Learning Sparse Representations for Audiovisual Signals

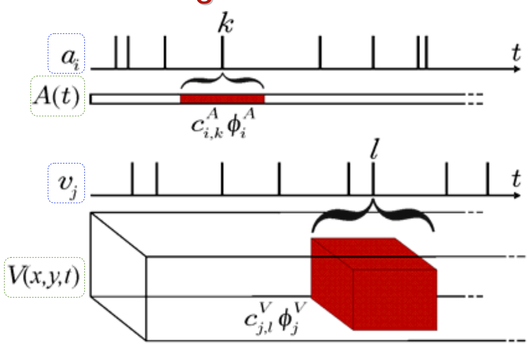
REDWOOD CENTER for Theoretical Neuroscience

Gianluca Monaci and Friedrich T. Sommer,
Redwood Center for Theoretical Neuroscience, University of California, Berkeley
Email: {gmonaci,fsommer}@berkeley.edu

Motivation

- **BACKGROUND**
  - Audio-visual interactions in the brain at many levels
    - *Perceptual evidence* : McGurk effect, bounce-stream illusion, sound-induced flashing... *[1]*
  - Evidence for **early fusion mechanisms**
    - *ERP, MEG, BOLD dynamics studies in humans [2]*
    - *Anatomical studies in monkeys [3]*
- **SPARSE CODING PARADIGM**
  - Successful in describing auditory and visual coding
    - *A1 : Smith&Lewicki 2006 - V1 : Olshausen&Field 1996*
- **OBSERVATION**
  - No computational model of early crossmodal interactions
- **GOAL**
  - Design a sparse, biologically plausible, audio-visual signal model accounting for early fusion mechanisms in the brain

Audio-visual signal model

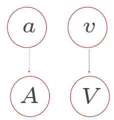


audio and visual signals — audio and visual spikes — audio and visual kernels

$$S = \begin{pmatrix} A(t) \\ V(x,y,t) \end{pmatrix} \approx \begin{pmatrix} \sum_i \sum_k a_{i,k} c^A_{i,k} \phi^A_i(t - k\cdot\tau^A) \\ \sum_j \sum_l v_{j,l} c^A_{j,l} T_{q_l,r_l} \phi^V_j(t - l\cdot\tau^V) \end{pmatrix}$$

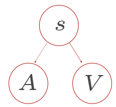
Adaptive crossmodal projections code

- **Learning sparse codes**



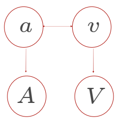
- Maximizing log-likelihood
  $\mathrm{O}(Z,\Phi,s) = \log(p(Z|\Phi)) = \log(\int p(Z|\Phi,s)p(s)ds)$
- Approximation: $\int p(Z|\Phi,s)p(s)ds = \int p(Z,s|\Phi)ds \approx p(Z,s^\star|\Phi)$
- Iteratively solving two nested steps
  - ▶ ***Code:*** *Find* $s^\star$ *optimizing* $\mathrm{O}(Z,\Phi,s)$ *w.r.t.* s
  - ▶ ***Learn:*** *Find* $\Phi^\star$ *optimizing* $\mathrm{O}(Z,\Phi,s^\star)$ *w.r.t.* Φ

- **Learning sparse audiovisual codes** [4]



- Enforce synchronicity : $a = v = s$
  $\mathrm{O} = \log(\int p(A,V|\Phi^A,\Phi^V,s)p(s)ds)$

- **Learning adaptive crossmodal projections**

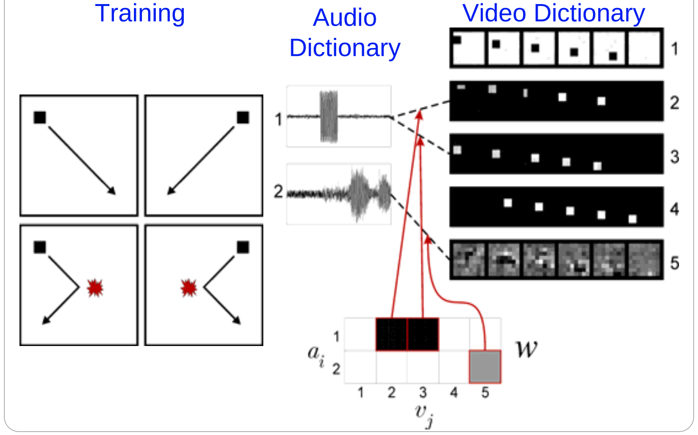


- Codes a and v are dependent

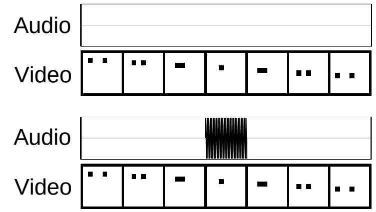
$$p(a,v) \propto \exp(-\lambda_a \underbrace{\|a\|_0}_{\#\text{ spikes in } a} - \lambda_v \underbrace{\|v\|_0}_{\#\text{ spikes in } v} + \sum\sum a_i \underbrace{w_{ij}}_{\text{crossmodal projections}} v_j)$$

$$\mathrm{O} = \log(\int\int p(A,V|\Phi^A,\Phi^V,a,v)p(a,v)dadv)$$

- **Learning done iteratively solving three nested steps**
  - ▶ ***Code:*** *greedy approximation, AV-Matching Pursuit [4]*
  - ▶ ***Learn Dictionary:*** *Gradient Descent on* Φ
  - ▶ ***Learn Projections:*** *Hebbian learning on* w



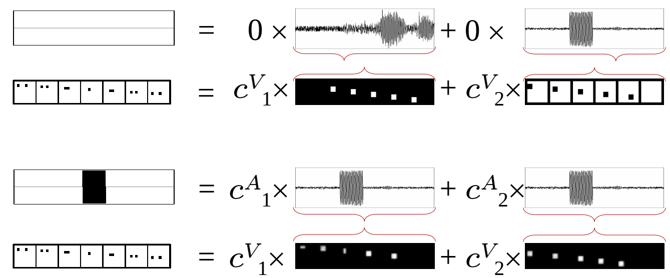
The bounce-stream illusion



~80% observers perceive STREAMING [1]

~70% observers perceive BOUNCING [1]

- **Encoding using the learned audiovisual dictionary with adaptive crossmodal projections**



Conclusions

- New model for early audio-visual fusion
- Model based on joint sparse coding
- New method to learn basis functions and cross-modal associations
- Model "suffers" from bounce-stream illusion

[1] **Sensory modalities are not separate modalities: plasticity and interactions**. Shimoho, S. and Shams, L., Current Opinion in Neurobiology, 2001, **11**:505-509
[2] **Sound alters activity in human V1 in association with illusory visual perception**. Watkins, S. *et al.*, NeuroImage, 2006, **31**:1247-1256
[3] **Multisensory convergence in calcarine visual areas in macaque monkey**. Rockland, K.S. and Ojima, H., Int. Journal of Psychophysiology **50**:19-26
[4] **Learning sparse generative models of audiovisual signals**. Monaci, G. and Sommer, F.T., submitted to European Signal Processing Conference (EUSIPCO), 2008

COSYNE Meeting 2008, 28 Feb - 2 Mar 2008, Salt Lake City, Utah

Acknowledgements : This work was supported by the Swiss NSF through the grant PBEL2-118742 to GM