Audio-Visual Speech Analysis & Recognition

Nassos Katsamanis

Institute of Communication and Computer Systems (ICCS),

National Technical University of Athens (NTUA),

School of E.C.E.,

Computer Vision, Speech Communication &

Signal Processing Group

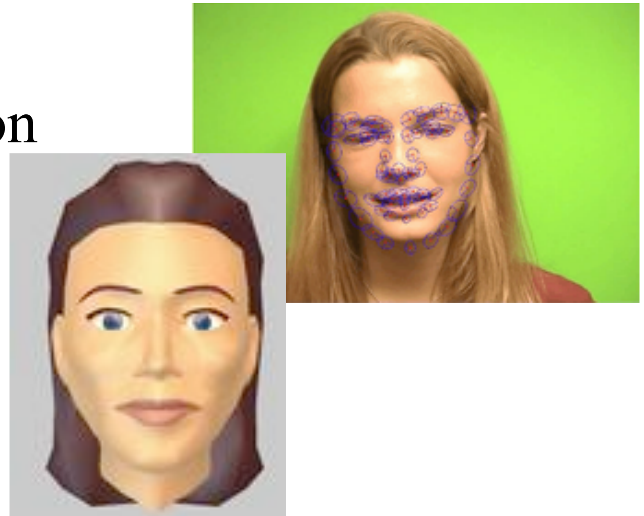
http://cvsp.cs.ntua.gr

Audio-Visual Speech

- Bimodal human speech perception
  - McGurk effect
- Bimodal human speech production
  - Face and articulatory motion higly correlated

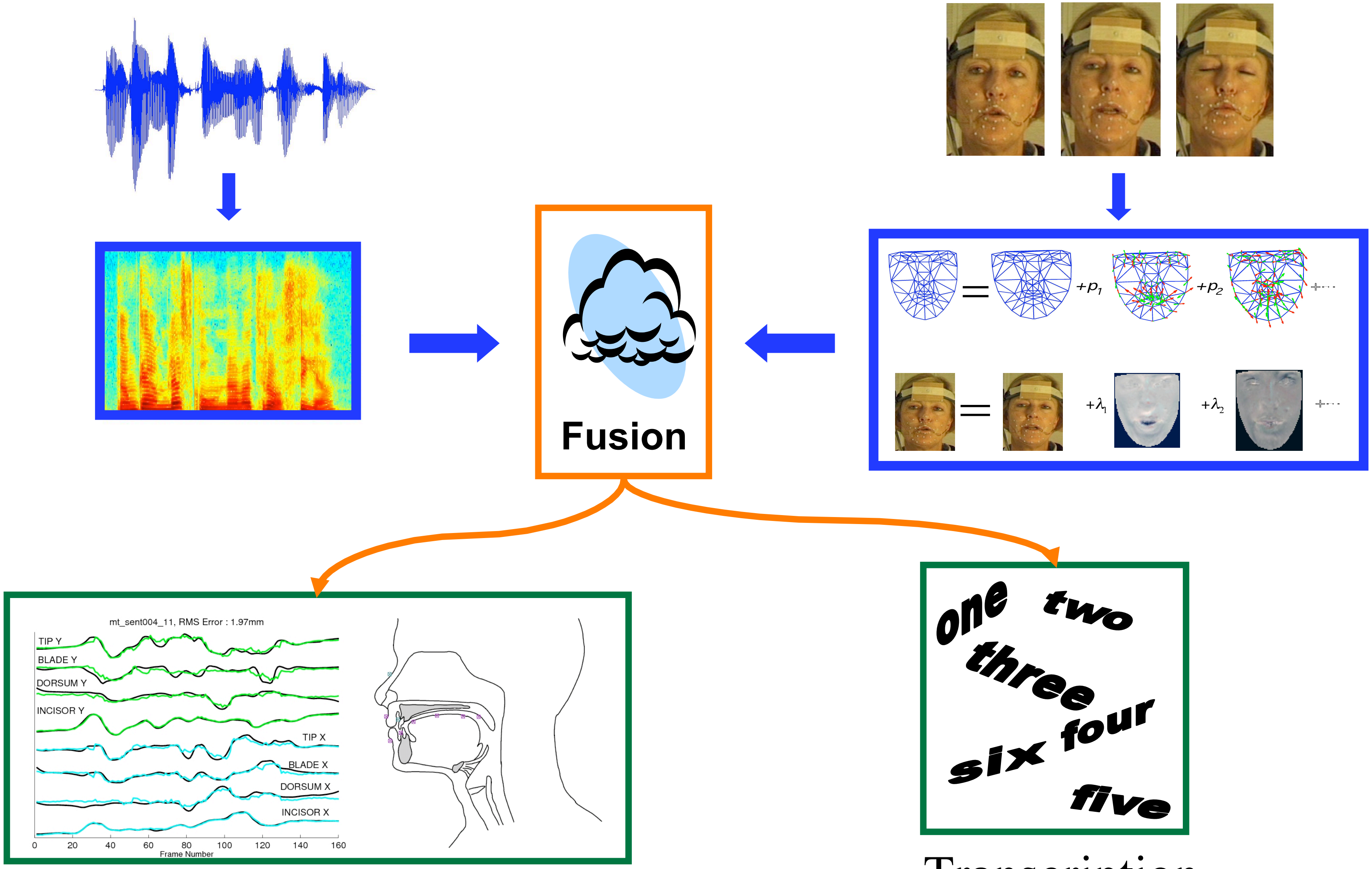
Found in Audio-Visual speech web-lab, University of California, Riverside

- Applications
  - Noise-Robust Automatic Speech Recognition
  - Natural and Intelligible Speech Synthesis



Found in Synface project page, KTH

Audiovisual Speech Analysis & Recognition



Transcription

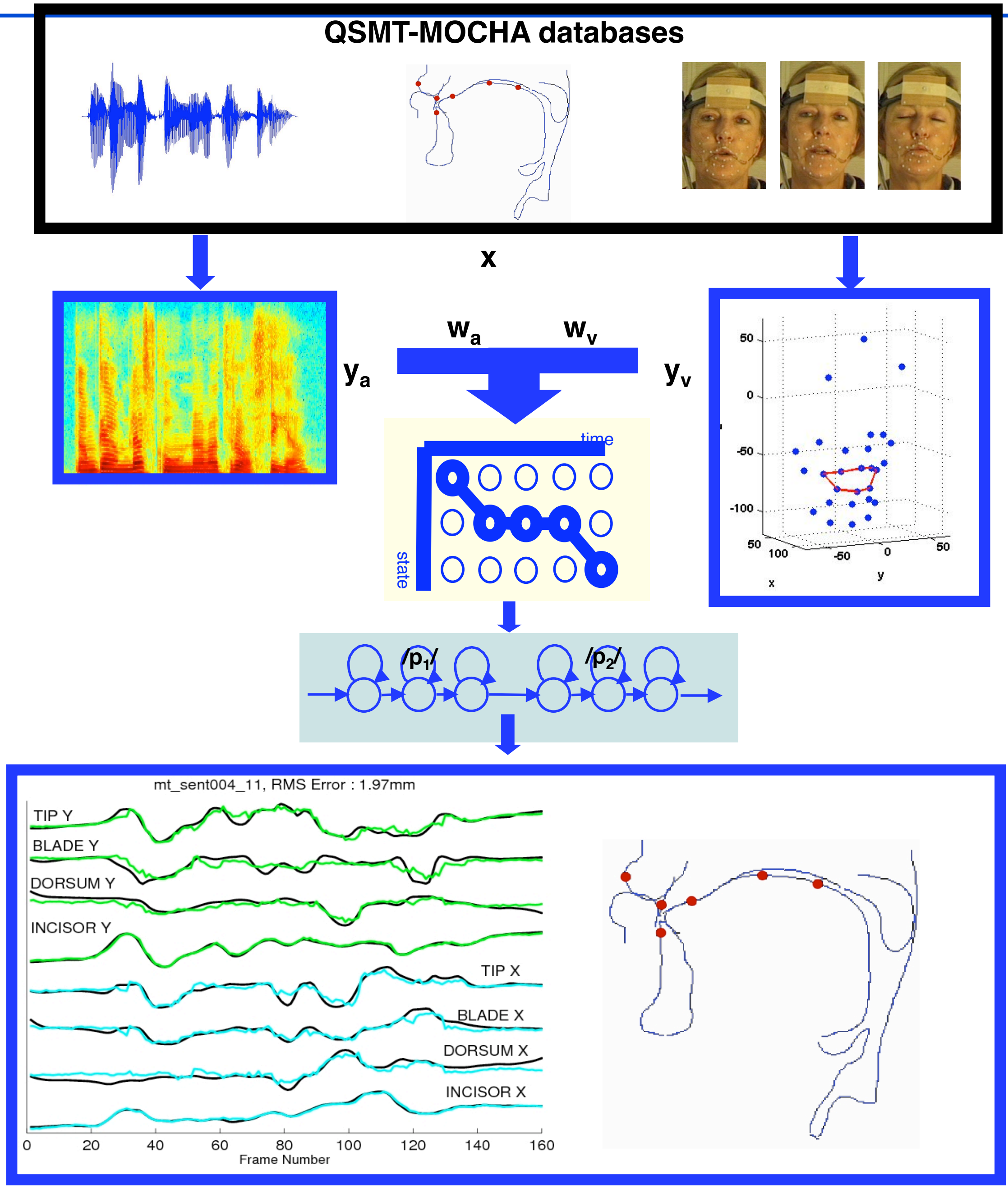
CVSP Group, ICCS-NTUA

- Involved Group Members
  - Prof. Petros Maragos — Group Leader
  - Nassos Katsamanis — PhD Student
  - George Papandreou — PhD Student
  - Dr. Vassilis Pitsikalis — Senior Research Associate
- Group's Research
  - Multimedia Analysis
    - Audio-Visual Speech Recognition and Inversion
    - Movie Summarization
    - Multimodal Integration-Fusion
  - Computer Vision
    - Biomedical, Geological and Astronomical Image Analysis
    - Reconstruction of Archeological Paintings
    - Sign Language Recognition
  - Speech Communication
    - Noise-Robust Speech Recognition
    - Audio Analysis, Event Detection
    - Speech Production Modeling

Audio-Visual Speech Analysis/Inversion

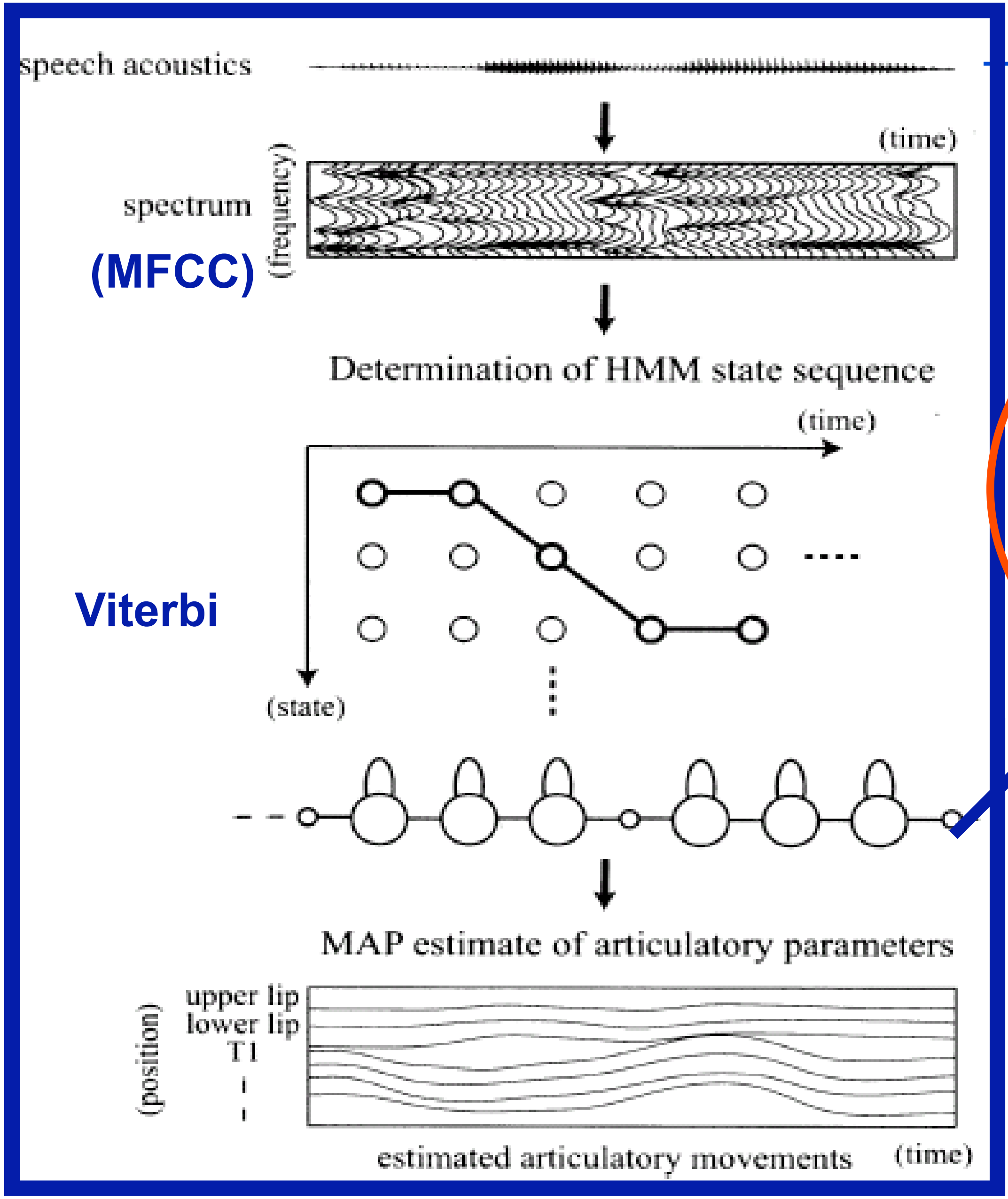
- Speech Inversion
  - Recover vocal tract properties, given speech
  - Applications in: Speech Synthesis, Recognition, Coding
  - Traditionally: Audio-only based inversion
  - Ill-posed problem

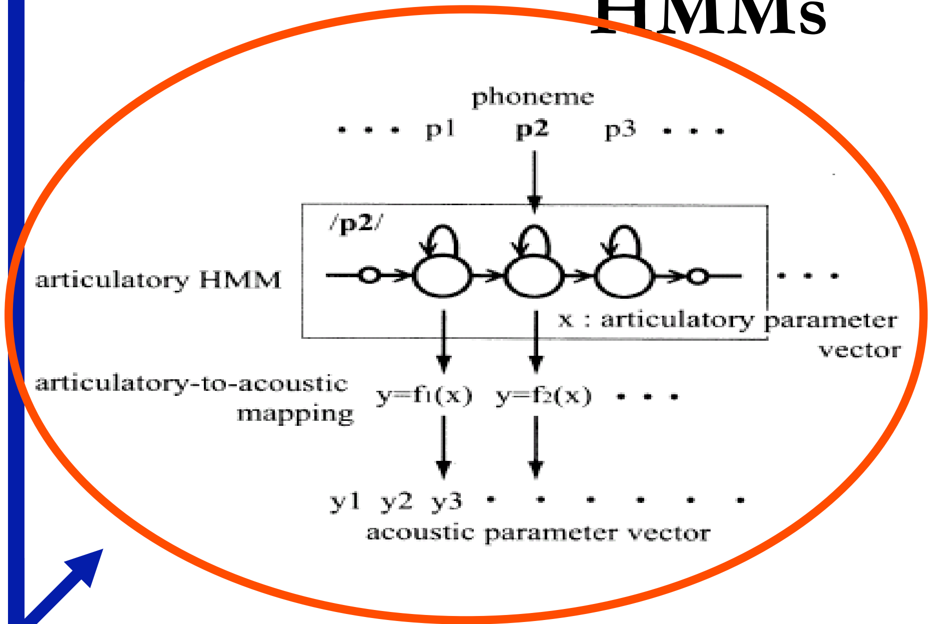
- Facial Information
  - Visible articulators, e.g., lips, jaw
  - Helps determining the location of the articulators
  - Constrains the articulatory space



Audio-Visual Speech Inversion Using HMMs

Speech Inversion using HMMs





- Piecewise linear approximation of the mapping between observed and articulatory data
- Stochastic approximation

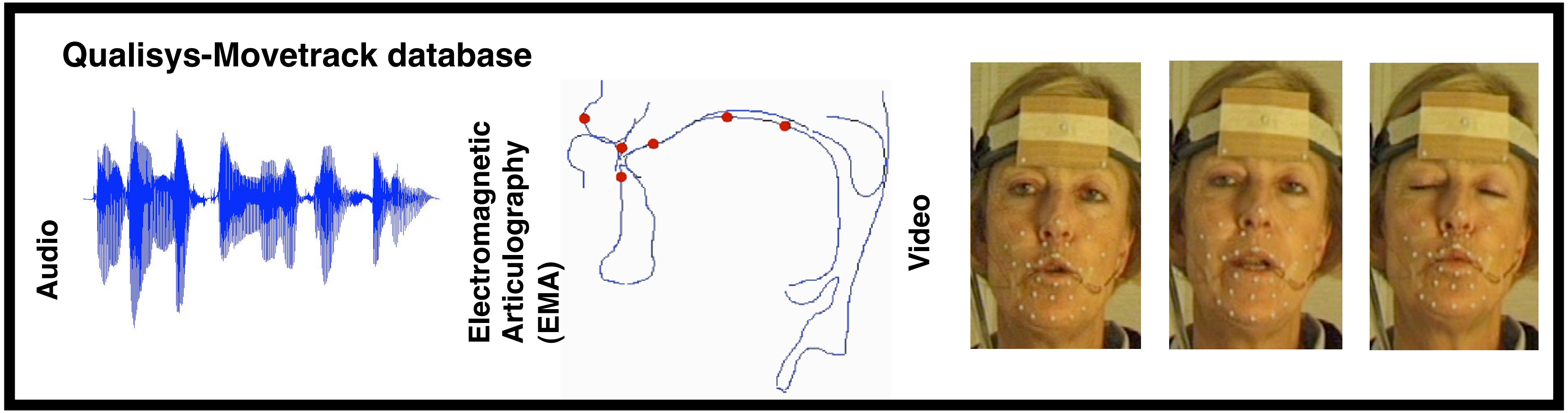
Audio-only initial framework proposed by Hiroya and Honda (IEEE TSAP 2004)

Audio-Visual Speech Inversion: Fusion

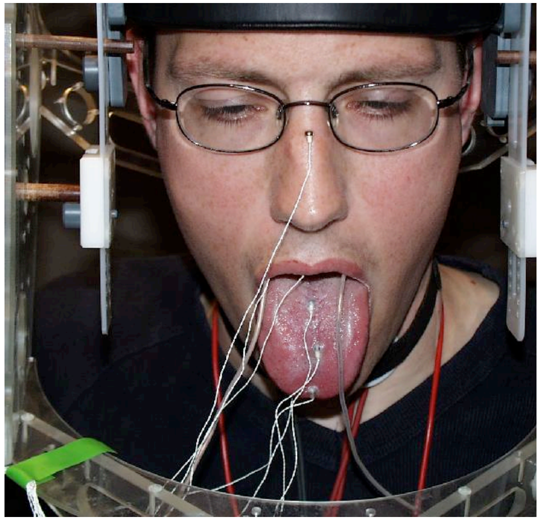
- Determine the switching process
  - Fully Synchronous Scenario
    - Independent streams, different weights
    - Multistream HMMs
  - Asynchronous Scenario
    - Constrained Asynchronicity
      - Product-, Asynchronous HMMs
    - Unconstrained Asynchronicity
      - Separate switching mechanism for each modality
      - Late-fusion
- Inversion, given the model-switching process
  - Maximum A Posteriori solution is a weighted average between prior, audio- and visual- based predictions
  - The audio and visual streams are weighted by their relative modeling reliability

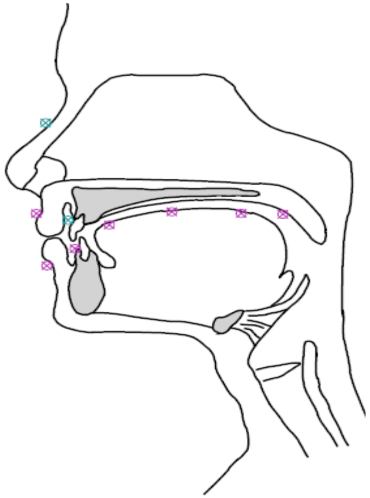
Audiovisual Speech Inversion: QSMT

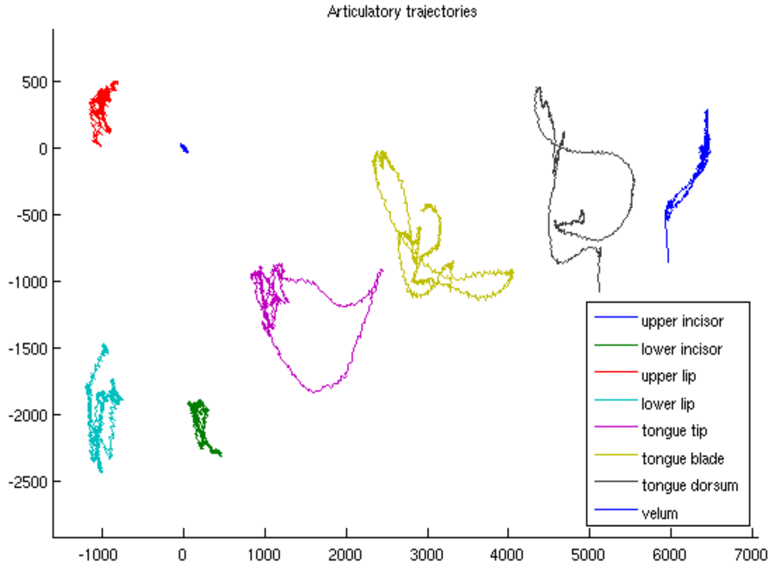
- QualiSys-MoveTrack dataset provided by KTH
- 438 Utterances, (VCVs, CVCs, simple Swedish phrases)
- Face Expression Y (3-D coords of 25 facial landmarks→75 params)
- Articulation X (2-D coords of 6 EMA coils→12 params)
- 15 minutes of usable data (1 female speaker)
- Needed Preprocessing and Synchronization between Video and other Streams (ICCS-NTUA)



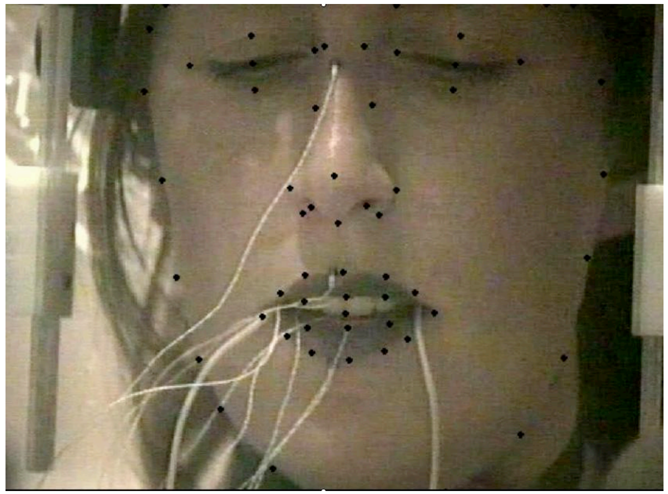
Audiovisual Speech Inversion: MOCHA





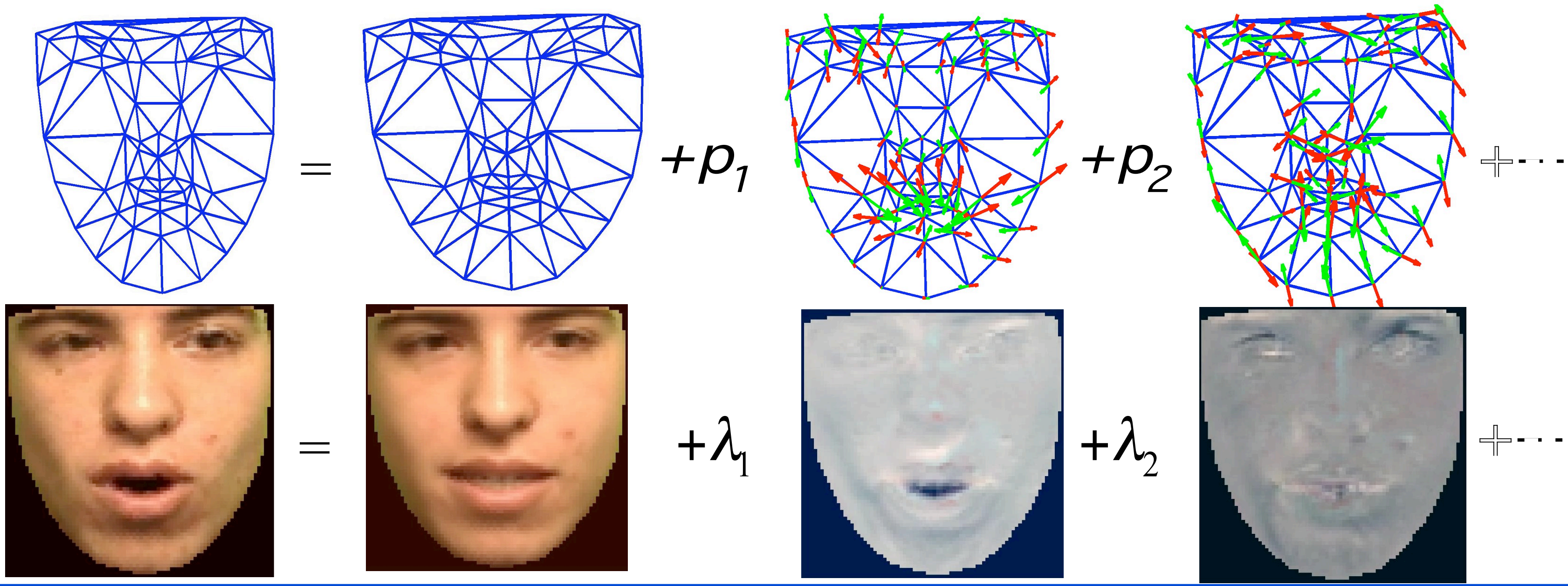


- Provided by CSTR, Univ. Edinburgh
- Two subjects (one male, one female), 460 British TIMIT Utterances each
- Articulation (2-D coords of 9 EMA coils)
- Video of the female speaker's face
- 30 minutes of usable data
- Needed Preprocessing-labeling Video (ICCS-NTUA)

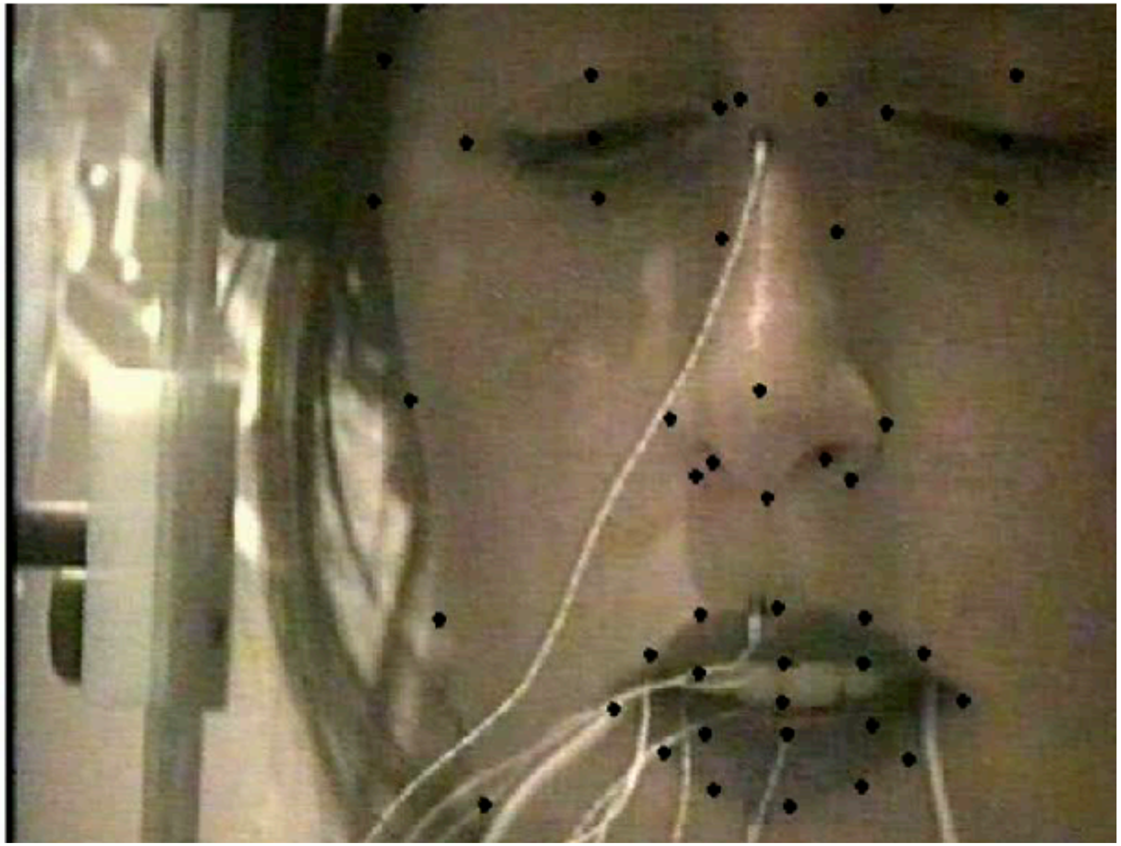


Visual Feature Extraction: Active Appearance Modeling of Visible Articulators

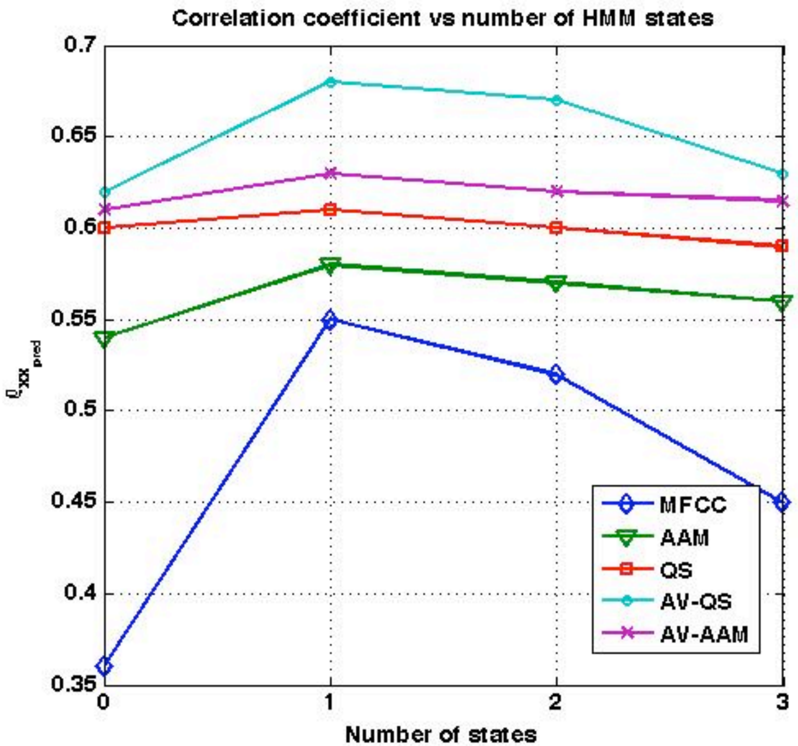
- Active Appearance Models for face modelling
- Shape & Texture related articulatory information
- Features: AAM Fitting (nonlinear least squares problem)
- Real-Time, marker-less facial visual feature extraction

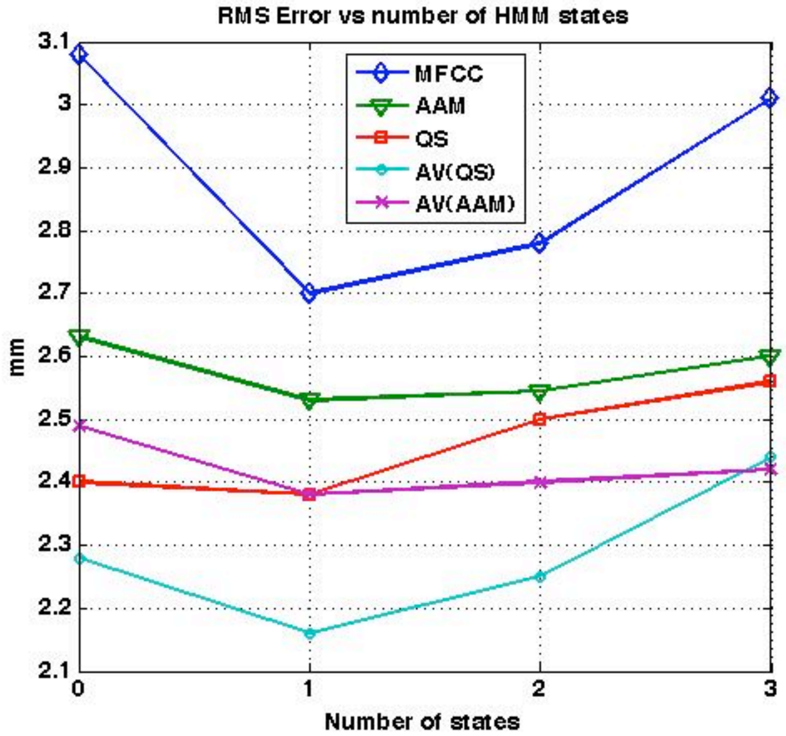


AAM Visual Feature Extraction – MOCHA



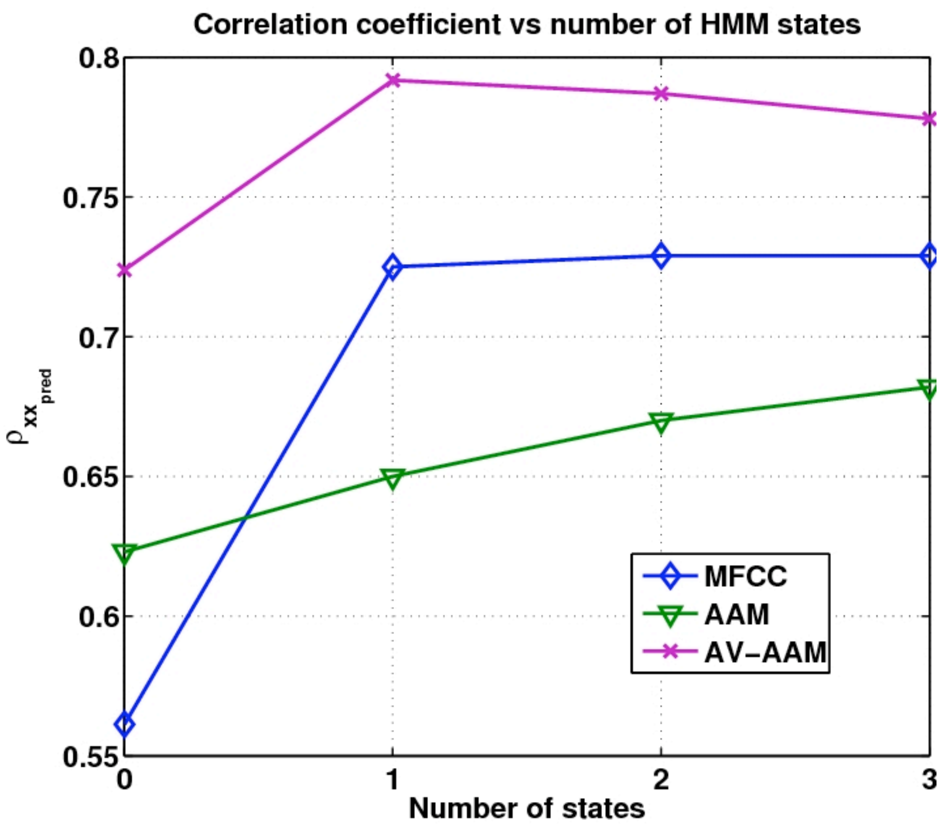
HMM Inversion Evaluation: QSMT

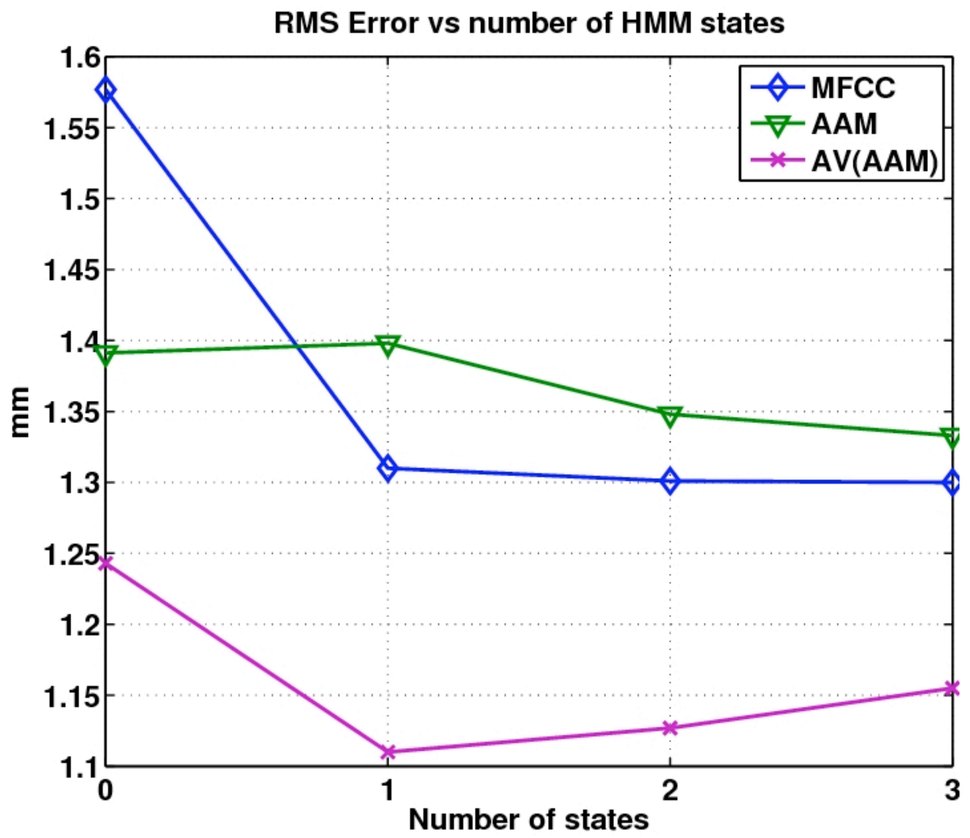




Features	Level	Type	States	RMS (mm)	$\rho_{x\hat{x}}$
Audio	P	HMM	2	2.56	0.60
QS	P	HMM	2	2.30	0.65
QS	V	HMM	3	2.24	0.66
A-QS	P	HMM	2	2.16	0.69
A-QS	P-P	HMM+LF	2-2	2.02	0.71
A-QS	P-V	HMM+LF	2-2	1.99	0.72
A-QS	**P**	**MS-HMM**	**2**	**1.95**	**0.74**

HMM Inversion Evaluation: MOCHA

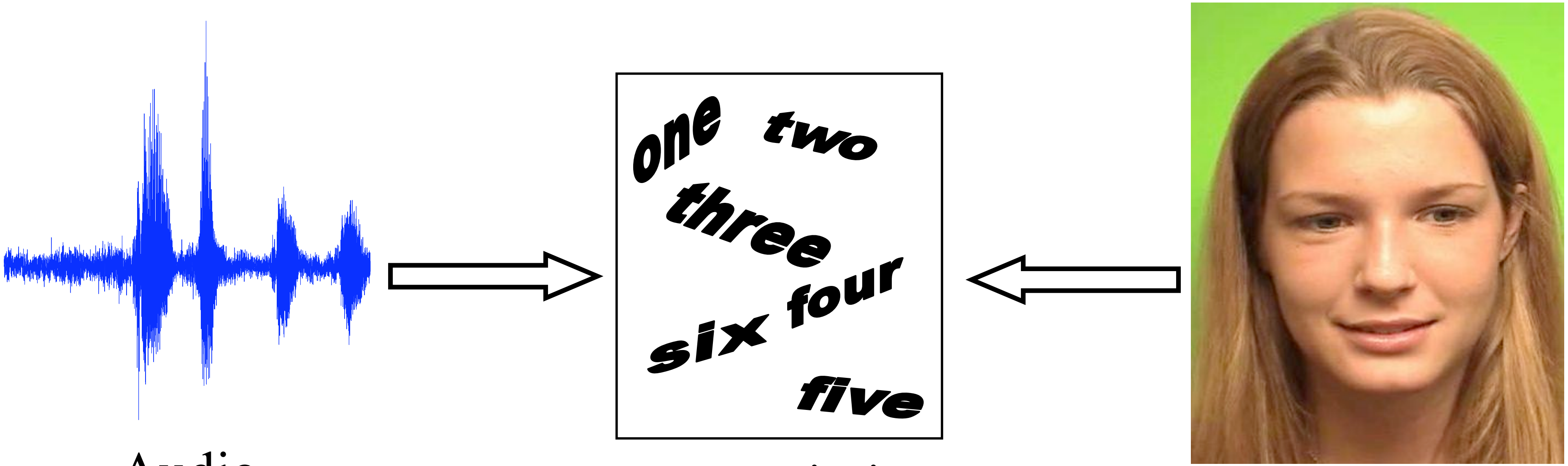




Audio-Visual Speech Inversion: Discussion

- Active Appearance Modeling of the Face
  - Automatic Visual Feature Extraction from the frontal view
  - Single camera, no markers needed
- Piecewise Linear Approximation of Audio-Visual to Articulatory Mapping
  - Switching governed by hidden Markov process
  - Fusion possible at various synchronization levels
- Incorporation of the visual modality is clearly beneficial to inversion

Audio-Visual Automatic Speech Recognition



Audio

Transcription

Video

- Improves ASR performance in adverse conditions
  - Noise
  - Interference

Feature measurement *uncertainty*

Conventional View: Unlimited feature precision

$$x = \begin{pmatrix} 1.0456874064446654 8 \\ 0.8785646865349865 5 \\ 0.1895498057354365 6 \\ 0.5499839983549594 3 \\ 1.0433040984554566 5 \end{pmatrix}$$

SNR= 20dB

$$x = \begin{pmatrix} 1.046 \pm 0.001 \\ 0.879 \pm 0.001 \\ 0.190 \pm 0.001 \\ 0.550 \pm 0.001 \\ 1.043 \pm 0.001 \end{pmatrix}$$

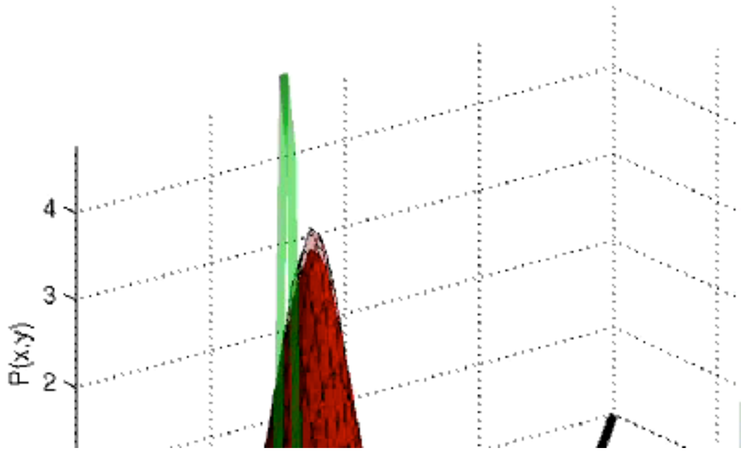
Our View: Finite, SNR-dependent feature precision

SNR= 5dB

$$x = \begin{pmatrix} 1.0 \pm 0.1 \\ 0.9 \pm 0.1 \\ 0.2 \pm 0.1 \\ 0.6 \pm 0.1 \\ 1.0 \pm 0.1 \end{pmatrix}$$

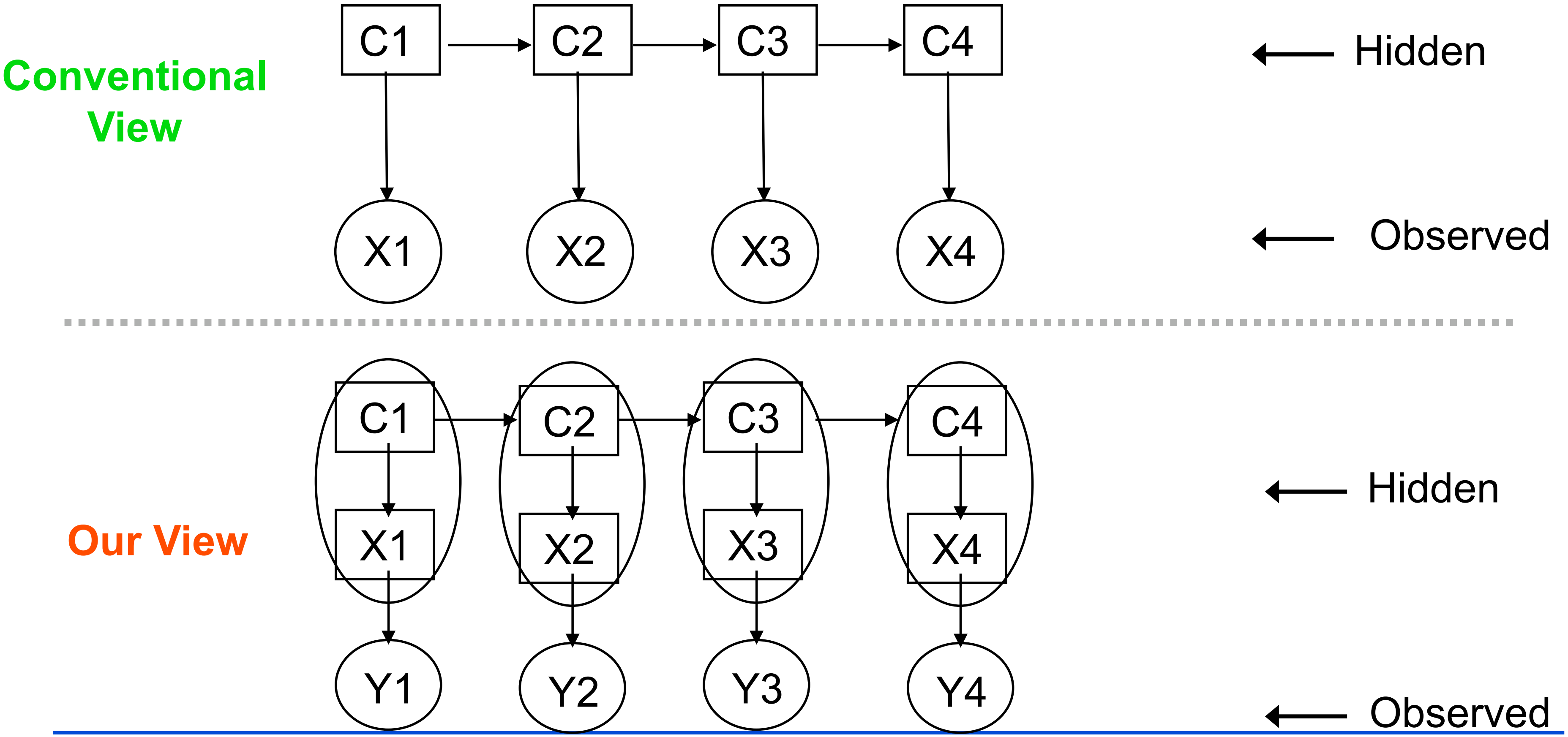
GMM Classification – Bimodal Case

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



Hidden Markov Models & Uncertain Data

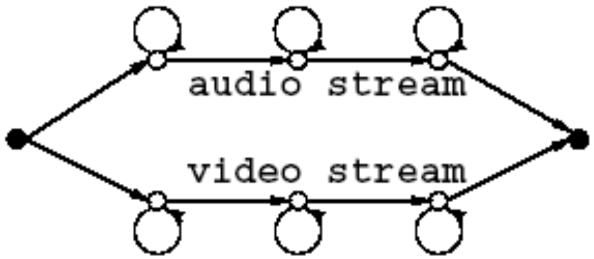
- Use measurement *variance compensated scores* in the **Viterbi** (decoding) and **α-β** (estimation) algorithms
- Adaptation at the finest time resolution (frame-level)



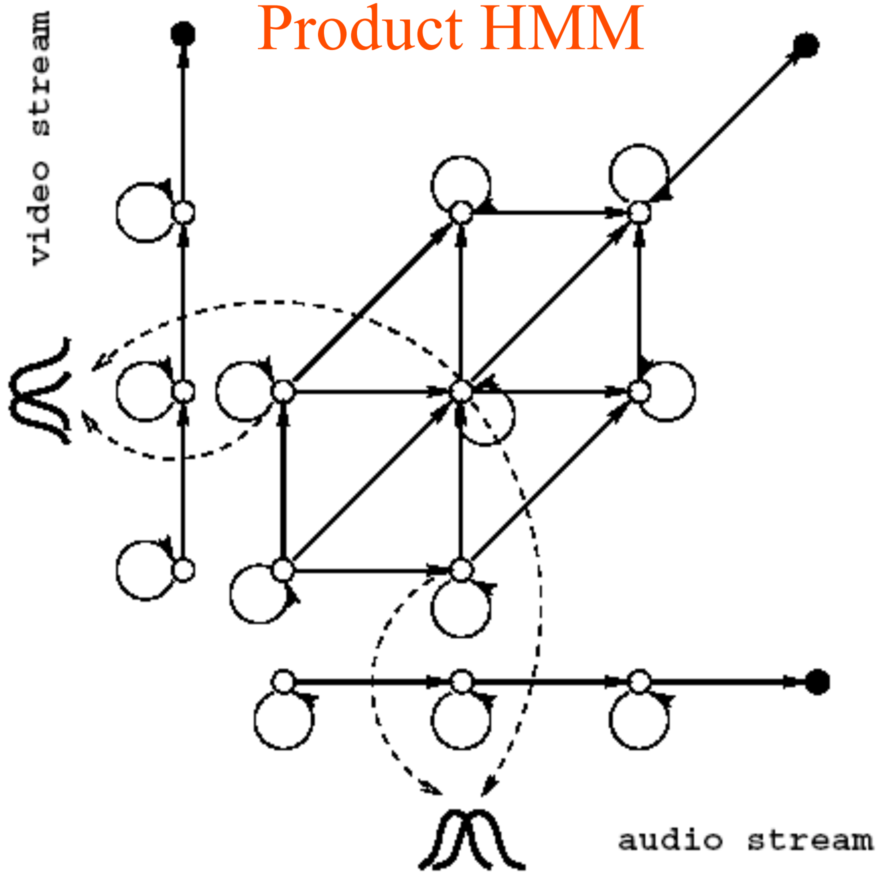
Asynchronous Sequence Models

- Seamless integration with asynchronous sequential models

Word-Boundary Synchronous HMM



Product HMM



Ref: Gravier et al., Proc. HLT 2002

Audio-Visual ASR: Database & Setup

- We used a subset of the CUAVE database:
  - 36 speakers (30 training, 6 testing)
  - 5 sequences of 10 connected digits per speaker
  - Training set: 1500 digits (30x5x10)
  - Test set: 300 digits (6x5x10)
- Task: Classification of isolated digits under noise
- Artificial “babble” noise from NOISEX database
- Word-level HMMs (left-to-right, 8 states, 1 mixture, diagonal covariance matrices)
- Use of HTK (Cambridge Univ.) and BNT (K. Murphy)

Ack: CUAVE was provided by the Clemson University

Audio Front End, Uncertain Features

- Log Mel Filterbank Energies (FBANK)
- Speech Enhancement Methods (e.g. SPLICE, ALGONQUIN)
- Model for FBANK degradation under noise (VTS)

$$X_{noisy} = f(X_{clean}, N)$$

Clean Speech FBANK

Noise FBANK

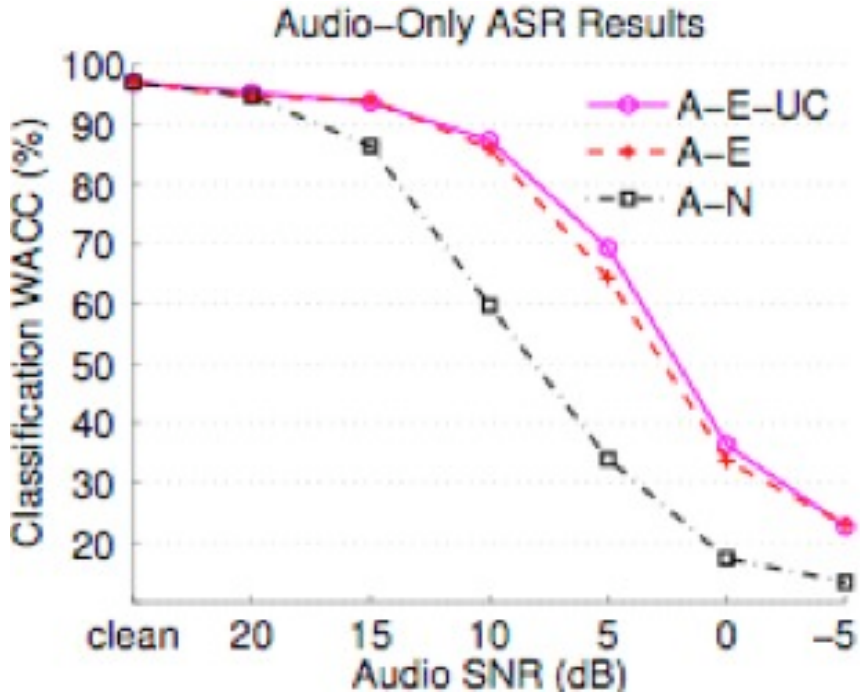
- Feature measurement + uncertainty

$$X_{clean} = \hat{X} + E$$

Feature Uncertainty

Deng, Droppo, Acero: "Dynamic compensation of HMM variances…", IEEE TSAP 2005

Audio Front-end, Evaluation



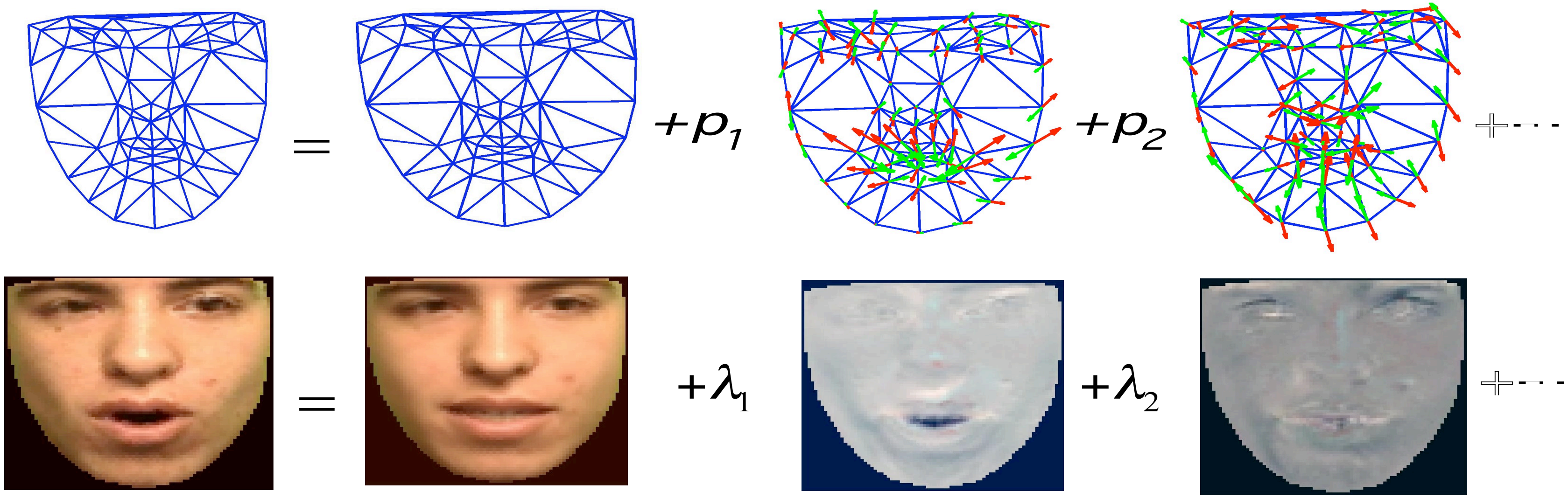
- Log Mel-scale Filterbank Energies
- Vector Taylor Series approximation and clean speech model (GMM) to get clean feature estimates and uncertainty

Visual Front End

- Both shape & texture can assist lipreading
- Active Appearance Models for face modeling
  - Shape and texture of faces "live" in low-dimensional manifolds
- Features: AAM Fitting (nonlinear least squares problem)

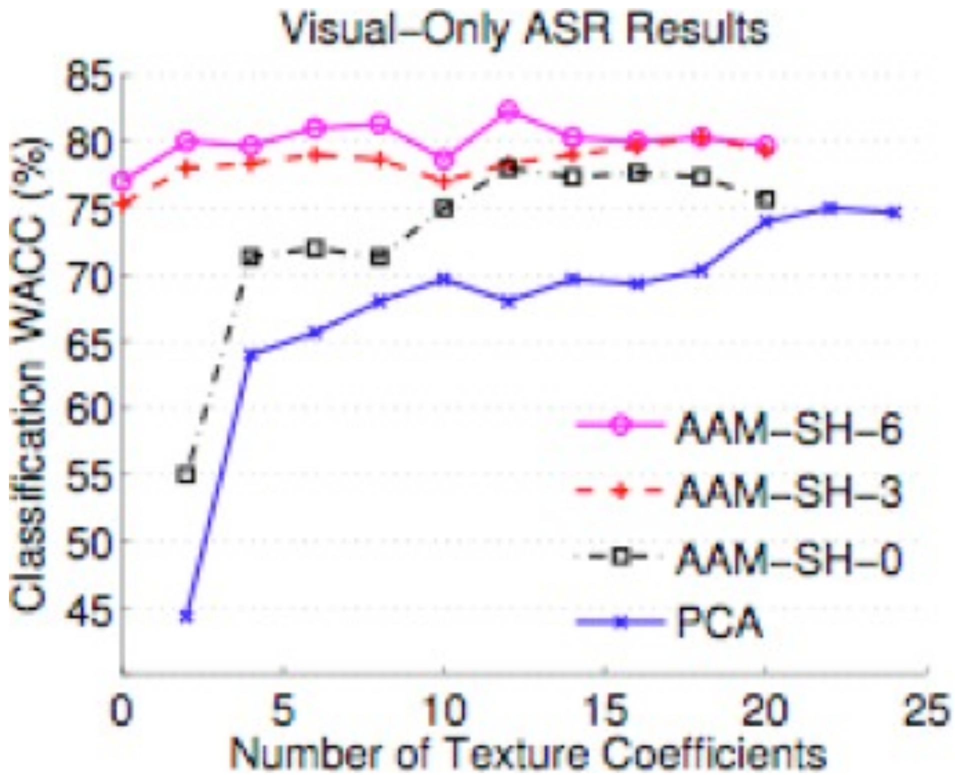
$$x = (p^T, \lambda^T)^T$$

- Visual feature uncertainty related to the sensitivity of the least-squares solution



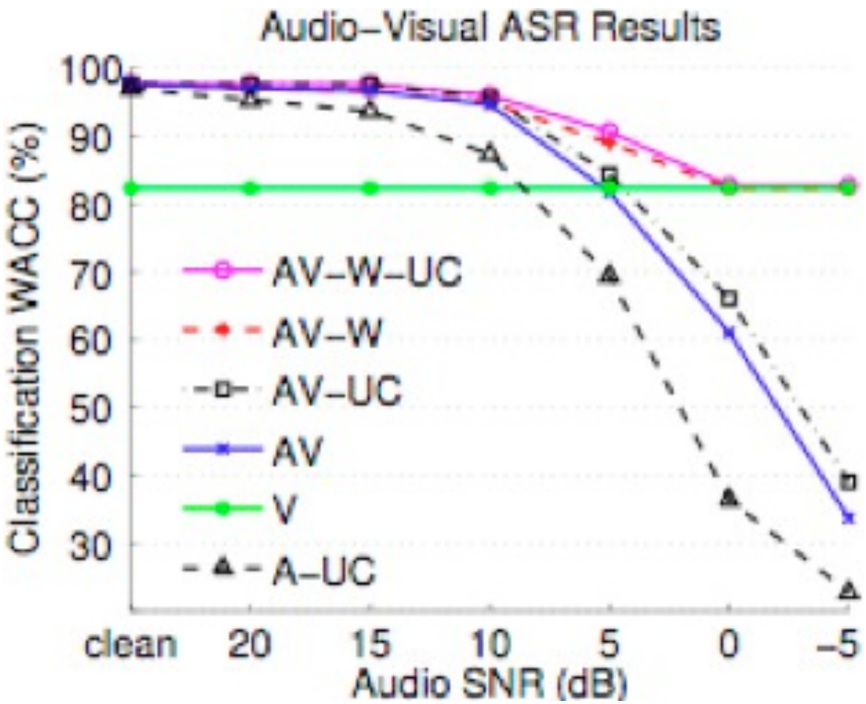
Matthews et al., PAMI 2002

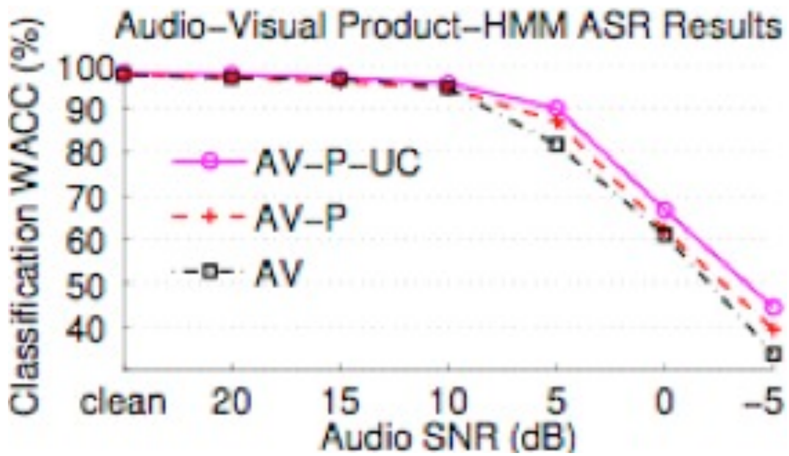
Visual Front-end, Evaluatio



- Active Appearance Modeling vs PCA

AV-ASR Results





- Weights and Uncertainty Compensation
  - Hybrid Fusion Scheme

Highlights - Conclusions

- Audio-Visual Speech Inversion
  - AAM Face Modeling, MFCC extracted from audio
  - Fusion at various synchronization levels
- Audio-Visual Speech Recognition
  - AAM Face Modeling, FBANK extracted from audio
  - Observation uncertainty estimation
  - Fusion by Uncertainty Compensation
- Bimodal Speech Processing clearly benefits both inversion and recognition

CVSP Group, ICCS-NTUA

- Involved Group Members
  - Prof. Petros Maragos — Group Leader
  - Nassos Katsamanis — PhD Student
  - George Papandreou — PhD Student
  - Dr. Vassilis Pitsikalis — Senior Research Associate
- Group's Research
  - Multimedia Analysis
    - Audio-Visual Speech Recognition and Inversion
    - Movie Summarization
    - Multimodal Integration-Fusion
  - Computer Vision
    - Biomedical, Geological and Astronomical Image Analysis
    - Reconstruction of Archeological Paintings
    - Sign Language Recognition
  - Speech Communication
    - Noise-Robust Speech Recognition
    - Audio Analysis, Event Detection
    - Speech Production Modeling

http://cvsp.cs.ntua.gr