ELL 788
Computational Perception & Cognition

July – November 2015

Module 8

Audio and Multimodal Attention

Audio Scene Analysis

- Two-stage process
  - Segmentation: decomposition to time-frequency segments
  - Grouping and segregation: based on perceived source
    - Auditory cues: feature (pitch, loudness ...), spatial (location, trajectory)
    - *Cross-modal cues (Visual)*
- Masking
  - Energy masking: (Lower level process)
    - Competing sounds interfereing in frequency and in time
  - Information masking: (Higher level process)
    - Difficulties to detect a target that cannot be explained with energy masking

Grouping (sreaming)

- Primitive grouping (Bottom-up: data driven)
  - Gestalt principle of perceptual organization
    - Frequency / Temporal proximity
    - Good continuation: Slowly varying frequency
    - Common fate: *Common start and stop time, amplitude / frequency modulation*
- Schema-based grouping (Top-down: knowledge-driven)
  - *E.g. human word recognition*
  - A schema is activated by adequate sensory information regarding one particular regularity in environment
  - Scemas can activate or suppress other schemas

Detection *vs*. Identification of sound

- Detection: asserting the presence
  - Signal to noise ratio
  - Saliency: Contextual incongruency with background
- Identification: associating a meaning
  - Familiarity (knowledge) of the listener
  - Physical components (unique features)
    - Standard deviation of the spectrum
    - Number of bursts or peaks
  - Information content
    - Not readily expected within a given environment

Audio attention

- Selective reception of an audio group (stream)
  - Perceived the same source
- Multiple cues
  - Auditory feature cues, Auditory spatial cues
  - Other cues
- Involuntary: A sudden gunshot (saliency-driven)
- Voluntary: Cocktail party effect (task-driven)
  - location, lip-reading, mean pitch differences, different speaking speeds, male/female speaking voices, distinctive accents

Computational models

- Studied more in visual domain
- Analogies between auditory and visual perception established in neuroscience literature
- Common models have been proposed for audio and visual domains
- Bottom-up (saliency-driven) vs. Top-down (task-driven)
  - *Attention on specific sound pattern (what is being said?)*
  - *Attention on specific sound source(s) (a bird singing amidst noise)*
  - *Attention on direction of sound (where that loud bang came from?)*

Experimental methods

- Psychological experiments
  - Users made to listen tones amidst white background noise
  - Originating at different spatial locations
  - Are asked to identify the tone listened
  - Accuracy / response time recorded
- Neurological experiments
  - Measure brain activities using EEG
  - Identify / characterize activations

Some observations

- Both short and long tones on a noisy background are salient
- Also the gaps (absence of a tone)
- Long tones / gaps accumulate more saliency than short tones
- A time gap (~ 1 – 1.5 s) between two tones leads to faster response
- Temporally modulated tones are more salient than stationary tones
- In a sequence of two closely spaced tones (within critical band), the second is less salient
- We can selectively attend to a piece of conversation when there are many overlapping conversations (cocktail party effect)
- In a noisy environment, people respond when their own names are uttered in an unattented stream

Audio Saliency Model

Kalini & Narayanan

Audio-scene
- 20 ms frames; shifted by 10 ms
- 128 filter-banks
- STFT used to compute spectogram

Feature extraction
- 8 features
  - 1 each for I, F and T
  - 2 for $O(\theta)$ $\theta=45^\circ,135^\circ$
  - 3 for P
- 8-level Gaussian pyramids (σ=1..8)

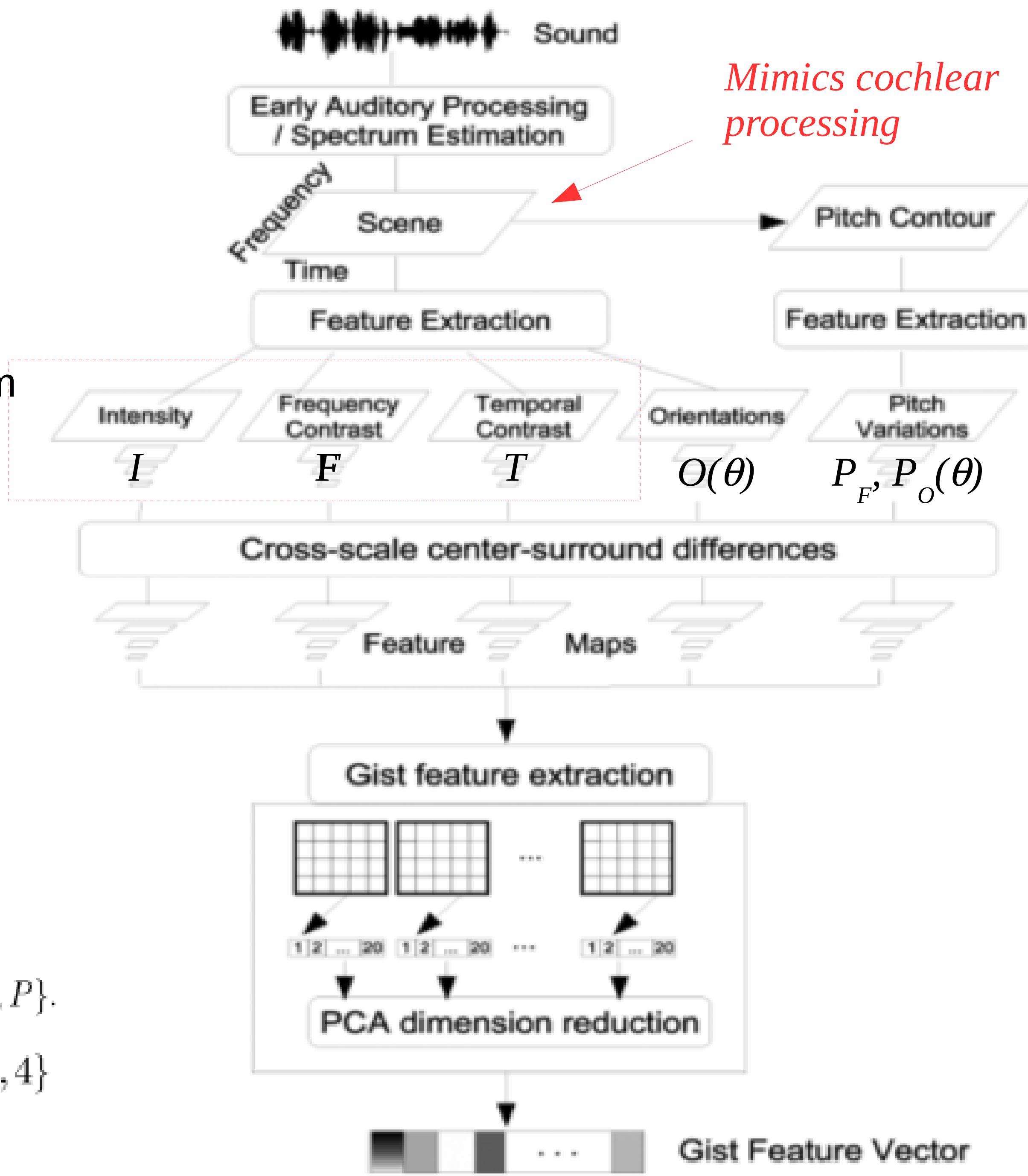
  (if time duration > 1.28 s)
  - 1:1 – 1:128

Center-Surround differences

$$\mathcal{M}(c,s) = |\mathcal{M}(c) \ominus \mathcal{M}(s)|, \quad \mathcal{M}\epsilon\{I,F,T,O_\theta,P\}.$$

$$c = \{2,3,4\}, s = c+\delta \text{ with } \delta\epsilon\{3,4\}$$

$\rightarrow$ 48 (6 x 8) feature maps



Gist features

- A low resolution representation of the feature maps for quick understanding of the audio scene
- The feature-maps are divided into m x n grids (m = 4, n = 5)
- Gist feature vectors are computed as averages in each cell of the grid

$$G_i^{k,l} = \frac{mn}{vh} \sum_{u=\frac{kv}{n}}^{\frac{(k+1)v}{n}-1} \sum_{z=\frac{lh}{m}}^{\frac{(l+1)h}{m}-1} \mathcal{M}_i(u,z), \text{ for}$$
$$k = \{0, \ldots, n-1\}, \quad l = \{0, \ldots, m-1\}$$

v = width
h = height of the cells
i = feature map index {1 .. 48}

- 20 (4 x 5) gist values for each of 48 feature maps
- Combined; PCA used for dimensionality reduction
- $\rightarrow$ Auditory gist feature $F = \{ f_i \}, i = 1 .. d$

Saliency in terms of audio features

Task dependent biasing of auditory cues

- Given a task
  - Enhance specific dimensions of the gist feature that are related to the task
  - Attenuate specific dimensions of the gist feature that are not related to the task
- Feature dimensions (to enhance / attenuate) are learnt with supervised training

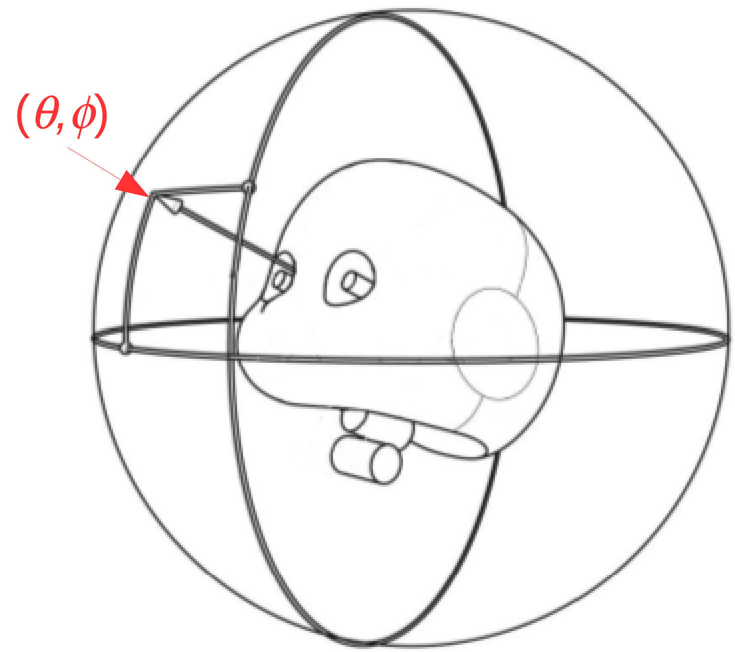
Example in speech understanding

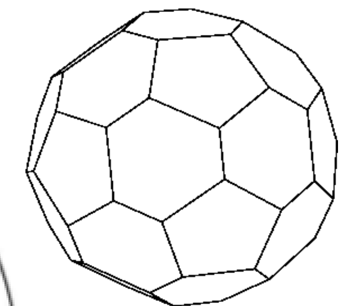
- An *audio scene* refers to utterance of a syllable
  - Probability (prominence) of a syllable uttered can be found from the gist features for an audio scene, p (c_i | F), i = {0|1}
- Lexical knowledge comes to play
  - The probabilities for sequences of utterings of the syllabi are learnt
    - Bi-gram / Tri-gram models
- Use a probabilistic model to adjust gist feature dimension weights
  - Tune the attention to the next expected syllabii depending on the previous syllabii uttered

Multimodal attention model
(Audio + Visual)

[Ruesch][Schauerte]

iCub (INRIA)
Six DoF

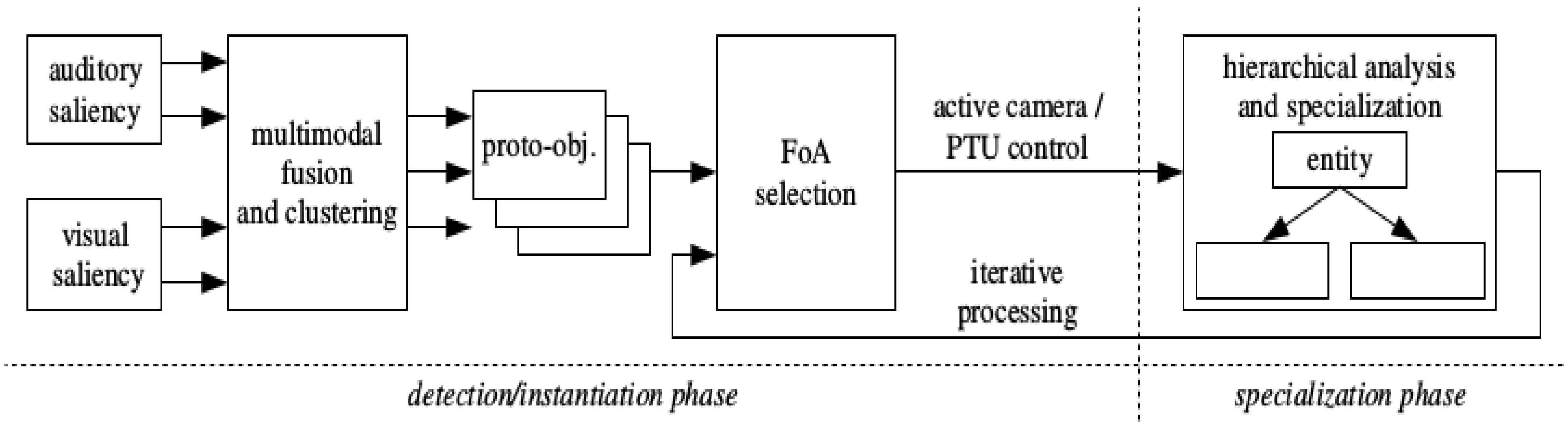




Ego-sphere (θ, ϕ): Fixed with respect to the torso
Reference does not change unless the robot moves
Hexagonal or pentagonal cells

Audio-visual saliency of a cell on the ego-sphere determines where the robot should look at (Bottom-up saliency model)

Schematic overview



Saliency features

- Visual saliency: similar to Itti, et al.
  - *Intensity, color and orientations*
  - (θ,ϕ) representation mapped to (x,y)
    - *There is some distortion*
- Audio saliency
  - STFT used to create spectogram for each “ear”
  - Inter-aural Time Difference (ITD) and Inter-aural Spectral Difference (ISD) used to for locating sound
    - Design of pinna (outer “ear”) for creating ISD
  - Spatio-temporal clustering to eliminate outliers (noise)
  - *Saliency determined based on “surprise element” [Schauerte]*

Computed with respect to current head-orientation of the robot

Surprise Element (Audio)

Following Itti & Baldi (2005)

Spectogram: $G(t,\omega)=|STFT(t,\omega)|^2$

Assume that $G(t,\omega)$ is caused by a GMM with parameters g

Prior probability : $P^{\omega}_{Prior}=P(g|G(t-1,\omega),G(t-2,\omega),\ldots G(t-N,\omega))$

Posterior probability : $P^{\omega}_{Posterior}=P(g|G(t,\omega),G(t-1,\omega),\ldots G(t-N,\omega))$

Surprise element : $S_A(t,\omega)=D_{KL}(P^{\omega}_{Posterior},P^{\omega}_{Prior})$

D_{KL} is the Kullback Leibler Divergence between two GMMs

Overall audio surprise element :

$$S_A(t)=\frac{1}{|\Omega|}\sum_{\omega\in\Omega}S_A(t,\omega)$$

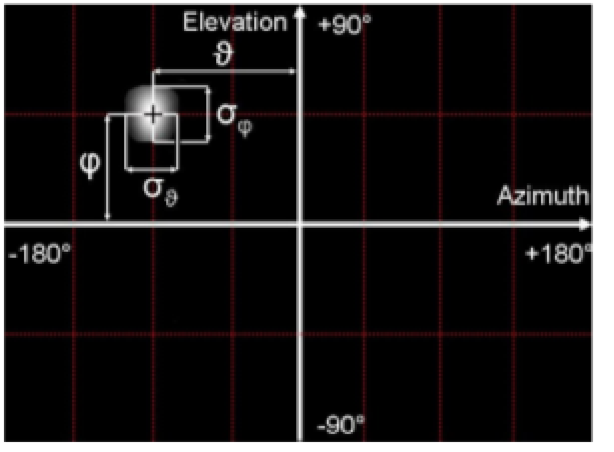
Ego-centric saliency determination

Basic steps

1. Convert stimulus orientation to torso-based, head-centered coordinates using head kinematics,
2. Represent orientation in spherical coordinates, project stimulus (saliency) intensity onto modality specific rectangular egocentric map,
3. Aggregate multimodal sensory information.

Multimodal saliency aggregation

- Project saliencies from (θ,ϕ) coordinates to (x,y) coordinates for visualization
  - There is some distortion

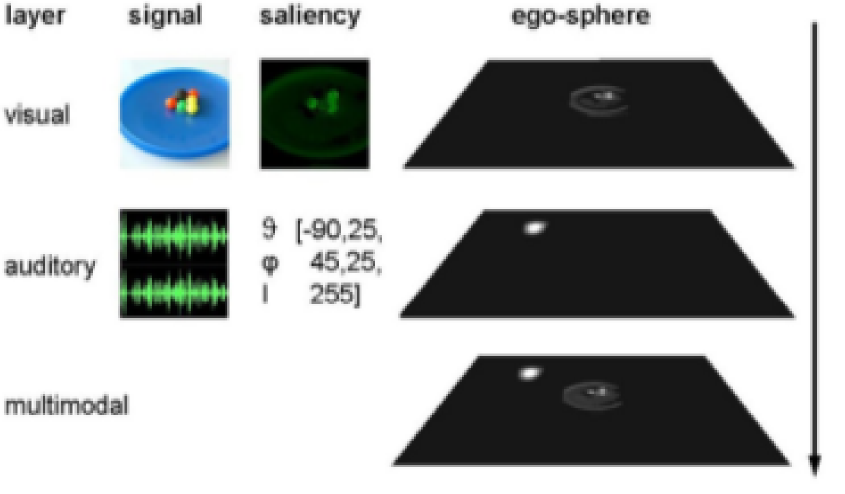


- Take union of visual and audio saliencies

$$S(\theta,\phi;t)=max(S_V(\theta,\phi;t),S_A(\theta,\phi;t))$$

- Proto-object regions determined by analysis of isophote curvatures
  - Isophote = contours of equal saliencies on the saliency map

- Center of the proto-object regions (θ_0,ϕ_0) are considered as proto-object locations



Attention and exploration

- The FoA goes to the proto-object region with highest saliency
- Habituation map
  - Initialized to zero
  - Updated recursively according to a Gaussian weighing function that favors the regions closer to the current FoA

$$H(\theta,\phi;t)=(1-d_H)H(\theta,\phi;t-1)+d_h G_h(\theta-\theta_{0,}\phi-\phi_0;t)$$

$\sigma \approx 6^{\circ}$

- While attending to a salient point, the Habituation Map at that location will asymptotically tend to 1
  - Rate of convergence depends on $d_h \in [0,1]$
- Whenever the habituation value exceeds a predefined level, the system becomes attracted to novel salient points.

Inhibition map

- Initialized to 1
- When the habituation of current FoA (θ_0,ϕ_0) exceeds a threshold t_h = 0.85
  - the Inhibition Map is modified by adding a scaled Gaussian function, G_a ,
  - with amplitude −1, centered at FoA (θ_0,ϕ_0) and
  - variance σ ≈ 6º
- The resulting effect is that G_a adds a smooth ”cavity” at (θ_0,ϕ_0) in the inhibition map

$$A(\theta,\phi;t)=(1-d_a)A(\theta,\phi;t-1)+d_a G_a(\theta-\theta_{0,}\phi-\phi_0;t)$$

- For attention selection
  - Multiply: S(θ,ϕ;t) × A(θ,ϕ;t)
  - (Combine instantaneous saliency and the memory of recently attended locations)

References

- Wrigley (2000). A model of auditory attention
- Kayser, et al. (2005). Mechanisms for Allocating Auditory Attention ...
- Kalinli & Narayanan (2009). Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information
- Oldoni, et al (2013). A computational model of auditory attention...
- Ruesch, et al. Multimodal Saliency-Based Bottom-Up Attention ...
- Schauerte, et al. Multimodal Saliency-based Attention for Object-based Scene Analysis ...
- Itti & Baldi (2005). Bayesian Surprise Attracts Human Attention (Visual)
- Lichtenauer, et al. Isophote Properties as Features for Object Detection