhilali, *A temporal saliency map for modeling auditory attention*, CISS 2012]
[E.M. Kaya and M. Elhilali, *Modelling auditory attention*. Phil. Trans. R. Soc., 2016]

Statistical-based Approach



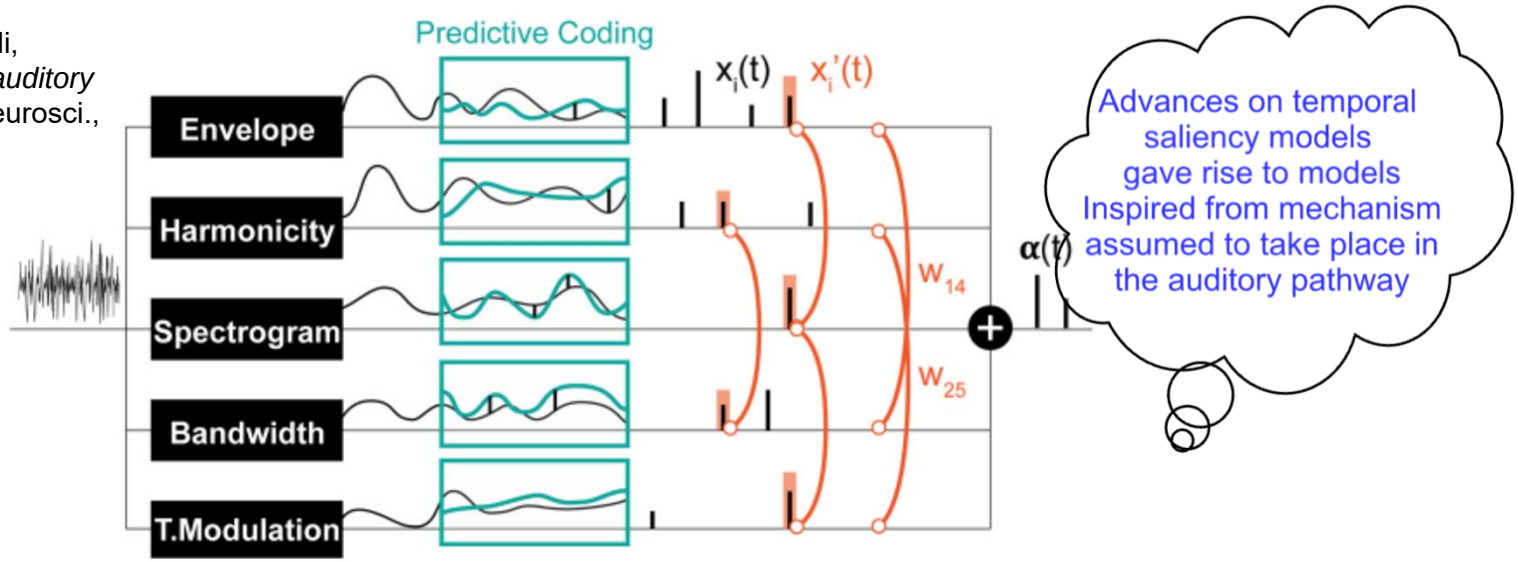
Spectrogram and saliency map (saliency values summed over frequency axis) for single and paired tones.

- Statistical approach adapted from vision*
- Combination of long-term statistics computed using natural sounds with short-term, temporally local, rapidly changing statistics of the incoming sound
- A sound is flagged as **salient** if it is determined to be unusual relative to learned statistics
- Cochleogram was used instead of a spectrogram (for computational efficiency) and PCA for dimensionality reduction

[T. Tsuchida and G. Cottrell, *Auditory saliency using natural statistics*, Society for Neuroscience Meeting, 2012]
[*L. Zhang, M.H. Tong, T.K. Marks, H. Shan and G.W. Cottrell, *SUN: A Bayesian framework for saliency using natural statistics,* J. Vis., 2008]

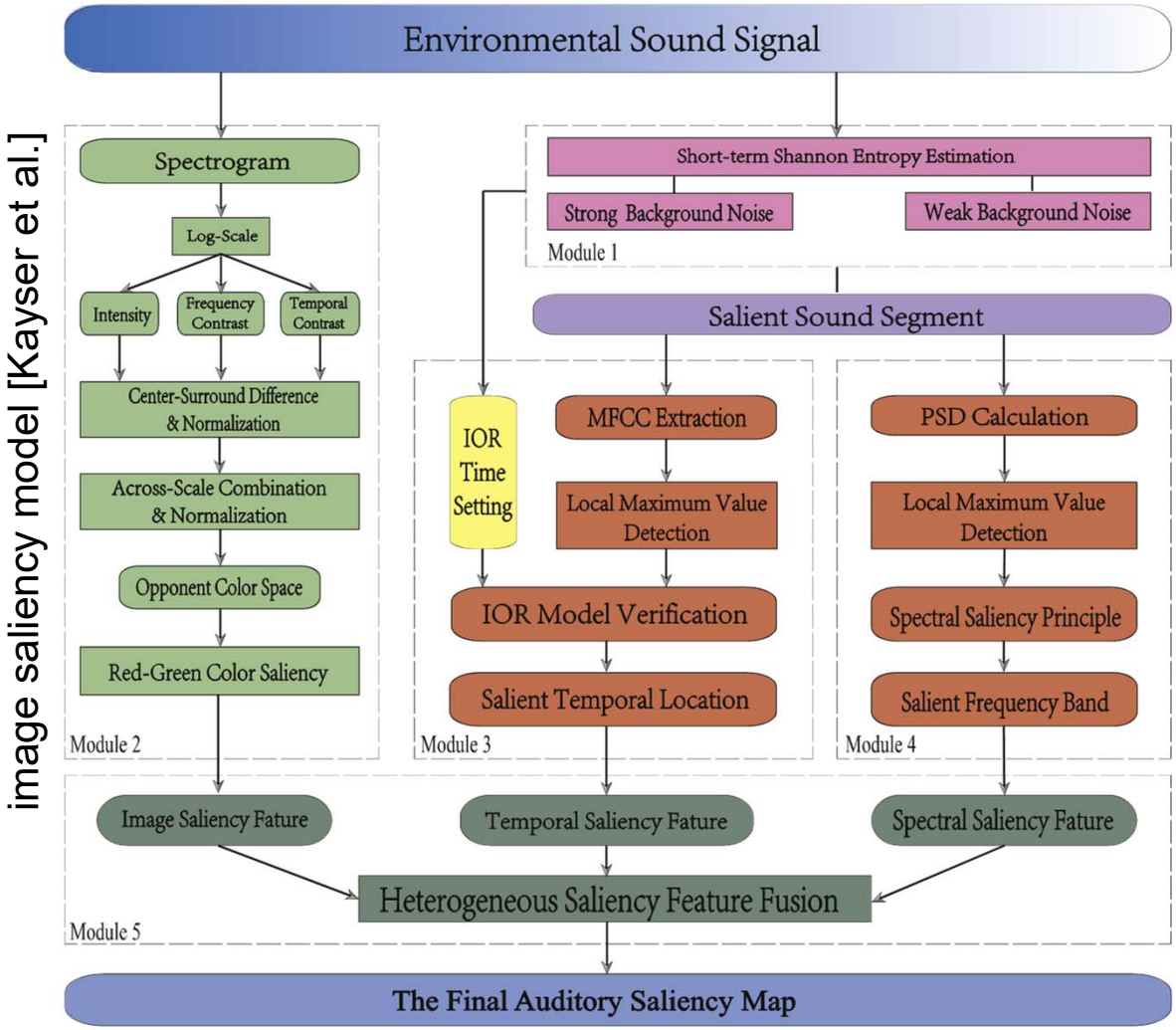
Predictive Coding

[E.M. Kaya and M. Elhilali, *Investigating bottom-up auditory attention*, Front. Hum. Neurosci., 2014]



- Emphasis on the role of processing events over time and shaping neural responses of current sounds based on their preceding context
- Mapping of the acoustic waveform onto a high-dimensional auditory space, encoding perceptual loudness, pitch and timbre of the incoming sound, building upon evolving temporal features
- Collect feature statistics over time and make predictions about future sensory inputs
- Regularities are tracked, and deviations from regularities are flagged as **salient**
- Nonlinear interaction across features, using asymmetrical weights between pairwise features

Bio-inspired Saliency Detection



Composite system:

- Shannon Entropy for global saliency: measure the sound's informational value
- MFCCs for acoustic saliency: temporal analysis of sound characteristics (using IOR model for saliency verification)
- Spectral saliency: analysis of the power spectral density of the stimulus
- Kayser's image model
- Robustness to saliency estimation especially in noisy environements

[J. Wang, K. Zhang, K. Madani and C. Sabourin, *Salient environmental sound detection framework for machine awareness*, Neurocomp. 2015]

Audio Processing and Saliency Computation

Motivation for our Saliency Computational Method

- Bottom up attention is based on sensory temporal & spectral cues of the acoustical stimuli
  - i.e., loudness, frequency, direction and their temporal or spatial contrast

In our case:

- Saliency computation is approached as a problem of assigning a measure of interest to audio frames, based on spectro-temporal cues
- The importance of **amplitude** and **frequency** changes has motivated a variety of studies
- Amplitude and frequency modulations are related to temporal acoustic micro-properties of sounds
  - Useful for auditory grouping, recognition of audio sources/events

[J. B. Fritz et al., *Auditory attention–focusing the searchlight on sound,* Current opinion in neurobiology 2007.]
[M. Elhilali et al., *Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene,* PLoS biology 2009]
[E. R. Hafter et. al., *Auditory Attention and Filters*, Auditory Perception of Sound Sources, 2007]

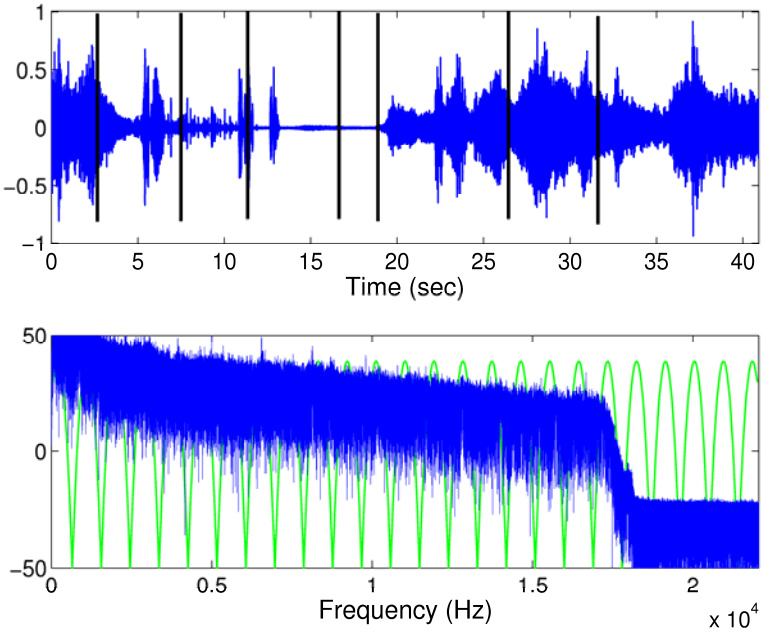
AM-FM Model

- AM-FM model for audio signals

  (speech, music environmental sounds)

$$s(t) = \sum_{i=1}^{K} \alpha_i(t) \cos(\varphi_i(t))$$

instantaneous amplitude — phase



Modeling of each resonance component as an amplitude and frequency modulated sinusoid (AM-FM signal), and the whole signal as a sum of such AM-FM components

- Nonlinear energy tracking
  - Teager-Kaiser energy operator
  - ESA demodulation
- Multiband filtering with:
  - K Gabor filters h_k, narrowband components

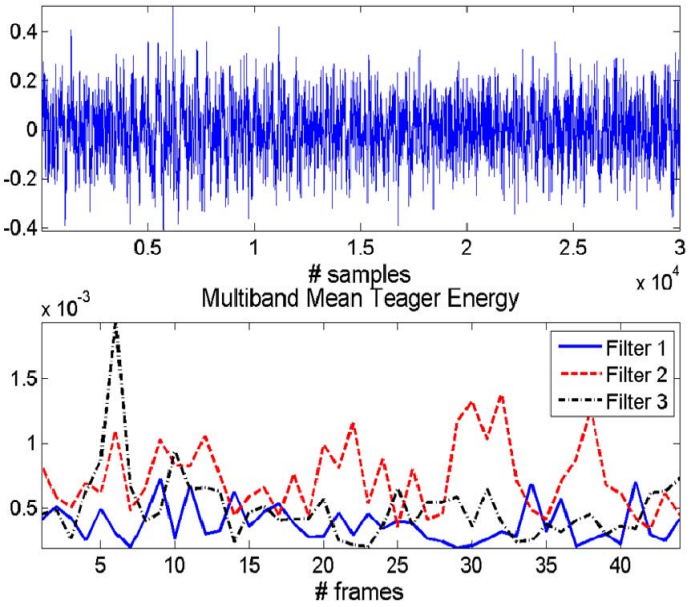
[P. Maragos, J.F. Kaiser and T.F. Quatieri, *Energy Separation in Signal Modulations with Application to Speech Analysis,* IEEE Trans. on Signal Process., 1993]

Teager-Kaiser Operator

Teager-Kaiser Energy Operator:

$$\Psi[x] = \dot{x}^2 - x\ddot{x}, \text{ where } \dot{x} = dx/dt$$

- For energy estimation and AM-FM demodulation, using ESA (Energy Separation Algorithm)
- Captures amplitude and frequency variation information
- Ψ can detect robustly & discriminate various acoustic events due to its sharp time resolution and lowpass behavior
- Important for auditory scene analysis
- Robust to noise compared to the squared energy operator
- Multiband TECC (Teager Energy Cepstrum Coefficients) successful in speech recognition

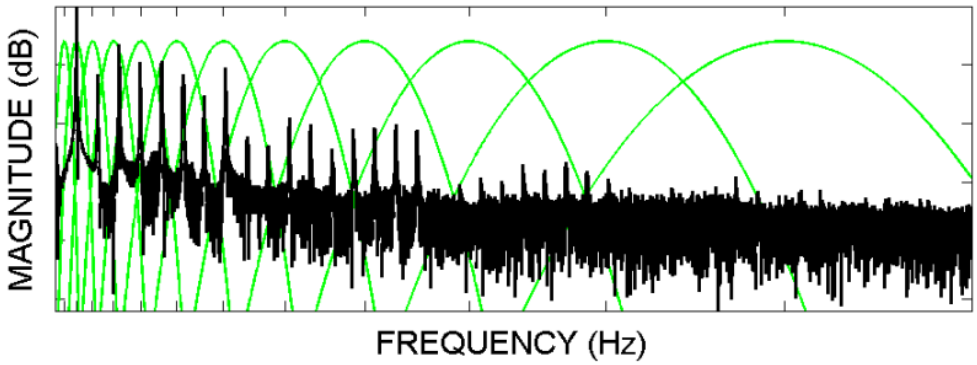


[J.F. Kaiser, *On a simple algorithm to calculate the energy of a signal*, ICASSP 1990]
[D. Dimitriadis, P. Maragos and A. Potamianos, *On the effects of filterbank design and energy computation on robust speech recognition*, TASLP 2011]

Multiband Filtering

Teager Energy: only meaningful in narrowband signals

- **Multiband filtering** of the signal with **Gabor filters**
  - Gabor filtering for isolation of narrowband components
  - Gabor filters: exhibit good joint time-frequency resolution



ESA Demodulation

ESA demodulation

(Energy Separation Algorithm)

$$|\alpha(t)| \approx \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \qquad f(t) \approx \frac{1}{2\pi}\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}}$$

- **Dominant modulation features**

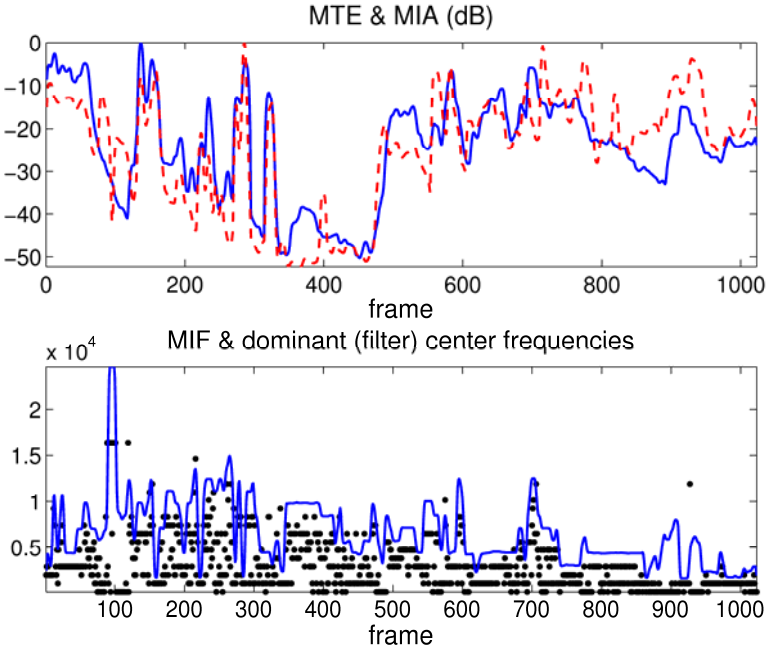
k-th filter response

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

Teager-Kaiser Energy Operator

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

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



MIA: parameterizing the resonance amplitudes and captures part of the nonlinear behavior of the signal

MIF: models the time varying frequency of the resonance and provides information about the signal's fine structure

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

Audio Analysis: Fusion and Saliency

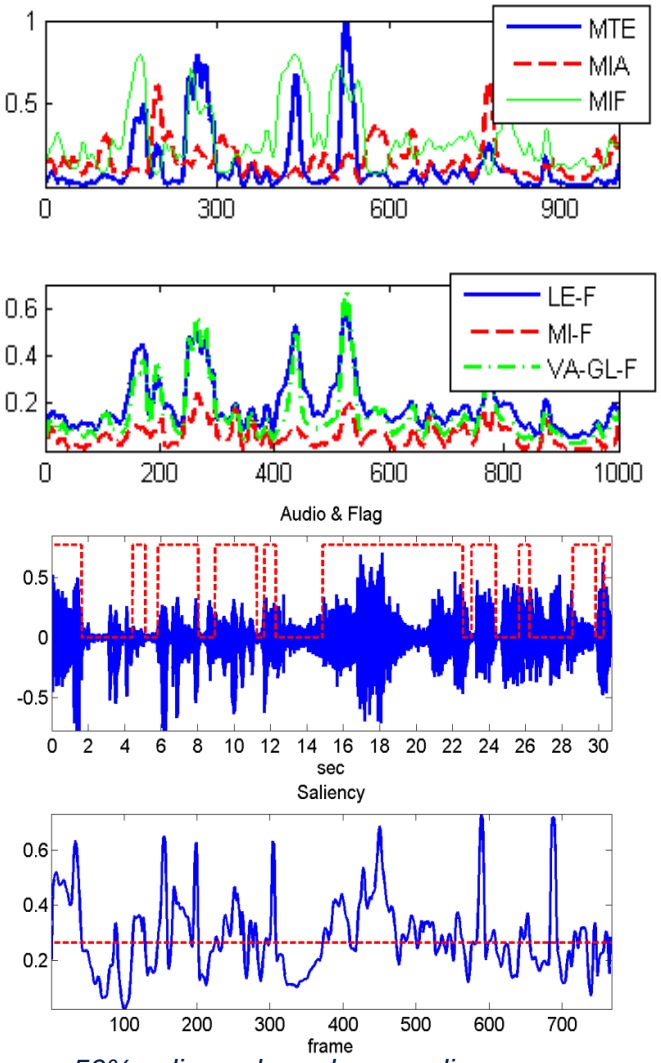
- Audio saliency cues
  - quantify multifrequency modulations
  - extracted through nonlinear operators
  - energy tracking
- 3D Feature vector formation

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

- Audio saliency curve (cue fusion) using:
  - Linear with constant weights

$$S_{\mathrm{a}}[m]=(w_1\mathrm{MTE}+w_2\mathrm{MIA}+w_3\mathrm{MIF})[m]$$

  - Non-linear, i.e., min, max, weighted min
  - Adapted linear
  - Continuous-valued indicator of salient events in [0,1]



50% saliency-based raw audio summarization

[G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas and Y. Avrithis, *Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention,* IEEE T-MM 2013]

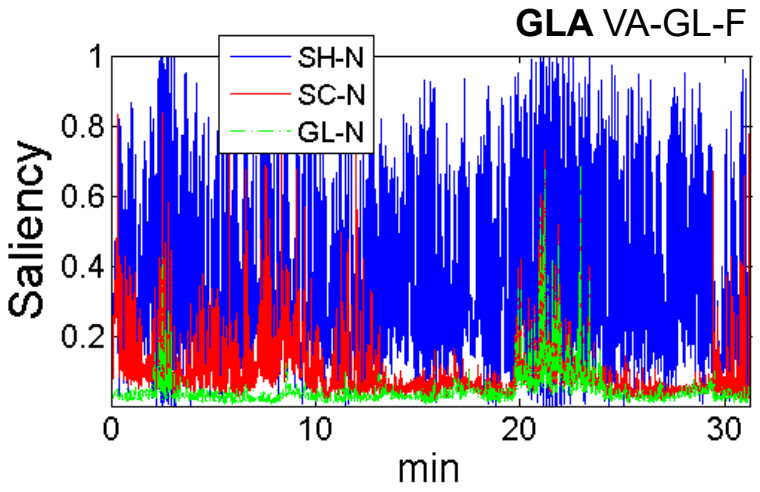
Fusion I: Scalar Operations

- 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 $S_{MIVA} = \min(S_1 - w_1, S_2 - w_2, S_3 - w_3) + \max(w_1, w_2, w_3)$

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

Fusion II: Normalization

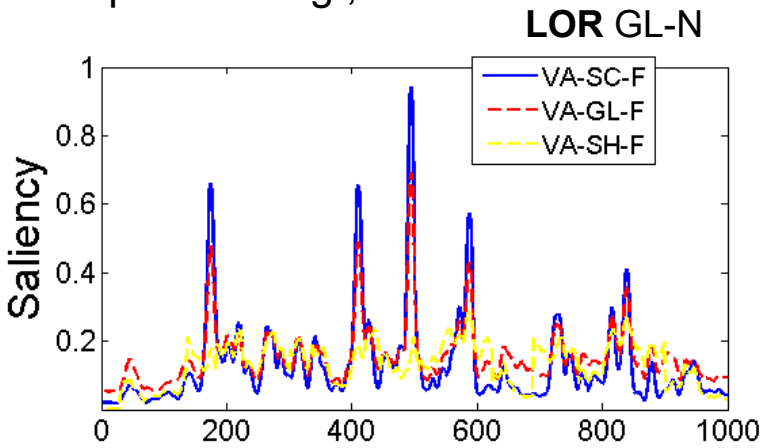
- Normalization intervals
  - Global linear normalization (GL)
  - Scene-based linear normalization (SC)
  - Shot-based linear normalization (SH)



- 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)



Audio Analysis: Perceptual Features

- **Roughness** (or sensory dissonance)
  - Associated to human attention, expressing the “stridency” of a sound due to rapid fluctuations in the amplitude
  - Related to the beating phenomenon whenever pair of sinusoids are closed in frequency
  - Computation of the peaks of the spectrum followed by averaging among all possible pair-wise combination of peaks
- **Loudness** (perceived sound pressure level)
  - Loudness model for time-varying sounds by Zwicker & Fastl (1999)
  - Temporal masking is taken into account in the model
  - Overall loudness computed by summing specific loudness on the bark scale
  - Loudness as a function of time

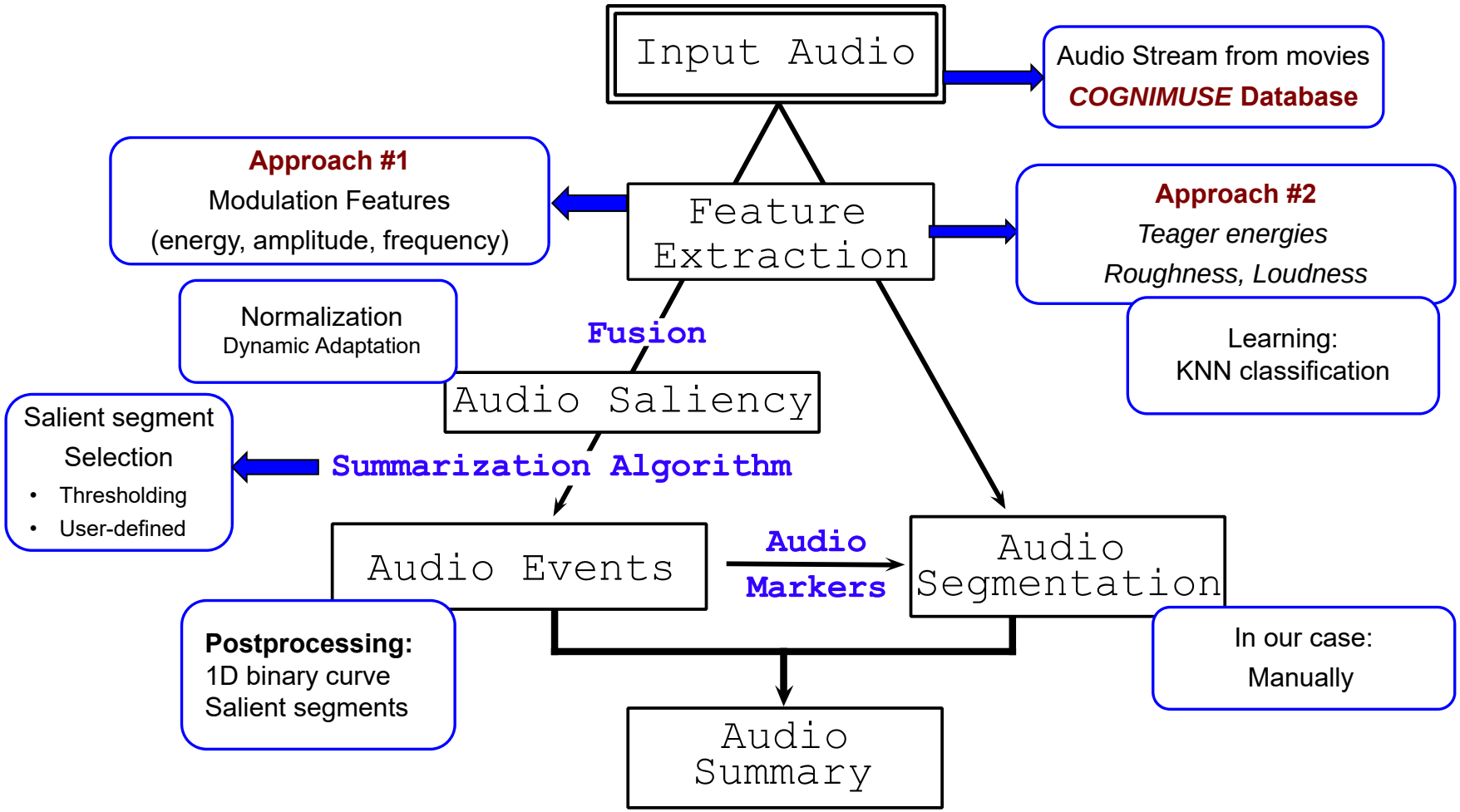
[R. Plomp and W.J.M. Levelt, *Tonal consonance and critical band- width*, Jour. JASA 1965]
[P.N. Vassilakis, *Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance*, Ph.D. thesis 2001]
[E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, Springer 1999]

Two Approaches for Saliency Computation

Problem of assigning a measure of interest to audio frames:

- **Approach #1**: bottom-up based on fusion of spectro-temporal cues and specifically AM-FM model and ESA demodulation
  - Dominant Modulation Features: MTE, MIA, MIF
- **Approach #2**: improved frontend with learning, based on Teager Energy and other perceptual features that correlate to the functioning of the human auditory system
  - Teager Energies, Roughness, Loudness

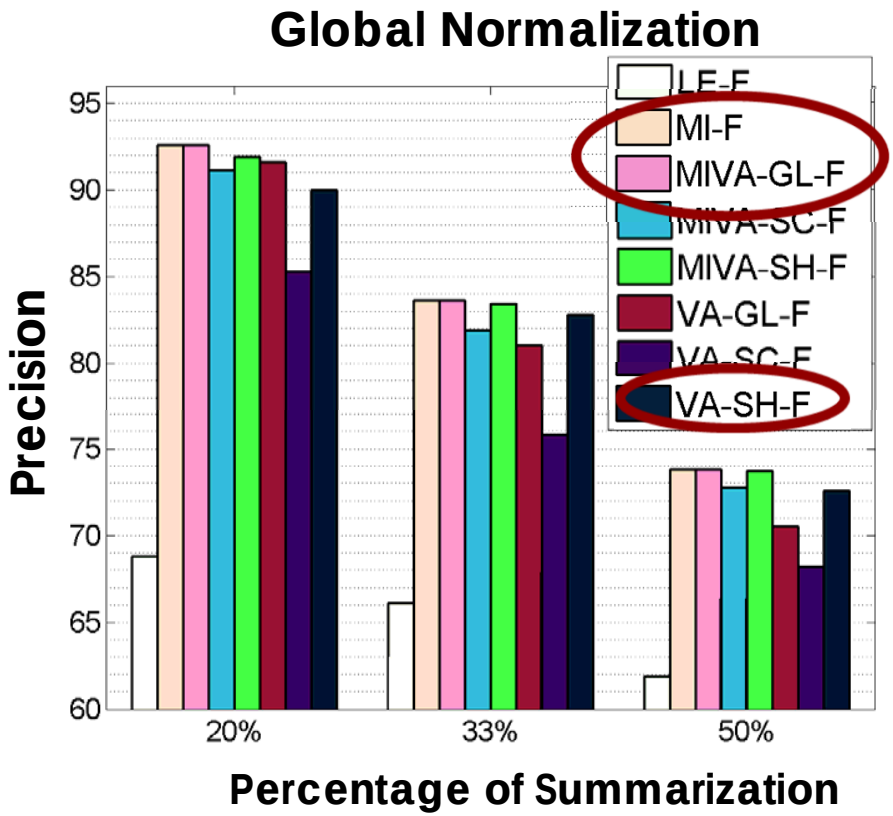
Audio Summarization System Overview

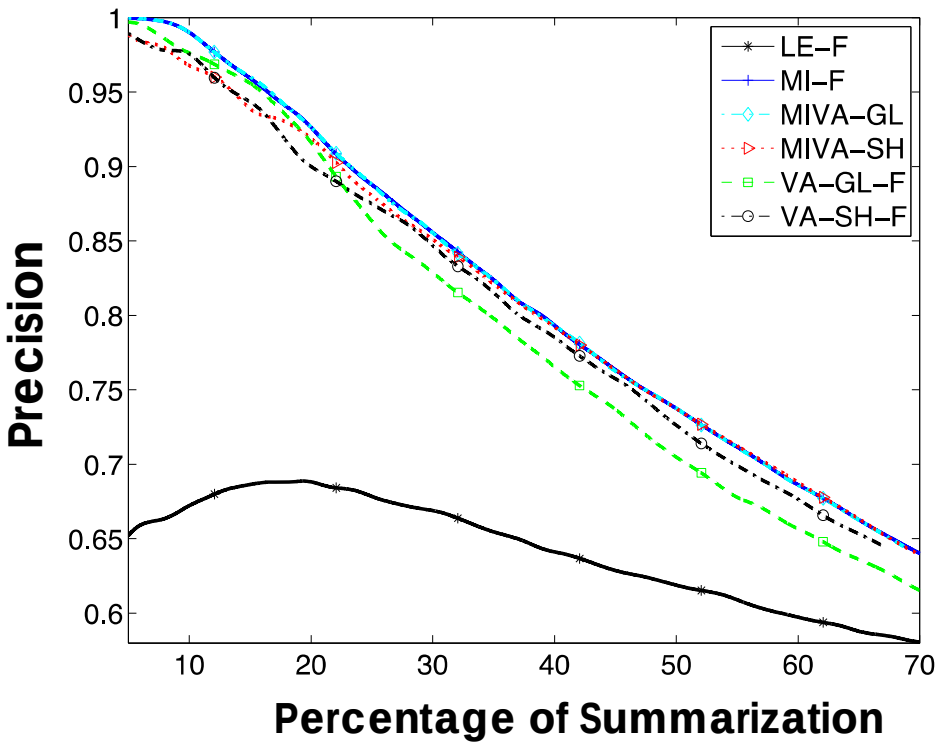


[A. Zlatintsi, E. Iosif, P. Maragos and A. Potamianos, *Audio Salient Event Detection And Summarization Using Audio And Text Modalities*, EUSIPCO 2015]
[P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos and A. Potamianos, *Predicting Audio-visual Salient Events based on A-V-T Modalities For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

Results for Approach #1

- Results in terms of frame-level precision, where various fusion methods were explored:
  - i.e., linear, min, weighted min, variance using adapted weights

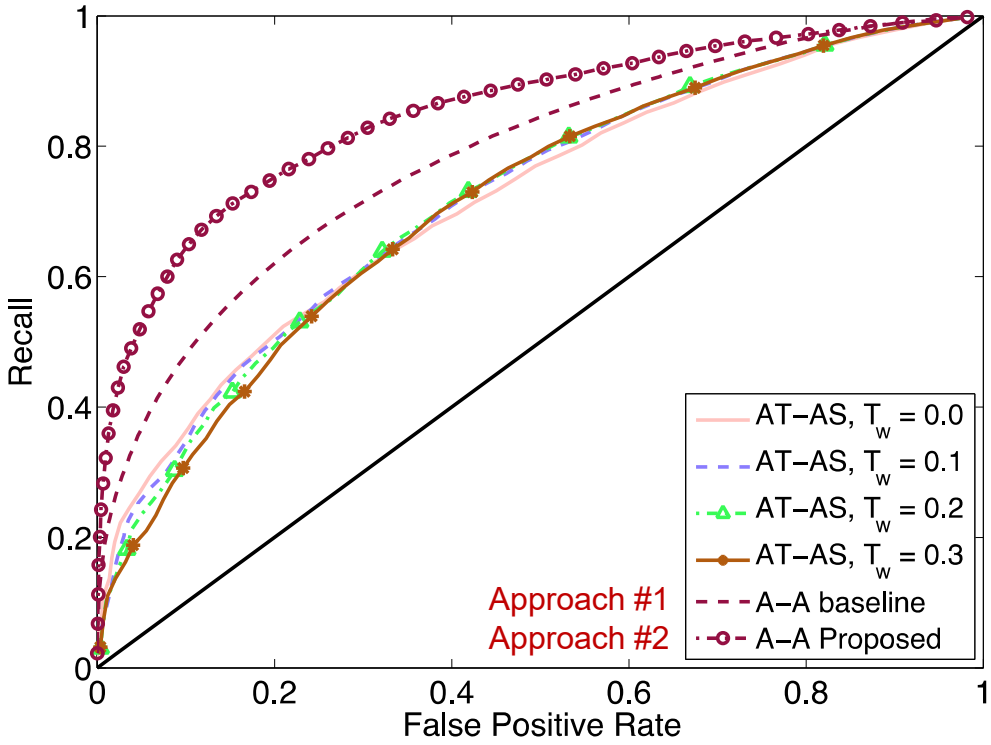


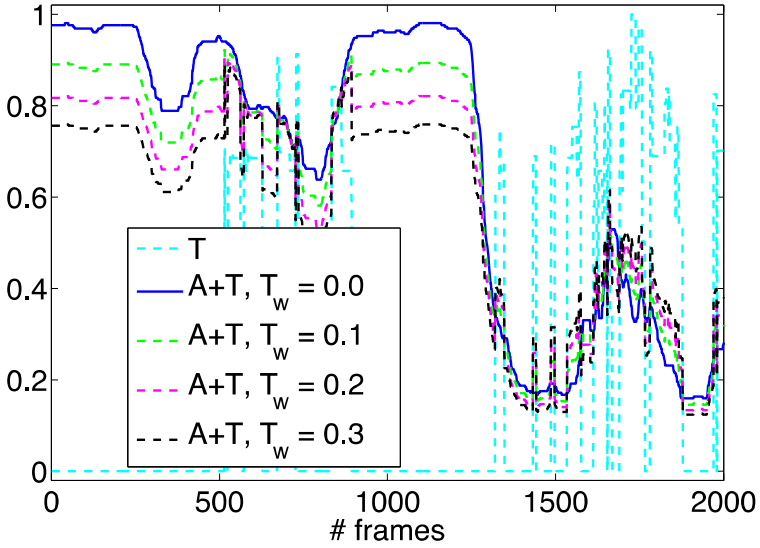


Results for Approach #2

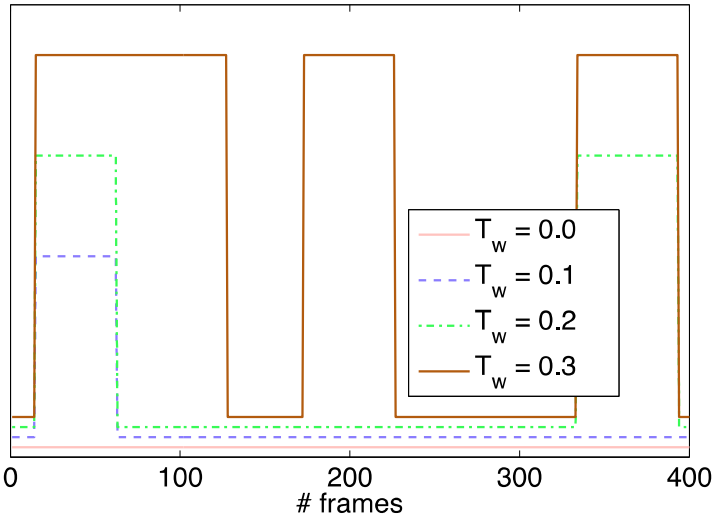
- Machine Learning Approach (using KNN)
- Comparison of approach #1 vs. #2
- Late Fusion of Audio with Text modality* exploring various weights for the text modality

*(discussed in Part 4 of the Tutorial presentation)





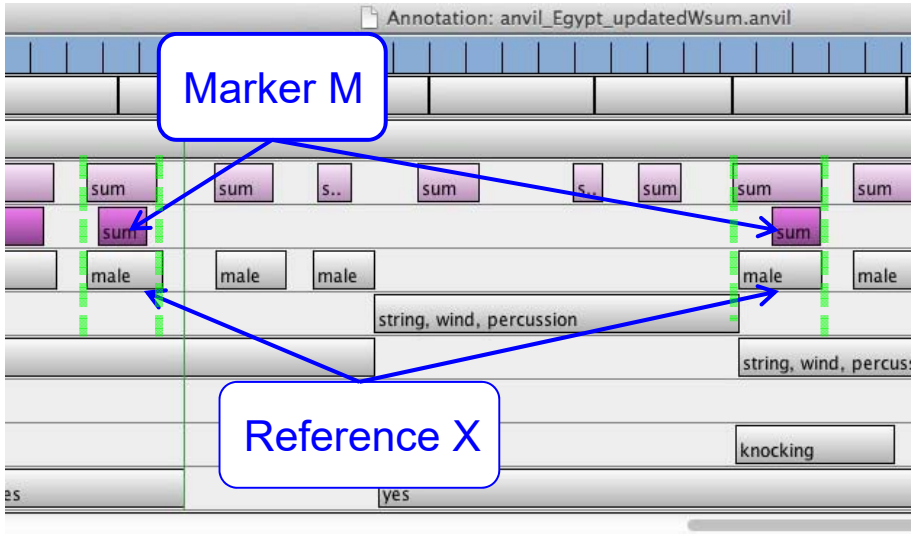
Late fusion of the audio and text modality



Effect of text weight in the selected segments

Audio Summarizer

- Choose segments salient and meaningful: perform boundary correction (to avoid word “clipping”),
  - thus “**speech reconstruction”** using part-of-speech tags indicating start and end times of each word
- Mathematical Morphology and specifically:
- 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]

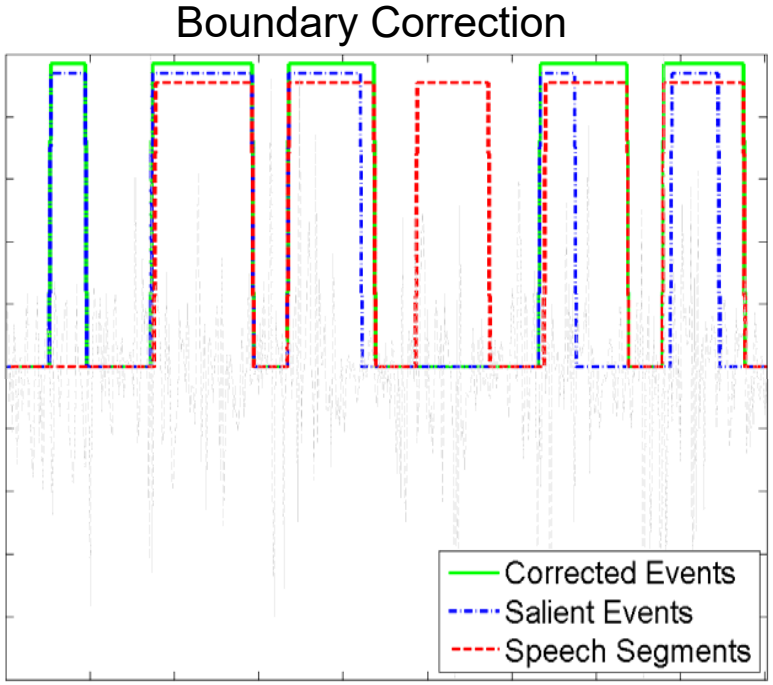
Audio Summary Demo

Audio Summary Demo

- 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



Part 3: Conclusions

- Two approaches for the detection of perceptually important audio events for the creation of summaries based on saliency models

Explored:

- A bottom-up based compact representation that tracks components with maximal energy contribution across frequency and time using the AM-FM model (and various fusion schemes)

- A carefully-designed audio saliency frontend using multiband Teager energies (in combination with perceptual features), which can detect robustly & discriminate various acoustic events due to its sharp time resolution and lowpass behavior

Tutorial slides: http://cognimuse.cs.ntua.gr/icassp17