Computer Vision, Speech Communication & Signal Processing Group,
Intelligent Robotics and Automation Laboratory
Institute of Communication and Computer Systems (ICCS)
National Technical University of Athens, Greece (NTUA)

Part 2:
Audio-Visual Processing, Fusion and Perception

Petros Maragos and Petros Koutras

Tutorial at IEEE International Conference on Image Processing 2019,
Taipei, Taiwan, September 22, 2019

Part 2: Outline

- A-V Perception
- Bayesian Formulation of Perception & Fusion Models
- Application: Audio-Visual Speech Recognition
- Application: Emotion-Expressive Audio-Visual Speech Synthesis
- Application: Multimodal (Vision+Text) Concept Learning in Videos
- Other Applications: Multimodal Video Summarization, A-V Music Synthesis

Audio-Visual Perception and Fusion

Perception: the sensory-based inference about the world state

Human versus Computer Multimodal Processing

- Nature is abundant with multimodal stimuli.
- Digital technology creates a rapid explosion of multimedia data.
- Humans perceive world multimodally in a seemingly effortless way, although the brain dedicates vast resources to these tasks.
- Computer techniques still lag humans in understanding complex multisensory scenes and performing high-level cognitive tasks. **Limitations**: inborn (e.g. data complexity, voluminous, multimodality, multiple temporal rates, asynchrony), inadequate approaches (e.g. monomodal-biased), non-optimal fusion.
- **Research Goal**: *develop truly multimodal approaches that integrate several modalities toward improving robustness and performance for anthropo-centric multimedia understanding*.

Multicue or Multimodal Perception Research

- ***McGurk effect**: Hearing Lips and Seeing Voices* [McGurk & MacDonald 1976]

- *Modeling Depth Cue Combination using Modified Weak Fusion* [Landy et al. 1995]
  - scene depth reconstruction from multiple cues: motion, stereo, texture and shading.
- *Intramodal Versus Intermodal Fusion of Sensory Information* [Hillis et al. 2002]
  - shape surface perception: intramodal (stereopsis & texture), intermodal (vision & haptics)

- *Integration of Visual and Auditory Information for Spatial Localization*
  - Ventriloquism effect
  - Enhance selective listening by illusory mislocation of speech sounds due to lip-reading [Driver 1996]
  - Visual capture [Battaglia et al. 2003]
  - Unifying multisensory signals across time and space [Wallace et al. 2004]

- *AudioVisual Gestalts* [Monaci & Vandergheynst 2006]
  - temporal proximity between audiovisual events using Helmholtz principle

- *Temporal Segmentation of Videos into Perceptual Events by Humans* [Zacks et al. 2001]
  - humans watching short videos of daily activities while acquiring brain images with fMRI

- *Temporal Perception of Multimodal Stimuli* [Vatakis and Spence 2006]

McGurk effect example

- [ba – audio] + [ga – visual] → [da] (fusion)
- [ga – audio] + [ba – visual] → [gabga, bagba, baga, gaba] (combination)
- Speech perception seems to also take into consideration the visual information. Audio-only theories of speech are inadequate to explain the above phenomena.
- Audiovisual presentations of speech create fusion or combination of modalities.
- One possible explanation: *a human attempts to find common or close information in both modalities and achieve a unifying percept.*

Attention

- ***Feature-integration theory of attention*** [Treisman and Gelade, CogPsy 1980]:
  - " Features are registered early, automatically, and in **parallel** across the visual field, while objects are identified separately and only at a later stage, which requires focused attention.
  - This theory of attention suggests that attention must be directed **serially** to each stimulus in a display whenever conjunctions of more than one separable feature are needed to characterize or distinguish the possible objects presented. "
- ***Orienting of Attention*** [Posner, QJEP 1980]:
  - Focus of attention shifts to a location in order to enhance processing of relevant information while ignoring irrelevant sensory inputs.
  - **Spotlight Model**: focus visual attention to an area by using a **cue** (a briefly presented dot at location of target) which triggers "formation of a spotlight" and reduces RT to identify target. Cues are *exogenous* (low-level, outside generated) or *endogenous* (high-level, inside generated).
  - Overt / **Covert** orienting (with / without eye movements): "Covert orientation can be measured with same precision as overt shifts in eye position."
- *Interplay between Attention and Multisensory Integration:* [Talsma et al., Trends CogSci 2010]: "Stimulus-driven, bottom- up mechanisms induced by crossmodal interactions can automatically capture attention towards multisensory events, particularly when competition to focus elsewhere is relatively low. Conversely, top-down attention can facilitate the integration of multisensory inputs and lead to a spread of attention across sensory modalities."

Perceptual Aspects of Multisensory Processing

Multisensory Integration: unisensory auditory and visual signals are combined forming a new, unified audiovisual percept.

Goal: *Perceiving Synchronous and Unified Multisensory Events*

Principles: Multisensory integration is governed by the following rules:

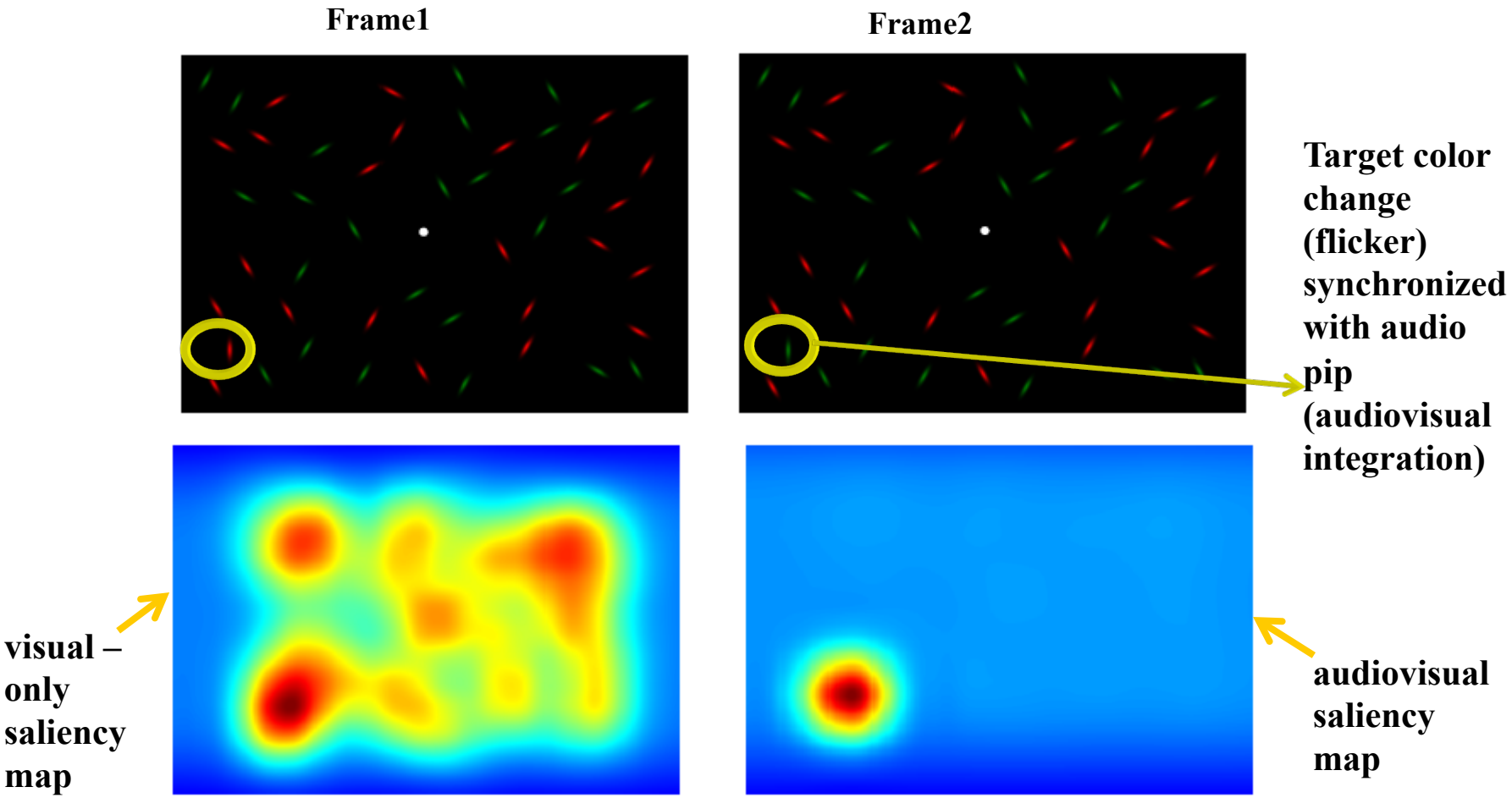
- **Spatial rule**,
- **Temporal rule**,
- **Modality Appropriateness**:
  - Visual dominance of spatial tasks.
  - Audition is dominant for temporal tasks.
- **Inverse effectiveness law**:
  - In multisensory neurons, multimodal stimuli occurring in close space-time proximity evoke supra-additive responses. The less effective monomodal stimuli are in generating a neuronal response, the greater relative percentage of multisensory enhancement.
  - Is this the case for behavior? Recent experiments indicate that inverse effectiveness accounts for some behavioral data.

Synchrony and Semantics are two factors that appear to favor the binding of multisensory stimuli, yielding a coherent unified percept. Strong binding, in turn, leads to higher stream asynchrony tolerance.

[E. Tsilionis and A. Vatakis, "*Multisensory Binding: Is the contribution of synchrony and semantic congruency obligatory*?", COBS 2016.]

Computational audiovisual saliency model

- Combining audio and visual saliency models by proper fusion
- Validated via behavioral experiments, such as pip & pop:



[A. Tsiami, A. Katsamanis, P. Maragos and A. Vatakis, ICASSP 2016.]

Bayesian Formulation of Perception

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)}$$

S : configuration of auditory and/or visual scene of world
D : mono/multi-modal data or features.

P(S): Prior Distribution, *P(D/S)*: Likelihood, *P(D)*: Evidence

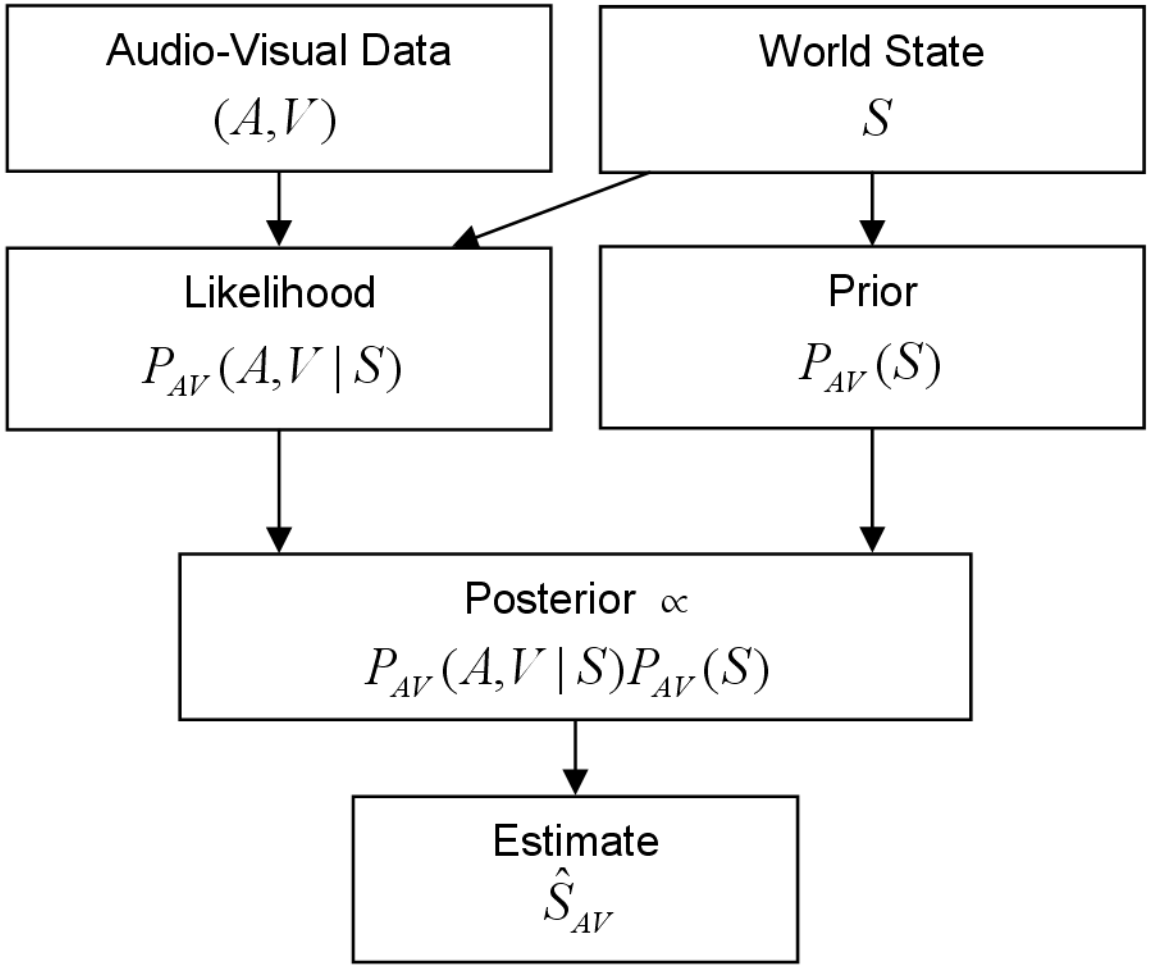
P(S/D): Posterior conditional distribution

S → D : World-to-Signal mapping

Perception is an ill-posed inverse problem

$$\hat{S}_{MAP} = \underset{S}{\operatorname{argmax}}\, P(D|S)P(S)$$

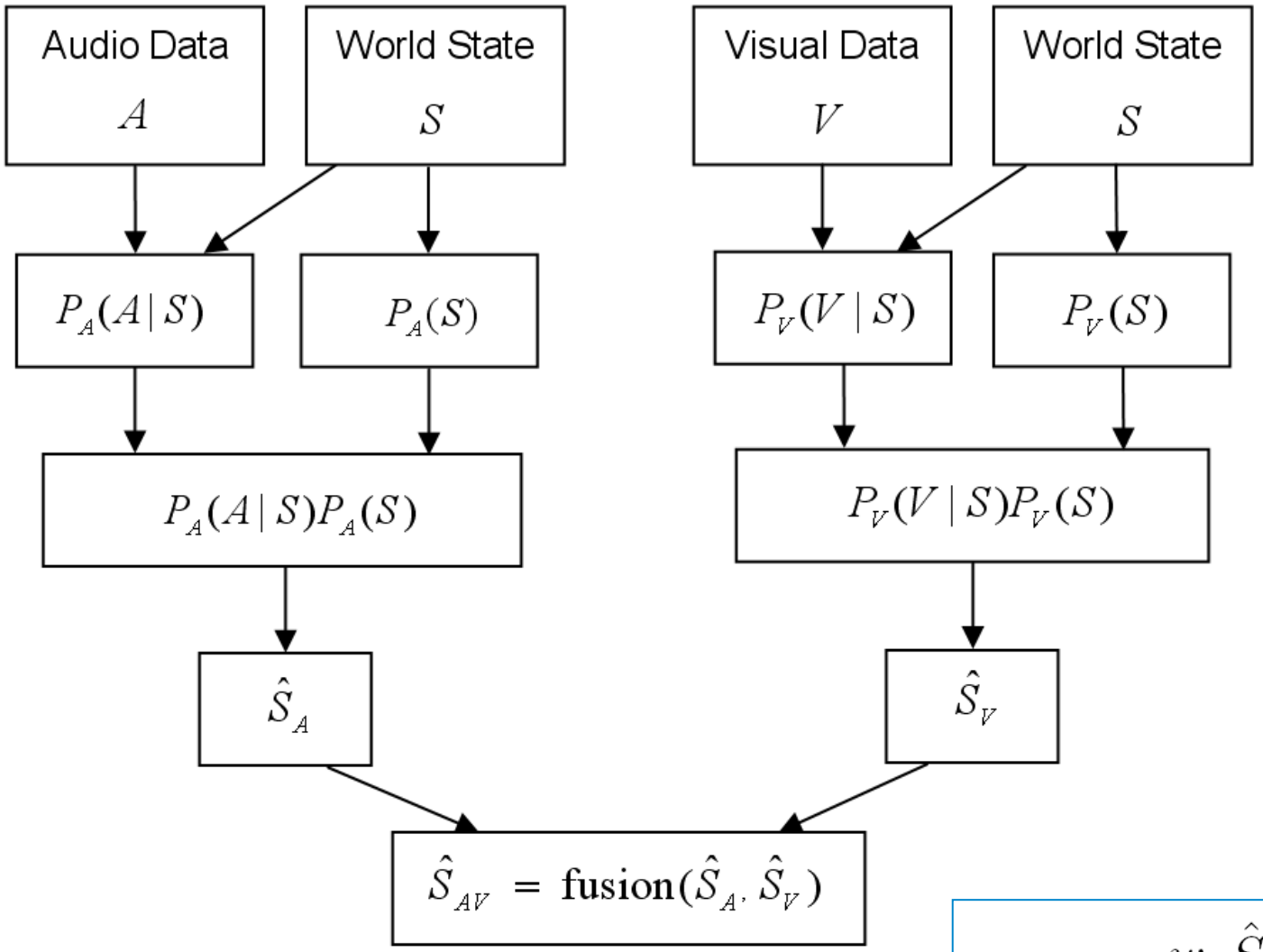
Strong Fusion: Bayesian formulation



$$P_{AV}(S|D_A, D_V) = \frac{P_{AV}(D_A, D_V|S)P_{AV}(S)}{P_{AV}(D_A, D_V)}$$

[Clark & Yuille 1990]

Weak Fusion: Bayesian formulation



For Gaussian distributions, or if the two single monomodal MAP estimates are close, their fusion is weighted average [Yuille & Bulthoff,1996]

$$\hat{S}_{AV} = \frac{w_{\mathrm{a}}\hat{S}_A + w_{\mathrm{v}}\hat{S}_V}{w_{\mathrm{a}} + w_{\mathrm{v}}}$$

Models for Multimodal Data Integration

Levels of Integration:

- *Early* integration (as in strong fusion)
- *Intermediate* integration
- *Late* integration (as in weak fusion)

Time dimension:

- *Static*: CCA- Canonical Correlation Analysis: e.g. “cocktail-party effect”
  Max Mutual Information
  SVMs- Support Vector Machines: kernel combination
- *Dynamic*: HMMs (Hidden Markov Models)
  DBNs (Dynamic Bayesian Nets)
  DNNs (Deep Neural Nets)
  Multimodal Hypothesis Rescoring

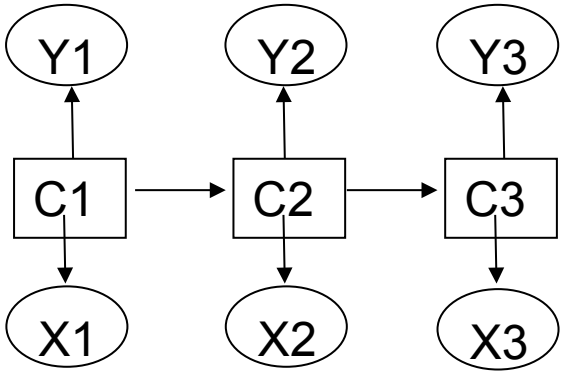
Multi-stream Weights for Audio-Visual Fusion

$$B(S|D_A, D_V) = [P_A(D_A|S)]^{q_1} [P_V(D_V|S)]^{q_2} \frac{P(S)}{P(D)}$$

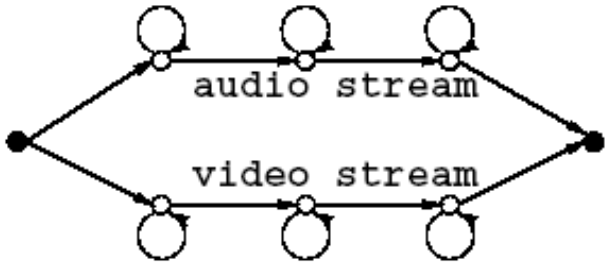
- Intermediate case between weak and strong fusion
- Select exponents q1, q2 for aural and visual streams
- Work in the LogProb domain → Weighted Linear combination

Multi-Stream HMM Topologies for Audio-Visual (A-)Synchrony

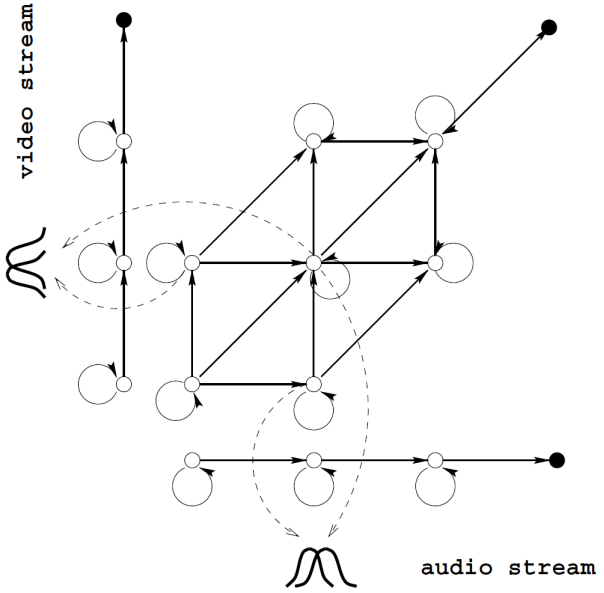
[G. Potamianos, C. Neti, G. Gravier, A. Garg and A. Senior, "*Advances in Automatic Recognition of AudioVisual Speech*", Proc. IEEE 2003]



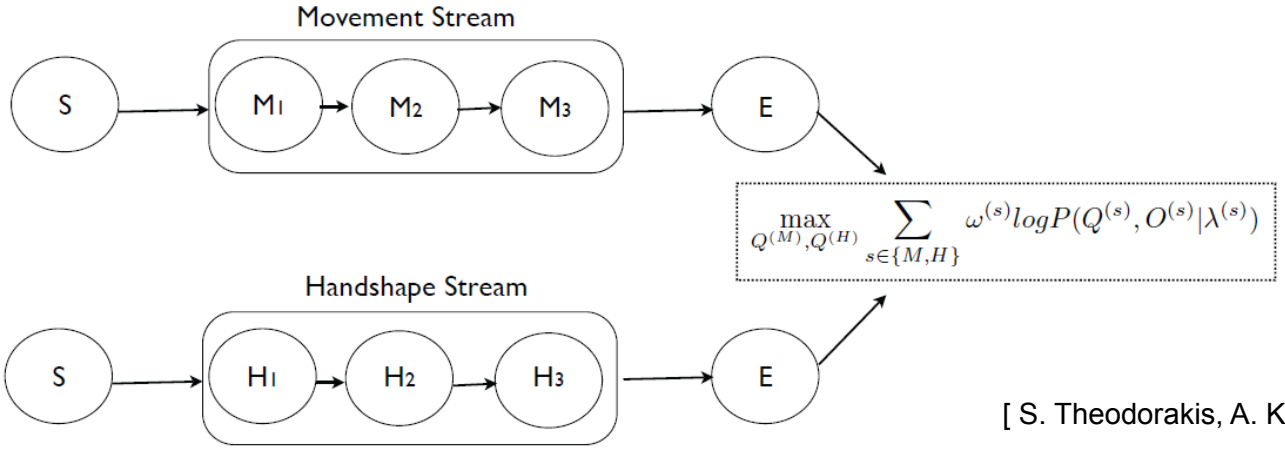
Synchronous HMMs
Synchrony at each state



Two-Stream HMMS
Phone-synchronous
State-asynchronous



Product-HMMs: Controlled synchronization freedom



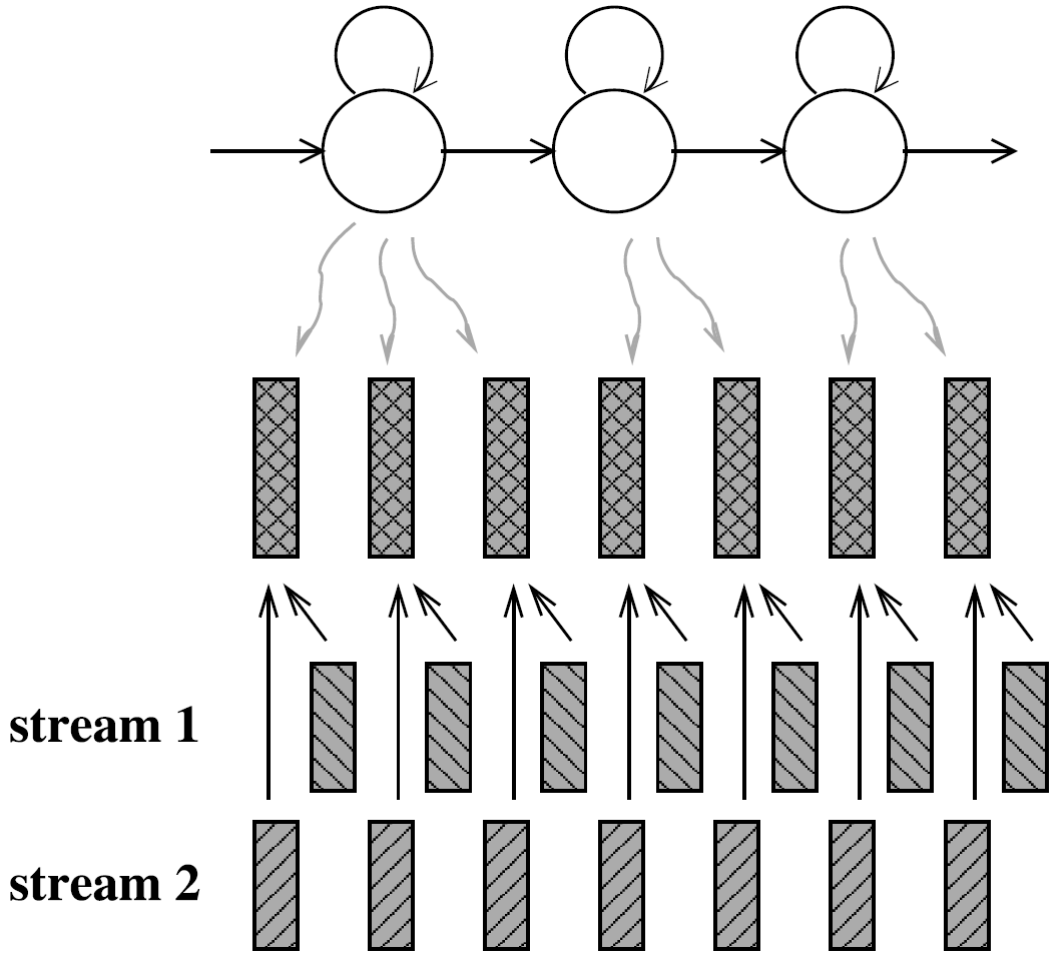
Parallel-HMMs for Sign Recognition

[C.Vogler & D. Metaxas, CVIU 2001]

[S. Theodorakis, A. Katsamanis & P. Maragos, ICASSP 2009]

Synchronous Multi-Stream HMMs

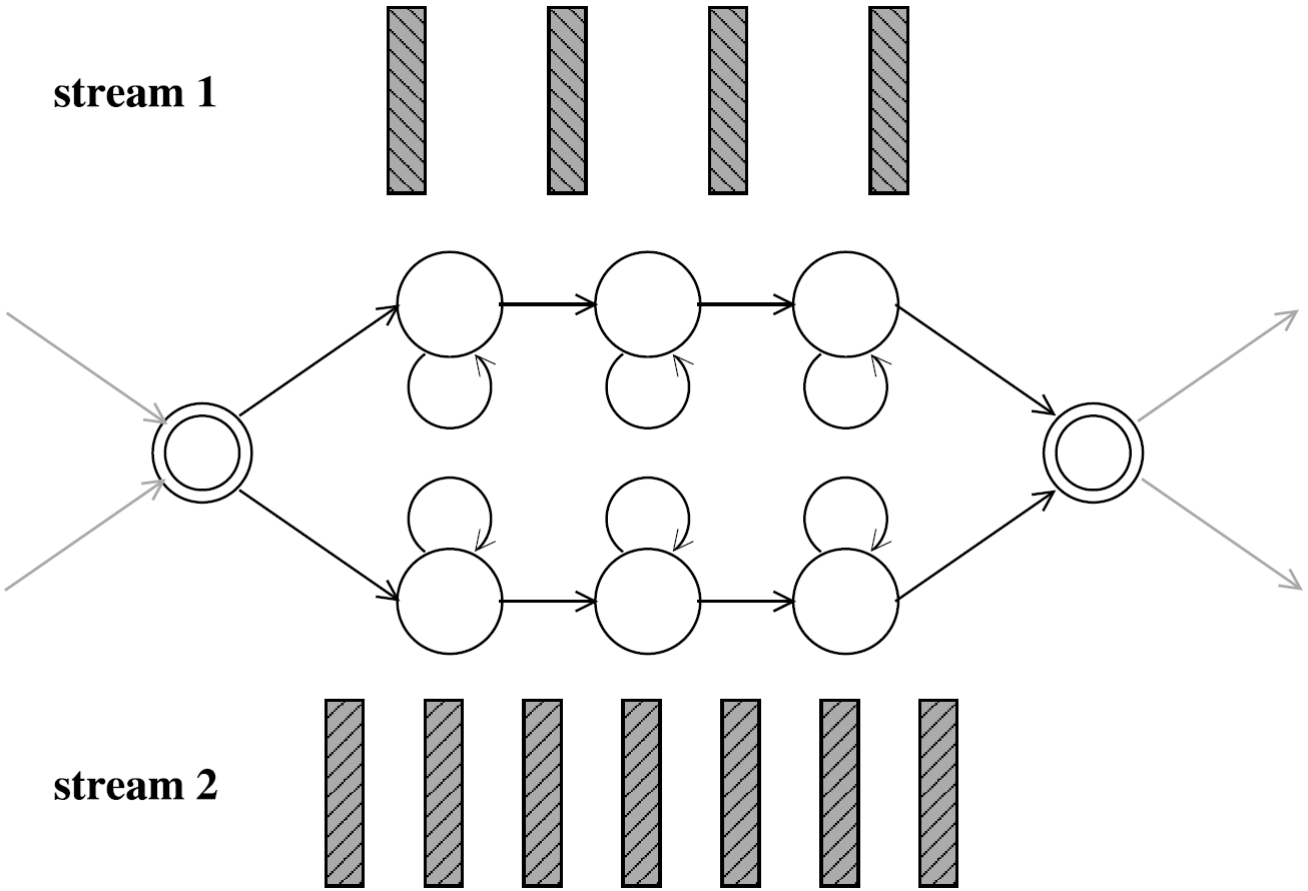
$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{x}) = p(x_0) \prod_{t=1}^{T} p(x_t|x_{t-1}) p(y_t^{(1)}|x_t) p(y_t^{(2)}|x_t)$$



[Fig. Credit: G. Gravier]

Asynchronous Multi-Stream HMMs

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = p(x_0^{(1)}, x_0^{(2)}) \cdot \prod_{t=1}^{T} p(x_t^{(1)}, x_t^{(2)} | x_{t-1}^{(1)}, x_{t-1}^{(2)}) p(y_t^{(1)}, y_t^{(2)} | x_t^{(1)}, x_t^{(2)})$$

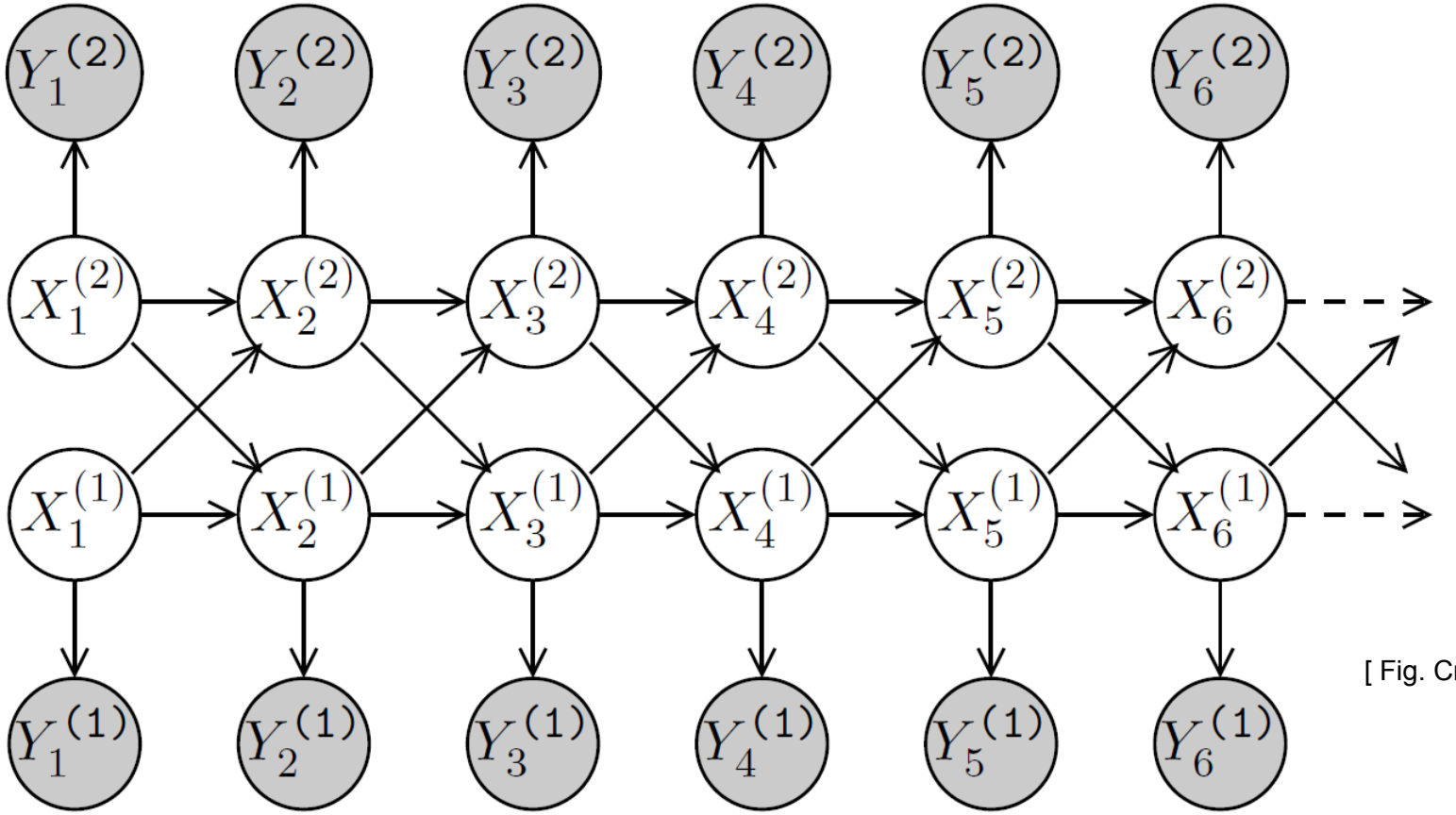


[Fig. Credit: G. Gravier]

DBNs: Coupled HMMs

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = p(x_0^{(1)}) p(x_0^{(2)}) \cdot$$

$$\prod_{t=1}^{T} p(x_t^{(1)} | x_{t-1}^{(1)}, x_{t-1}^{(2)}) p(x_t^{(2)} | x_{t-1}^{(1)}, x_{t-1}^{(2)}) p(y_t^{(1)} | x_t^{(1)}) p(y_t^{(2)} | x_t^{(2)})$$



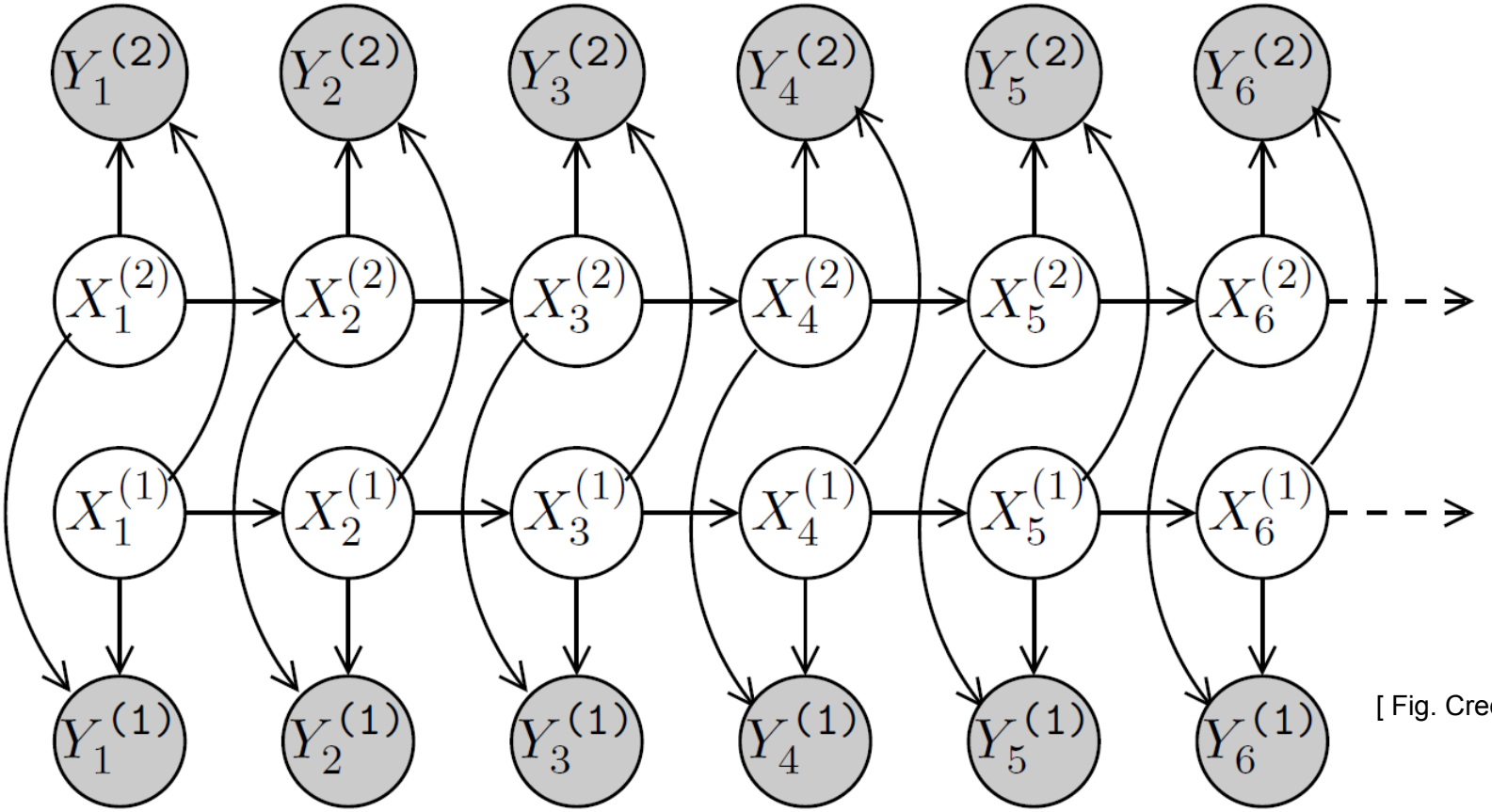
[Fig. Credit: G. Gravier]

[A. Nefian, L. Liang, X. Pi, X. Liu and K. Murphy, “Dynamic Bayesian Networks for Audio-Visual Speech Recognition”, EURASIP J. ASP 2002]

DBNs: Factorial HMMs

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = p(x_0^{(1)}) p(x_0^{(2)}) \cdot$$

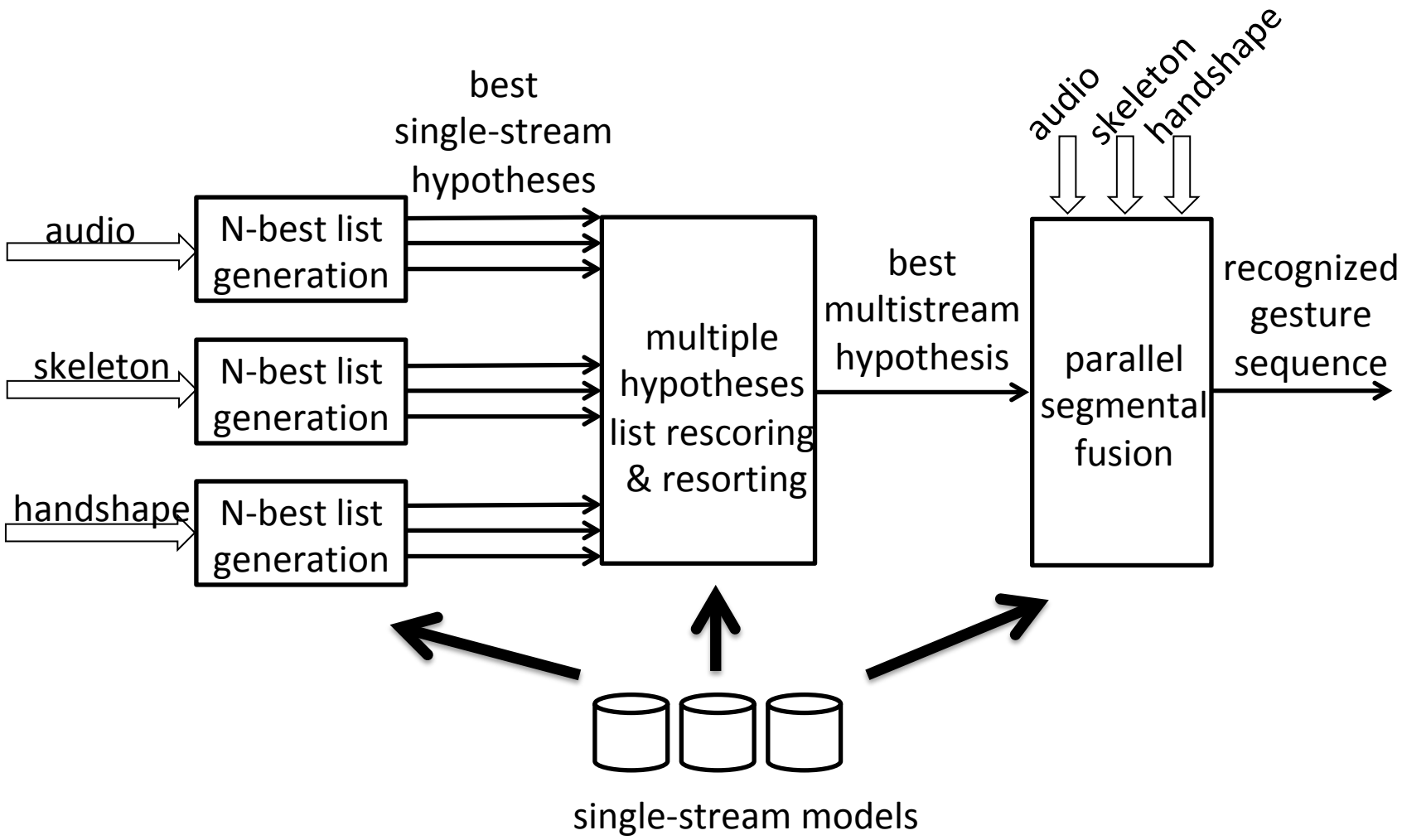
$$\prod_{t=1}^{T} p(x_t^{(1)} | x_{t-1}^{(1)}) p(x_t^{(2)} | x_{t-1}^{(2)}) p(y_t^{(1)} | x_t^{(1)}, x_t^{(2)}) p(y_t^{(2)} | x_t^{(1)}, x_t^{(2)})$$



[Fig. Credit: G. Gravier]

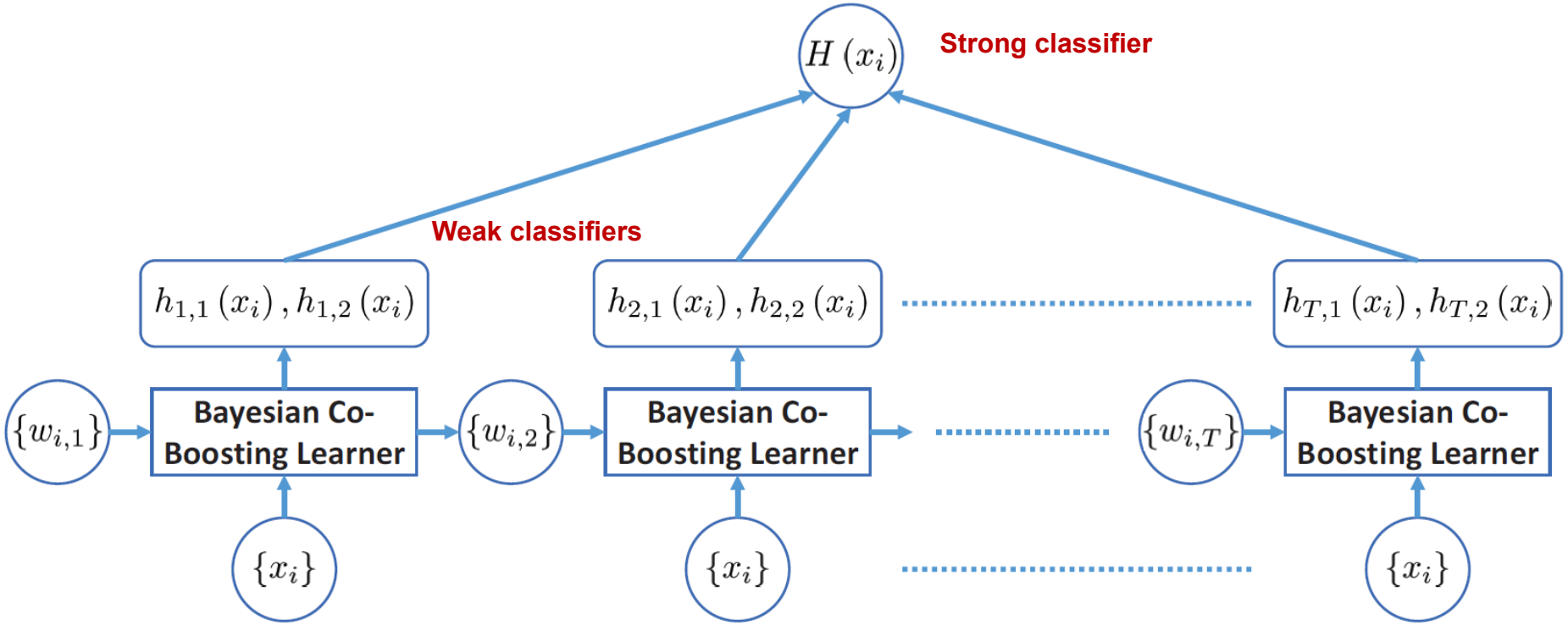
[A. Nefian, L. Liang, X. Pi, X. Liu and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition", EURASIP J. ASP 2002]

Multimodal Hypothesis Rescoring + Segmental Parallel Fusion



[V. Pitsikalis, A. Katsamanis, S. Theodorakis & P. Maragos, "*Multimodal Gesture Recognition via Multiple Hypotheses Rescoring*", JMLR 2015]

Bayesian Co-Boosting for Multimodal Gesture Recognition

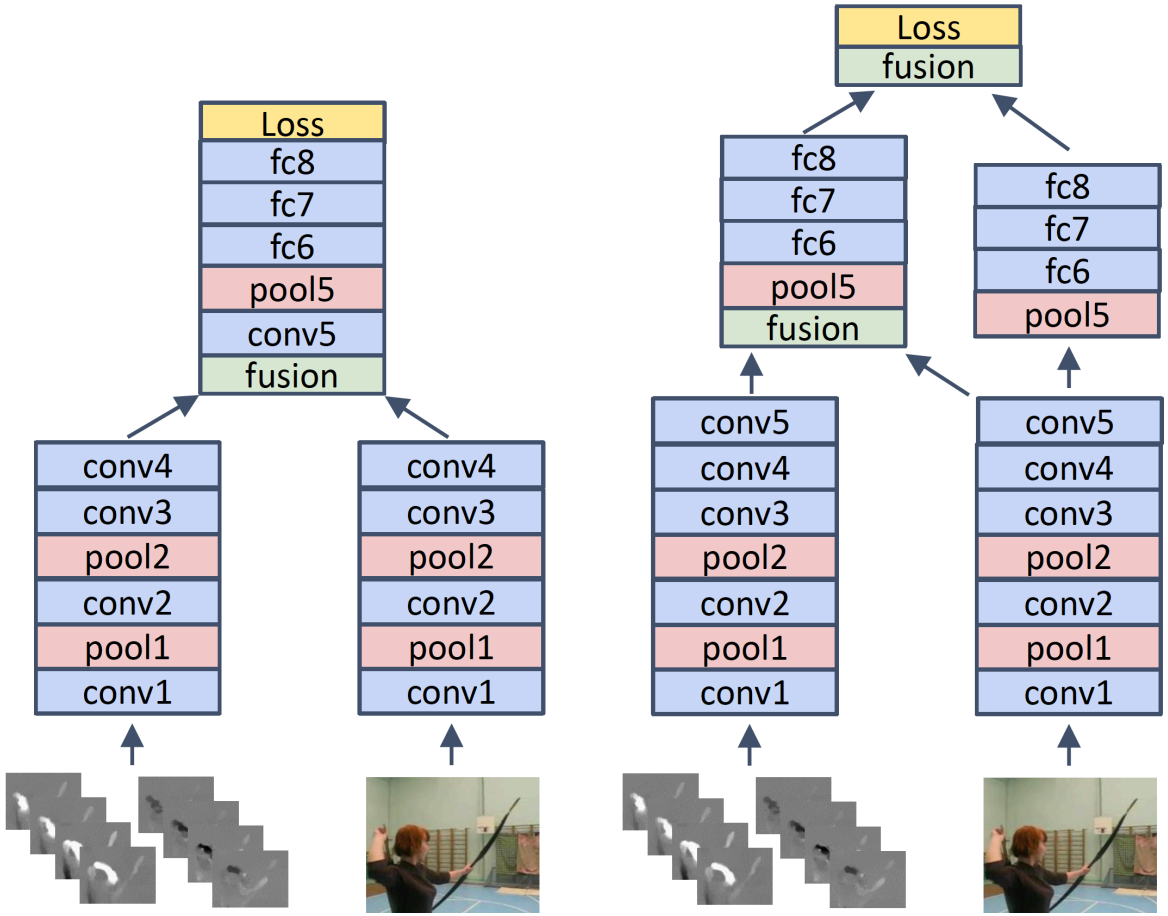


x_i: training instance; $w_{i,t}$: training instance x_i's weight at the t-th iteration; $h_{t,v}(x_i)$: weak classifier learnt from modality v at the t-th iteration; $H(x_i)$: final strong classifier.

[J. Wu and J. Cheng, "*Bayesian Co-Boosting for Multi-modal Gesture Recognition*", JMLR 2014]

Two-Stream CNN-based Fusion for Action Recognition

- Two-Stream CNN
  - RGB
  - Optical Flow
- Fusion after conv4 layer
  - single network tower
- Fusion at two layers (after conv5 and after fc8)
  - both network towers are kept
  - one as a hybrid spatiotemporal net
  - one as a purely spatial network



[C. Feichtenhofer, A. Pinz and A. Zisserman, "*Convolutional two-stream network fusion for video action recognition*", CVPR 2016.]

Audio-Visual Speech Recognition

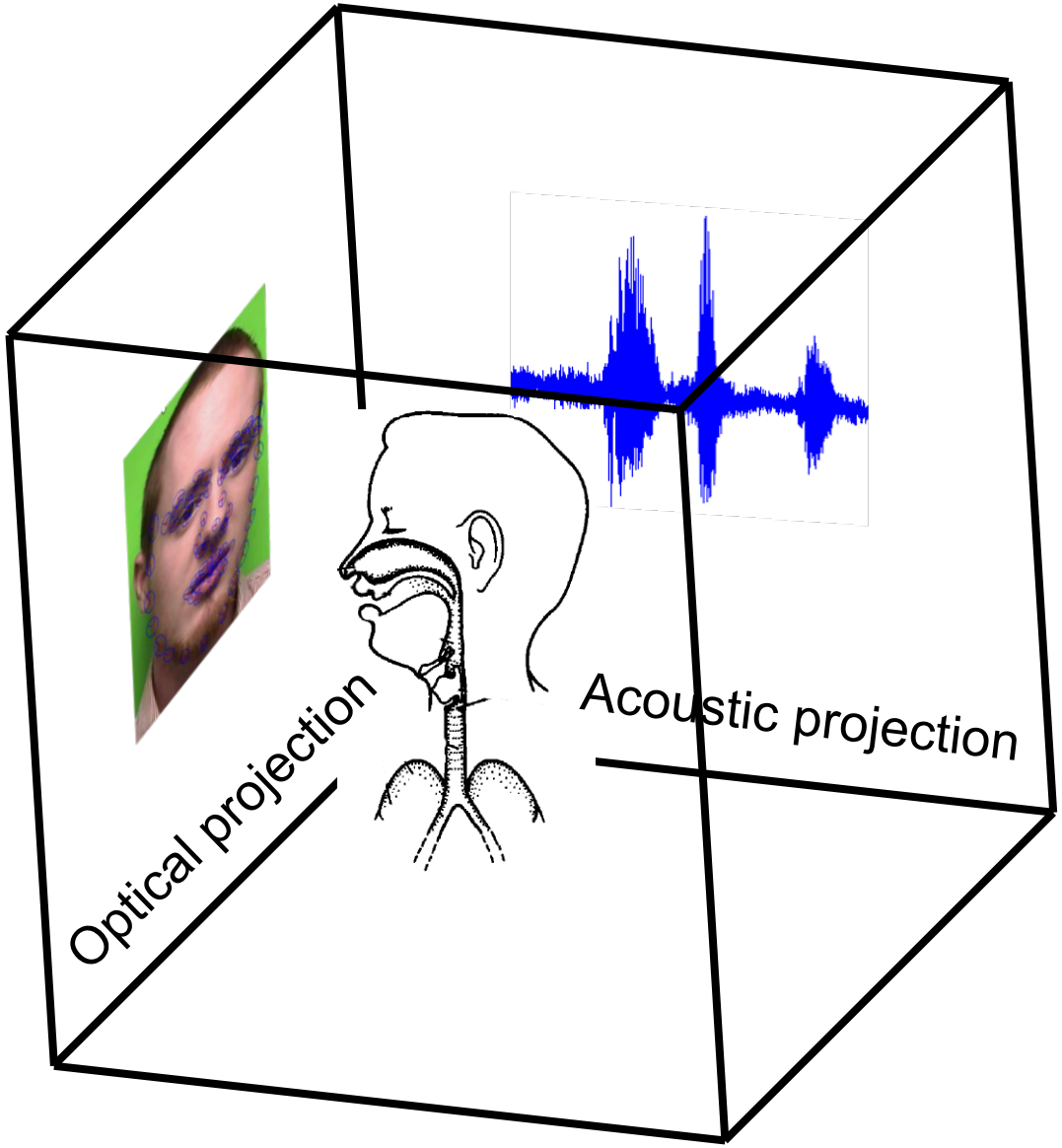
Main reference:

- [G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "*Adaptive Multimodal Fusion by Uncertainty Compensation with Application to Audio-Visual Speech Recognition*", IEEE Trans. Audio, Speech & Lang. Proc., 2009.]

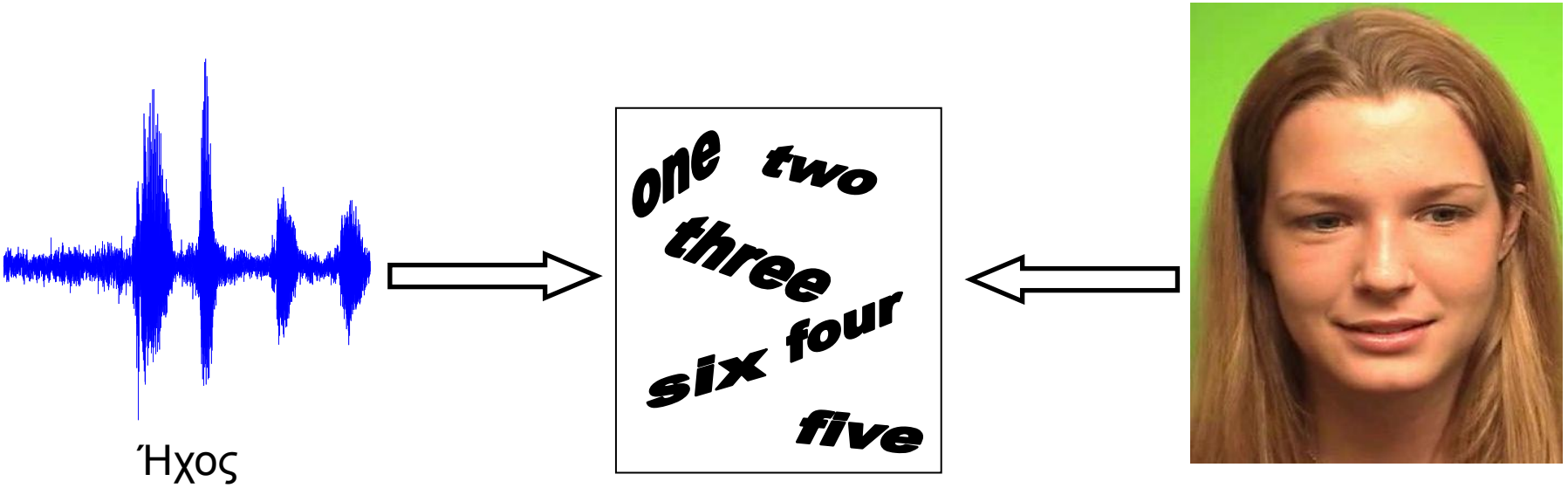
General References:

- [G. Potamianos, C. Neti, G. Gravier, A. Garg and A. Senior, "*Recent Advances in the Automatic Recognition of Audiovisual Speech*", Proc. IEEE 2003.]
- [P. Aleksic and A. Katsaggelos, "*Audio-Visual Biometrics*", Proc. IEEE 2006.]
- [P. Maragos, A. Potamianos and P. Gros, *Multimodal Processing and Interaction: Audio, Video, Text,* Springer-Verlag, 2008.]
- [D. Lahat, T. Adali and C. Jutten, "*Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects*", Proc. IEEE 2015.]
- [A. Katsaggelos, S. Bahaadini and R. Molina, "*Audiovisual Fusion: Challenges and New Approaches*", Proc. IEEE 2015.]
- [G. Potamianos, E. Marcheret, Y. Mroueh, V. Goel, A. Koumbaroulis, A. Vartholomaios, and S. Thermos, "*Audio and visual modality combination in speech processing applications*", In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Kruger, eds., The Handbook of Multimodal-Multisensor Interfaces, Vol. 1: Foundations, User Modeling, and Multimodal Combinations. Morgan Claypool Publ., San Rafael, CA, 2017.]

Speech: Multi-faceted phenomenon

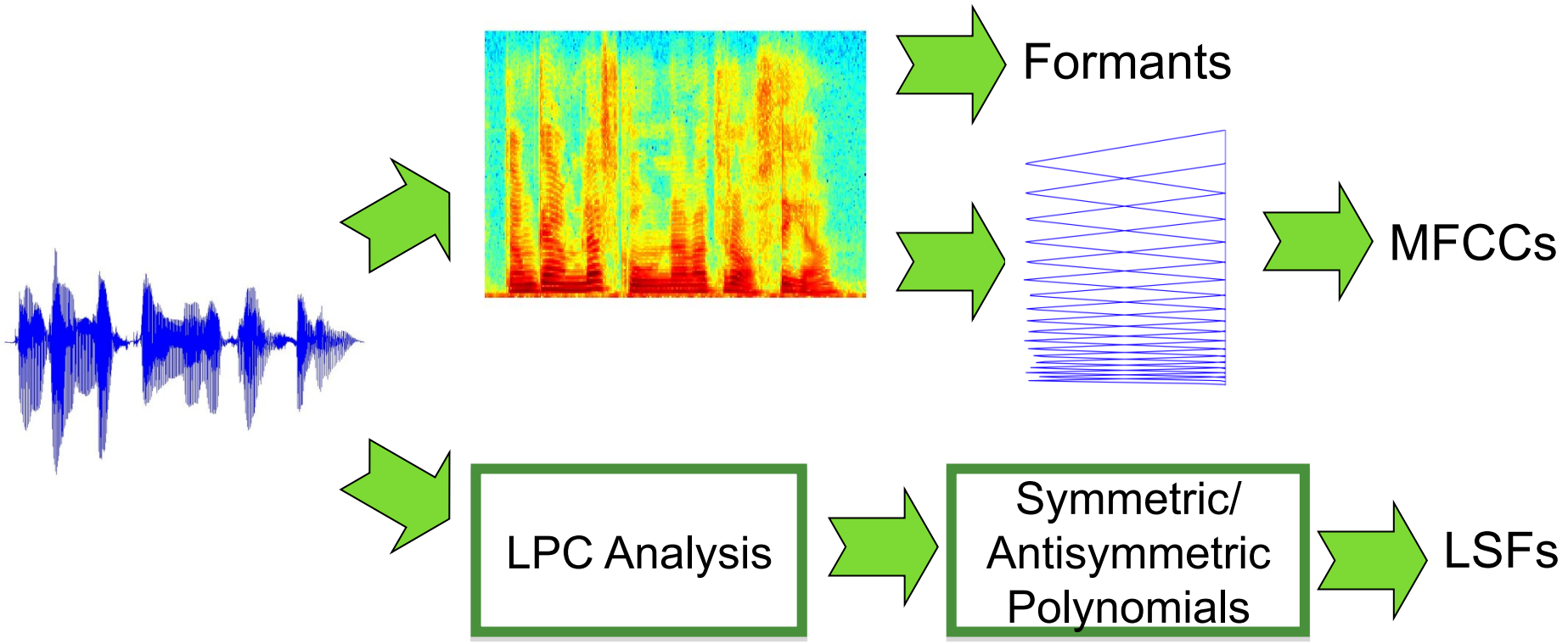


Recognizing Speech from Audio and Video



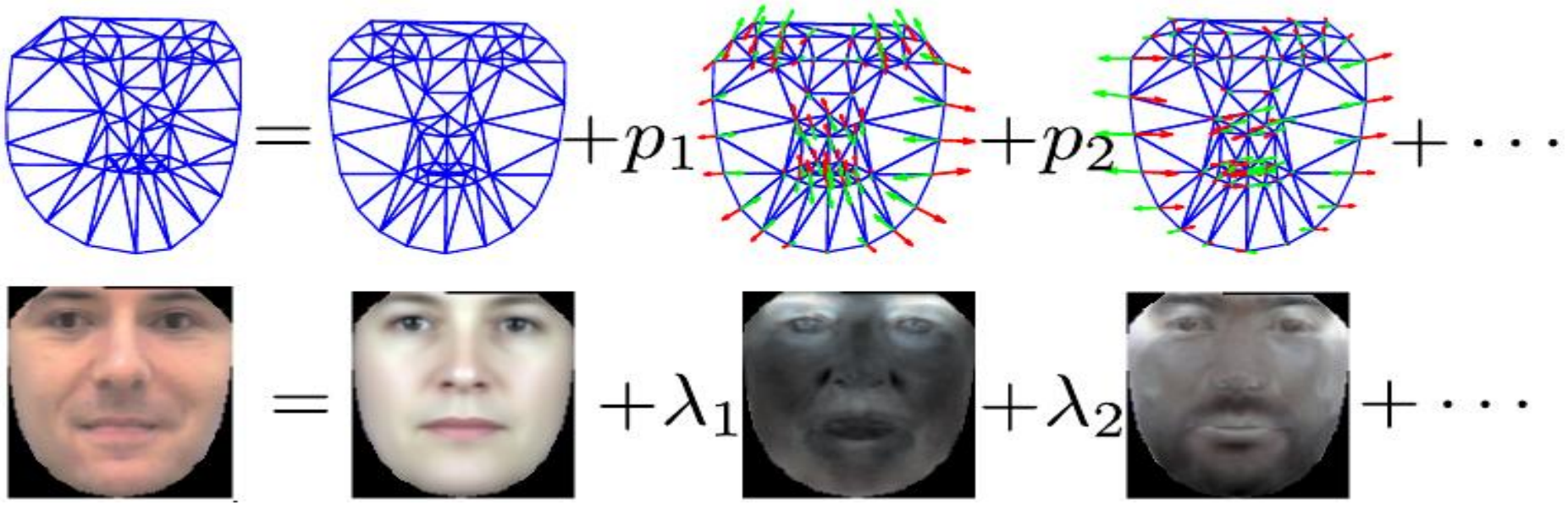
- A fundamental phenomenon in speech perception (McGurk & MacDonald)
- Improving Automatic Speech Recognition (ASR) systems performance in adverse acoustical conditions:
  - Noise, Interferences

Audio Feature Extraction

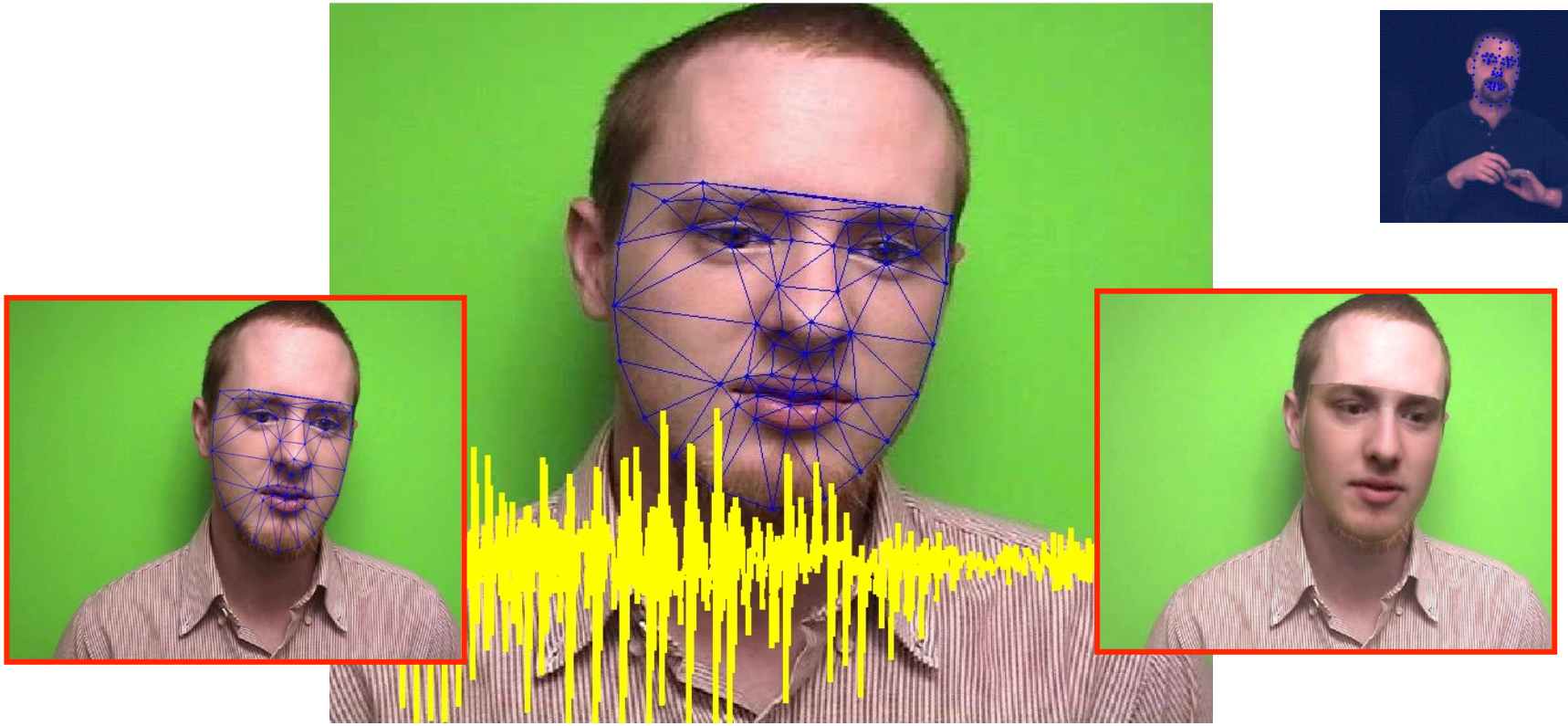


Visual Feature Extraction: Active Appearance Modeling of Visible Articulators

- Active Appearance Models for face modelling
- Shape & Texture related articulatory information
- Features: AAM Fitting (nonlinear least squares problem)
- Real-Time, marker-less facial visual feature extraction



Example: Face Analysis and Tracking Using AAM



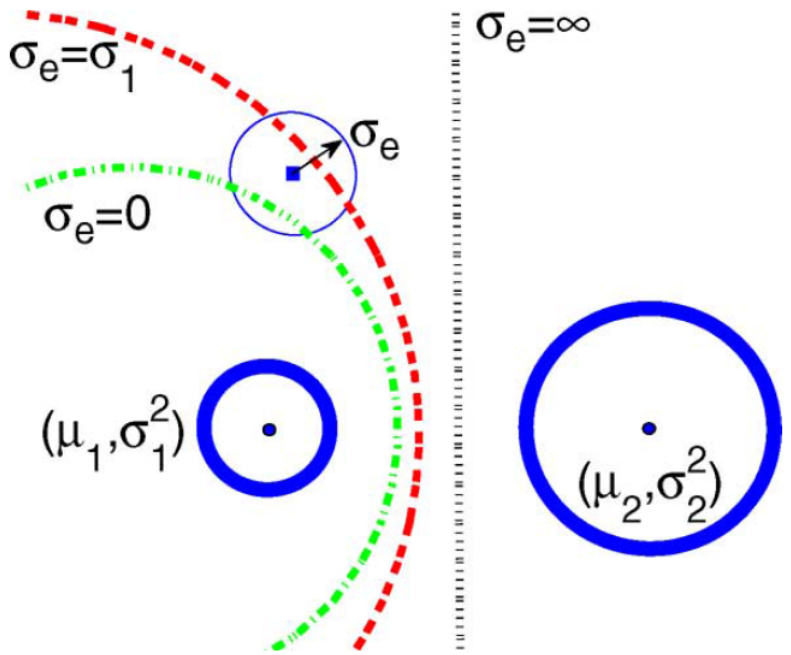
original

shape tracking

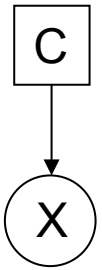
reconstructed face

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

Measurement Noise and Adaptive Fusion

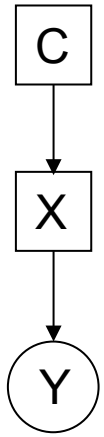


Conventional View: Features are directly observable



$$p(c|x_{1:S}) \propto p(c)\prod_{s=1}^{S} p(x_s|c)$$

Our View: We can only measure noise-corrupt features

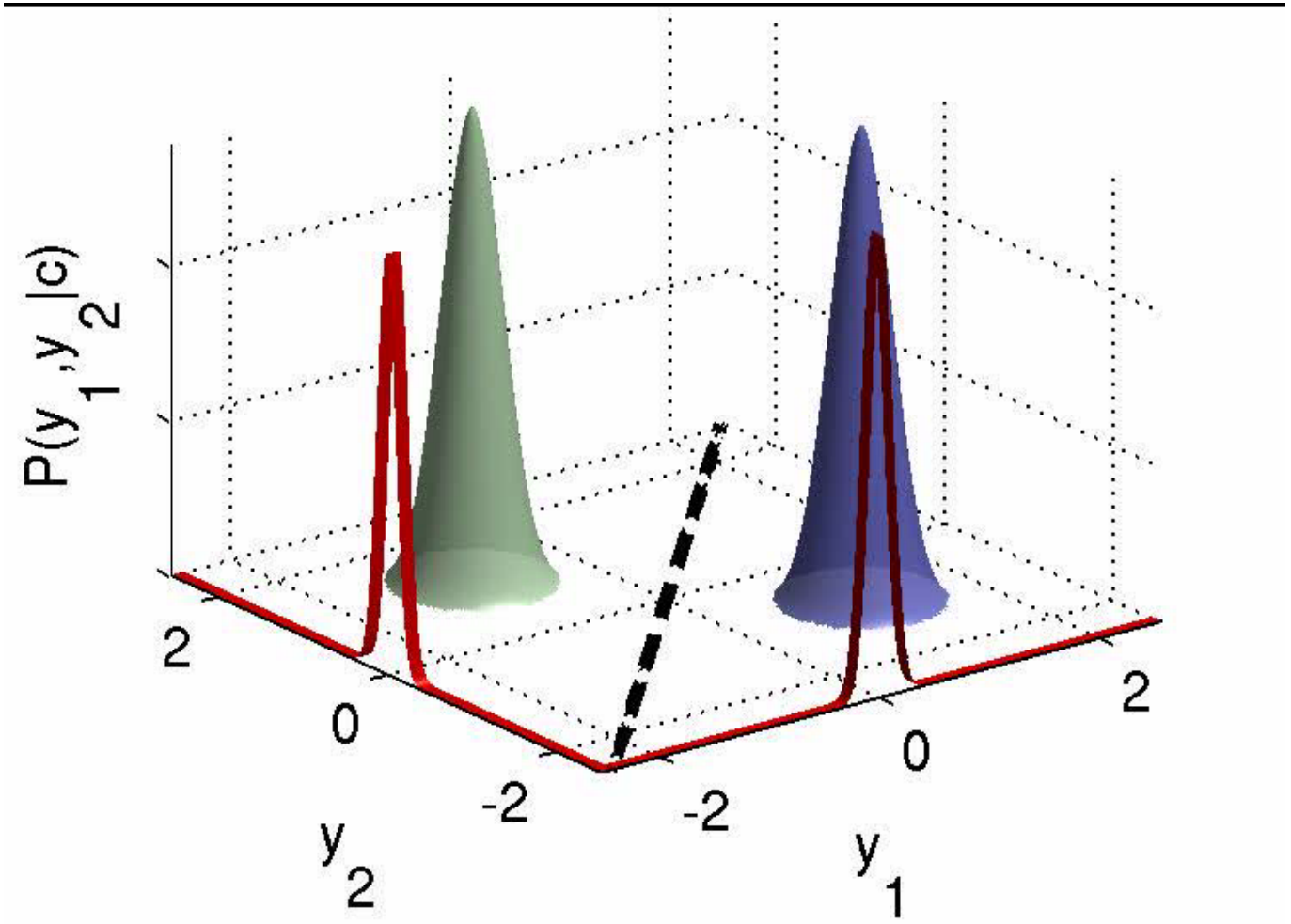


$$p(c|y_{1:S}) \propto p(c)\prod_{s=1}^{S}\int p(x_s|c)p(y_s|x_s)dx_s$$

$$p(c\,|\,y_{1:s}) \propto p(c)\prod_{s=1}^{S}\sum_{m=1}^{M_{s,c}} \rho_{s,c,m} N\left(y_s;\mu_{s,c,m}+\mu_{e,s},\Sigma_{s,c,m}+\Sigma_{e,s}\right)$$

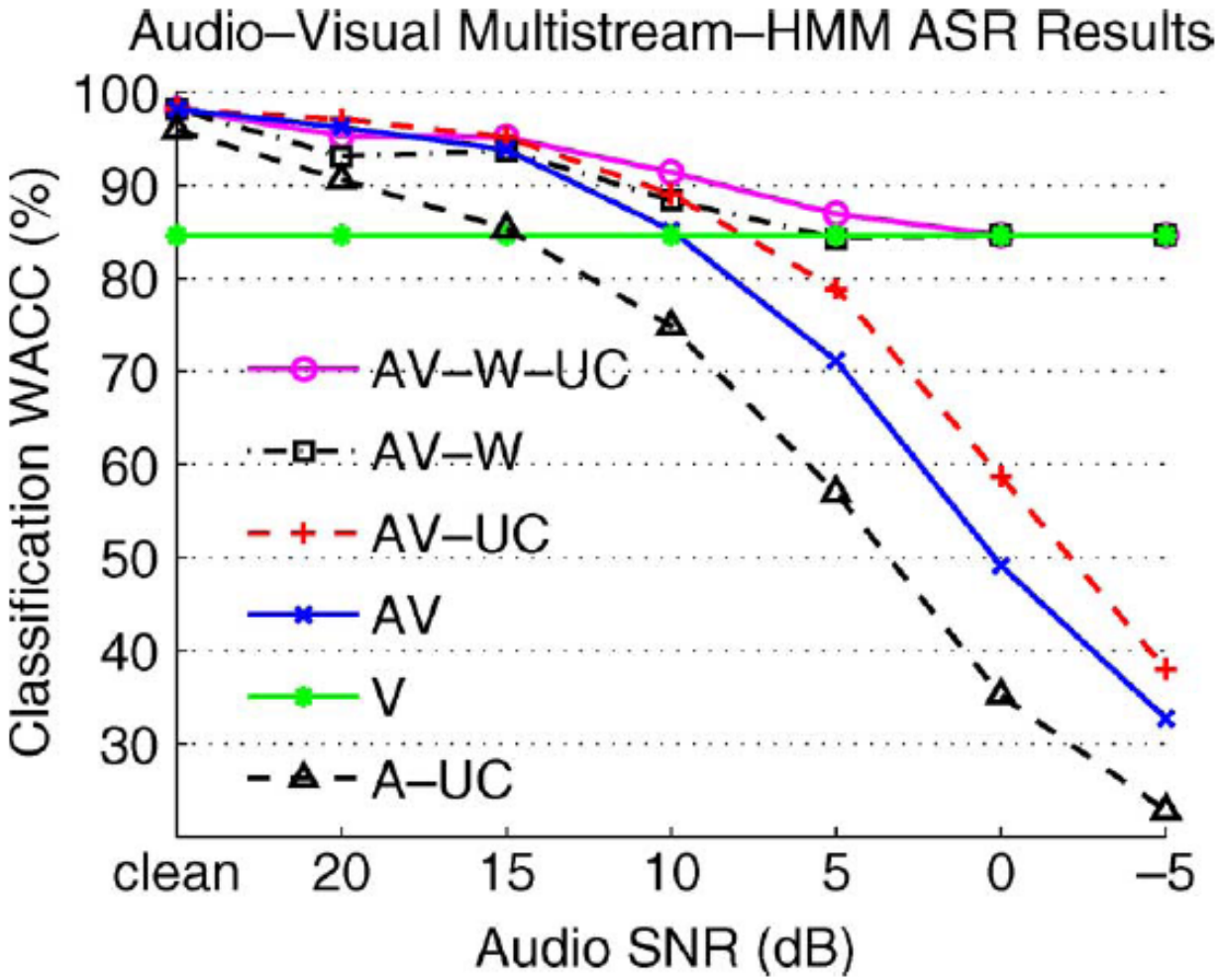
Demo: Fusion by Uncertainty Compensation

- Classification decision boundary w. increasing uncertainty
  - Two 1D streams (y1 and y2-streams), 2 classes



AV-ASR Evaluation on CUAVE Database

Audio-Visual Recognition



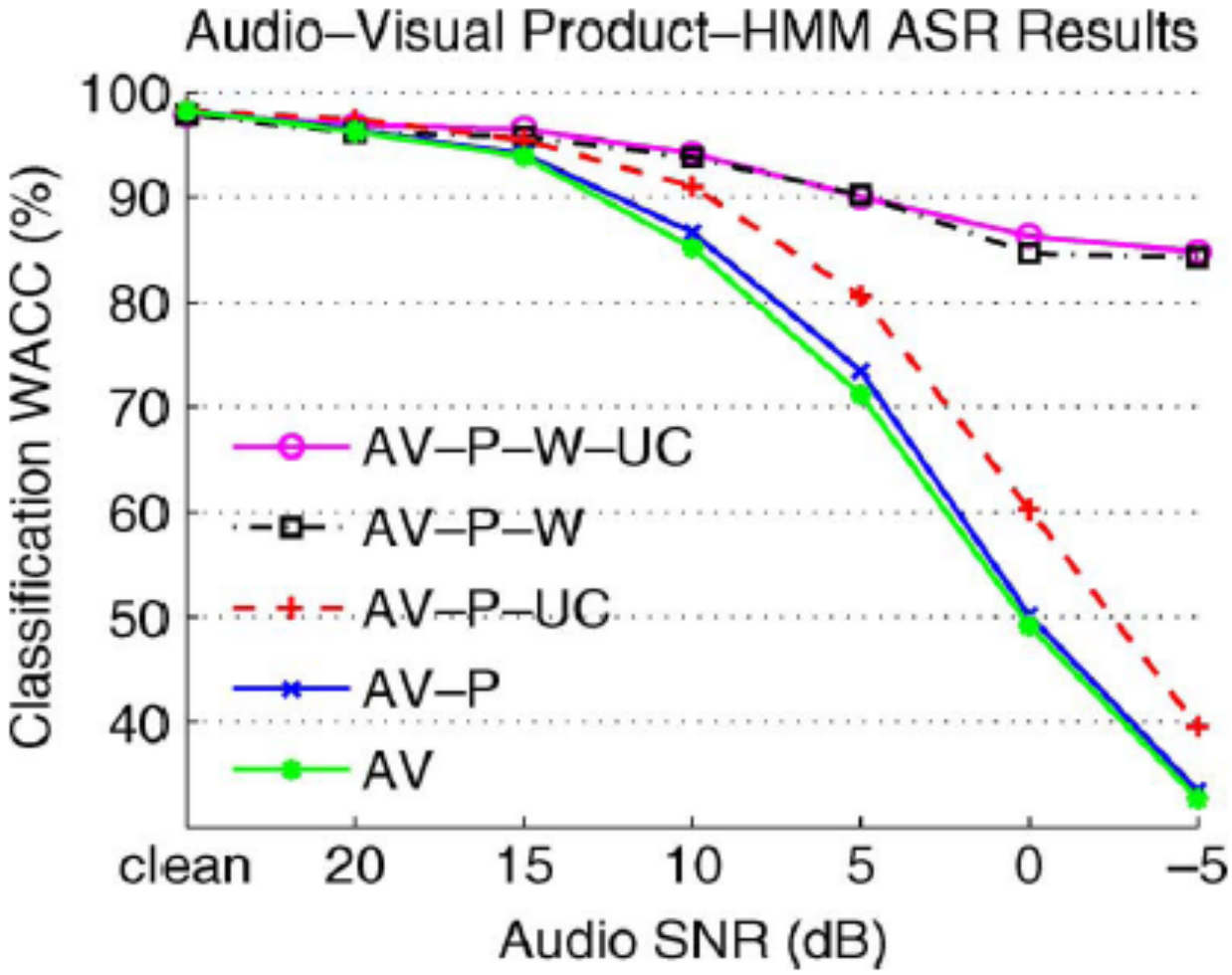
Average Absolute Improvement due to Visual information

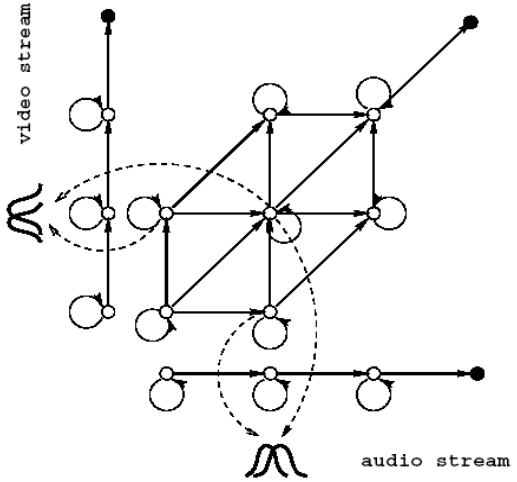
AV-W-UC vs. A-UC

28.7 %

- Weights and Uncertainty Compensation
- Hybrid Fusion Scheme

Asynchrony Modeling with Product-HMMs

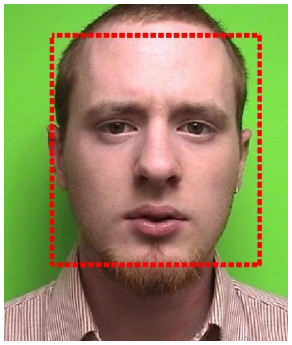




Average absolute improvement due to modeling with Product-HMM vs. Multistream-HMM

1.2 %

A Real-Time AV-ASR Prototype



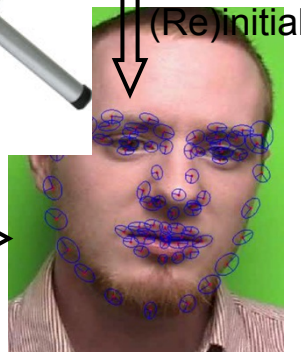
Face detector

Adaboost-based, @5 fps

Image Acquisition

Firewire color camera, 640x480 @25 fps

(Re)initialization



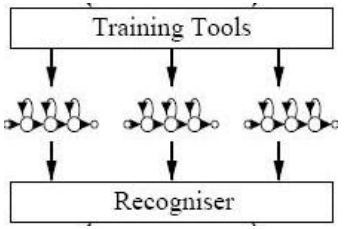
Face tracking & feature extraction

Real-time AAM fitting algorithms

System Overview

GPU-accelerated processing

OpenGL implementation



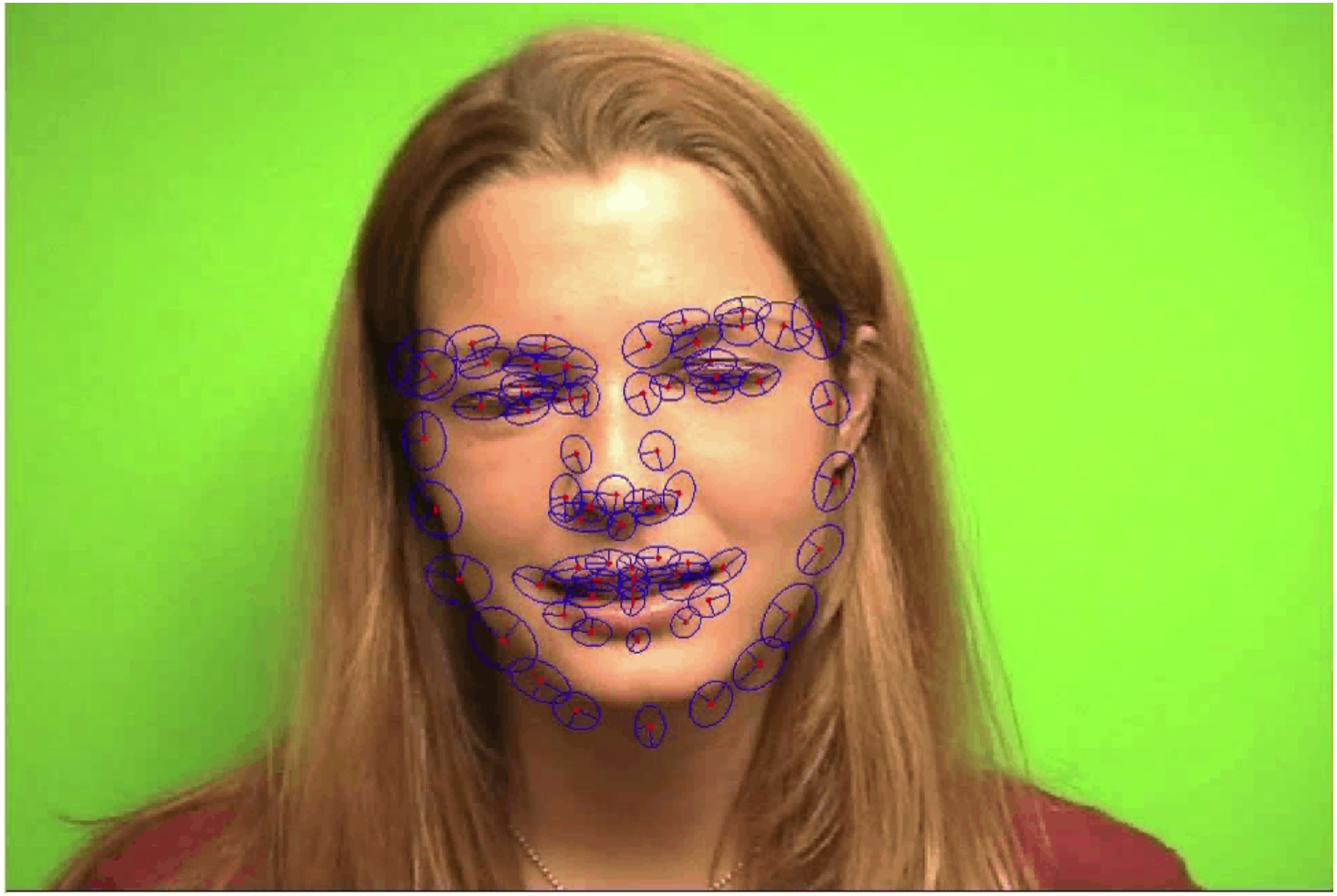
HMM-based backend

Transcription

Audio-Visual Speech Recognition Demo

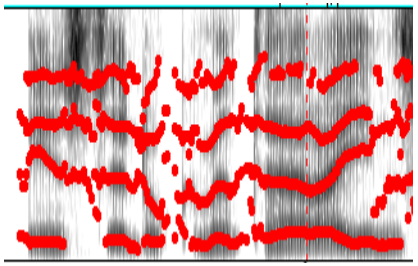
(WACC: AV=89%, A=74% at 5 dB SNR babble noise)

AV

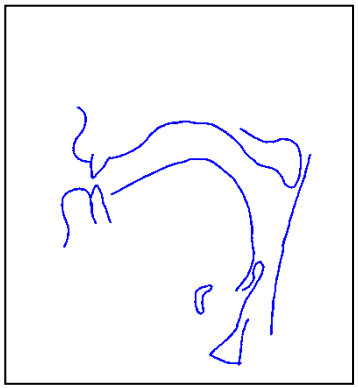


A

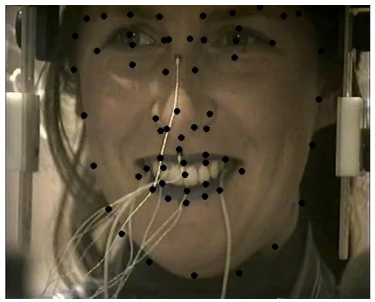
Audio-Visual Recovery of Vocal Tract Geometry



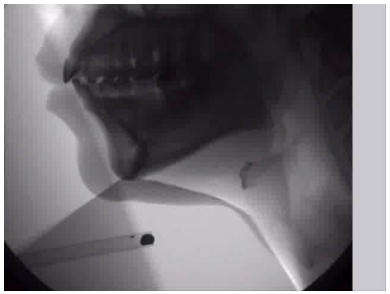
Acoustics



Vocal tract Geometry



Images



- Applications:
  - Speech Mimics
  - Articulatory ASR
  - Speech Tutoring
  - Phonetics

[A. Katsamanis, G. Papandreou, and P. Maragos, "*Face Active Appearance Modeling and Speech Acoustic Information to Recover Articulation*", IEEE Trans. ASLP 2009.]

Emotion-Expressive Audio-Visual Speech Synthesis

References:

- [P.P. Filntisis, A. Katsamanis and P. Maragos, "*Photo-realistic Adaptation and Interpolation of Facial Expressions Using HMMs and AAMs for Audio-visual Speech Synthesis"*, ICIP 2017.]

- [P.P. Filntisis, A. Katsamanis, P. Tsiakoulis and P. Maragos, "*Video-Realistic Expressive Audio-Visual Speech Synthesis for the Greek Language*", Speech Communication, 2017.]

Expressive Audio-Visual Speech Synthesis (EAV-TTS)

- A virtual/physical agent employing expressive speech is more natural
- [SpeCom 2017]: Given a text to be synthesized we use DNNs to find the corresponding output visual and acoustic features.
- HMM adaptation to adapt EAV-TTS system to unseen **emotions** [ICIP 2017]
- HMM interpolation to generate speech with **mixed** expressions [ICIP 2017]

HMM-based EAV-TTS [ICIP-2017]

Linguistic Features

494-dim feature vector with lexicological info: phoneme, vowel, # of syllables of sentence, relative location, etc.

Visual Features

Face shape $s = s + \sum_{i=1}^{n} p_i s_i$

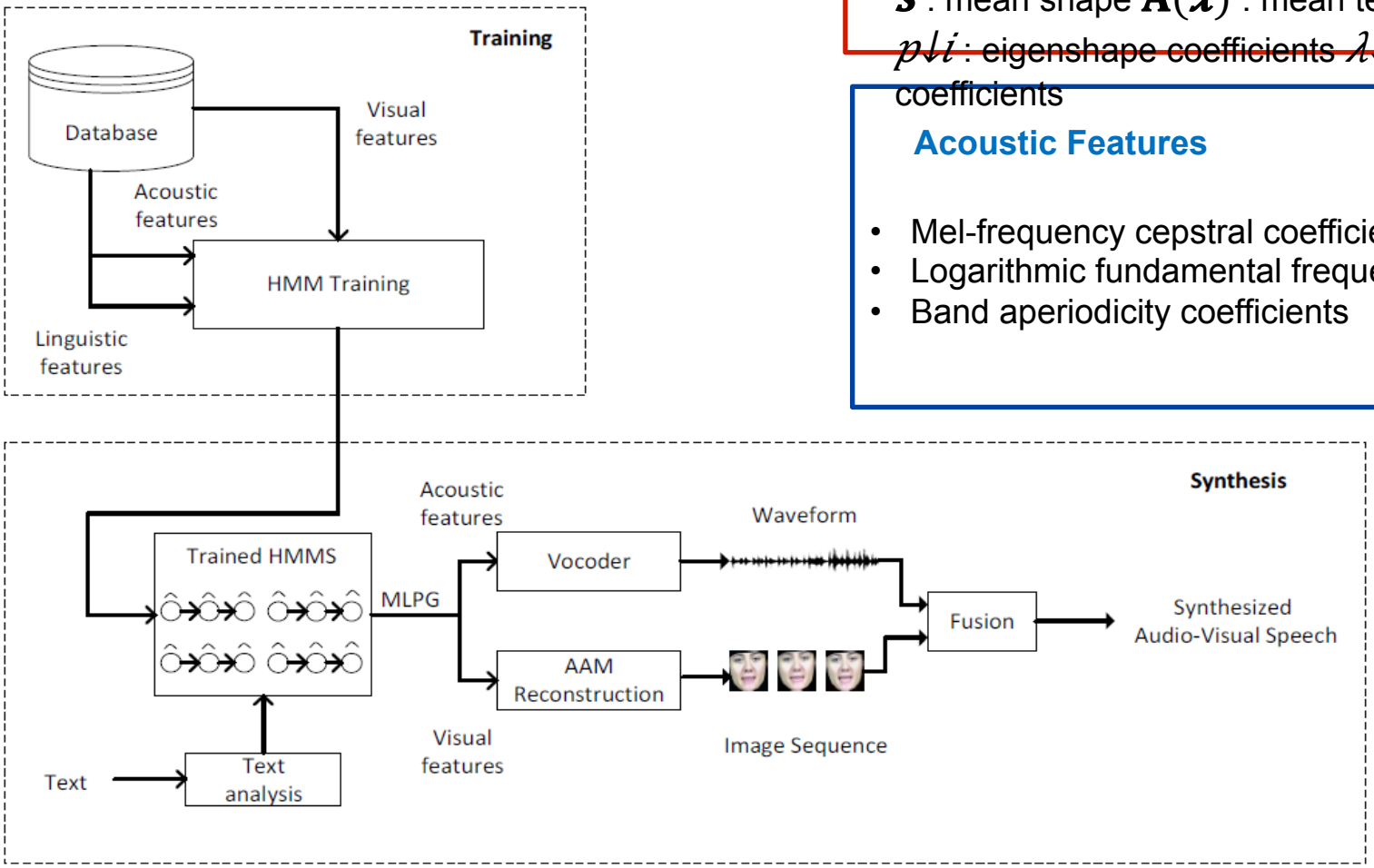
Face texture $A(x) = A(x) + \sum_{i=1}^{m} \lambda_i A_i(x)$

s : mean shape $\mathbf{A}(x)$: mean texture

p_i : eigenshape coefficients λ_i : eigentexture coefficients

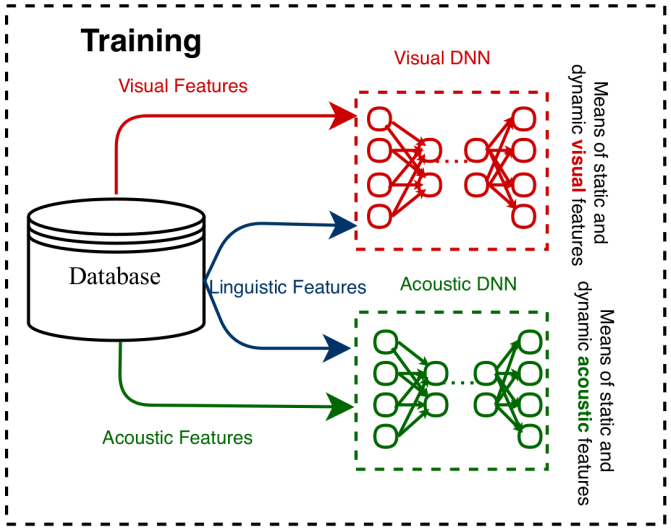
Acoustic Features

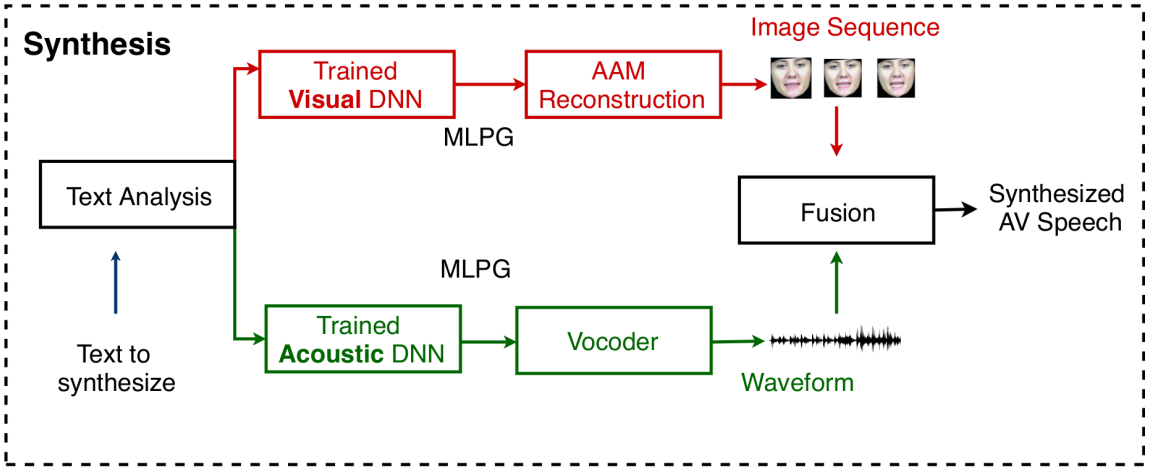
- Mel-frequency cepstral coefficients
- Logarithmic fundamental frequency
- Band aperiodicity coefficients



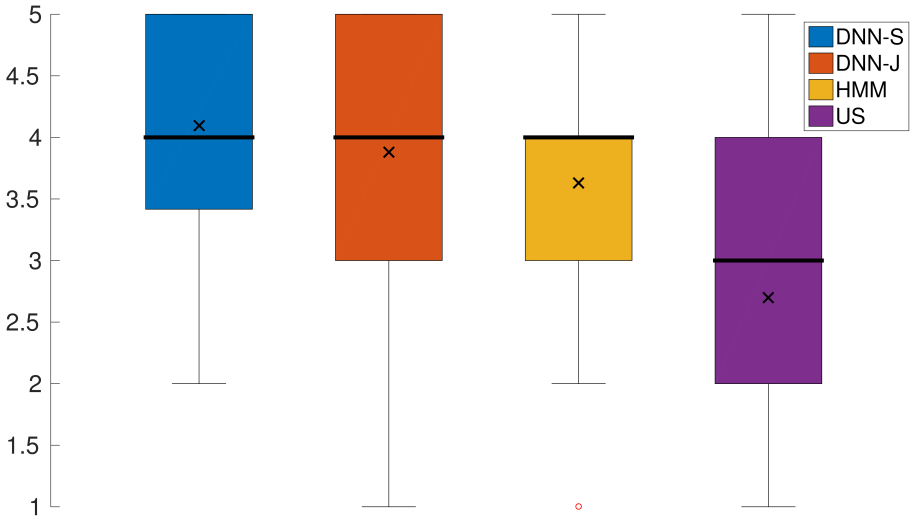
HMM Architecture

DNN-Based Audio-Visual Speech Synthesis [SpeCom 2017]





MOS Evaluation



Results show **significant preference** of DNN methods on audio-visual realism and **significant preference** of DNN-S method on audio-visual expressiveness

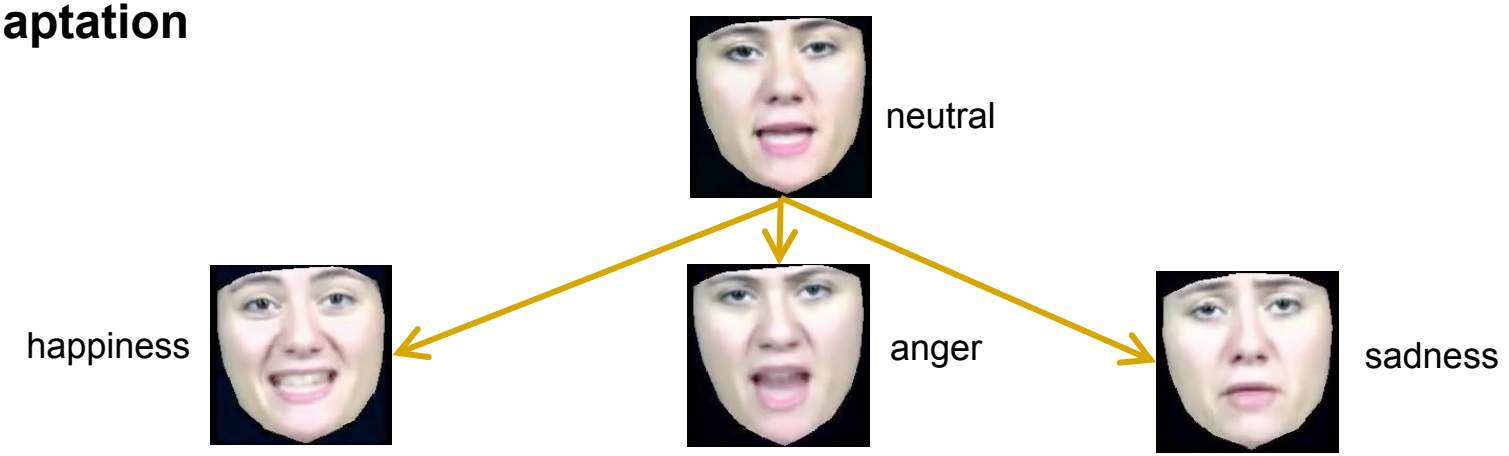
Two architectures:
joint modeling of acoustic/visual features (DNN-J)
separate modeling of acoustic/visual features (DNN-S)

Box plot of MOS tests of audio-visual realism

EAV-TTS: HMM Adaptation - Interpolation

Tackle data sparsity by using HMMs for **Audiovisual Adaptation** and **Interpolation**

Adaptation

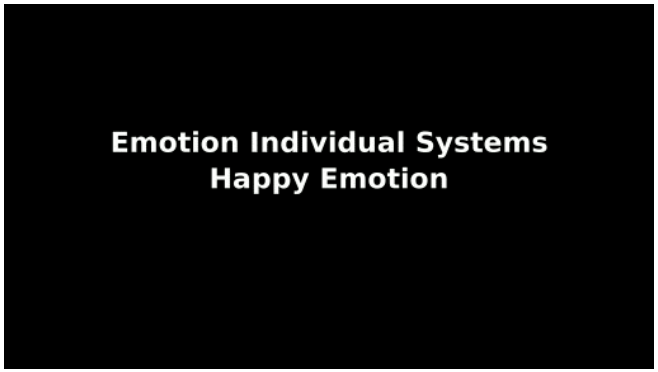


Interpolation

*Interpolating the **anger** and **happiness** HMM sets. (respective weights shown under each image).*

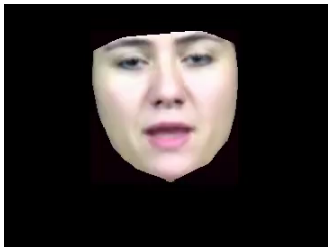
EAV-TTS: Example Videos (in Greek)

"You should have listened to my first album"

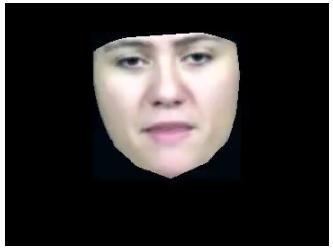


General Comparison

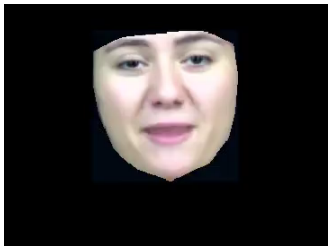
"He has all of Olympiacos dollars in front of him"



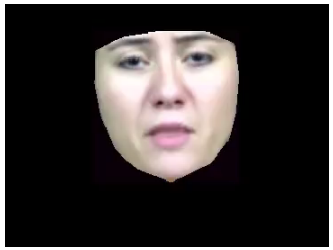
Neutral (DNN-S)



Anger (DNN-S)

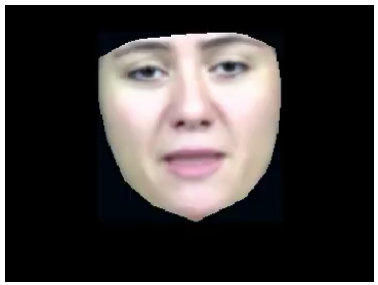


Happiness (DNN-S)



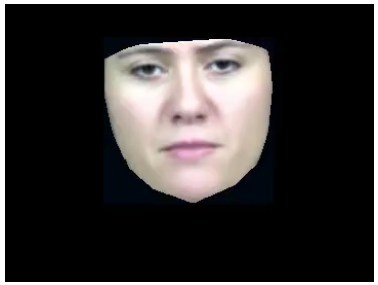
Sadness (DNN-S)

"*What are you talking about, why did he go to the doctor's office*"



Happy – Sad Interpolation

"I have learned to accept everything in my life"



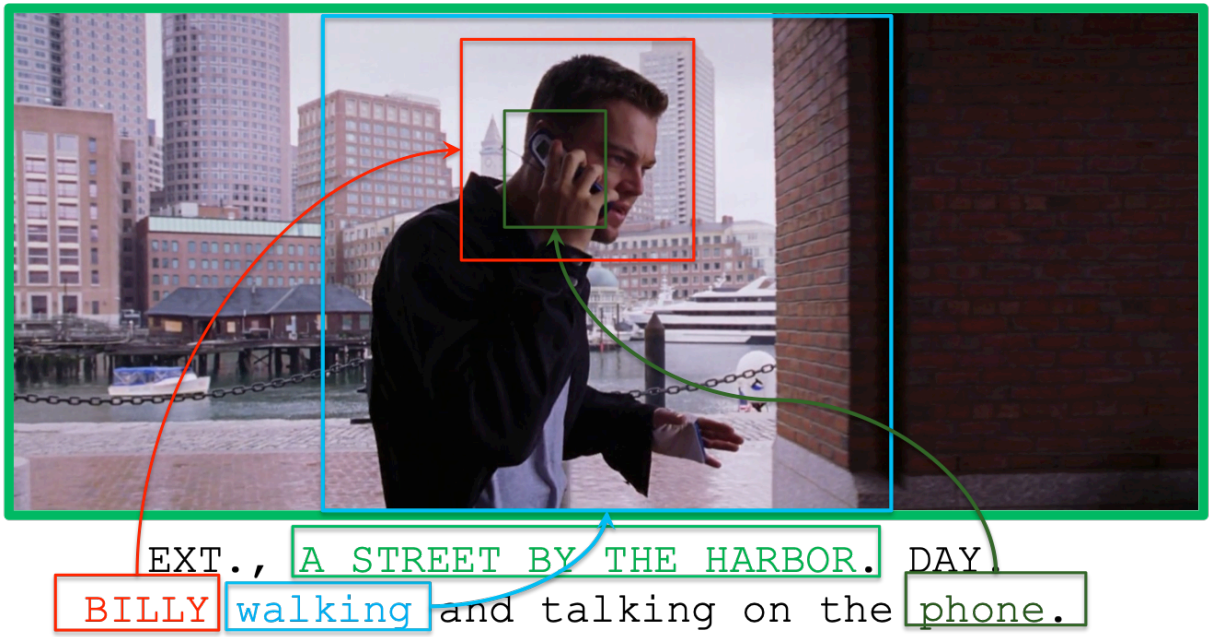
Neutral adapted to Anger with 50 sentences

Multimodal (Visual + Textual) Concept Learning in Videos with Weakly Supervised Techniques

Visual Concepts

- Detect and recognize visual concepts in videos in a weakly supervised manner, mining their labels from an accompanying descriptive text.
- **Visual Concepts**: *Spatio-temporally localized video segments that carry a specific structure in the visual domain.*

1. Faces
2. Actions
3. Scenes
4. Objects



G. Bouritsas, P. Koutras, A. Zlatintsi and P. Maragos, Multimodal Visual Concept Learning with Weakly Supervised Techniques, CVPR 2018

Weak Supervision with Natural Language

- **Motivation:**
  - **Why Natural Language?**
    - Rich semantics – interpretable – easy to extract.
  - **Why Weak Supervision?**
    - Reduce the time-consuming and costly procedure of manual annotation.

a) Achieve recognition in data annotated sparsely/imprecisely.

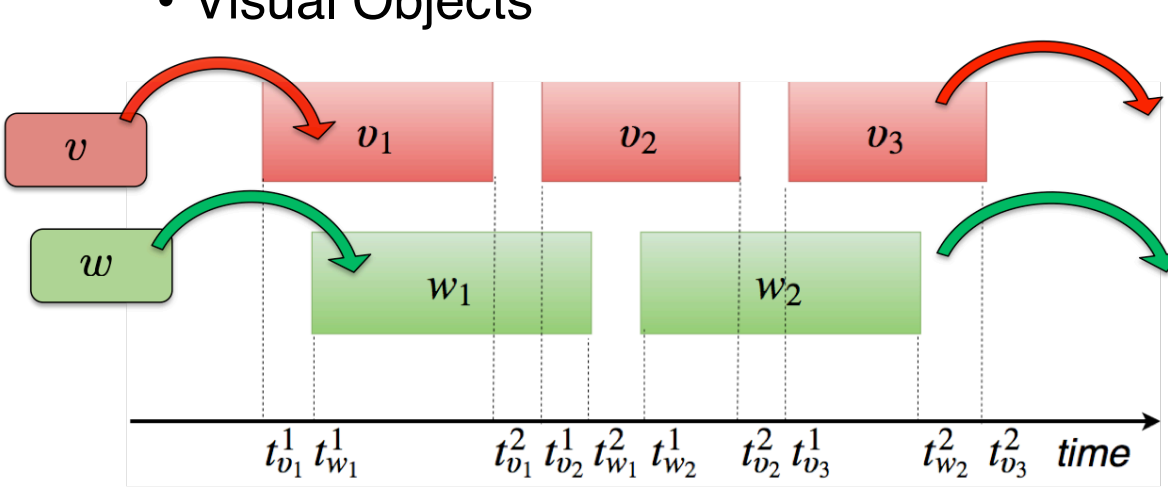
b) Collect new data to train fully supervised models.

- **Challenges:**
  - Spatio-Temporal ambiguity: absence of specific spatio-temporal correspondence between visual and textual objects.
  - Semantic ambiguity: Words/sentences may have several different meanings.

Multimodal Visual Concept Learning

- **Dual Modality scheme:** Two data streams flowing in parallel.

  - Visual Objects



  - Representation: features $\boldsymbol{x_v}$
  - Temporal interval: $T_v = [t_v^1, t_v^2]$
  - Class: latent variable $z_v = ??$
  - Implied label: $y_w = ??$
  - Temporal interval: $T_w = [t_w^1, t_w^2]$

  - Textual Objects

- Extend Discriminative clustering model (DIFFRAC)

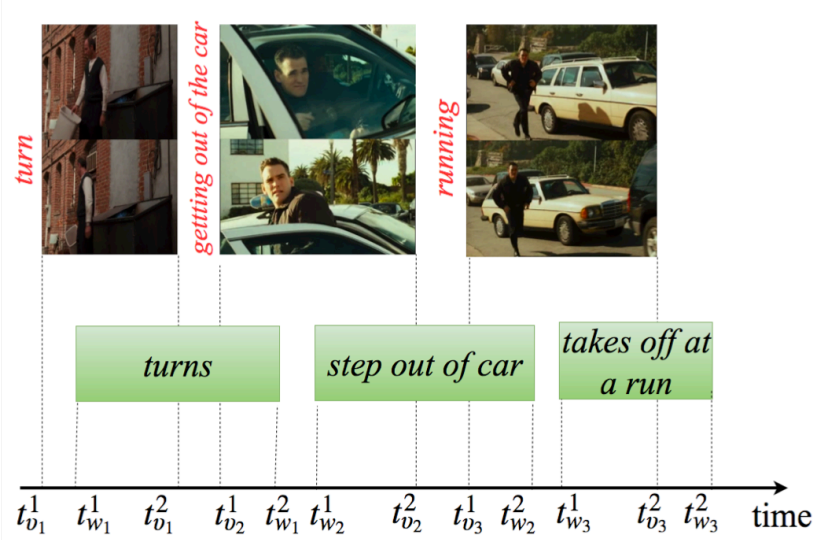
Weakly Supervised frameworks

- Fuzzy Sets MIL (**FSMIL**): Fuzzy bags of Multiple Instances.

$$\mathcal{V}_w = \{(v, \mu_w(v)) \mid v \in \mathcal{V}, \mu_w(v) = g(\frac{|T_w \cap T_v|}{|T_v|})\}$$

Visual objects overlapping with the textual one

Membership grade



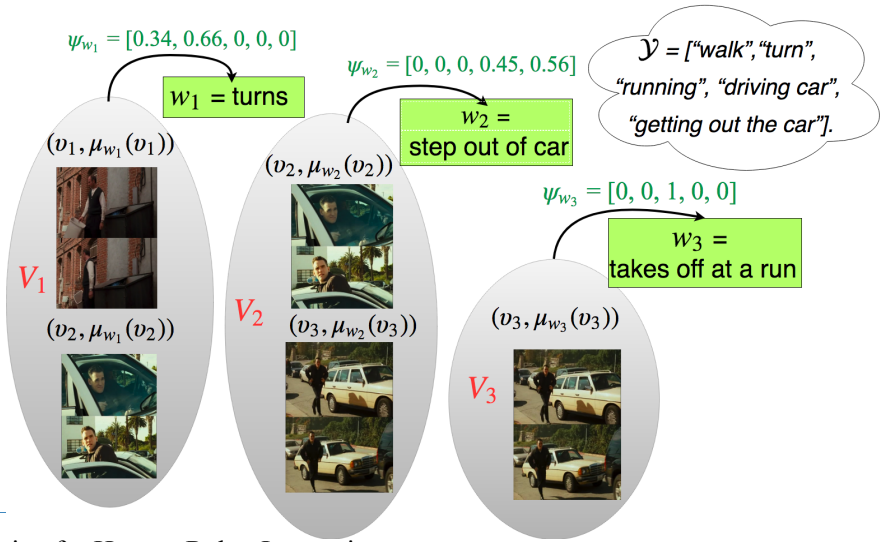
- Probabilistic Label MIL (**PLMIL**):

Each bag is assigned a probabilistic label.

Unsupervised estimation via semantic similarity.

$$\psi_w(y) = \mathbb{P}[y_w = y|w]$$

$$\psi_w(y) = s_{wy} / \sum_{\ell \in \mathcal{Y}} s_{w\ell}$$



Results: Face Recognition

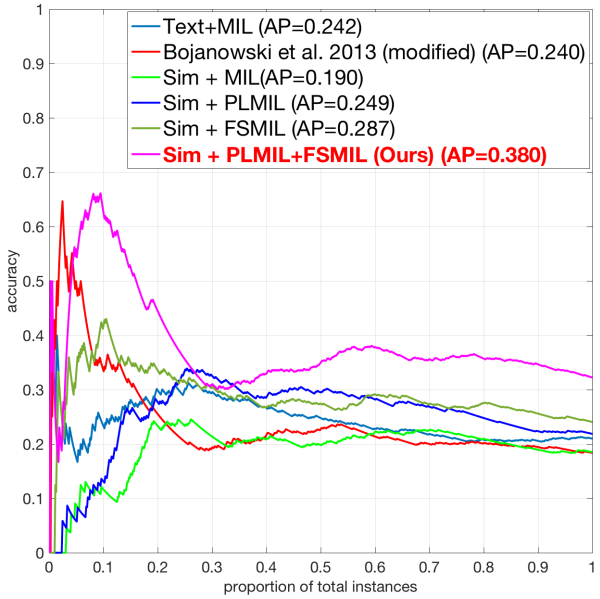
- COGNIMUSE Dataset: 5 movies + scripts

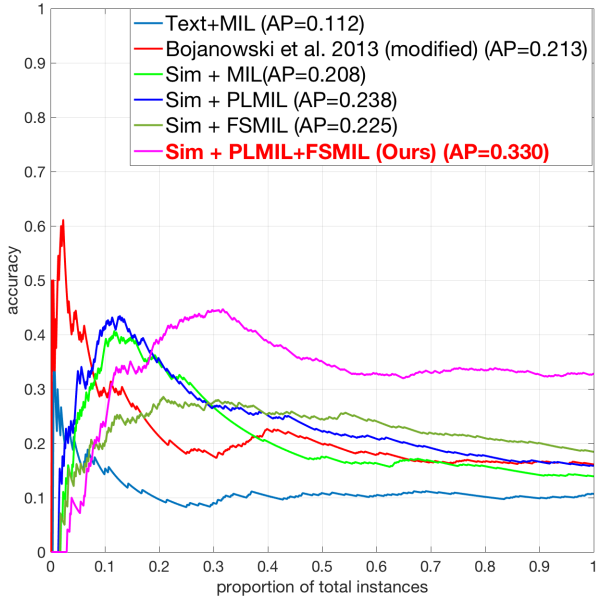
Set	Development			Test				All
	DEP	LOR	MAP	BMI	CRA	GLA	MAP	MAP
Text+MIL	0.433	0.656	0.544	0.551	0.434	0.437	0.474	0.502
SIFT+MIL [*Bojanowski et al. 2013*]	0.630	0.879	0.755	0.724	0.644	0.681	0.683	0.711
SIFT+FSMIL	0.693	0.881	0.787	0.770	0.691	0.746	0.736	0.756
VGG+MIL	0.834	**0.954**	0.894	0.825	0.696	0.830	0.784	0.828
VGG+FSMIL (Ours)	**0.864**	0.952	**0.908**	**0.857**	**0.731**	**0.901**	**0.830**	**0.861**
[*Miech et al. 2017*]+VGG: fg	0.788	0.898	0.843	0.666	0.479	0.577	0.574	0.682
[*Miech et al. 2017*]+VGG+FSMIL: fg	**0.810**	**0.913**	**0.862**	**0.696**	**0.505**	**0.651**	**0.617**	**0.715**
[*Miech et al. 2017*]+VGG: bg	**0.185**	**0.189**	**0.187**	**0.304**	0.047	**0.052**	0.134	0.155
[*Miech et al. 2017*]+VGG+FSMIL: bg	0.184	**0.189**	**0.187**	0.269	**0.278**	0.038	**0.195**	**0.192**

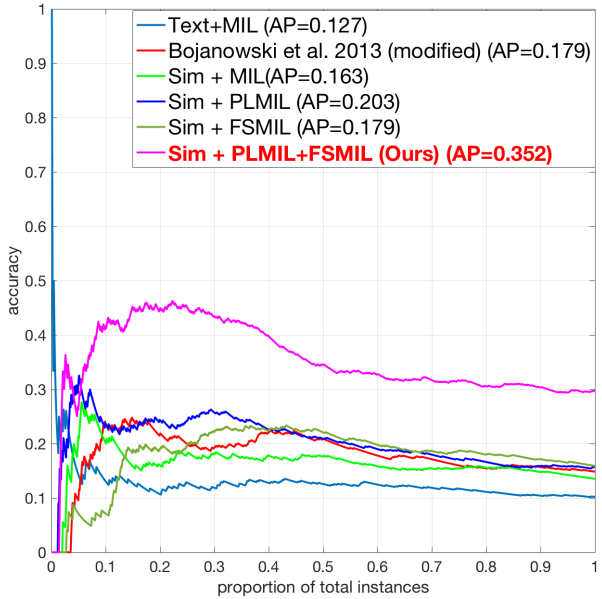
- Bojanowski et al. 2013: treats both ambiguities with hard constraints (MIL).
- Miech et al. 2017: extra constraint for background concepts.
- Bouritsas et al. 2018: FSMIL extension

Results: Action Recognition

- COGNIMUSE Dataset: 5 movies + scripts
  - mean per sample accuracy curves for 6, 8 & 10 action classes.



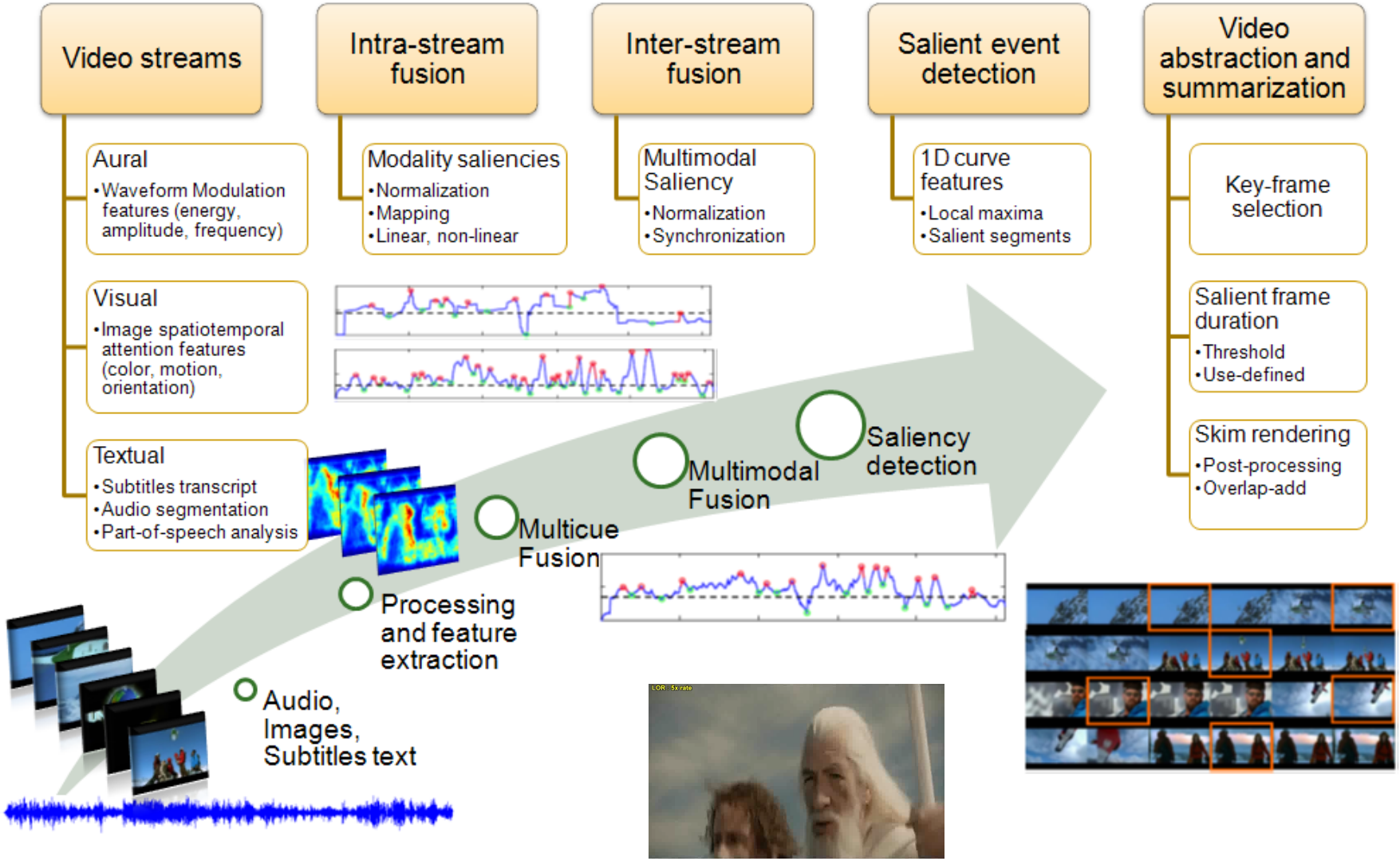




Multimodal Saliency &Video Summarization

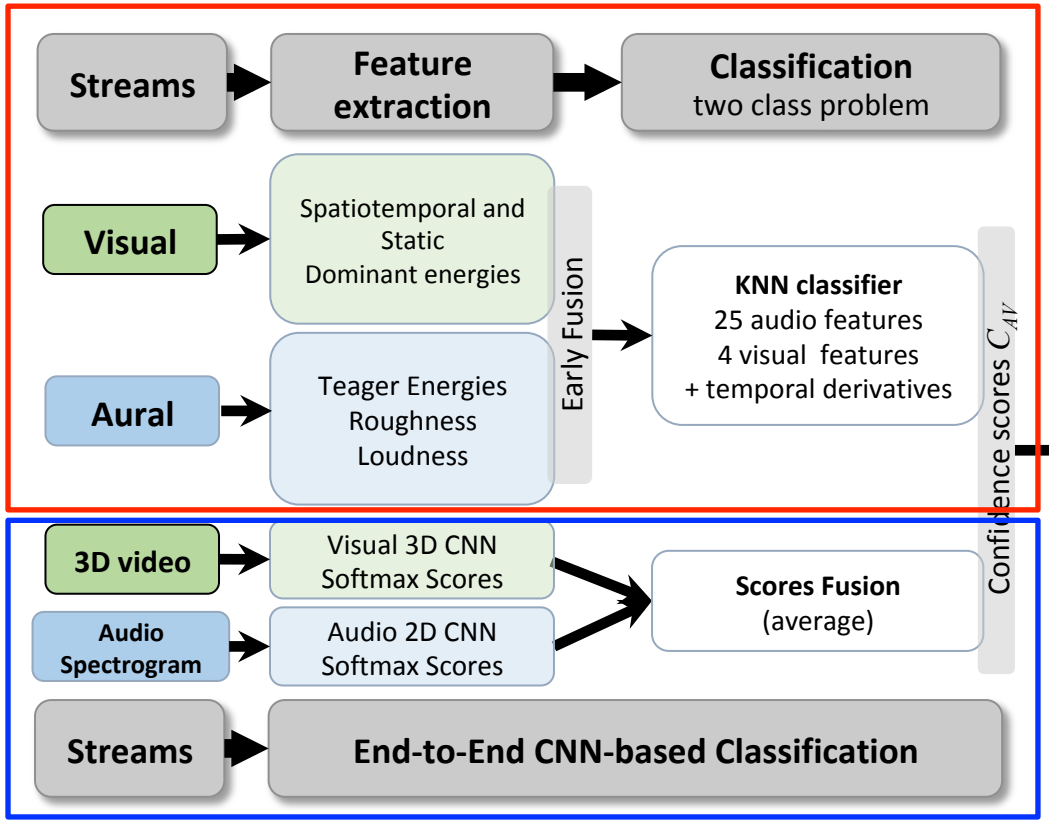
COGNIMUSE: Multimodal Signal and Event Processing In Perception and Cognition

website: http://cognimuse.cs.ntua.gr/

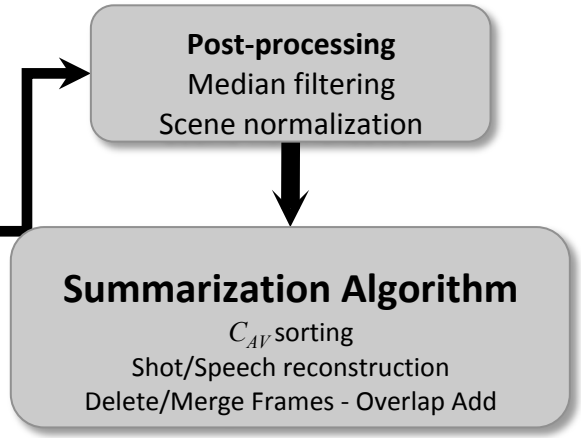


Multimodal Salient Event Detection:
Handcrafted vs. Multimodal CNN-based approach

handcrafted features + classification algorithms



multimodal CNN-based architectures for saliency detection

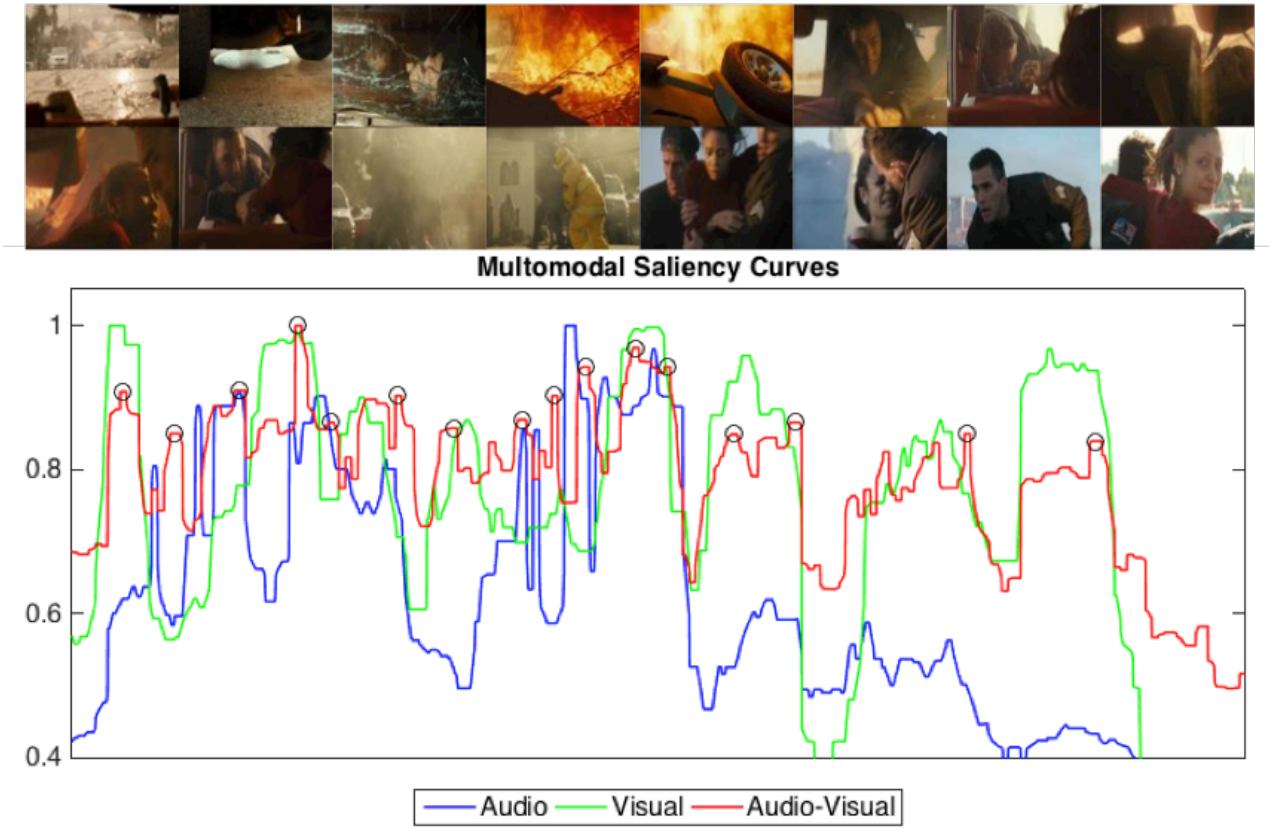


experimental evaluation

COGNIMUSE database:
- multimodal video database with multilevel saliency annotation
- *http://cognimuse.cs.ntua.gr/database*

[P. Koutras, A. Zlatintsi and P. Maragos, *Exploring CNN-based architectures for Multimodal Salient Event Detection in Videos*, IVMSP 2018.]

CNN Estimated Audio-Visual Saliency Curves



- Audio-Visual Saliency Curves
  - two-stream CNNs trained with the audio-visual annotation labels
  - average the softmax scores
- Keyframes extracted as local extrema of the audio-visual curve

COGNIMUSE Database
Saliency, Semantic & Cross-Media Events Database

http://cognimuse.cs.ntua.gr/database

Including:

- Saliency annotation on multiple layers
- Audio & Visual events annotation
- COSMOROE cross-media relations annotation
- Emotion annotation

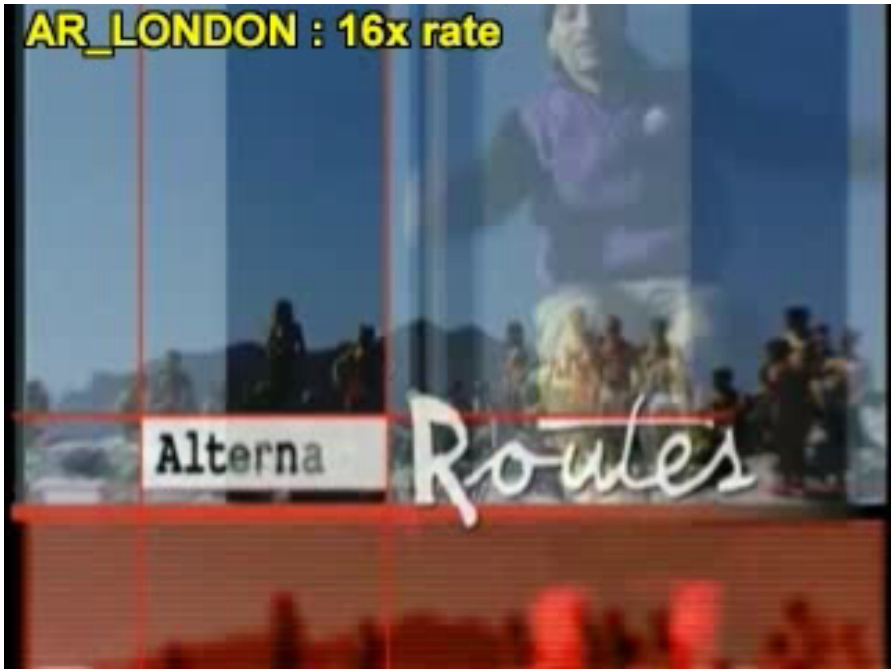
Database Content:

- **7** 30-min **movie clips** from: Beautiful Mind (BMI), Chicago (CHI), Crash (CRA), The Departed (DEP), Gladiator (GLA), Lord of the Rings III: The return of the king(LOR), Finding Nemo (FNE)
- **5** 20-min **travel documentaries**
- **1** 100-min **movie**: Gone with the Wind (GWTW)

[A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Marandrakis, N. Efhymiou, K. Pastra, A. Potamianos and P. Maragos, *COGNIMUSE: A Multimodal Video Database Annotated with Saliency, Events, Semantics and Emotion with Application to Summarization,* EURASIP Jour. on Image and Video Proc., 2017]
[A. Zlatintsi, P. Koutras, N. Efthymiou, P. Maragos, A. Potamianos and K. Pastra, *Quality Evaluation of Computational Models for Movie Summarization,* QoMEX 2015]

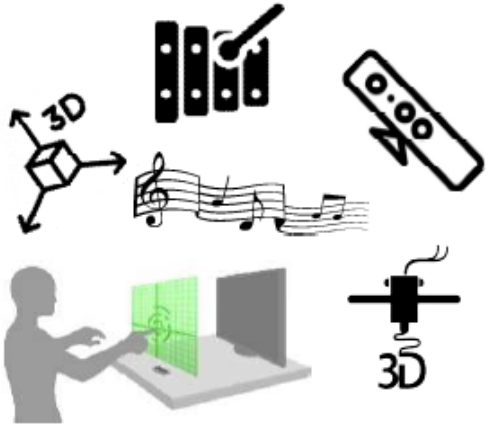
Video Summaries

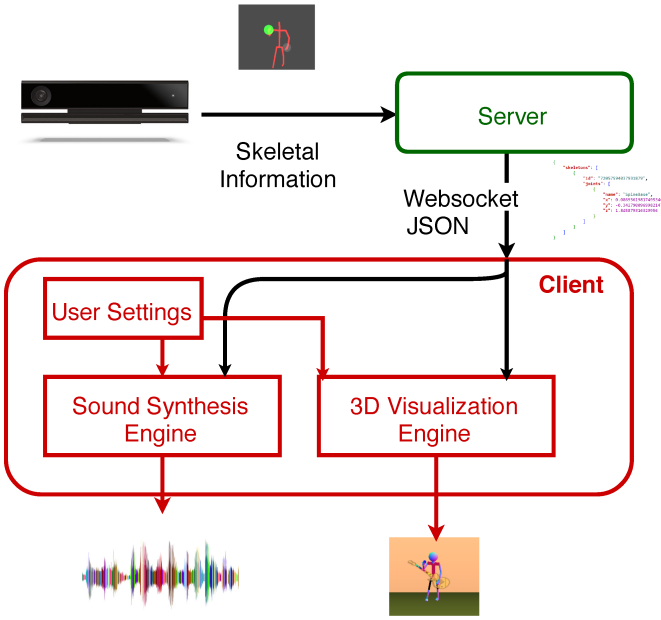
AR London ca 16% ca 3’40”

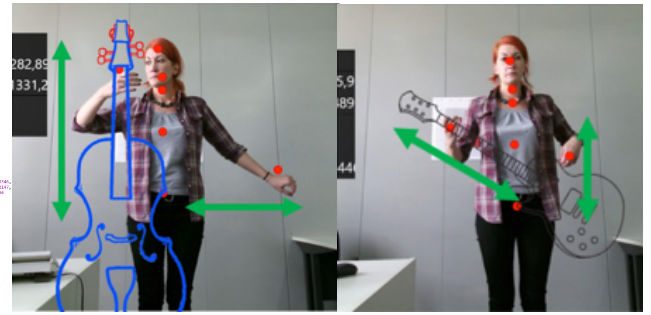


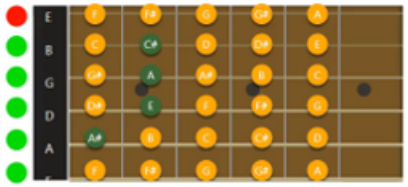
GWTW ca 3% ca 3’
(3min from full duration movie)

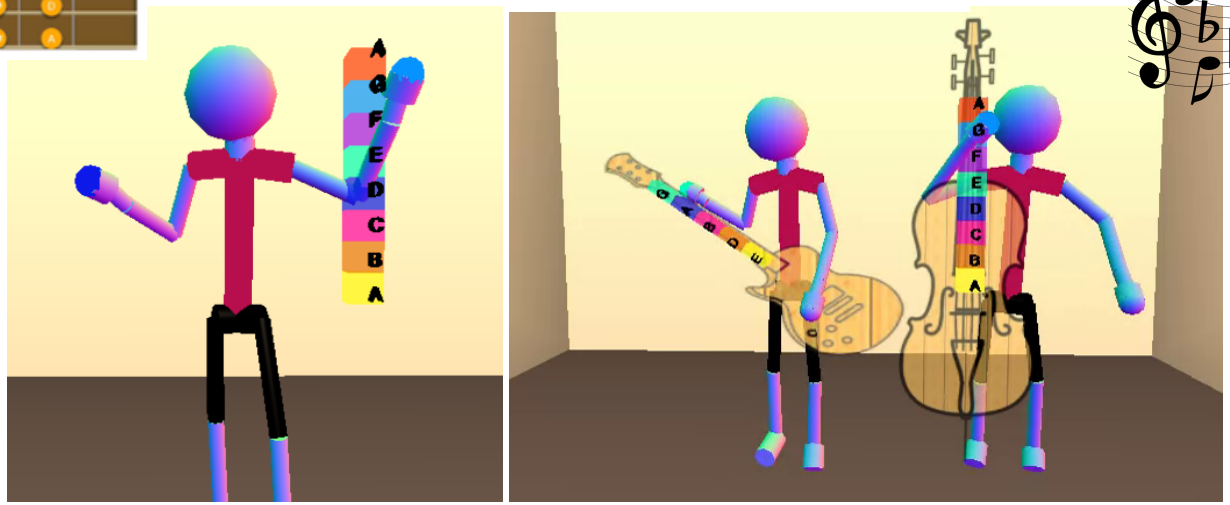
Audio-Gestural Music Synthesis











iMuSciCA Project: interactive Music Science Collaborative Activities

- New pedagogical methodologies and innovative educational tools to support active, discovery-based, personalized, and engaging learning
- Provide students and teachers with opportunities for **collaboration, co-creation** and **collective knowledge building**.
- Design and implement a suite of software tools and services that will deliver interactive **music** activities for teaching/learning **STEM**

STEM = **Science, Technology, Engineering** and **Mathematics** fields

- Bring **Arts** (A) at the heart of the academic curriculum

STEM + A =S TEAM

iMuSciCA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731861.
http://www.imuscica.eu

Gesture and Virtual Reality Interaction for Music Synthesis and Expression

- Virtual Music Instrument: analogous to a physical musical instrument, a *gestural interface*, that could provide for much greater freedom in the mapping of movement to sound.
- Innovative interactive and collaborative application (used for STEM) with advanced multimodal interface for musical co-creation and expression
  - Musically "air control" virtual instruments without any physical contact
- Web-based application: widely accessible to everyone
- Intuitive gestural control for triggering the sound

[A. Mulder, *Virtual Musical Instruments: Accessing the sound synthesis universe as a performer*. In Proc. Brazilian Symposium on Computer Music, 1994.]

Modes of Gestural Control and Interaction

i. Air Guitar interaction

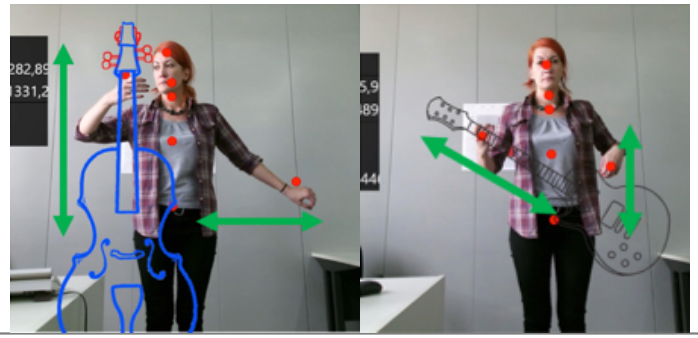
ii. Upright Bass interaction (using a virtual bow)

iii. Air Xylophone interaction

iv. Air Membrane/drums

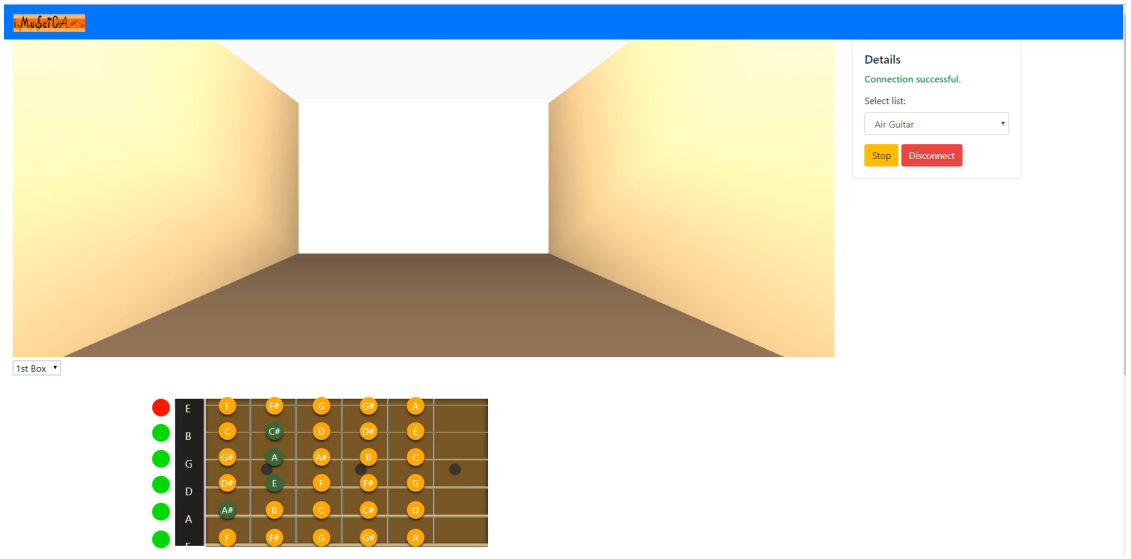
v. Conductor (two hands) interaction: each hand is assigned with one of the two previously named instruments

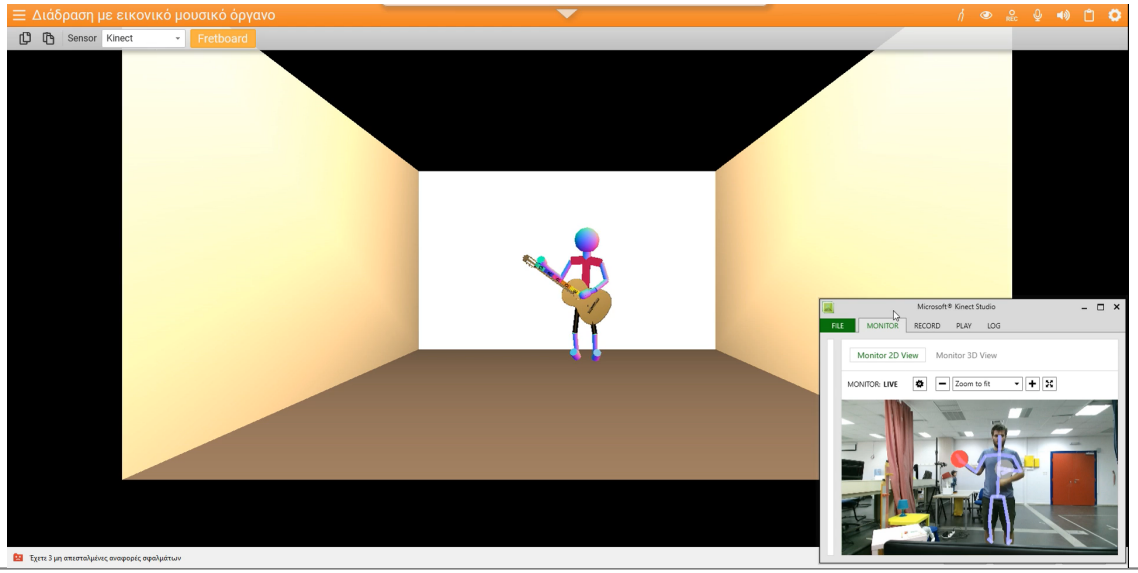
- Multiplayer interaction: for collaborative playing
- Using simple and more intuitive gestures:
  - Provide the users, especially those that are not musically educated, the ability to perform various virtual instruments without constraints.



- [A. Zlatintsi et al, *A Web-based Real-Time Kinect Application for Gestural Interaction with Virtual Musical Instruments,* Audio Mostly Conf., 2018.]
- [C. Garoufis et al, *An Environment for Gestural Interaction with 3D Virtual Musical Instruments as an Educational Tool*, EUSIPCO 2019].

Demo


Details
Connection successful.
Select list:
Air Guitar
Stop
Disconnect
1st Box
E
B
G
D
A


Διάδραση με εικονικό μουσικό όργανο
Sensor
Kinect
Fretboard
Microsoft® Kinect Studio
FILE
MONITOR
RECORD
PLAY
LOG
Monitor 2D View
Monitor 3D View
MONITOR: LIVE
Zoom to fit

Part 2: Conclusions

- Audio-Visual Fusion → Better Results (ASR, TTS, HRI, Saliency, Music).
- More Data → Big Databases → Better training algorithms (Training processes work better if we have significant amounts of training data).
- More Big Data → Needs for annotations and possibly summarization. Not only data compression or dimensionality reduction for storage or fast access.
- Multimodal Data (audio/speech, visual, depth, text):
  - Need for advanced signal processing algorithms for each modality (different nature of each modality).
  - Signal modalities or dimensions are complementary (i.e. microphones arrays enhance audio signal for distant ASR, audio-visual fusion improves speech/gesture understanding, video summarization).

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