Imperial College London

Audiovisual Classification of Vocal Outbursts in Human Conversation using Long-Short-Term Memory Networks

Florian Eyben, Stavros Petridis, Björn Schuller, George Tzimiropoulos, Stefanos Zafeiriou, Maja Pantic

Department of Computing,
Imperial College London, UK

Institute for Human-Machine Communication
Technische Universität München, Germany

Motivation - Objectives

Vocal outbursts = Non-linguistic vocalisations

(breathing, sighing, laughing, coughing, …)

Non-linguistic vocalisations play an **important role in spontaneous speech** and conversations

Video-based features in addition to audio features might **increase performance of current audio only recognition systems**

Database – TUM AVIC

- AVIC = Audiovisual Interest Corpus
- Interactive product presentations

- 21 subjects (10 female)
- 3,901 turns with speech from 10,5 h total

- Audio part used in the 2010 INTERSPEECH Paralinguistic Challenge

Database – TUM AVIC

- Non-linguistic vocalisations (NLV) labeled:

  Breath, Consent, Hesitation, Laughter, Garbage

- “Breath” is excluded here

[#]	Train	Eval
Garbage	420	161
Consent	218	91
Hesitation	731	403
Laughter	204	63

Example - Laughter

Example - Consent

Example - Hesitation

Method

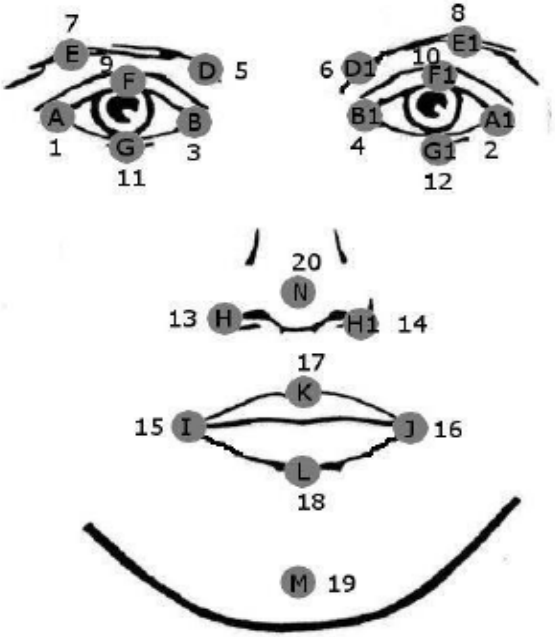
- Classification of isolated NLV
- NLV extracted by forced aligning speech

- Low-level feature fusion on frame level:
  - Video features
  - Audio features
- Modeling:
  - dynamic (LSTM-RNN)
  - static (SVM)

Video features

- Frame-rate: 25 fps
- Shape features
  - Based on Point Distribution Model
- Appearance features
  - PCA on gradient orientations

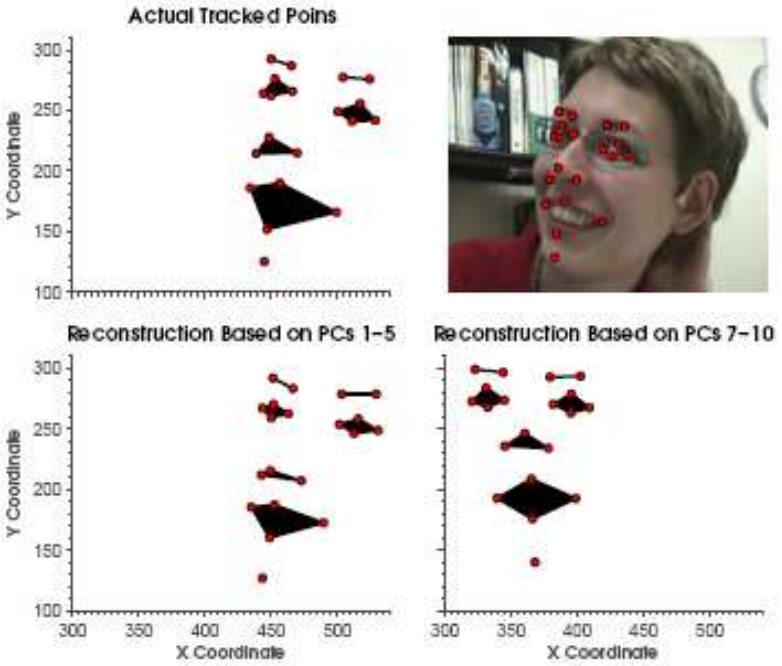
VISUAL FEATURES - TRACKING



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

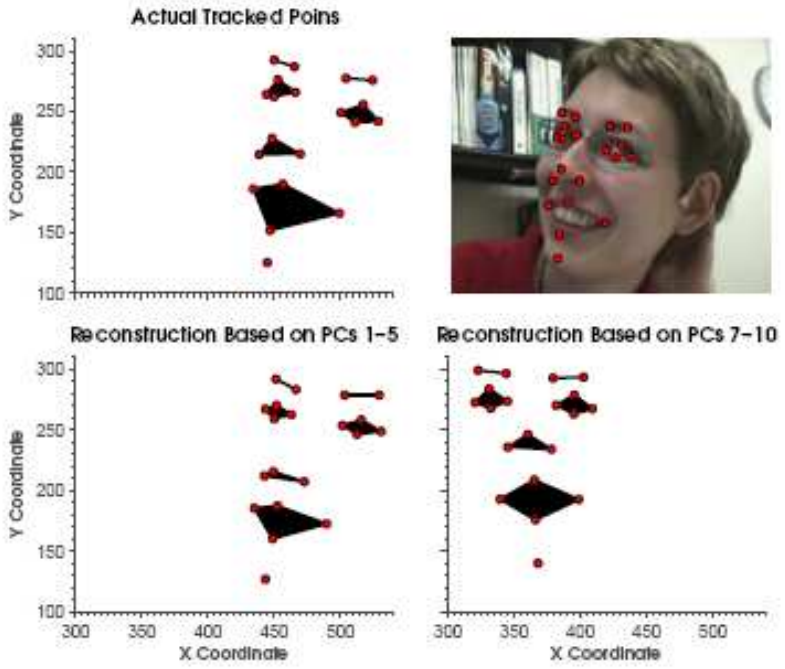
VISUAL FEATURES – DECOUPLING OF HEAD AND FACE

- 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



VISUAL FEATURES – DECOUPLING OF HEAD AND FACE

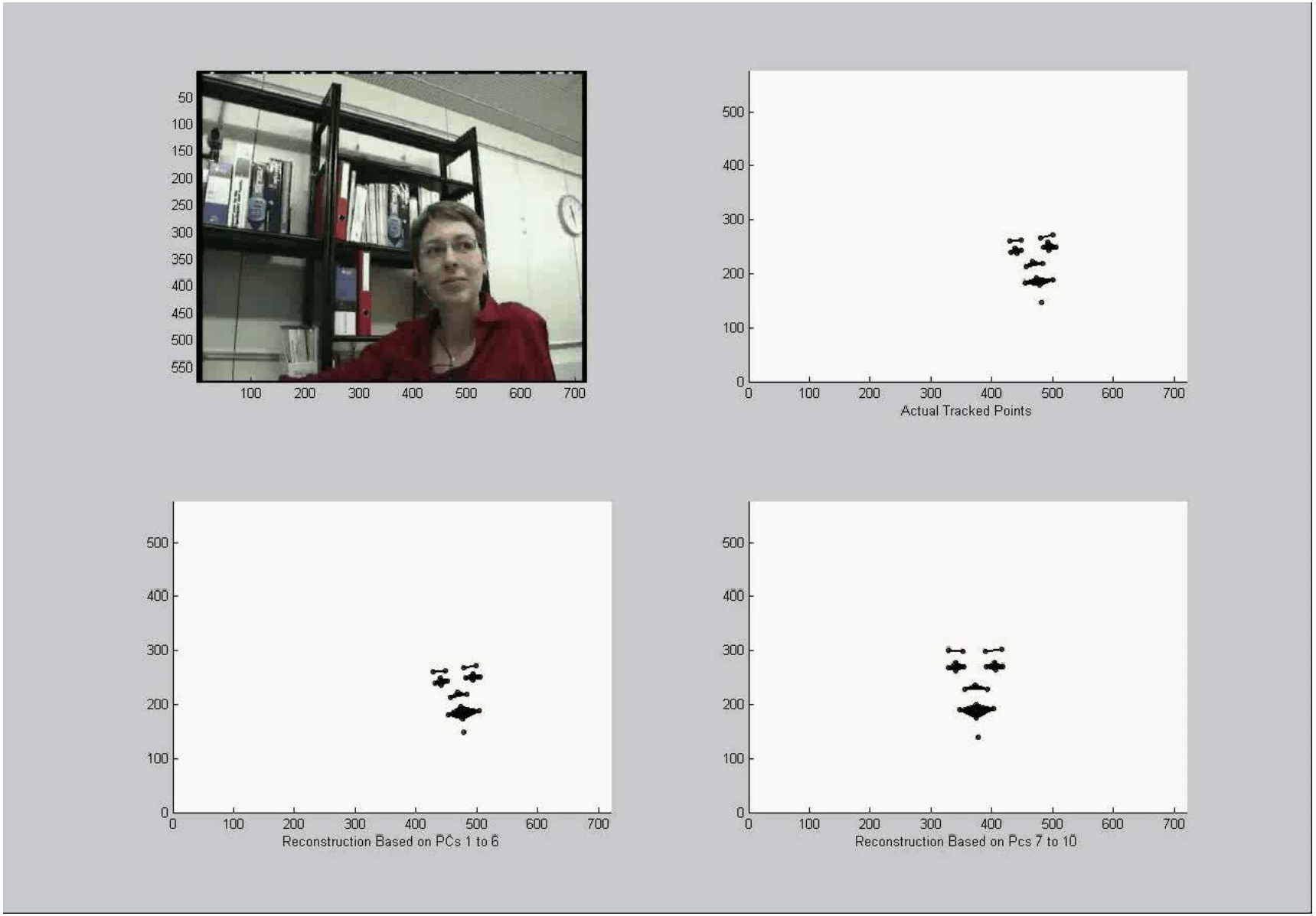
- Greatest variance of the data lies on the first PCs
- Greatest variance = head (rigid) movements
- First N(=4) PCs: reflect head movements (large variance)
- N+1...M(=5-10) PCs: reflect facial expressions (small variance)



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

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

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


Actual Tracked Points
Reconstruction Based on PCs 1 to 6
Reconstruction Based on Pcs 7 to 10

Appearance Features

- Compute the affine transformation between each frame and the reference frame (scale, translation, rotation)
- Warp each face to the reference frame
- Apply PCA to image gradient orientation

G. Tzimiropoulos, S. Zafeiriou, M. Pantic *Principal Component Analysis of Image Gradient Orientations for Face Recognition*. *IEEE FG'11*

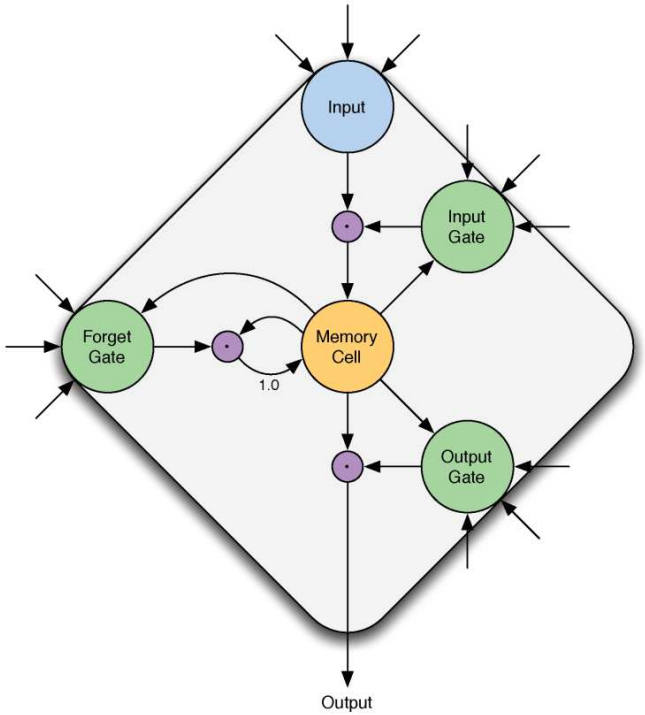
Fig. 1: Original frame (left) with tracked facial points and warped frontal face example (right). TUM AVIC corpus.

Audio features

- Computed with openSMILE
- Frame-rate: 100 fps
- 9 acoustic low-level descriptors:
  - Perceptual Linear Prediction Cepstral Coeff. 1-5
  - Logarithmic Energy
  - Loudness
  - Fundamental Frequency
  - Probability of Voicing
- First and second order delta coefficients

Dynamic approach: LSTM-RNN

- Frame wise classification with LSTM-RNN
- Topology:
  - 1 input unit for every feature
  - 125 LSTM cells in hidden layer
  - 4 output units (soft-max)
- Training: Resilient propagation
- Segment: Majority vote of frame-wise predictions



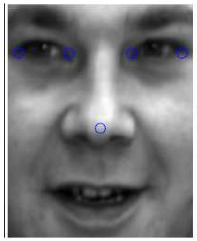
Static approach

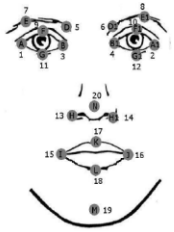
- Projection of low-level features via
  7 functionals:
  - Extremes (max, min)
  - Range (max – min)
  - Arithmetic mean
  - Standard deviation, skewness, kurtosis

→ static feature vector per NLV segment

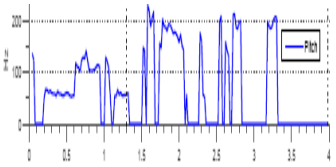
- Classification with Support-Vector Machines

System Overview





PLP

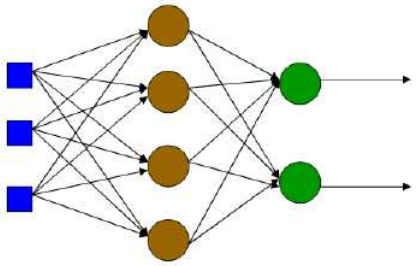


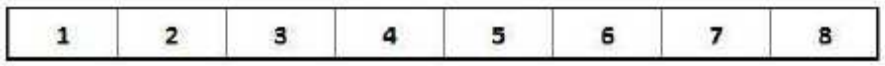
Shape
Appearance
PLP
F0, Energy
Loudness
Prob. Voicing

Functionals

SVM

Sequence Label





Majority

Sequence Label

Shape
Appearance
PLP
F0, Energy
Loudness
Prob. Voicing

Results

[%] Features	LSTM UAR	LSTM WAR	SVM UAR	SVM WAR
Appear	32.5	50.0	31.8	60.0
Shape	48.4	56.1	39.6	60.2
Shape+Appear	40.8	51.8	39.2	58.2
Audio	64.6	73.5	59.1	72.4
Audio+Appear	60.3	64.2	59.4	72.1
Audio+Shape	**72.0**	**73.5**	60.5	72.4
Audio+Shape+Appear	64.3	63.1	62.7	74.2

Results for multimodal non-lingusitic vocalisation classification on TUM AVIC. Low-level feature fusion.

Results

Confusion matrix of LSTM-RNN for Audio (left) and Audio+Shape (right):

[%] as →	GAR	CON	HES	LAU
GARBAGE	**62.1/65.2**	1.2/1.9	27.3/16.8	9.3/16.1
CONSENT	24.2/15.4	**47.3/65.9**	26.4/17.6	2.2/1.1
HESITATION	9.4/13.4	4.0/7.9	**85.6/77.7**	1.0/1.0
LAUGHTER	20.6/15.9	1.6/0.0	14.3/4.8	**63.5/79.4**

→ improvements for *consent* and *laughter*

Summary

- LSTM-based NLV modeling superior to static SVM modeling
- Shape features improve performance for *consent* and *laughter*

Outlook

- Spotting of NLV in continuous speech
  → inclusion into audiovisual ASR framework
- Bottle-neck LSTM topologies

http://ibug.doc.ic.ac.uk/

http://www.schuller.it/

Thank You.