Imperial College London

# AUDIOVISUAL DISCRIMINATION BETWEEN LAUGHTER AND SPEECH

**Stavros Petridis, Maja Pantic**
**sp104@doc.ic.ac.uk, m.pantic@imperial.ac.uk**

**Department of Computing**

Berlin 27/2/2009

Imperial College London

# Motivation

- Powerful affective and social signal

- Most frequently annotated non-verbal behaviour in recorded natural speech
  - 8.6% of the time a person vocalizes is spent on laughing
  - 0.8% is spent on laughing while talking

# Previous work

- Audio-only Laughter Detection
  - Audio features only (MFCC, prosody)
  - Hidden Markov Models (HMMs) / SVM /GMM

- Audiovisual laughter Detection
  - Audio features (MFCC)
  - Visual features ( lip lengths, lip angles)
  - Improved performance
  - Tested on 3 sequences

Imperial College London

# Objective

- To use standard tools to discriminate laughter from speech using an audiovisual approach

# Audiovisual Recognition

- Audiovisual speech recognition
  - Audio + Visual features (e.g. lip contours)
  - Hidden Markov Models (HMMs) + Variants of HMMs
  - Feature level
  - Enhanced performance

- Audiovisual affect recognition
  - Recognition of six basic + non basic emotions
  - HMMs, GMMs, RNN, Bayesian Networks
  - Decision level, model-based fusion

# Dataset

- AMI Meeting Corpus
- 10 subjects – spontaneous expressions
- 67 voiced laughters (100.88 sec)
- 48 unvoiced laughters (60.76 sec)
- 93 speech segments (177.86 sec)
- 720 x 576 pixels, 25 FPS

Imperial College
London

Imperial College
London

Imperial College London

# Audio Features

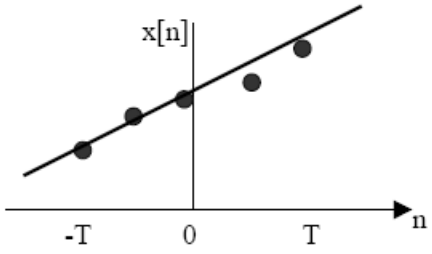
- Spectral Features

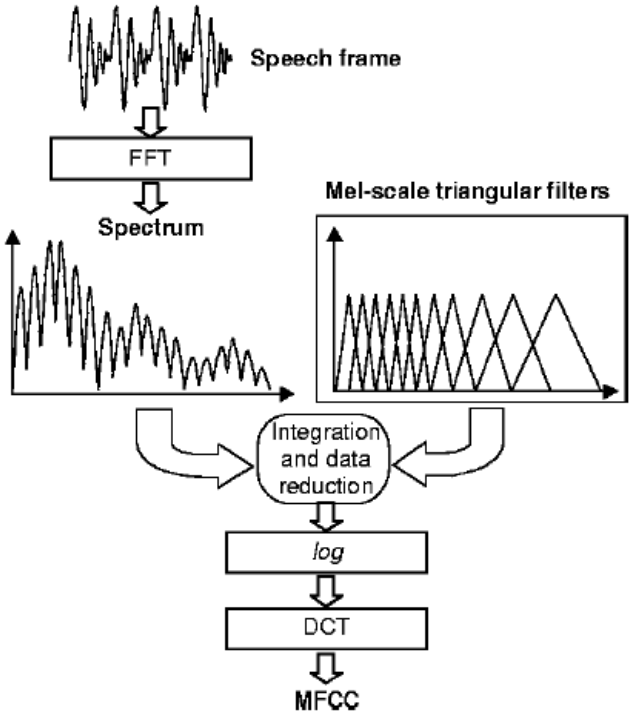
  - MFCC
  - PLP

- Non Spectral Features (Prosodic)

  - Pitch
  - Energy

- Delta Features



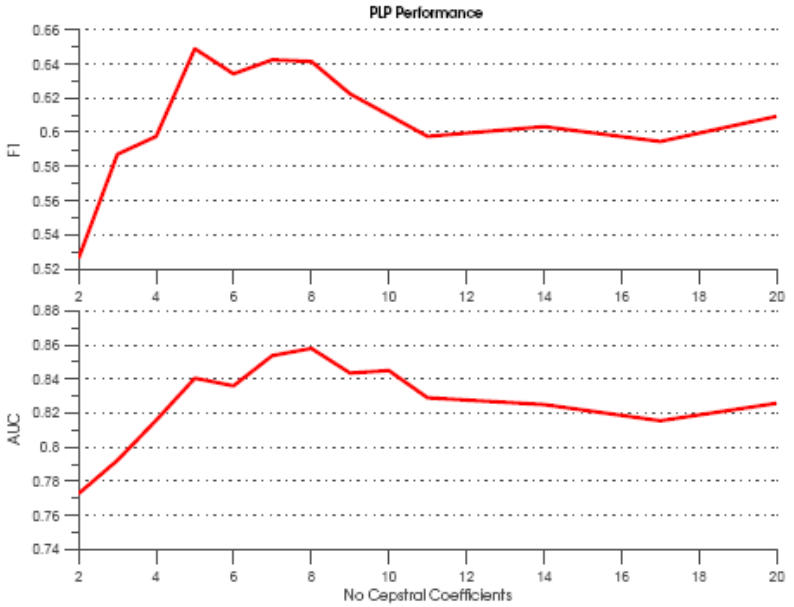



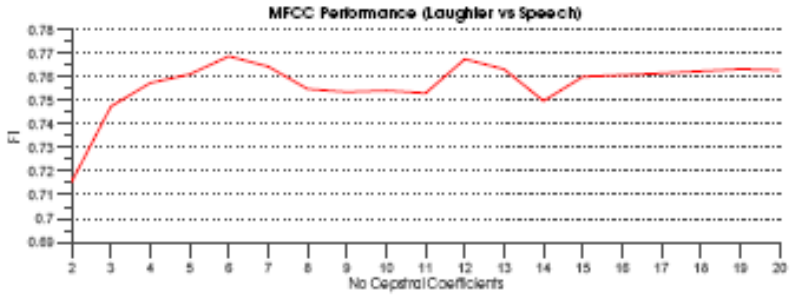
Imperial College London

# AUDIO FEATURES

- Spectral Features
  - 6 MFCC coeff. + 6 delta
  - Same results with PLP
  - Frame rate: 50 FPS, 50% overlap

- Prosodic Features
  - Pitch
  - Energy

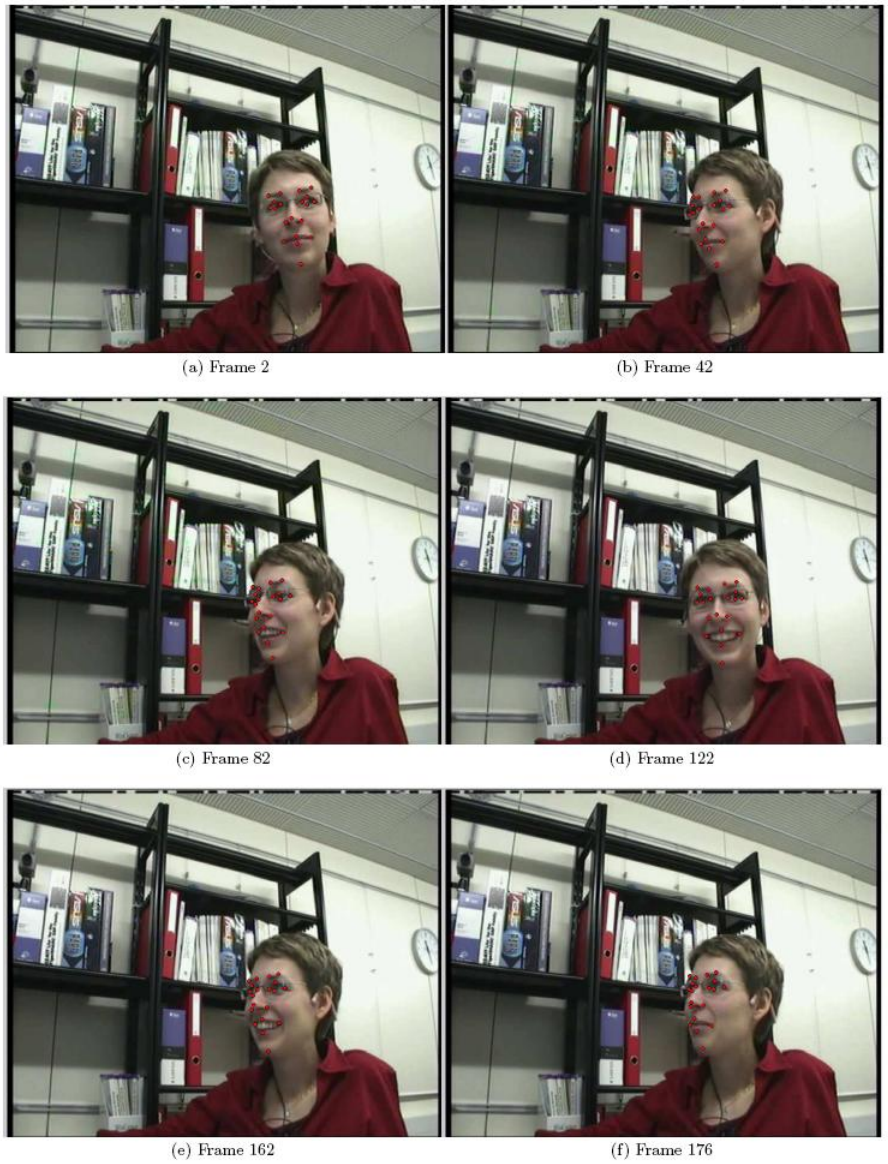


Imperial College London

# VISUAL FEATURES - TRACKING

- 20 facial points
- Eyes: 4 points each
- Eyebrows: 2 points each
- Nose: 3 points
- Mouth: 4 points
- Chin: 1 point



(a) Frame 2 (b) Frame 42

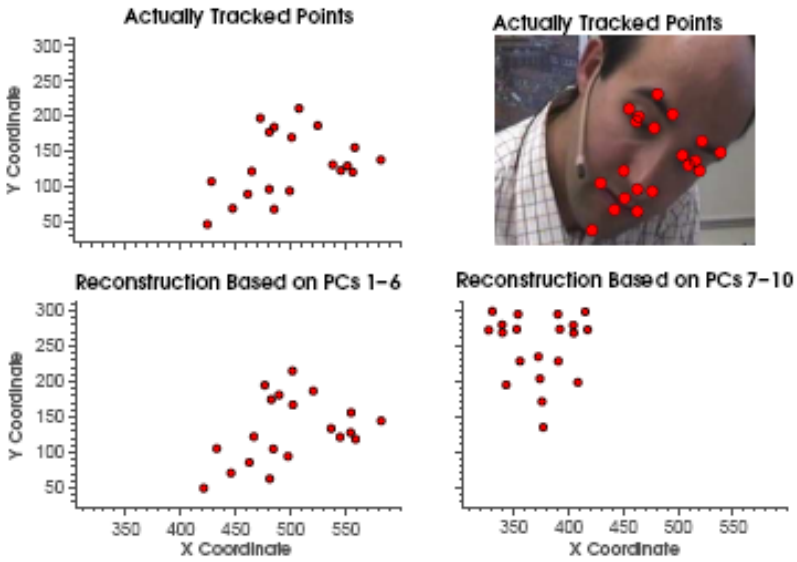
(c) Frame 82 (d) Frame 122

(e) Frame 162 (f) Frame 176

Imperial College London

# VISUAL FEATURES – DECOUPLING OF HEAD AND FACE

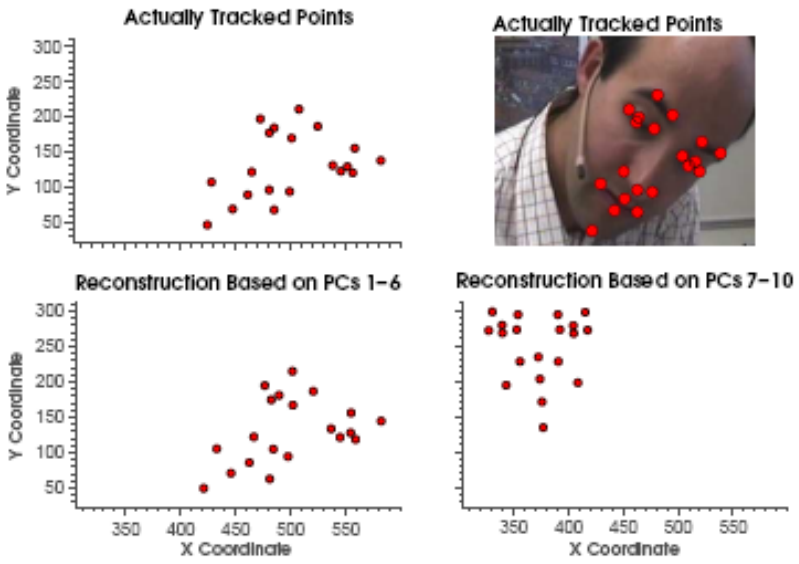
- Concatenate the (x,y) coord.
  → 40-dim vector
- k frames → k x 40 data matrix
- Find the PCs of the data matrix (PCA)
- Point Distribution Model with no alignment



Imperial College London

# VISUAL FEATURES – DECOUPLING OF HEAD AND FACE

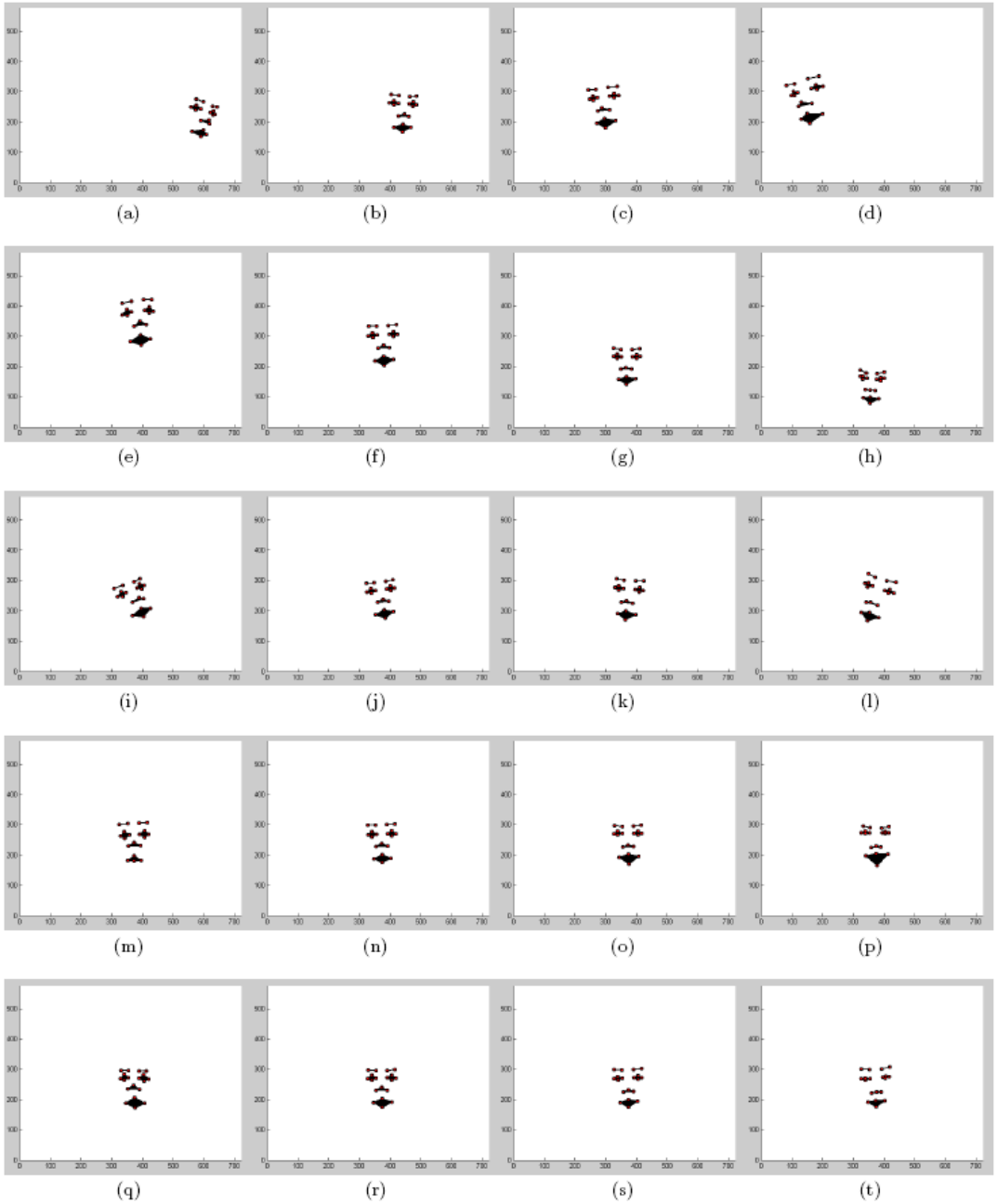
- Greatest variance of the data lies on the first PCs
- Greatest variance = head (rigid) movements
- First N(=5) PCs: reflect head movements (large variation)
- N+1...M(=6-10) PCs: reflect facial expressions (small variation)

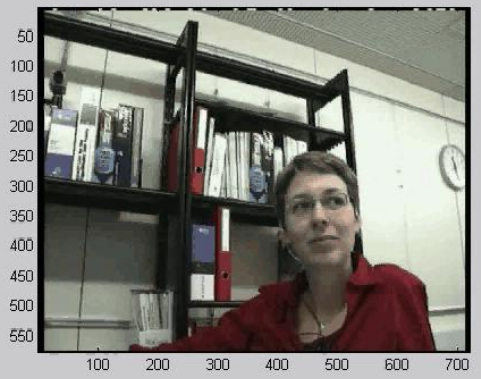


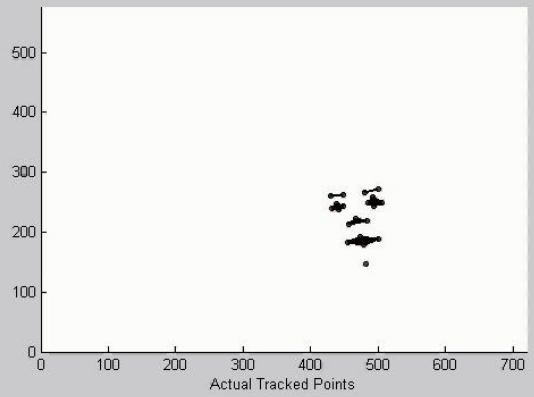

$$b = (x - \overline{x})P$$

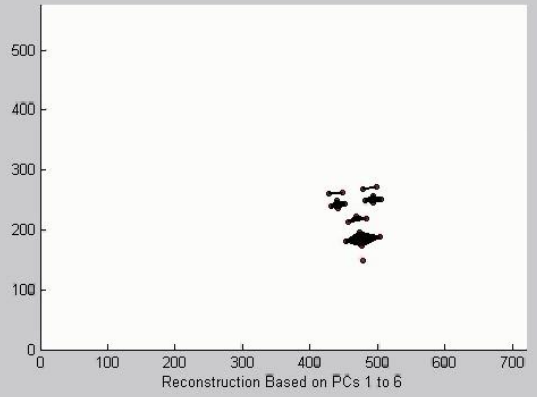
$$x \approx \hat{x} = \overline{x} + bP^T$$

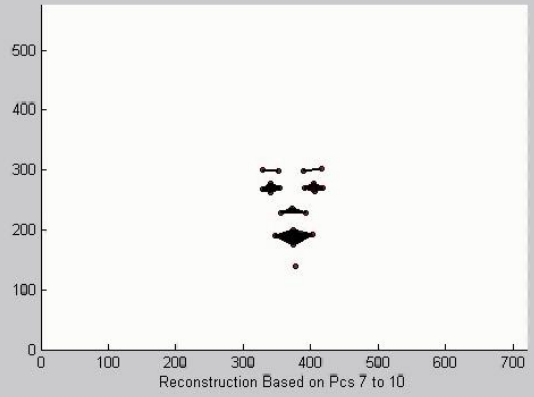
- P: eigenvector matrix (40xN)
- b: shape parameters, (1xN)


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)
(o)
(p)
(q)
(r)
(s)
(t)


50
100
150
200
250
300
350
400
450
500
550
100
200
300
400
500
600
700


500
400
300
200
100
0
0
100
200
300
400
500
600
700
Actual Tracked Points

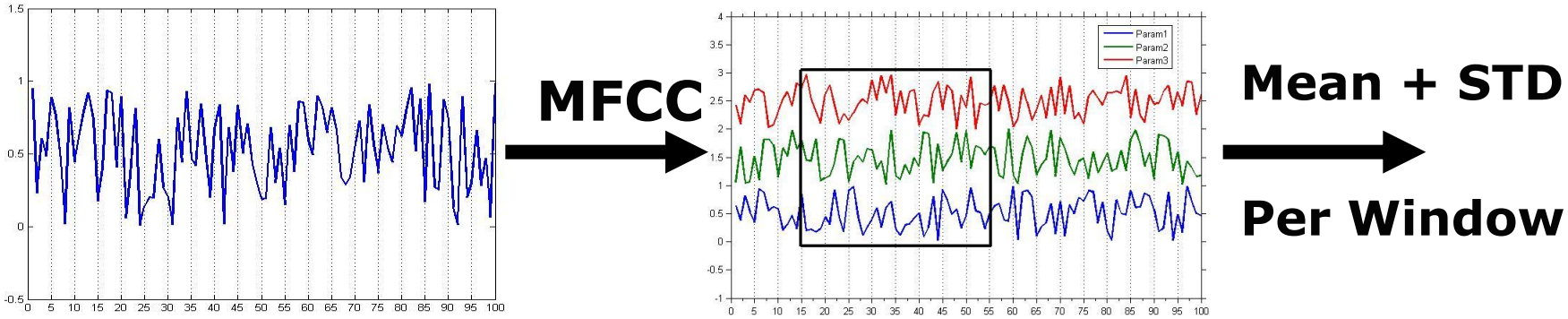

500
400
300
200
100
0
0
100
200
300
400
500
600
700
Reconstruction Based on PCs 1 to 6


500
400
300
200
100
0
0
100
200
300
400
500
600
700
Reconstruction Based on Pcs 7 to 10

Imperial College London

# FEATURE EXTRACTION

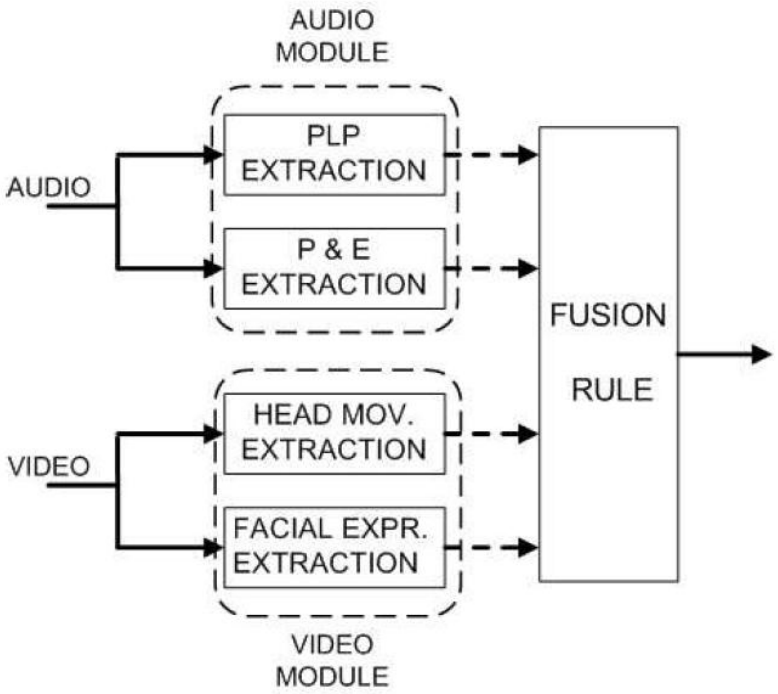
- Visual features
  - b1 – b5, b6 – b10 per frame
- Audio features
  - Mean + STD of MFCC, Pitch, Energy over 320ms window
  - Pitch Unvoiced Ratio



Imperial College
London

# SYSTEM OVERVIEW

- Leave-one-subject-out cross validation
- 2 / 3 classes: Laughter vs Speech
- Decision / Feature Level Fusion



Imperial College London

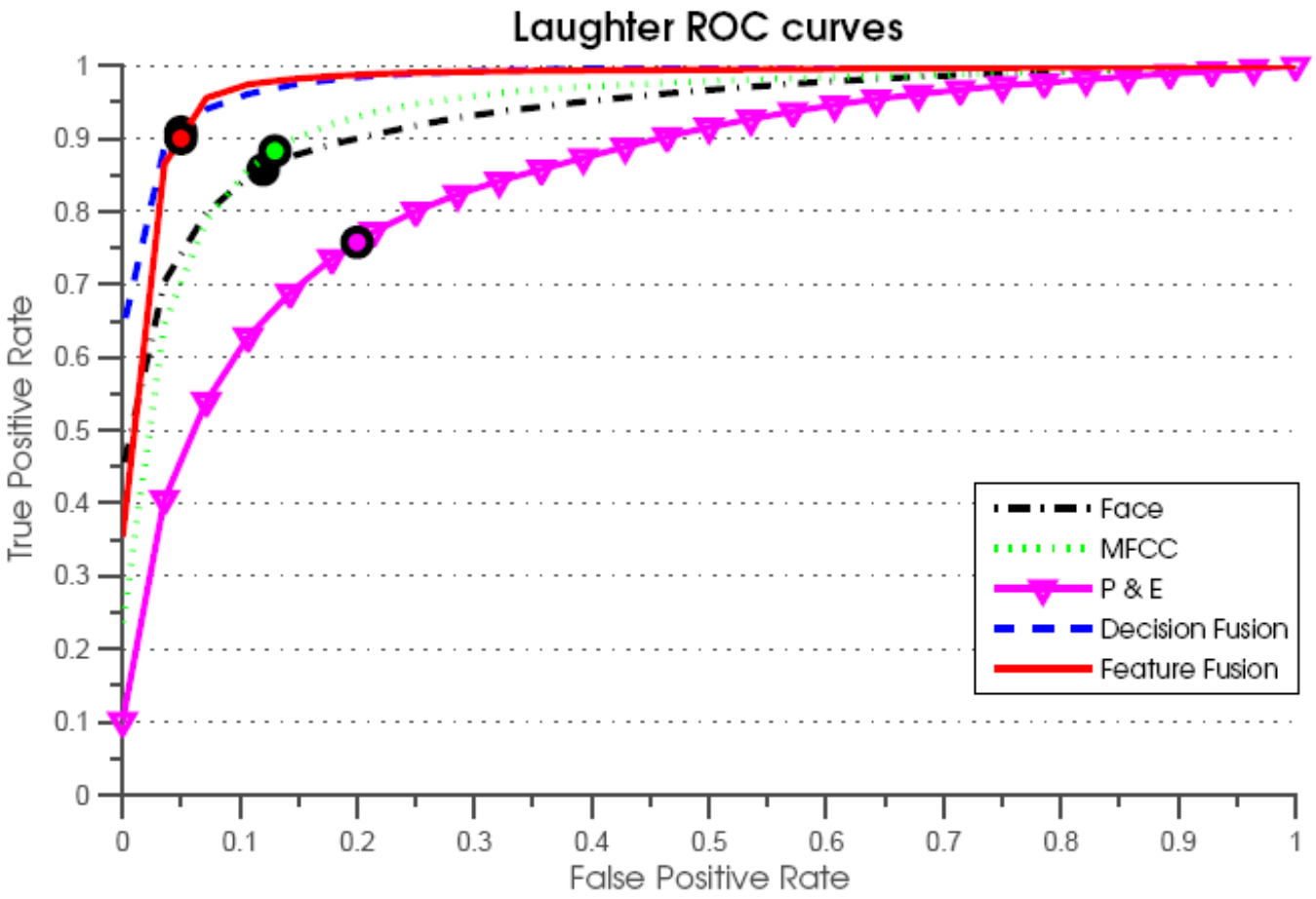
# RESULTS – LAUGHTER vs SPEECH

| *Cues* | *Features* | *F1* | *CR* |
|---|---|---|---|
| *FACE* | *b6 – b10* | *84.28* | *85.06* |
| *HEAD* | *b1 – b5* | *50.71* | *51.02* |
| *PROSODY* | *Pitch + energy* | *72.69* | *74.43* |
| *CEPSTRAL FEATURES* | *MFCC* | *86.16* | *86.69* |

| *Fusion* | *F1* | *CR* |
|---|---|---|
| *FACE + MFCC (DF)* | *93.17* | *93.37* |
| *FACE + MFCC + P&E (FF)* | *94.37* | *94.59* |

Imperial College
London

# RESULTS – AUDIOVISUAL



Imperial College London

# RESULTS – Voiced Laughter vs Unvoiced Laughter vs Speech

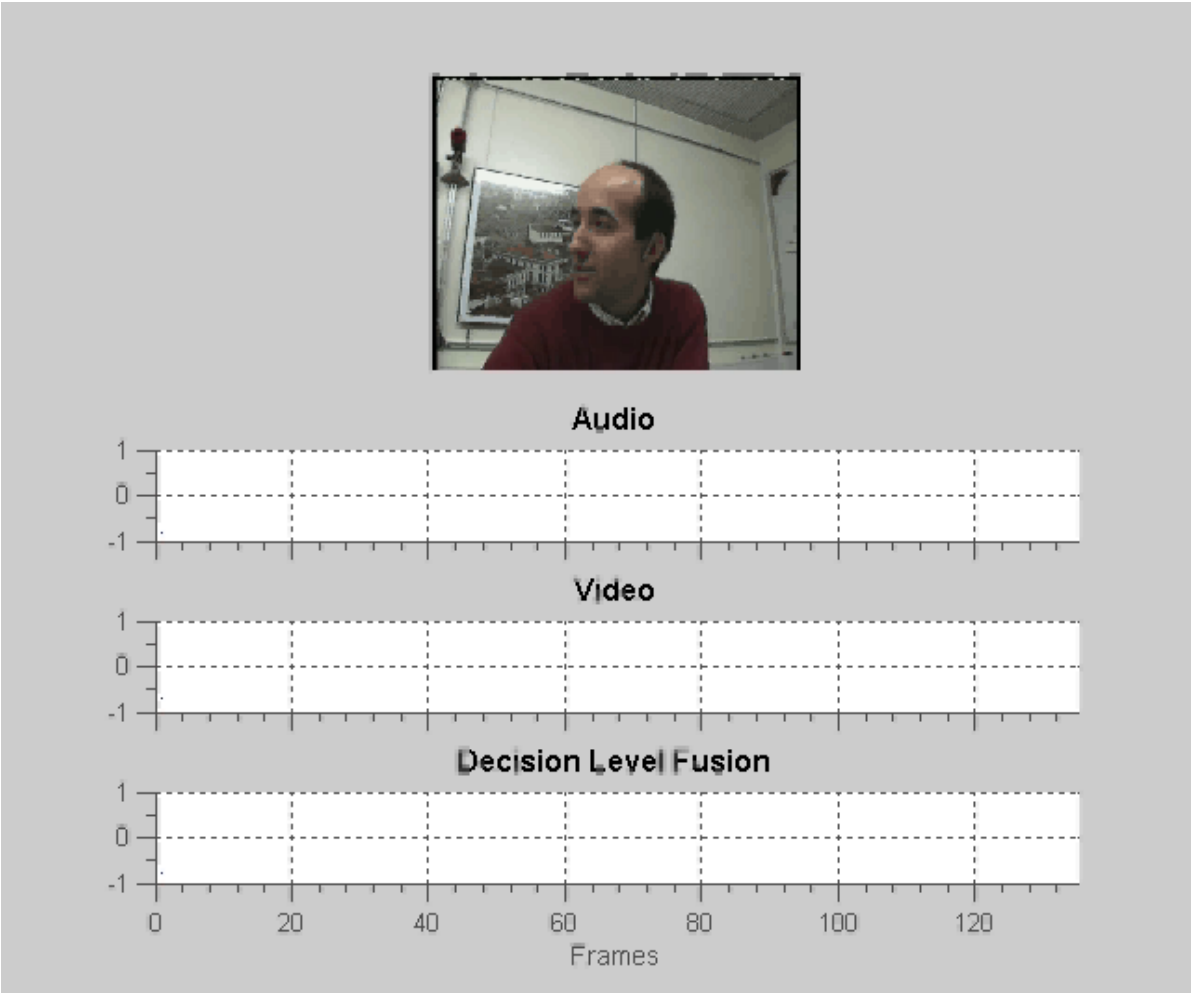
| *Cues* | *Features* | *F1 Unvoiced* | *F1 Voiced.* | *F1 Speech* | *CR* |
|---|---|---|---|---|---|
| *FACE* | *b6 – b10* | *37.21* | *57.48* | *84.39* | *67.03* |
| *HEAD* | *b1 – b5* | *19.32* | *34.56* | *49.72* | *37.27* |
| *PROSODY* | *Pitch + energy* | *57.24* | *59.65* | *74.39* | *66.20* |
| *CEPSTRAL FEATURES* | *MFCC* | *57.97* | *69.14* | *85.68* | *74.83* |

| *Fusion* | *F1 Unvoiced* | *F1 Voiced* | *F1 Speech* | *CR* |
|---|---|---|---|---|
| *FACE + MFCC + PE (DF)* | *66.14* | *74.62* | *91.89* | *81.00* |
| *FACE + MFCC + P&E + HEAD(FF)* | *66.80* | *78.86* | *94.14* | *83.51* |

Imperial College London

# FEATURE SELECTION

| Speech | Voiced | Unvoiced |
|---|---|---|
| B7 | B7 | Mean Energy |
| Mean MFCC2 | Std Energy | Mean MFCC6 |
| Std MFFC2 | Mean MFCC1 | Mean MFCC1 |
| Std MFCC4 | B10 | B7 |
| B10 | B8 | Mean MFCC2 |
| Std MFCC3 | B4 | B6 |
| Mean Pitch | Mean MFCC3 | Std Pitch |


Audio
1
0
-1
Video
1
0
-1
Decision Level Fusion
1
0
-1
0
20
40
60
80
100
120
Frames

# Conclusions

- Audiovisual fusion leads to improved performance
- Best cues: cepstral features, facial expressions
- High accuracy for speech vs laughter
- Discriminating 2 types of laughter is challenging

THANK YOU! ☺