

# Multimodal emotion recognition: Implementation, results and challenges



ERMIS partners  
(KCL, QUB, ICCS)

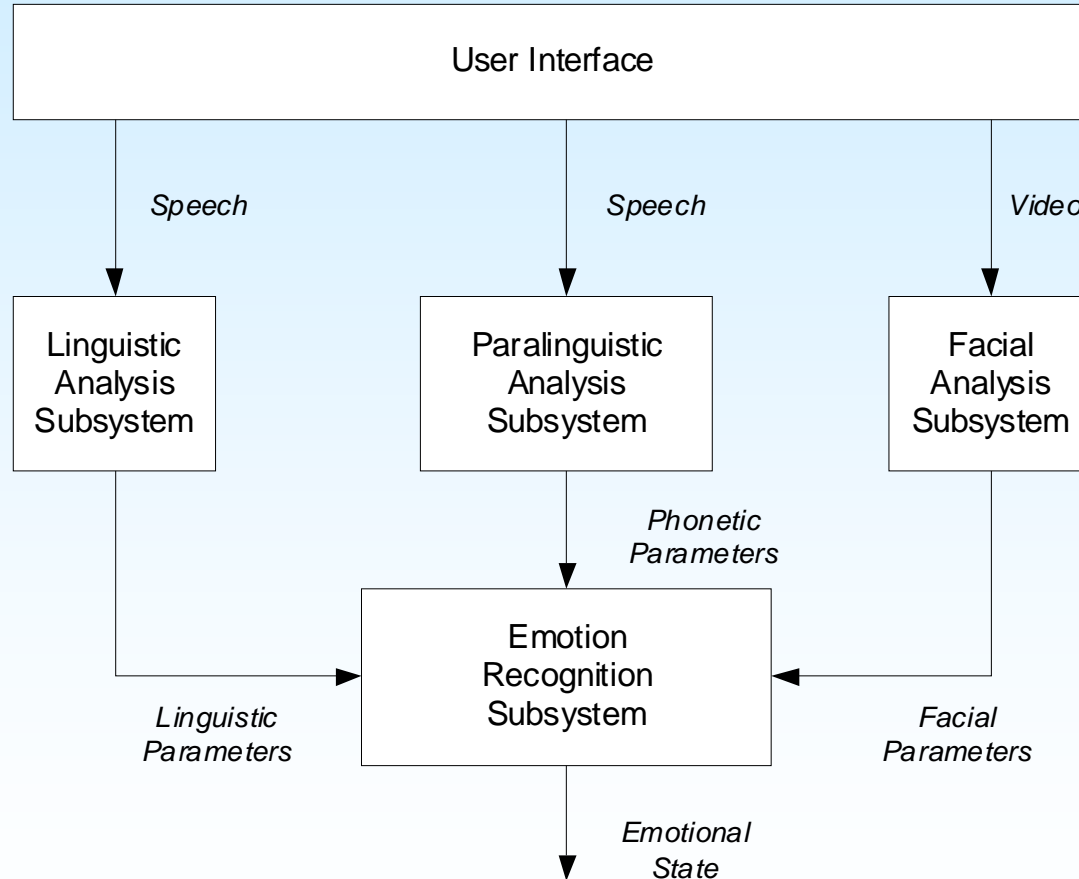
# Overview

- Introduction to the ERMIS system
- Basic system components
- System architecture presentation
- Project results
- Challenges identified

# ERMIS project overview

- Development of a prototype system for human computer interaction that can interpret its users' attitude or emotional state,
  - e.g., activation/interest, boredom, and anger, in terms of their speech and/or their facial gestures and expressions
- Both unimodal (visual) and multimodal (audio-visual) input approaches

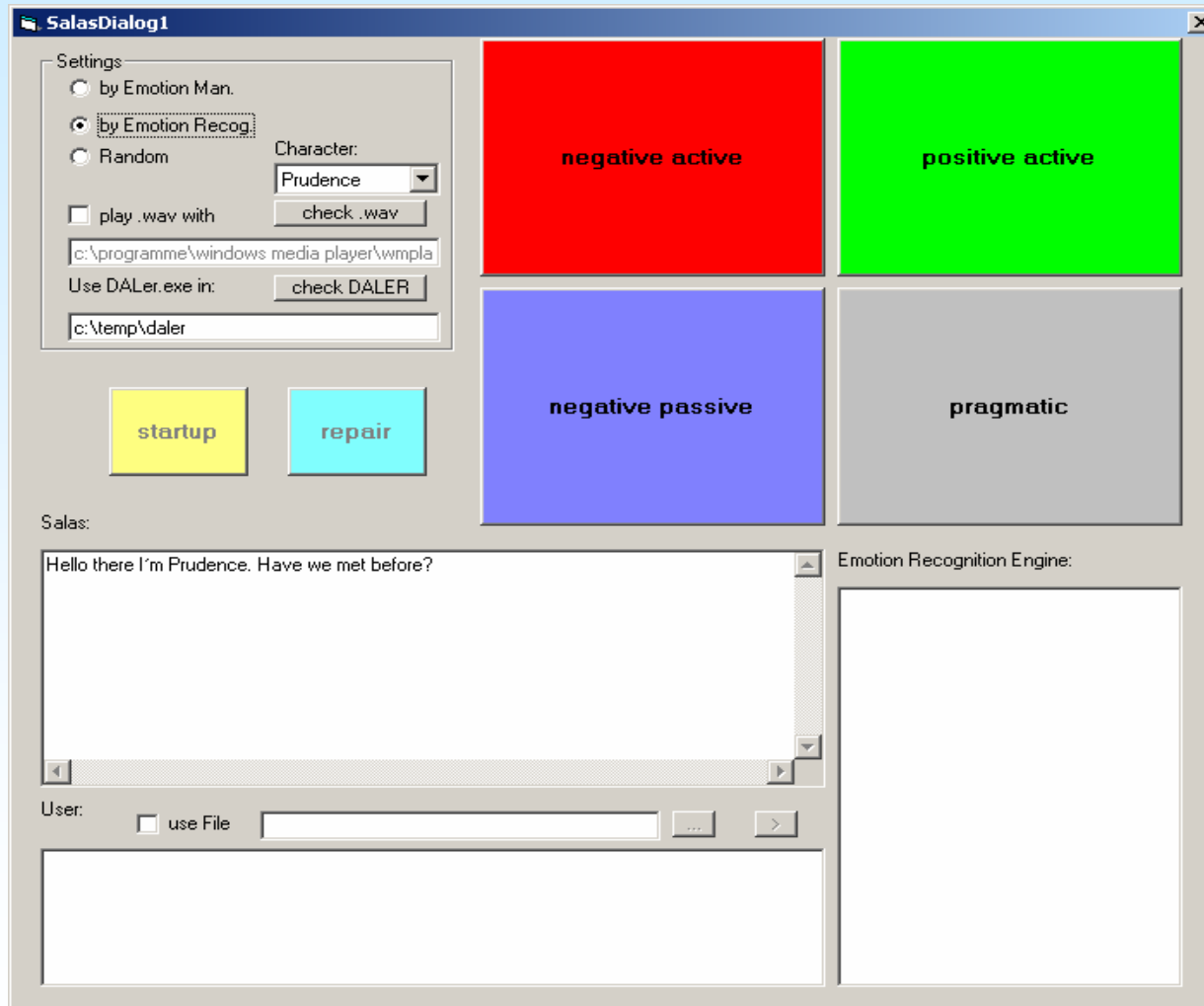
# Basic System Components



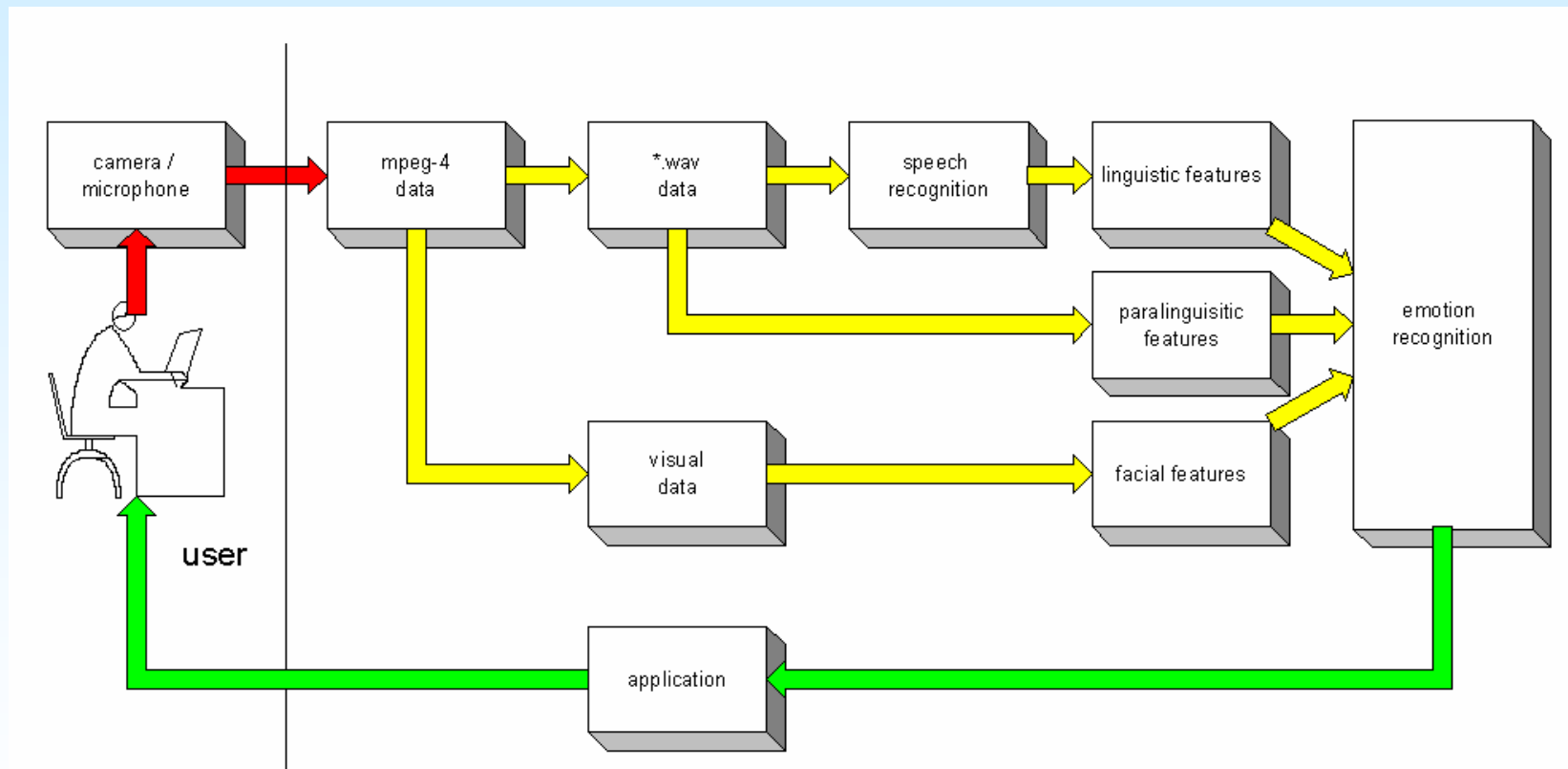
# ERMIS testbed application

- SALAS - Sensitive Artificial Listener - a software application designed to let a user work through various emotional states
- The system invokes and captures emotional input in the form of emotional speech and facial expressions from the user.
- "Wizard of Oz" (WoZ) - no automated emotion recognition module
- a WoZ identifies the emotional tone and uses a Windows-based application to select the system's response from a relevant subset of the available verbal cues.
- The user selects one of four "artificial listeners"/ personas to interact with at any given time.

# SALAS



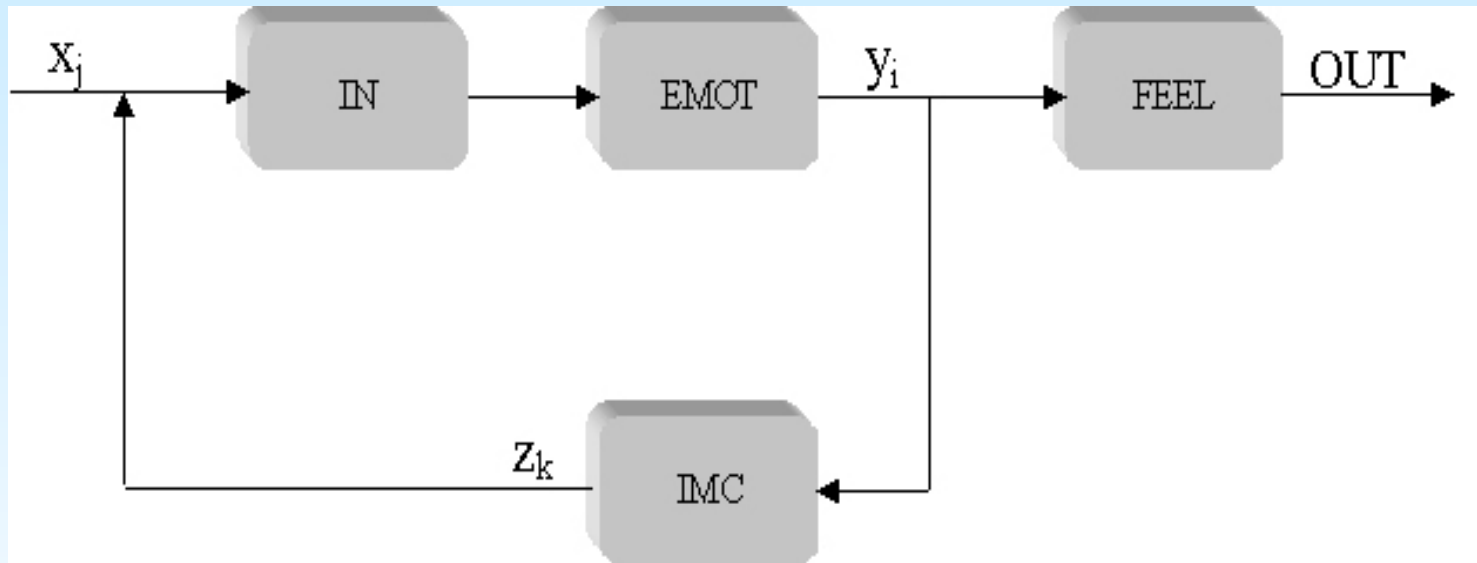
# System Architecture



# Anna - An Artificial Neural Network for Attention to Emotion Recognition

- An artificial emotion processor based on brain function knowledge
- Emotional experience has 2 components: 'automatic' and attended
- Standard feed-forward net, with additional attention feedback from the hidden layer
- Produces emotional classification output from a set of input features

# ANNA Architecture

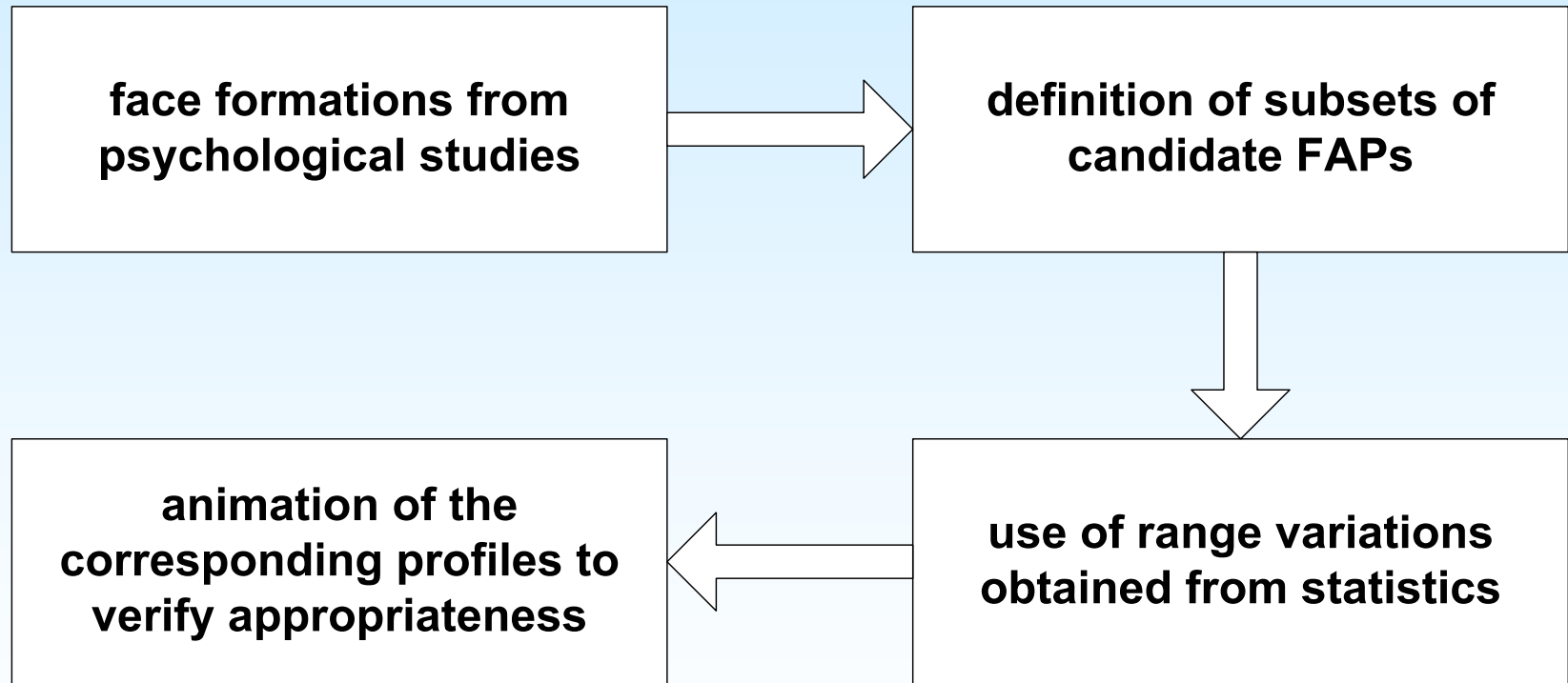


- Attention acts as a feedback modulation onto the feature inputs, so as to amplify or inhibit the various feature inputs, as they are or are not useful for the emotional state detection
- feedback layer (IMC= inverse model controller = attention movement signal generator) modulating the activity in the inputs to the hidden layer (EMOT)

# Emotion recognition system based on facial expression analysis

- Rule based system
- Face detection and face feature detection
- MPEG-4 facial animation parameter (FAPs) extraction
- Characterization using 6 universal/archetypal expressions: joy, surprise, fear, anger, disgust, sadness

# how rules were derived

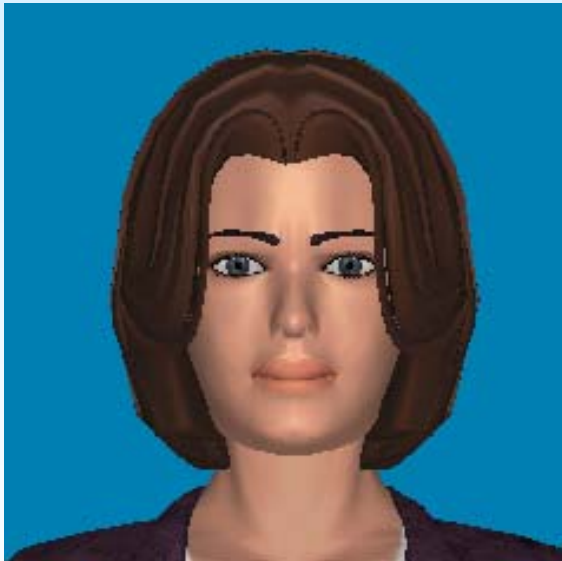


# Archetypal Expression Profiles

- Profile: set of FAPs accompanied by the corresponding range of variation

Surprise ( $P_{Su}^{(0)}$ )	$F_3 \in [569,1201]$ , $F_5 \in [340,746]$ , $F_6 \in [-121,-43]$ , $F_7 \in [-121,-43]$ , $F_{19} \in [170,337]$ , $F_{20} \in [171,333]$ , $F_{21} \in [170,337]$ , $F_{22} \in [171,333]$ , $F_{31} \in [121,327]$ , $F_{32} \in [114,308]$ , $F_{33} \in [80,208]$ , $F_{34} \in [80,204]$ , $F_{35} \in [23,85]$ , $F_{36} \in [23,85]$ , $F_{53} \in [-121,-43]$ , $F_{54} \in [-121,-43]$
$P_{Su}^{(1)}$	$F_3 \in [1150,1252]$ , $F_5 \in [-792,-700]$ , $F_6 \in [-141,-101]$ , $F_7 \in [-141,-101]$ , $F_{19} \in [-350,-324]$ , $F_{20} \in [-346,-320]$ , $F_{21} \in [-350,-324]$ , $F_{22} \in [-346,-320]$ , $F_{31} \in [314,340]$ , $F_{32} \in [295,321]$ , $F_{33} \in [195,221]$ , $F_{34} \in [191,217]$ , $F_{35} \in [72,98]$ , $F_{36} \in [73,99]$ , $F_{53} \in [-141,-101]$ , $F_{54} \in [-141,-101]$
$P_{Su}^{(2)}$	$F_3 \in [834,936]$ , $F_5 \in [-589,-497]$ , $F_6 \in [-102,-62]$ , $F_7 \in [-102,-62]$ , $F_{19} \in [-267,-241]$ , $F_{20} \in [-265,-239]$ , $F_{21} \in [-267,-241]$ , $F_{22} \in [-265,-239]$ , $F_{31} \in [211,237]$ , $F_{32} \in [198,224]$ , $F_{33} \in [131,157]$ , $F_{34} \in [129,155]$ , $F_{35} \in [41,67]$ , $F_{36} \in [42,68]$
$P_{Su}^{(3)}$	$F_3 \in [523,615]$ , $F_5 \in [-386,-294]$ , $F_6 \in [-63,-23]$ , $F_7 \in [-63,-23]$ , $F_{19} \in [-158,-184]$ , $F_{20} \in [-158,-184]$ , $F_{21} \in [-158,-184]$ , $F_{22} \in [-158,-184]$ , $F_{31} \in [108,134]$ , $F_{32} \in [101,127]$ , $F_{33} \in [67,93]$ , $F_{34} \in [67,93]$ , $F_{35} \in [10,36]$ , $F_{36} \in [11,37]$

# Example of Anger Profiles



# ..how rules were derived

- They embody common knowledge on emotion and facial expression - psychological findings about emotion representation
- They were verified on image databases
- And quantified in terms of FAP intensities

# ERMIS Project Results

- ERMIS Applications
  - Sensitive Artificial Listeners (SALAS) -data capture application
  - The speech recognition module, operating in English and Greek languages.
  - The linguistic feature extraction module, providing English terms that can deduce emotional content in speech.
  - The paralinguistic speech analysis and feature extraction module.
  - The facial detection, facial feature estimation and MPEG-4 expression related feature extraction.
  - The emotion recognition module based on facial expression analysis
  - The emotion recognition module based on audio-visual feature analysis.
- ERMIS emotional database
  - Samples of reasonably naturalistic emotional behavior
  - Both audio and visual modalities

# Challenges - Modality Interaction

- Difficult to catch the *interaction* between modalities,
  - a tune can contain a plethora of information and be coupled with a single "indicative" facial expression frame
  - The characteristic frame to a tune is the one that correlates the least with the others
    - Rapid emotion changes in a single tune are not caught
  - When the user was talking the FAP info on the mouth was not taken into account
  - When a subject is silently smiling, although no tune is associated to the frames, the facial stimuli is enough for emotion recognition

# Example frame



London, 5th July 2005

# Audiovisual Processing Challenge

- Different time base speech uses a window based approach, while facial analysis uses a frame based approach
  - Speech processing uses tunes i.e. that is the period between two pauses
  - No synchronization (only on a time code level and not on an event driven base)
- No combined audiovisual processing
  - pause/ viseme detection
  - Idea to fortify text extraction process using viseme information

# Suggested future steps

- Inclusion of richer multimodal information
  - Gesture analysis as well as facial expression analysis
  - Focus research on catching the modality interactions
  - Deeper linguistic feature extraction, semantic analysis of content
  - Dealing with uncertainty in facial emotion recognition / addition of flexibility

# Open issues in multimodal processing

- Controlled conditions in lighting, head direction etc => reduce naturalness
- Humans can discern emotional states of others with much poorer input
- Ideas for a multimodal approach that requires less strict data collection conditions?

# References

- ERMIS project web site:

[www.image.ntua.gr/ermis](http://www.image.ntua.gr/ermis)

