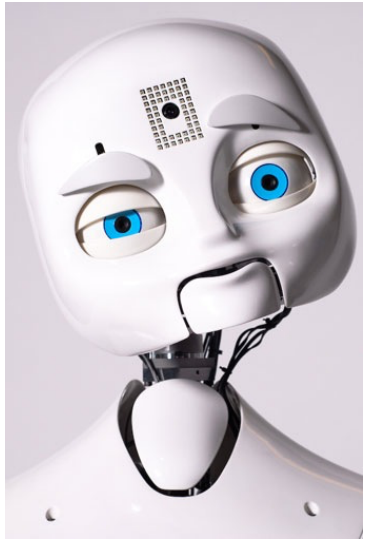
Affective computing from speech: Towards robust recognition of emotions in ecologically valid situations

Fabien Ringeval
Laboratoire d'Informatique de Grenoble
Université Grenoble-Alpes

Motivation

- **Interactive technology is getting more and more ubiquitous**

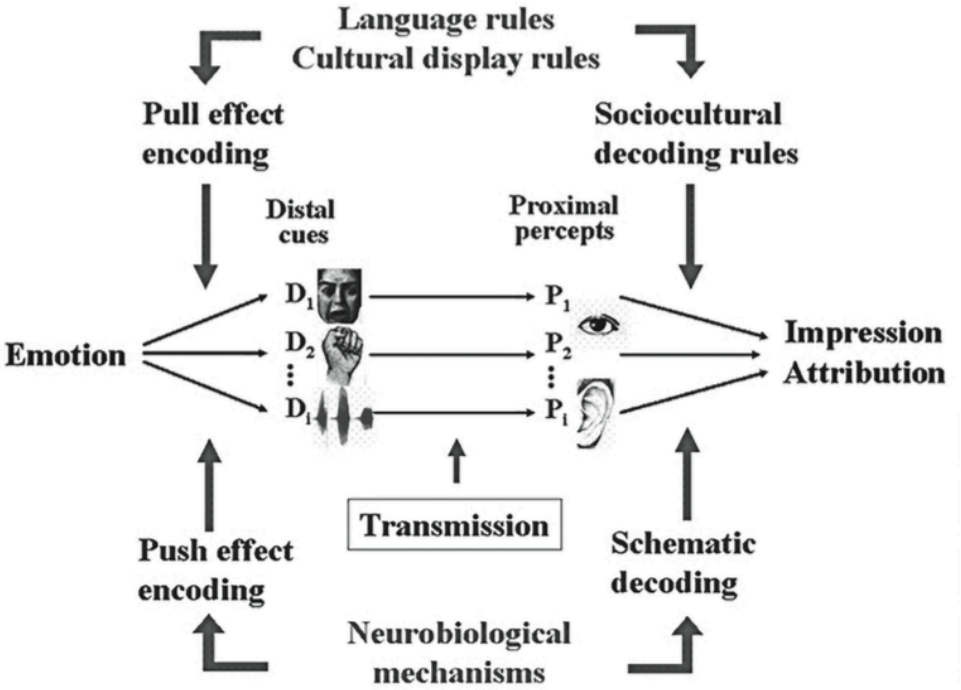


Motivation

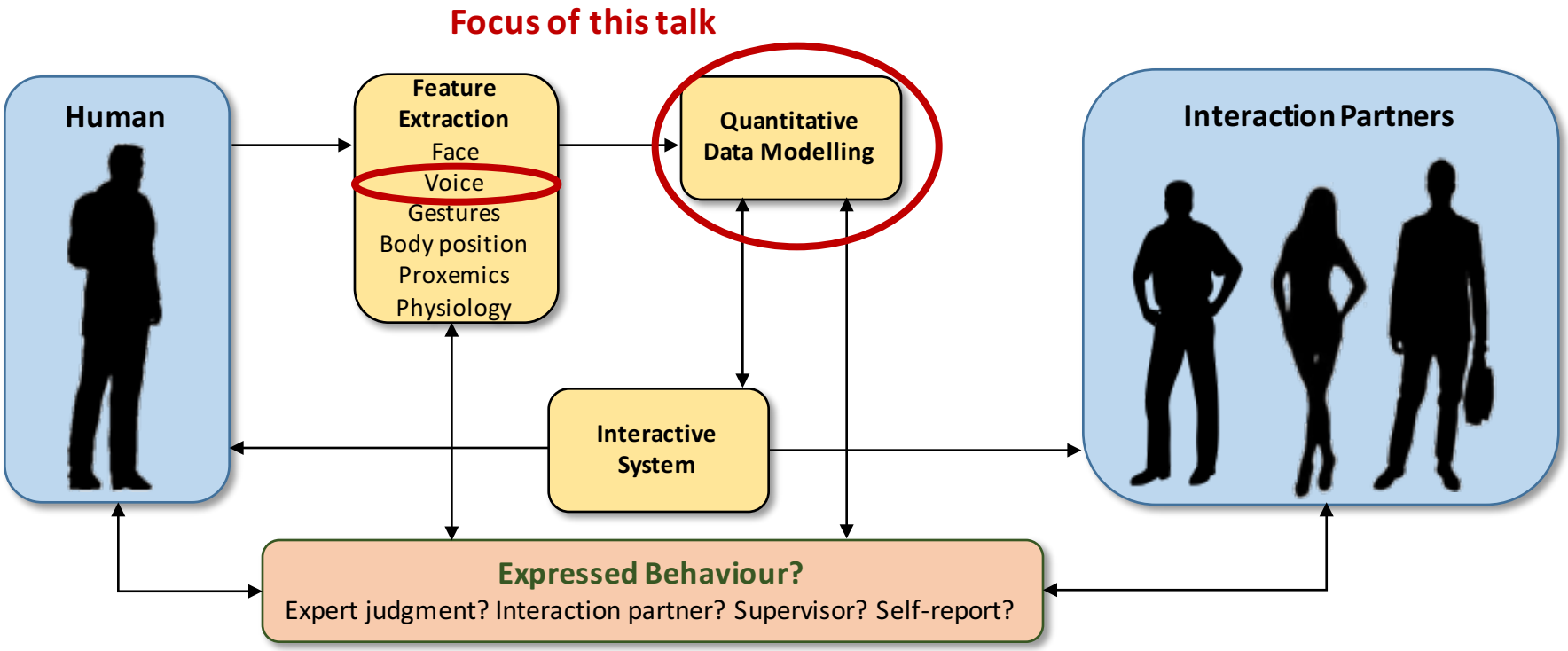
- **Engineering research focused on human-centered technology**
  - **Automatic human behaviour understanding**
    - **General interaction dynamics**
    - **Developmental, psychiatric disorders**
    - **Depression, suicide risk, marital therapy**
    - **Solitude, gerontechnology**
  - **Overlap with many research domains**
    - **Social & Human Sciences**
    - **Psychology, Psychiatry**
    - **Signal processing**
    - **Machine learning**

Human Behaviour

- **Push-Pull effect (Scherer) — WYSIWYG**

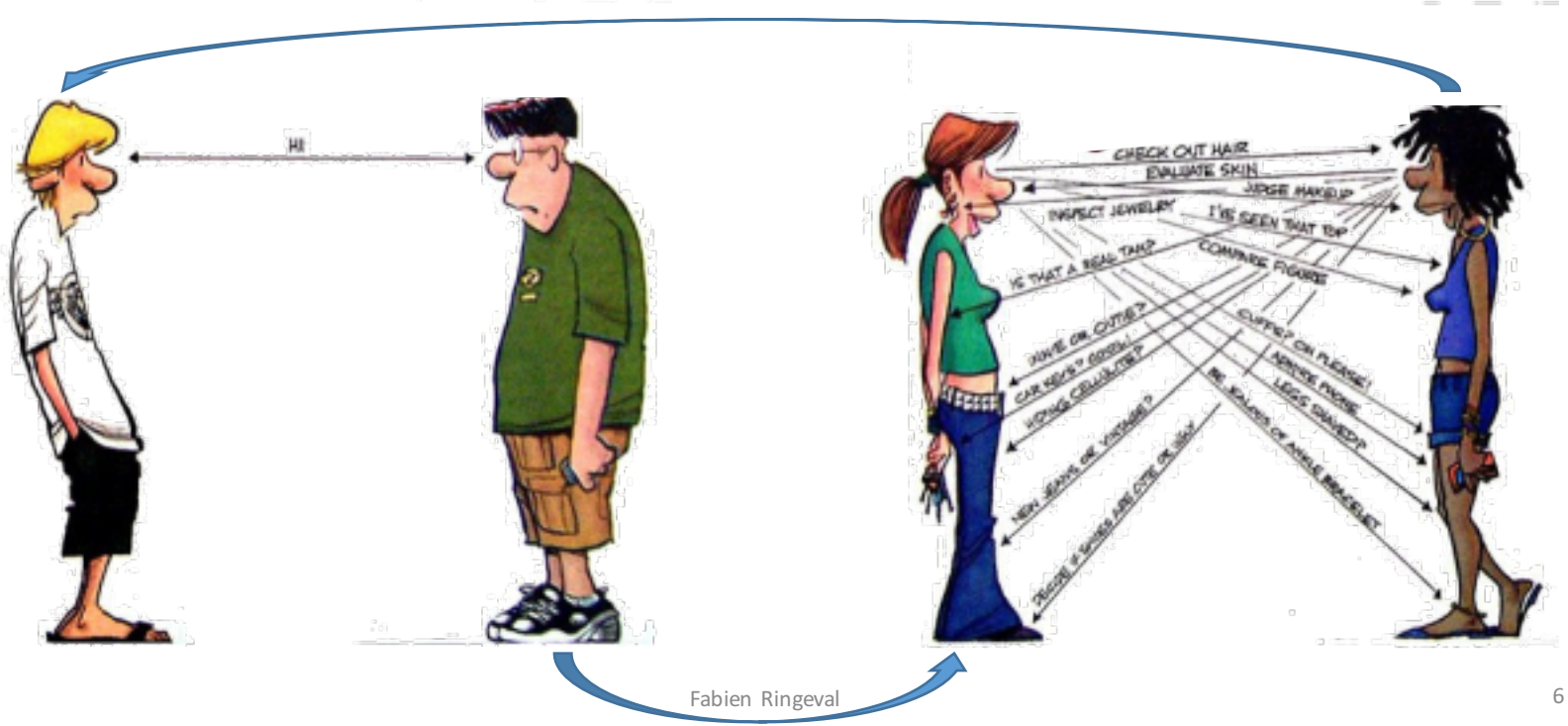


Human Behaviour Understanding



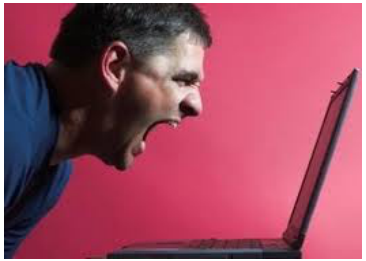
Human Behaviour Understanding from Speech

- **Paralinguistic – reading between the lines: « How » vs « What »**



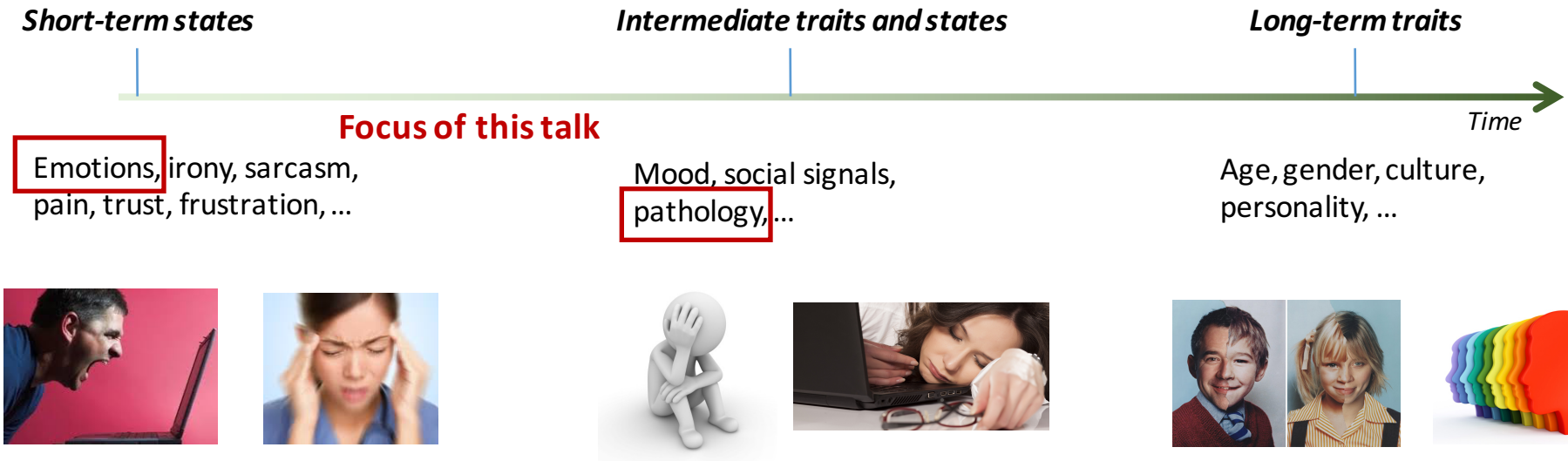
Human Behaviour Understanding from Speech

- **The voice tells a lot about ourselves**
  - *Age, gender, emotion, culture, personality, role, mental state, physical state, …*



Human Behaviour Understanding from Speech

- **Paralinguistic: from short-term states to long-term traits**



Collecting Affective Data

- **Elicitation methods**
  - **Acted**, certainty **+**, naturalness **–** (EMODB, **GEMEP**)
  - **Induced**, certainty **+–**, naturalness **+–** (eNTERFACE, **CPESD**)
  - **Spontaneous**: certainty **–**, naturalness **+** (SEMAINE, **RECOLA**)

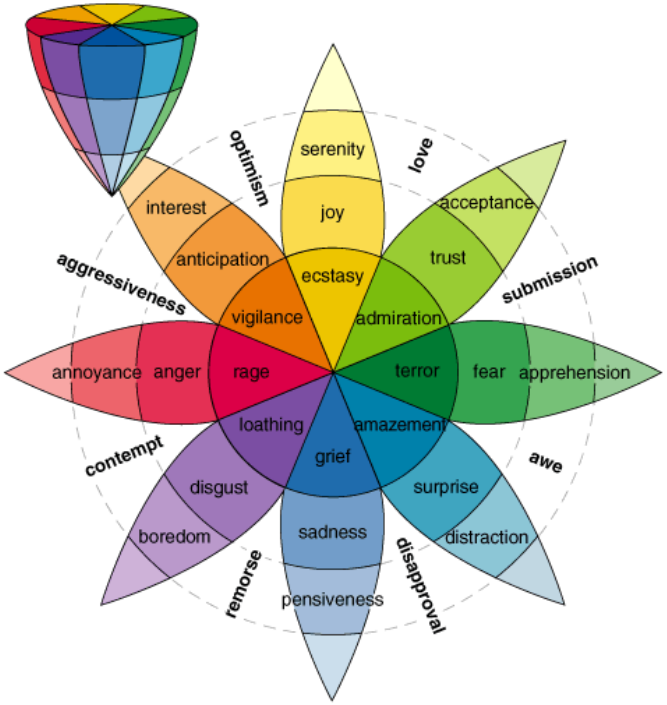
GEneva **M**ultimodal **E**motion **P**ortrayal

Child **P**athological & **E**motional **S**peech **D**atabase

REmote **COL**laborative and **A**ffective interaction

Quantifying Emotions

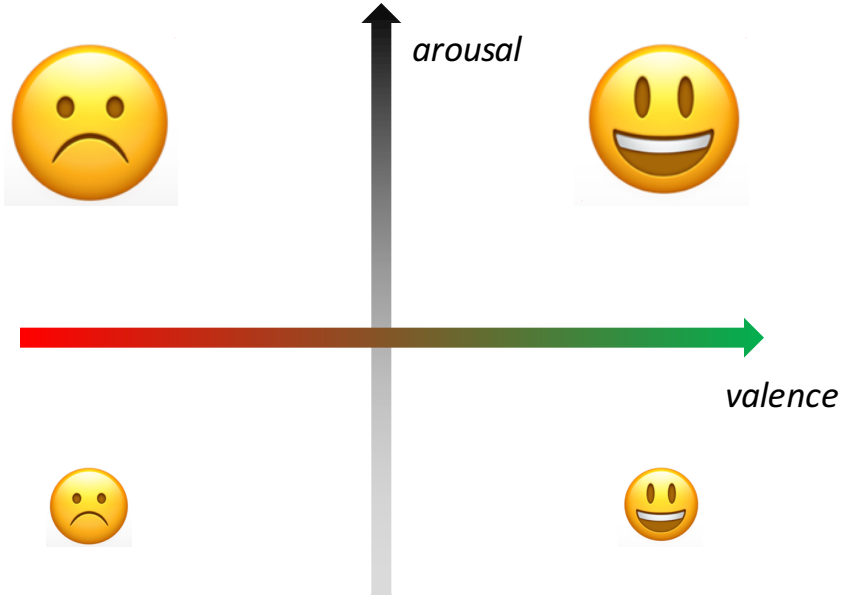
- **Discrete framework**
  - **Semantic descriptions**
  - **“Hard-coded” in the brain**
  - **Common categories:**
    - **Anger**
    - **Fear**
    - **Happiness**
    - **Sadness**



R. Plutchik, *Emotions: A general psychoevolutionary theory*, in K. R. Scherer and P. Ekman [Eds], *Approaches to Emotion*, Erlbaum, Hillsdale (NJ), 1984.

Quantifying Emotions

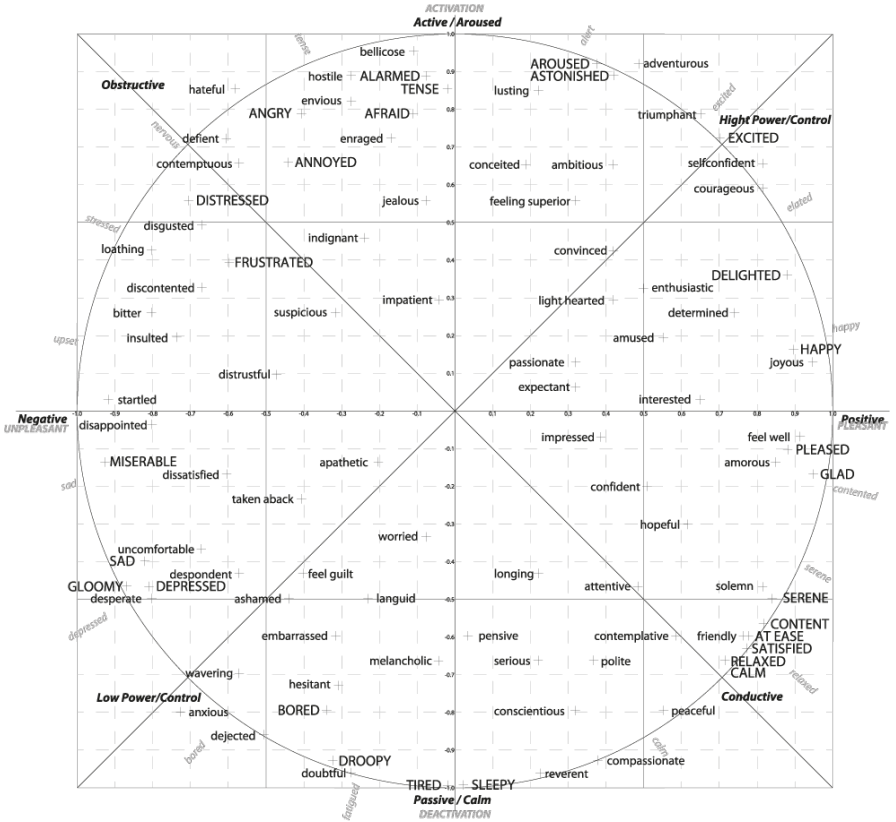
- **Continuous framework**
  - **Description of properties**
  - **Common dimensions**
    - **Arousal: calm vs. excited**
    - **Valence: negative vs. positive**



J. A. Russell. *A circumplex model of affect*. Journal of Personality and Social Psychology, 39(6):1161–1178, 1980.

Quantifying Emotions

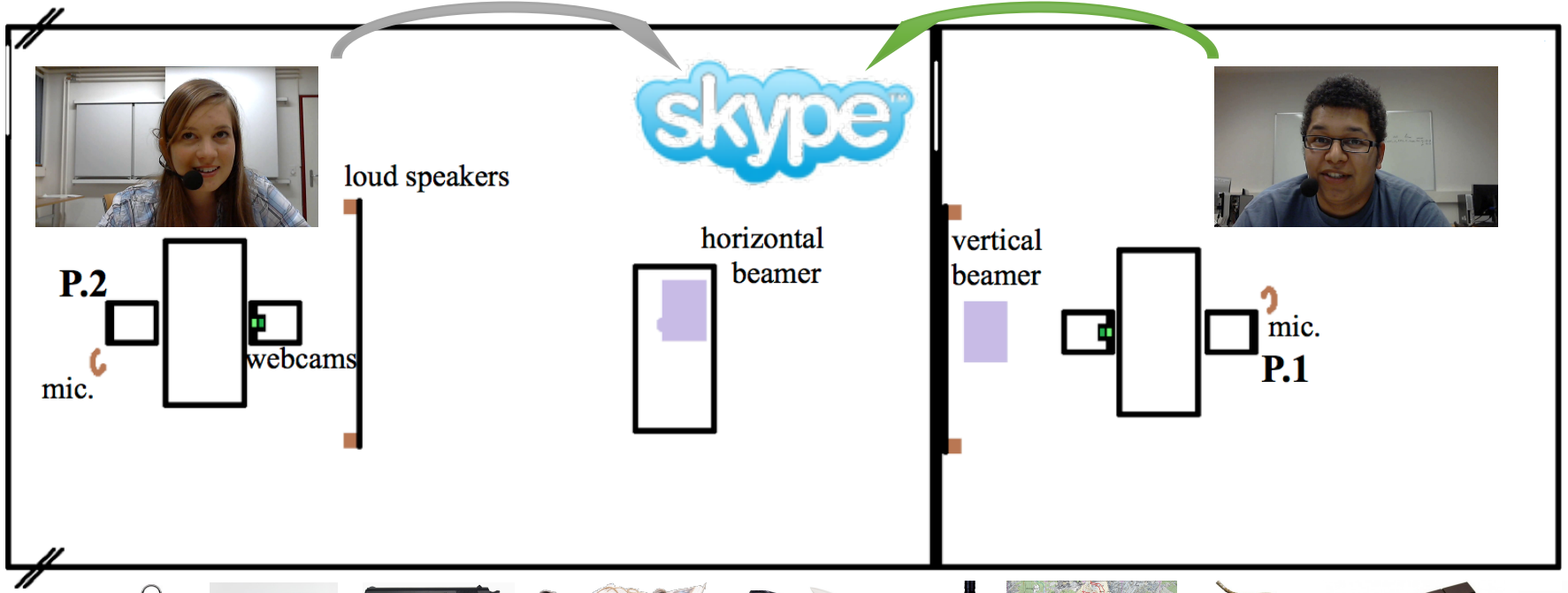
- **Continuous framework**
  - **Description of properties**
  - **Common dimensions**
    - **Arousal: calm vs. excited**
    - **Valence: negative vs. positive**
  - **Many categories can be mapped onto the VA space**



G. Paltoglou and M. Thelwall. *Seeing stars of valence and arousal in blog posts*. IEEE TAFC, 4(1):116–123, 2013.

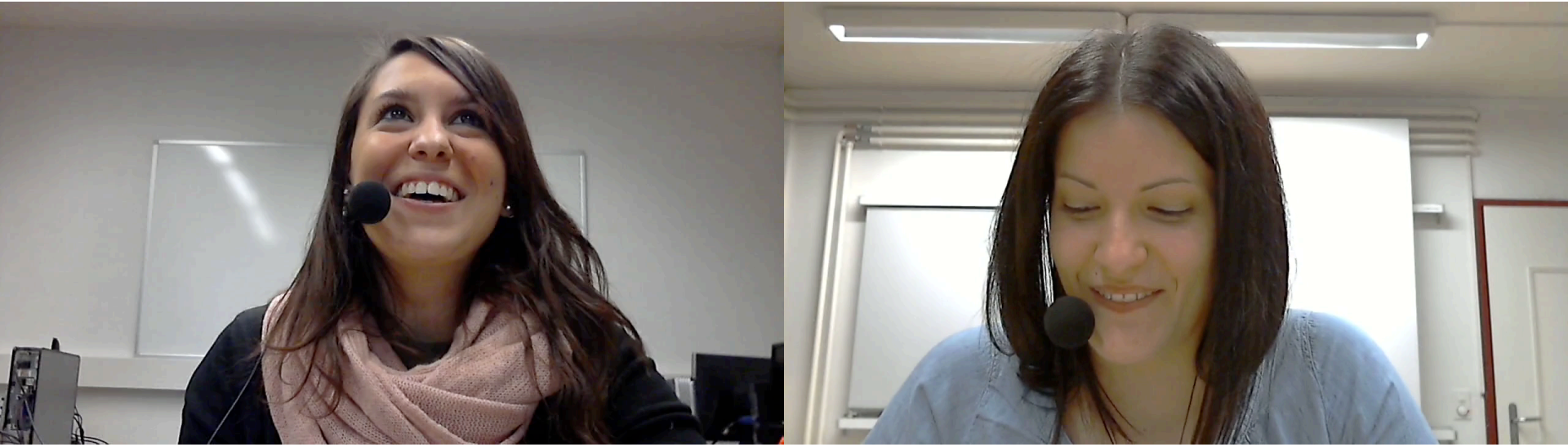
Data Collection: RECOLA database

- Collect **multimodal data** in **ecologically valid** situation (collaboration)



Data Collection: RECOLA database

- First dyad (P13 — P14)



F. Ringeval, *Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions*, in Proc. of EmoSPACE, IEEE International Conference on Face & Gestures 2013.

Data Annotation: ANNEMO

- **Web-based** annotation interface (open source); 6 raters (3F/3M)

arousal - SEQUENCE_2.mp4
valence - SEQUENCE_2.mp4
social dimension annotation

arousal - SEQUENCE_3.mp4
valence - SEQUENCE_3.mp4
social dimension annotation

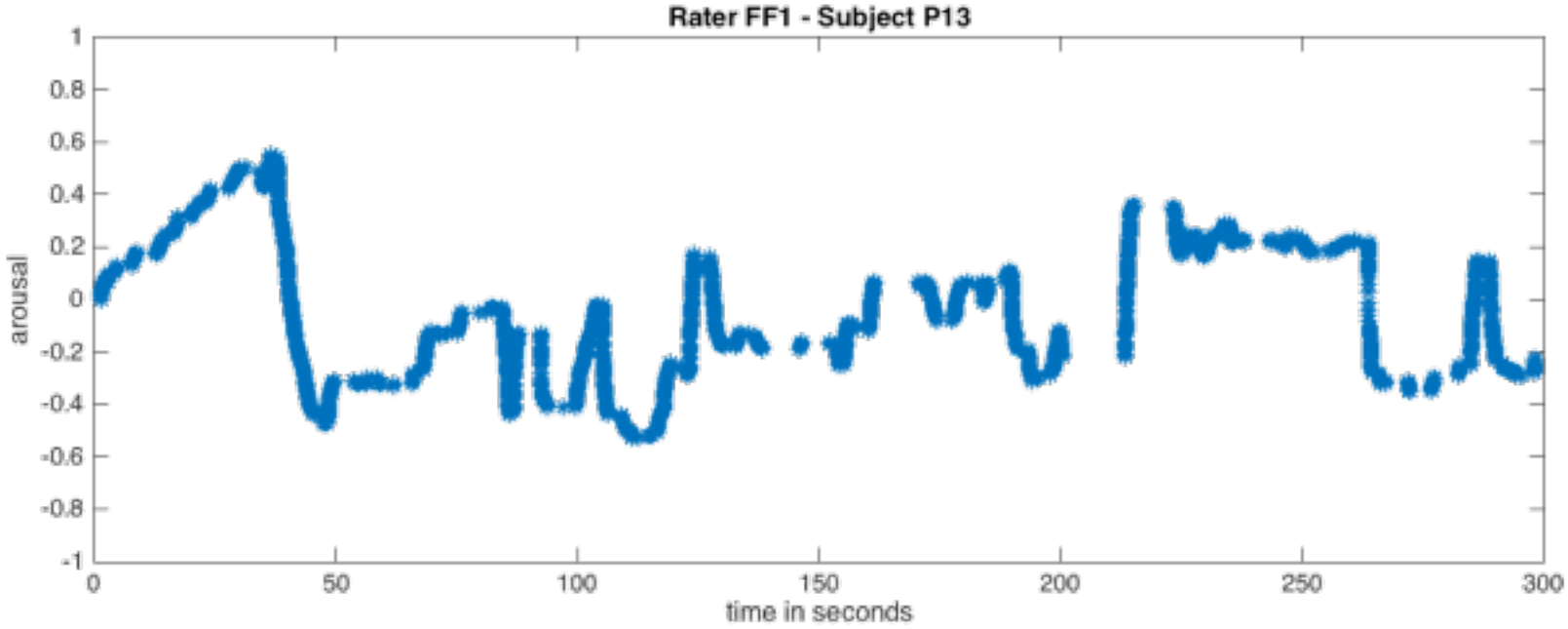
arousal - SEQUENCE_4.mp4
valence - SEQUENCE_4.mp4
social dimension annotation

arousal - SEQUENCE_5.mp4
valence - SEQUENCE_5.mp4
social dimension annotation

arousal - SEQUENCE_6.mp4
valence - SEQUENCE_6.mp4
social dimension annotation

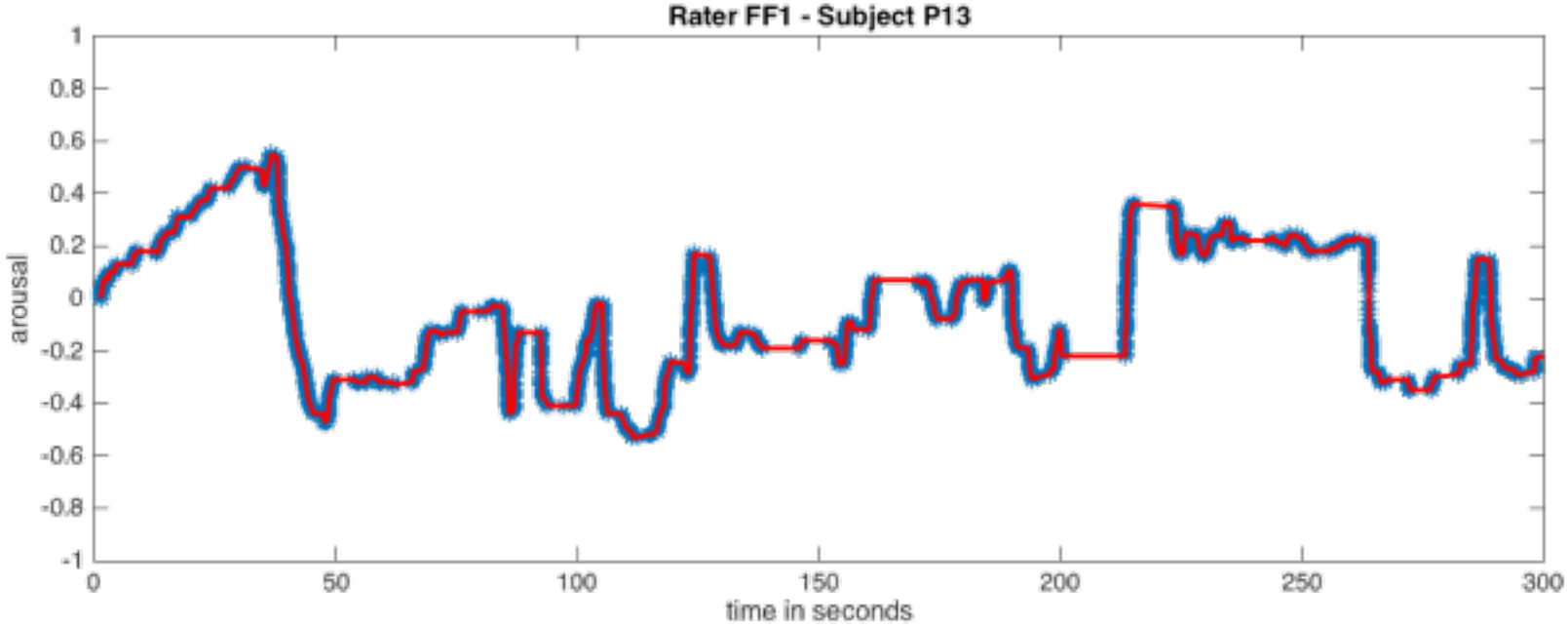
Data Annotation: ANNEMO

- **Raw data**



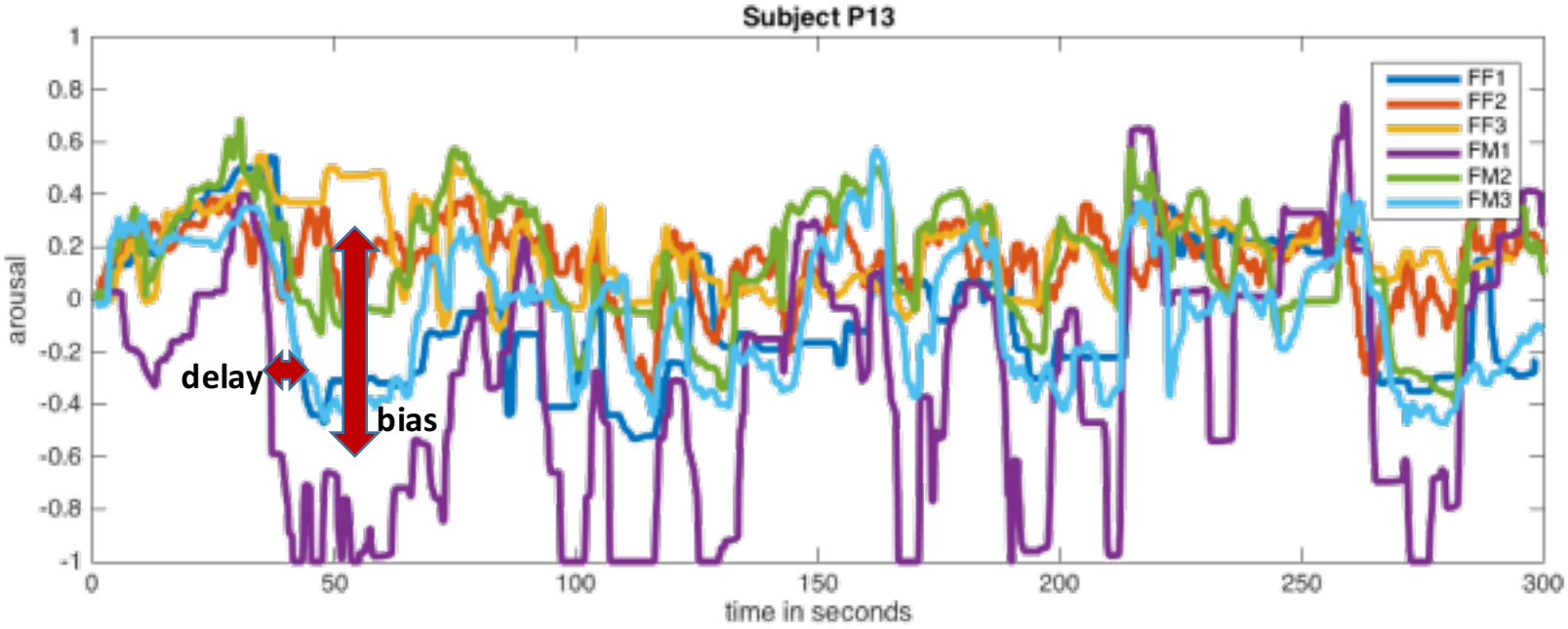
Data Annotation: ANNEMO

- **Interpolated data** (piecewise Hermite interpolation; $d^2/\ dt$ not continuous)



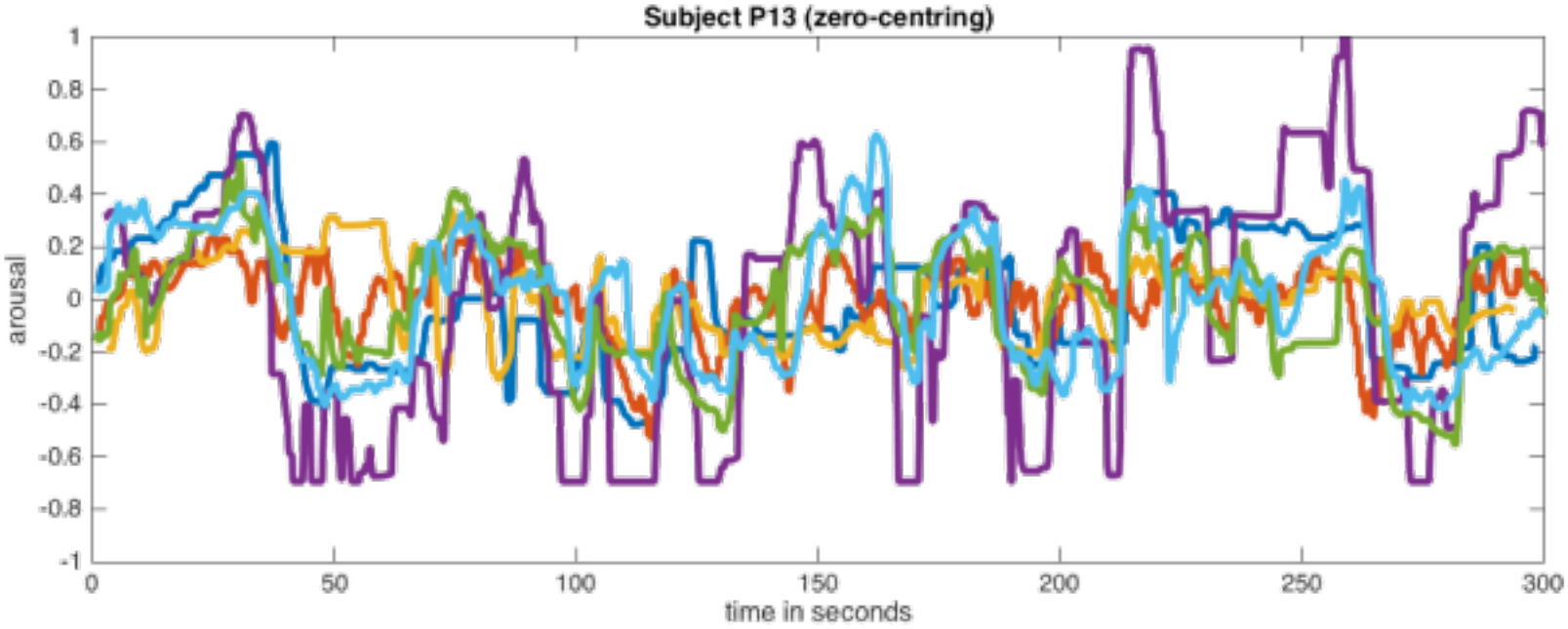
Data Annotation: ANNEMO

- **Interpolated data for all raters**



Data Annotation: post-processing

- **Can we improve by post-processing? (E.g., zero-centring)**



Data Annotation: post-processing

- **Can we improve by post-processing? (E.g., zero-centring)**
  - Looks great and **improves Inter-Rater Agreement** (averaged ρ_c)
  - **BUT,** mean perceived emotion is **forced to neutral**

	Arousal	Valence
raw; ρ_c	0.28	0.37
raw; % pos.	59.0	70.5
zero-m.; ρ_c	0.33	0.43
zero-m.; % pos.	50.8	44.8

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

Data Annotation: post-processing

- **Can we improve by post-processing?**
  - **Proposed method**: centre to mean weighted by the **inter-rater agreement**
    1. Compute inter-rater agreement $\bar{\rho}_d(i)$ for a dimension d and an evaluator $i \in [1, N_e]$
    2. Average ratings $y_d^{e_i}$ of all evaluators e_i and weight by $\bar{\rho}_d(i)$ to obtain the centring value $\bar{y}_d$
    3. Centre ratings and average to obtain a **gold-standard** trace $g_d(t)$

$$\bar{\rho}_d(i) = \frac{1}{N_e - 1} \sum_{i=1, j \neq i}^{N_e} \rho_d(i,j) \qquad (1)$$

$$\bar{y}_d = \frac{1}{\sum_{i=1}^{N_e} \bar{\rho}_d(i)} \sum_{i=1}^{N_e} \frac{1}{T} \sum_t y_d^{e_i}(t) \bar{\rho}_d(i) \qquad (2)$$

$$g_d(t) = \frac{1}{N_e} \sum_{i=1}^{N_e} y_d^{e_i}(t) + \bar{y}_d - \frac{1}{T} \sum_t y_d^{e_i}(t) \qquad (3)$$

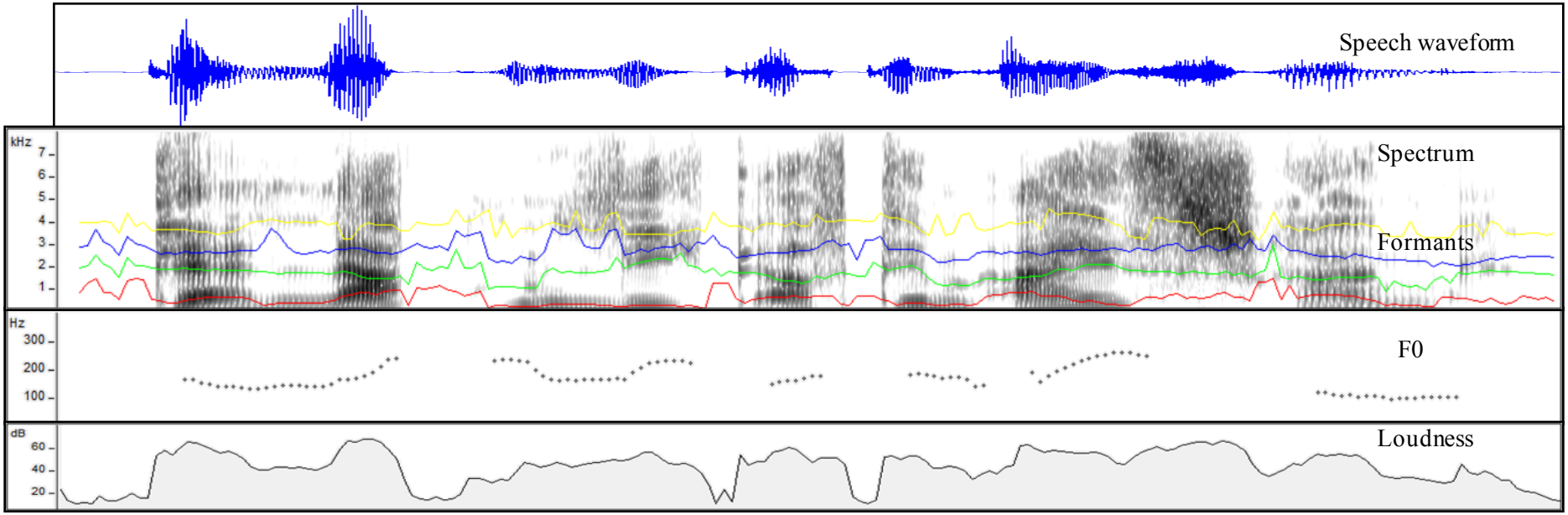
Data Annotation: post-processing

- **Post-processing by centring with IRA**
  - Same improvement as zero-centring for IRA
  - Original (positive) skew better preserved

	Arousal	Valence
raw ; ρ_c	0.28	0.37
raw ; % pos.	59.0	70.5
zero-m. ; ρ_c	0.33	0.43
zero-m. ; % pos.	50.8	44.8
wgt-m. ; ρ_c	0.33	0.43
wgt-m. ; % pos.	48.5	74.1

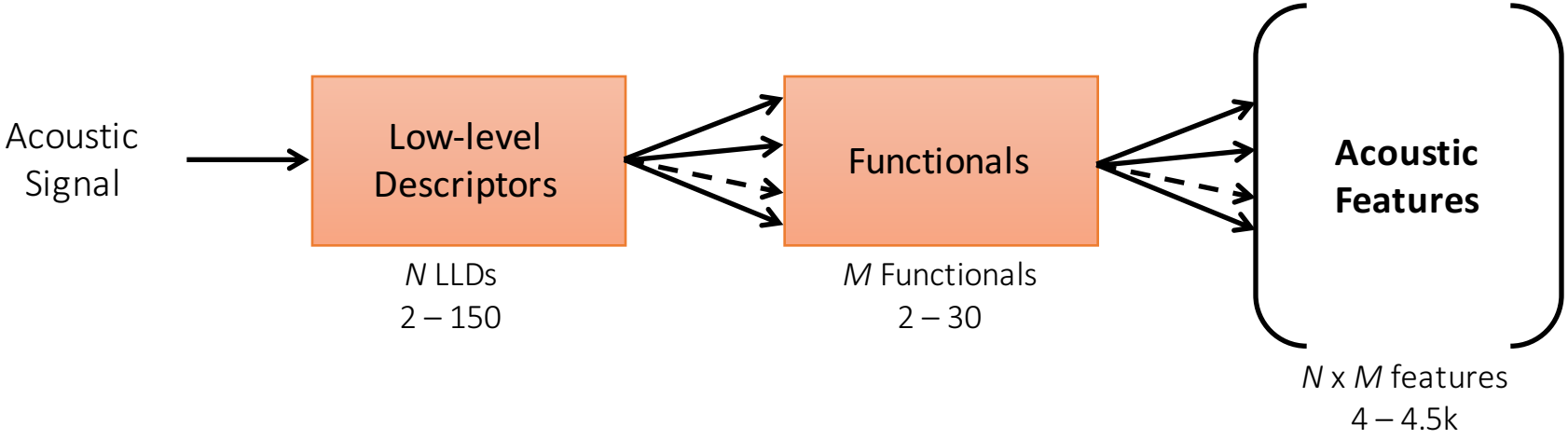
Emotion Recognition from Speech

- **Speech conveys many relevant cues**



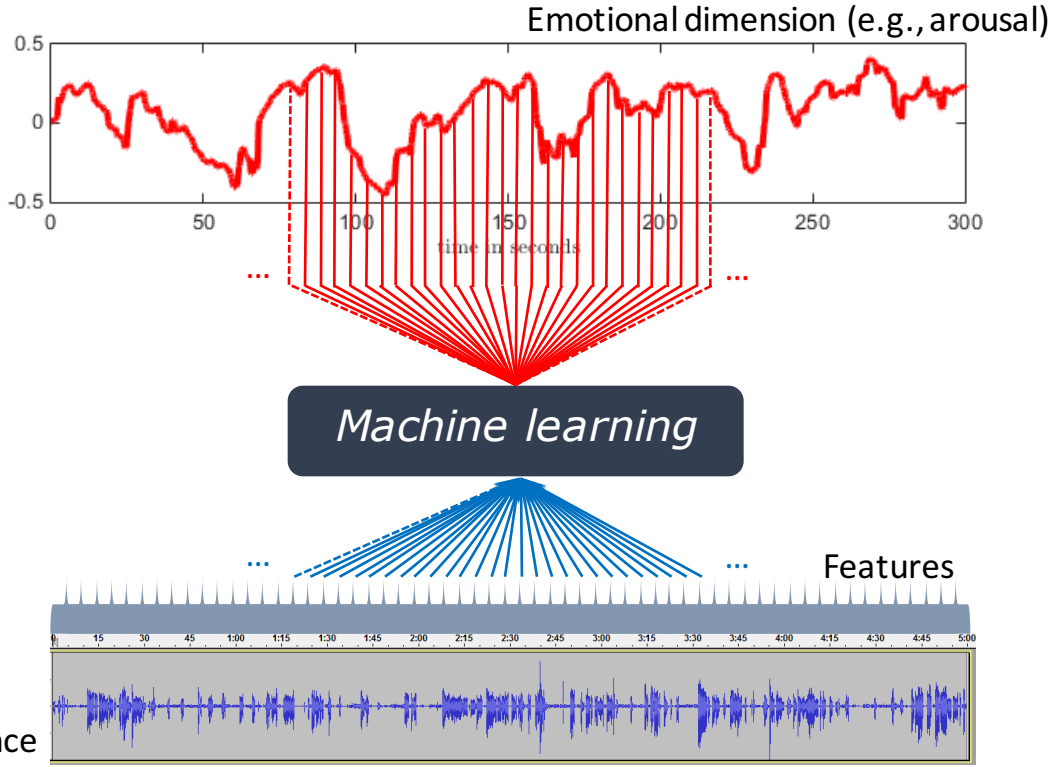
Emotion Recognition from Speech

- **Features extraction from speech —** openSMILE:) by audEERING™
  - Low-level descriptors and functionals
  - Sliding window shifted forward at a constant rate (25 Hz)



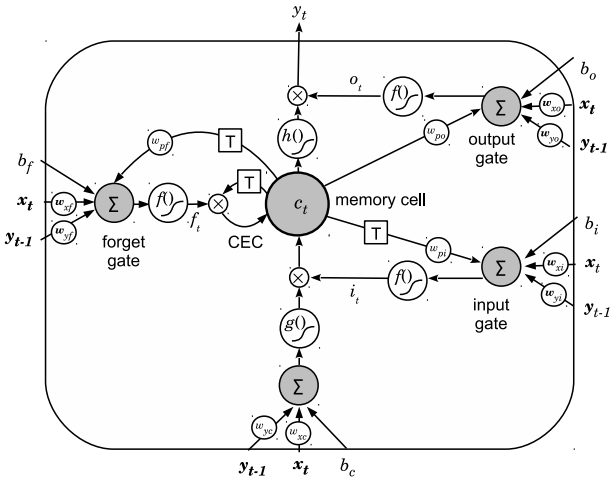
Emotion Recognition from Speech

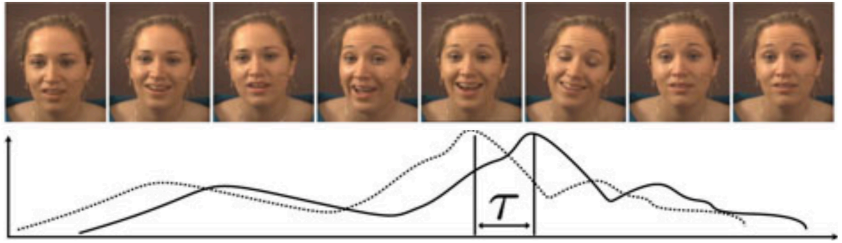
- **Machine learning: sequential learning problem**



Experiment: Asynchronous modelling

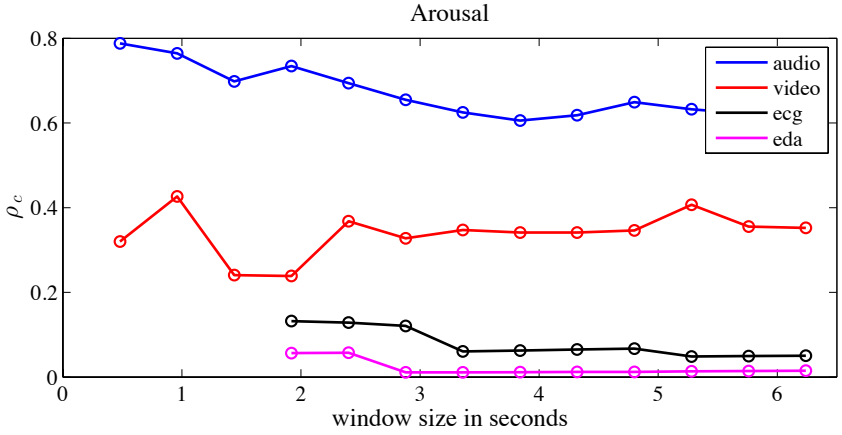
- **Modelling asynchronous ratings of emotion with (B)LSTM-RNN**
  - LSTM-RNN can capture long-range contextual dependencies
  - Exploit this to model delay between raters (multi-task learning)

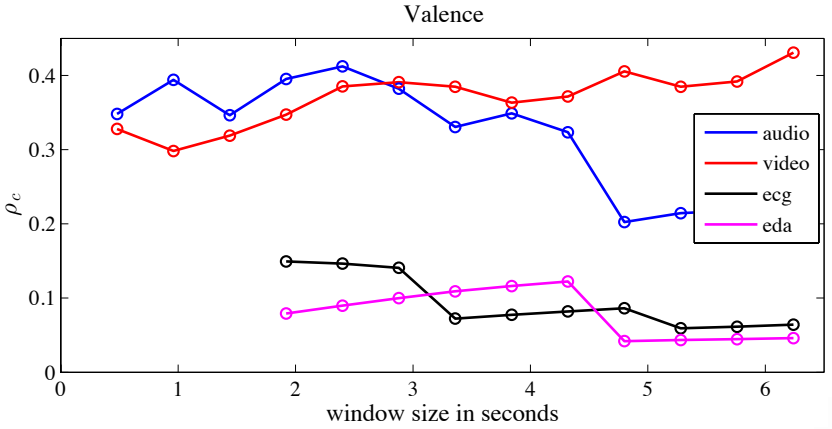




Experiment: Asynchronous modelling

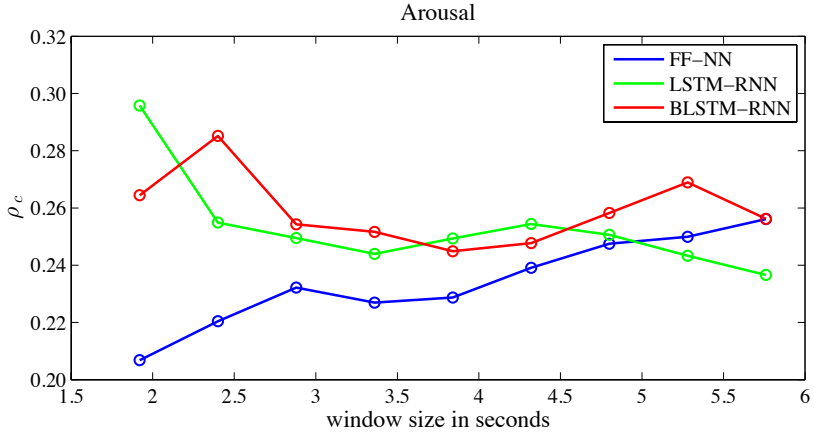
- **Setup**
  - **RECOLA**: 27 subjects (9/9/9) x 5 min. **Features**: ComParE 130 LLDs x 5 func.
  - **Window size**: from 0.5 s (micro-expressions) to 6 s (information loss)
  - (FF, (B)LSTM) **DNN**: $\sigma_i = 0.1$; 2 h.l. (128); 100 epochs; early stop. MSE
  - **Effect of window size** (best network topology, single/multi-task)

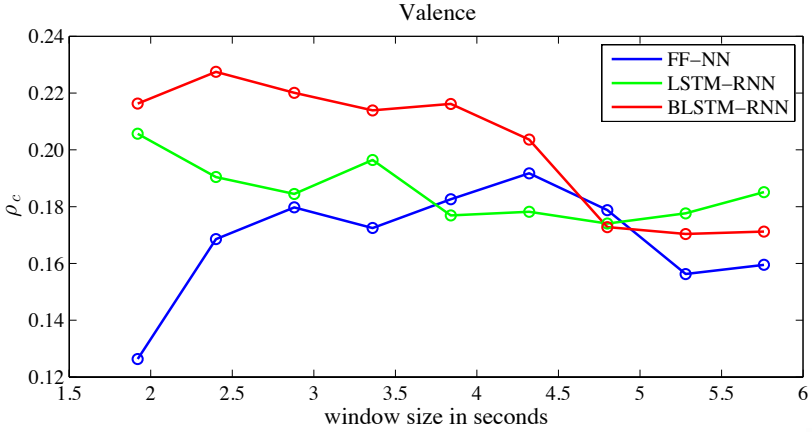




Experiment: Asynchronous modelling

- **Setup**
  - **RECOLA**: 27 subjects (9/9/9) x 5 min. **Features**: ComParE 130 LLDs x 5 func.
  - **Window size**: from 0.5 s (micro-expressions) to 6 s (information loss)
  - (FF, (B)LSTM) **DNN**: $\sigma_i = 0.1$; 2 h.l. (128); 100 epochs; early stop. MSE
  - **Impact of contextual modelling** (averaged over all modalities)





Experiment: Asynchronous modelling

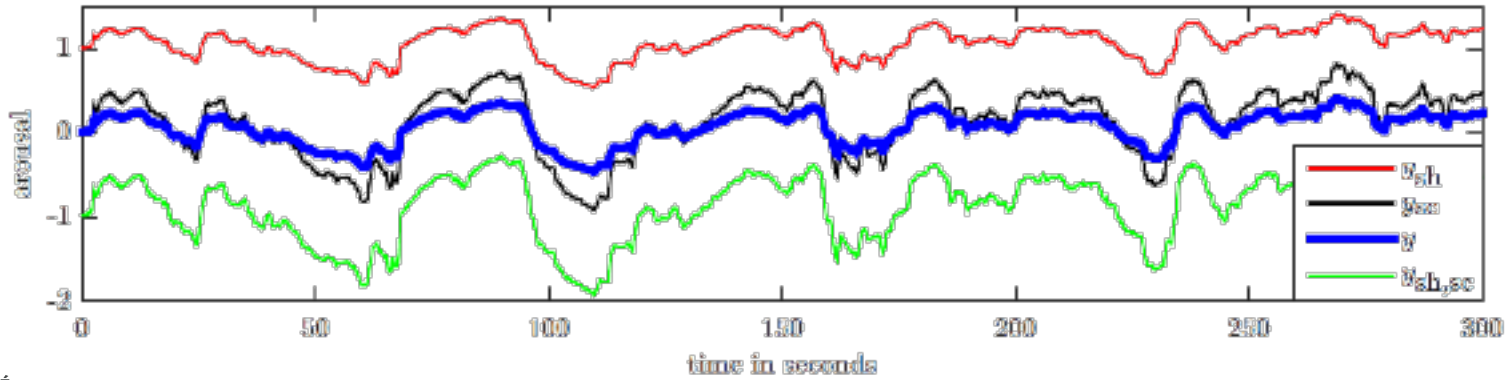
- **Setup**
  - **RECOLA**: 27 subjects (9/9/9) x 5 min. **Features**: ComParE 130 LLDs x 5 func.
  - **Window size**: from 0.5 s (micro-expressions) to 6 s (information loss)
  - (FF, (B)LSTM) **DNN**: $\sigma_i = 0.1$; 2 h.l. (128); 100 epochs; early stop. MSE
  - **Effect of multi-tasking** (mean vs. individual ratings)

	Arousal	Valence
AUDIO - SINGLE	.732	.412
AUDIO - MULTI	.738	.343
VIDEO - SINGLE	.403	.339
VIDEO - MULTI	.427	.349

Experiments: CCC as objective function

- Neural networks typically trained using **SSE-related** criteria
- **Goal**: Directly optimize predictor for the evaluation metric instead
- ρ as evaluation metric **not directly usable** for optimization
  - Scale (y_{sc}) and shift (y_{sh}) invariant (!!!)
  - Leads to **ill-posed** optimisation problem

$$\mathcal{O} = -\sum_{i \in \beta, f \in \mathcal{F}} \rho_c$$

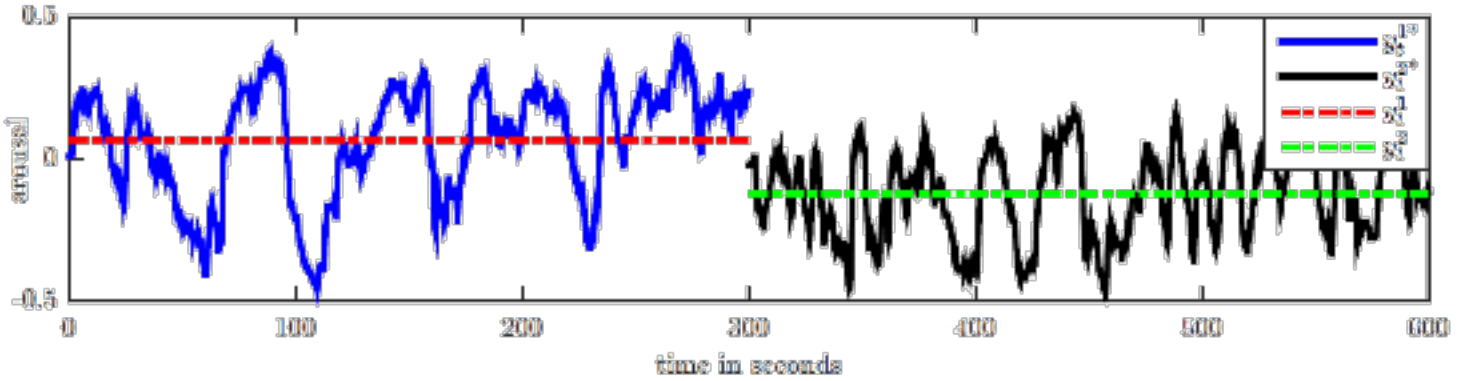


ρ_c=1.07
ρ_c=1.78
ρ_c=1.15

Felix Weninger, Fabien Ringeval, Erik Marchi, and Björn Schuller. *Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio*. In *Proc. of IJCAI 2016*, pp. 2196–2202, New York City (NY), USA, July 2016

Experiments: CCC as objective function

- $\boldsymbol{\rho_c}$: CCC over the whole evaluation set
  - Expect good CCC when predicting the correct mean for each sequence
- $\boldsymbol{\Sigma\rho_c}$: Sum the CCC in each sequence
  - Need to get the variation within each sequence right, but sensitive to noise in the gold-standard
- ρ_c = 0.37, $\Sigma\rho_c$ = 0



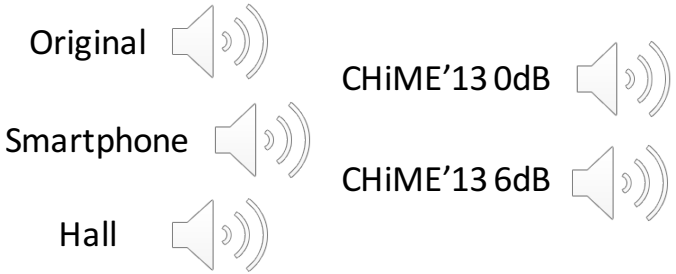
Experiments: CCC as objective function

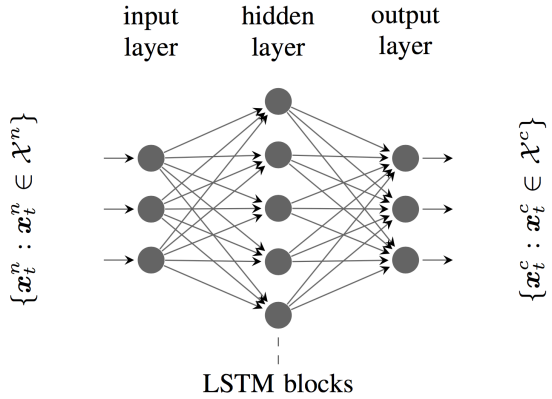
- **Setup**
  - **RECOLA**: 46 subjects (16/15/15) x 5 min. **Features**: eGeMAPS (AV+EC 2015)
  - (FF, (B)LSTM) **DNN**: $\sigma_i = 0.1$; 2 h.l. (128); 100 epochs; early stop. MSE
  - Significant improvements with $\boldsymbol{\Sigma\rho_c}$ objective; **trade-off** between RMSE and $\boldsymbol{\rho_c}$

	Arousal			Valence		
$\mathcal{O}$	RMSE	ρ_c	E$\{\rho_c\}$	RMSE	ρ_c	E$\{\rho_c\}$
SSE	**.128**	.097	.161	**.108**	.131	.052
$\boldsymbol{\rho_c}$	.193	.254	.212	.130	.155	.080
$\boldsymbol{\Sigma\rho_c}$	.200	**.350**	**.268**	.192	**.199**	**.139**

Experiments: Robustness to noise

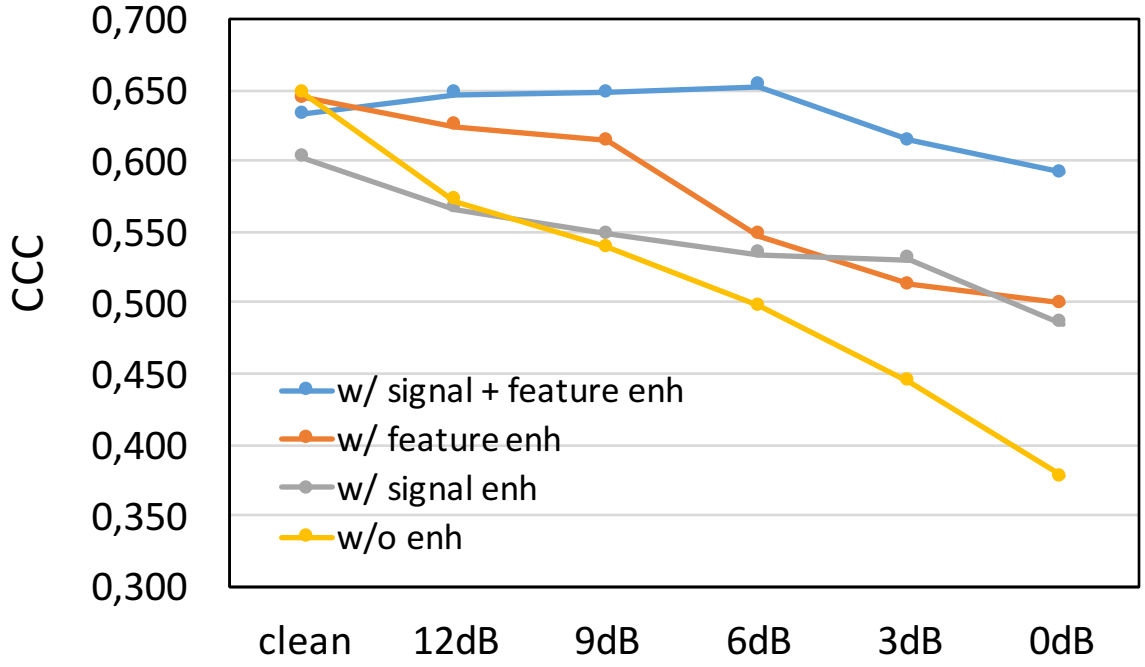
- **How non-stationary noise and reverberation degrade performance?**
- **Noises**: CHiME'15, Smartphone (Nexus One), Hall, Train
- **Various SNR** (0dB, 3dB, 6dB, 9dB, 12dB)
- **Method:** signal/feature enhancement with auto-encoders





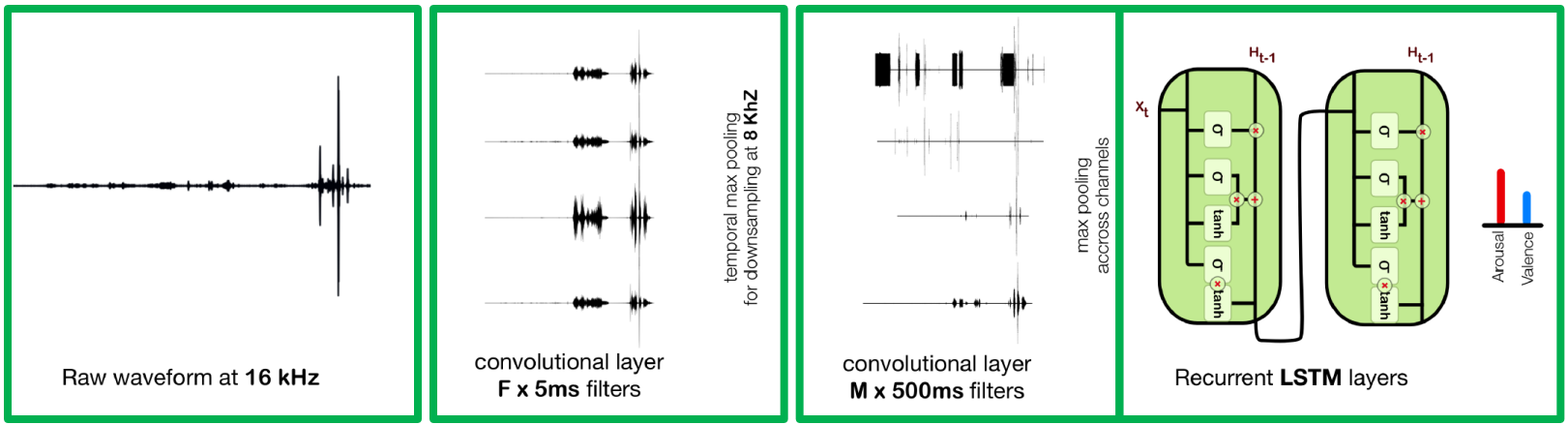
Experiments: Robustness to noise

- **Results on Arousal / CHiME'15 (MFCCs, SVR for prediction – AVEC'16)**



Experiments: Adieu features?

- **CNN-DNN** can learn **appropriate representation from raw input**
- How well can it perform for **emotion recognition**?



- Temporal convolution (F=40 filters of size 200 @ 5ms)
- Max-pooling across time (pool size 2) to downsample @ 8kHz

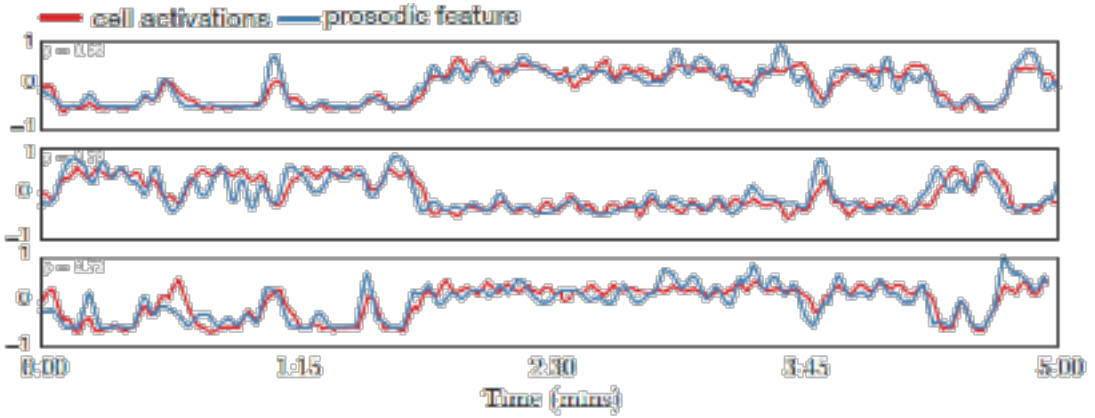
Experiments: Adieu features?

- **CNN-DNN** can learn **appropriate representation from raw input**
- How well can it perform for **emotion recognition**?

RMS energy range (.81)

loudness (.73)

F0 mean (.72)

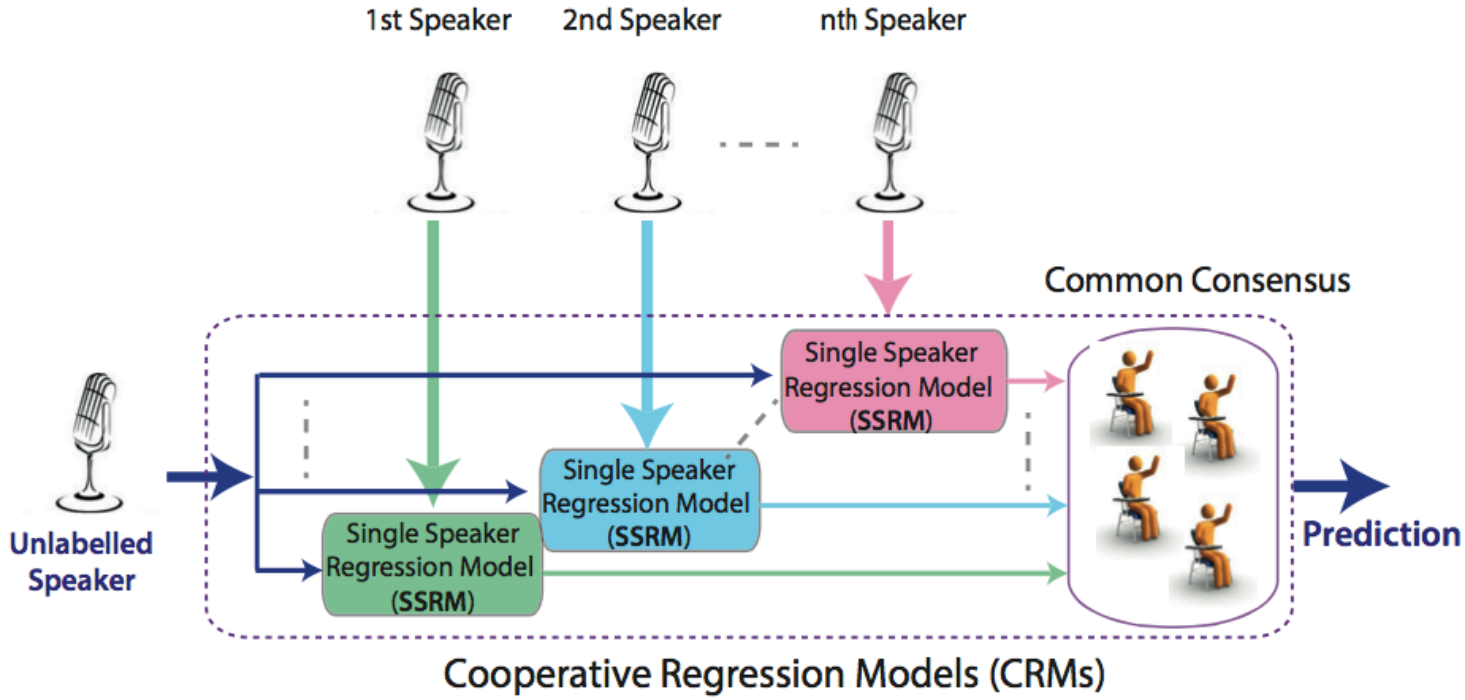


Experiments: Adieu features?

- **Results**

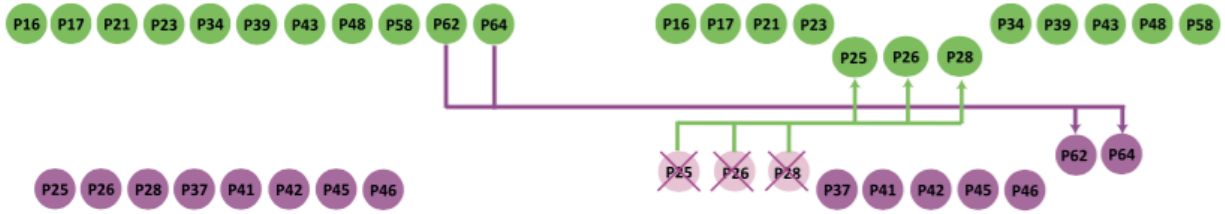
Predictor	Features	Arousal (ρ_c)	Valence (ρ_c)
a. Mean squared error objective			
SVR	eGeMAPS	.318 (.489)	.169 (.210)
SVR	ComParE	.366 (.491)	.180 (.178)
BLSTM	eGeMAPS	.300 (.404)	.192 (.187)
BLSTM	ComParE	.132 (.221)	.117 (.152)
Proposed	raw signal	.684 (.728)	.249 (.312)
b. Concordance correlation coefficient objective			
BLSTM	eGeMAPS	.316 (.445)	.195 (.190)
BLSTM	ComParE	.382 (.478)	.187 (.246)
Proposed	raw signal	**.686 (.741)**	**.261 (.325)**

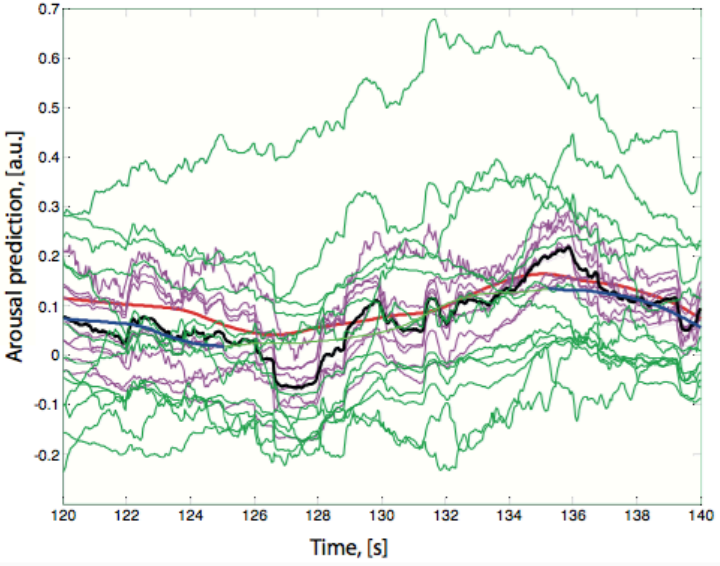
Experiments: Adaptive cooperation

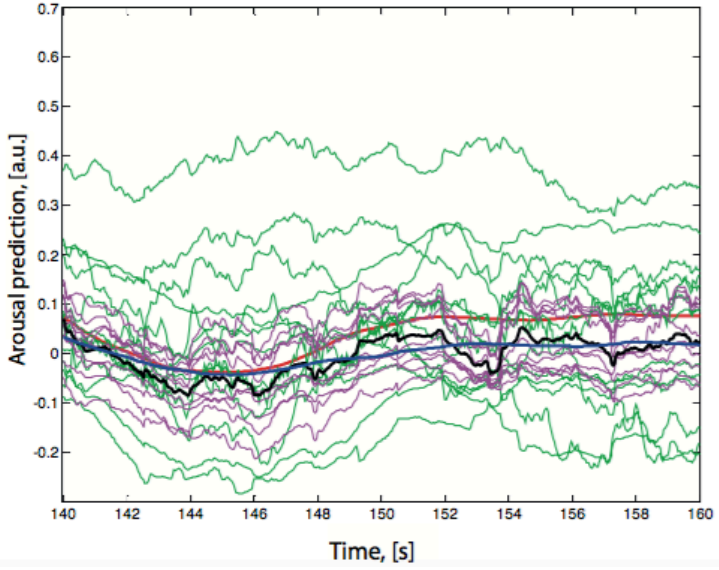


Arianna Mencattini, Eugenio Martinelli, Fabien Ringeval, Björn Schuller, and Corrado Di Natale. Continuous estimation of emotions in speech by dynamic cooperative speaker models. *IEEE Transactions on Affective Computing*, 2016.

Experiments: Adaptive cooperation

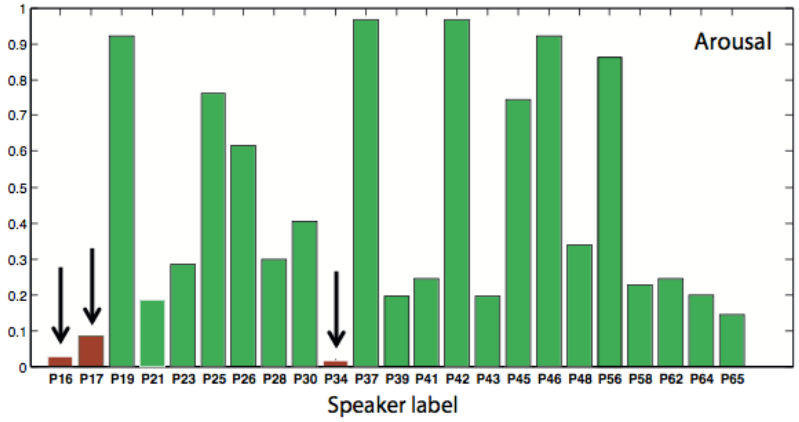


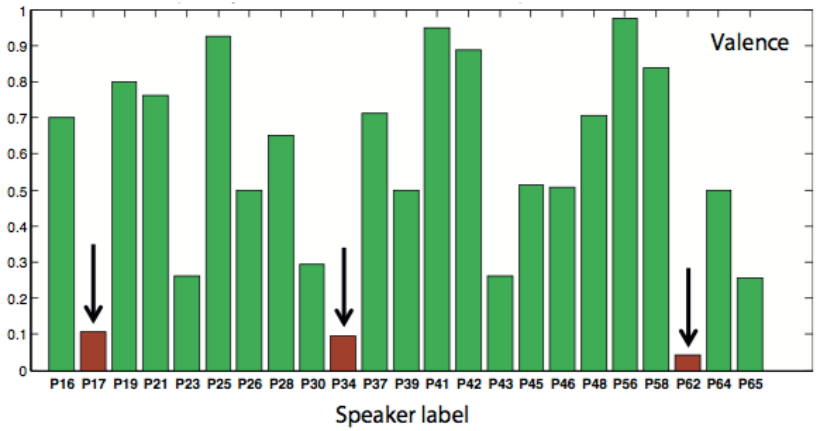




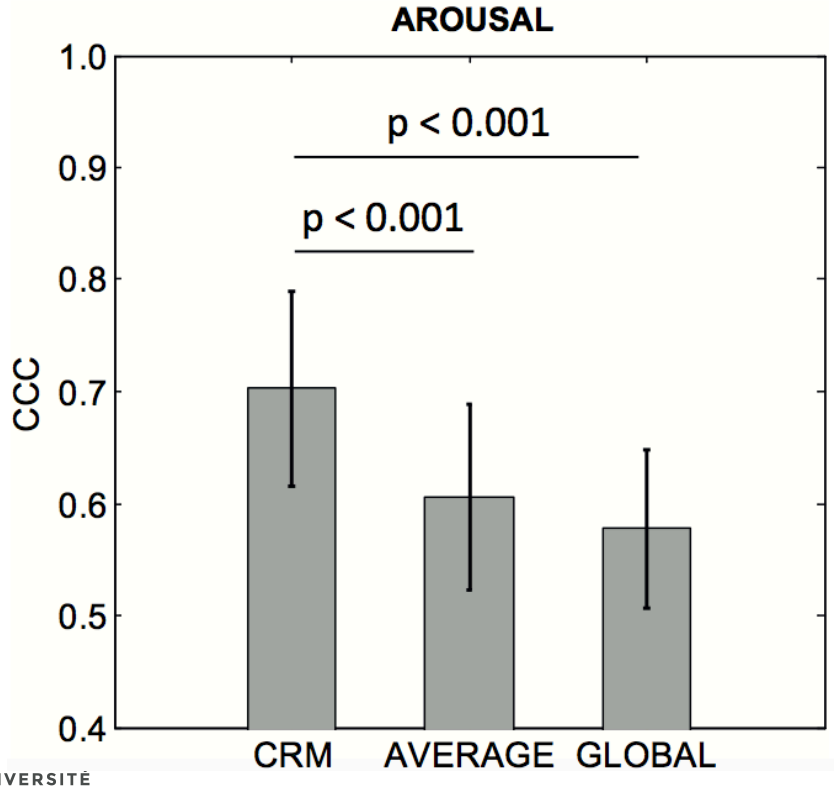
Experiments: Adaptive cooperation

- Frequency of Single Speaker Regression Model inclusion

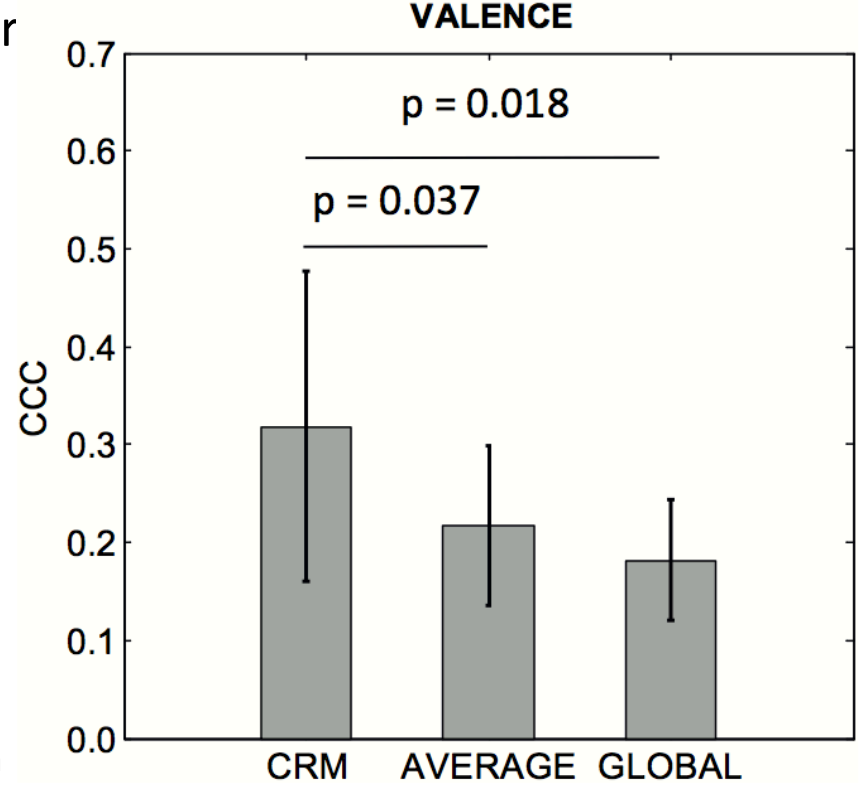




Experiments: Adaptive cooperation

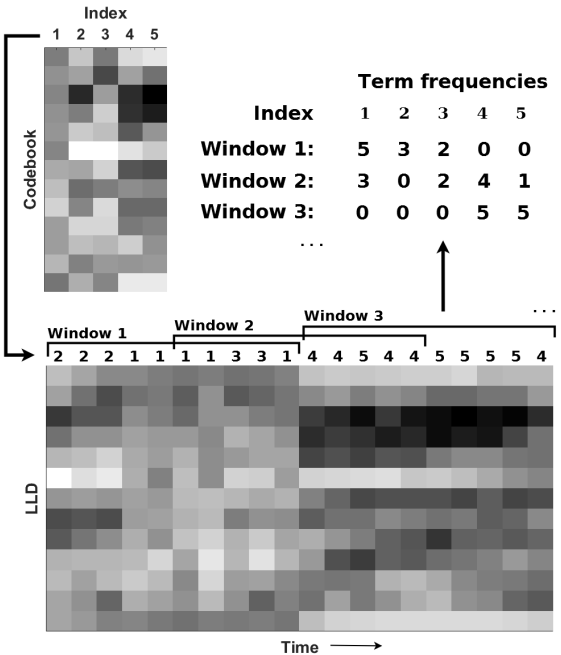


r



Experiments: Bag of audio words

- Learn dictionary of acoustic words instead of using functionals (MFCC)



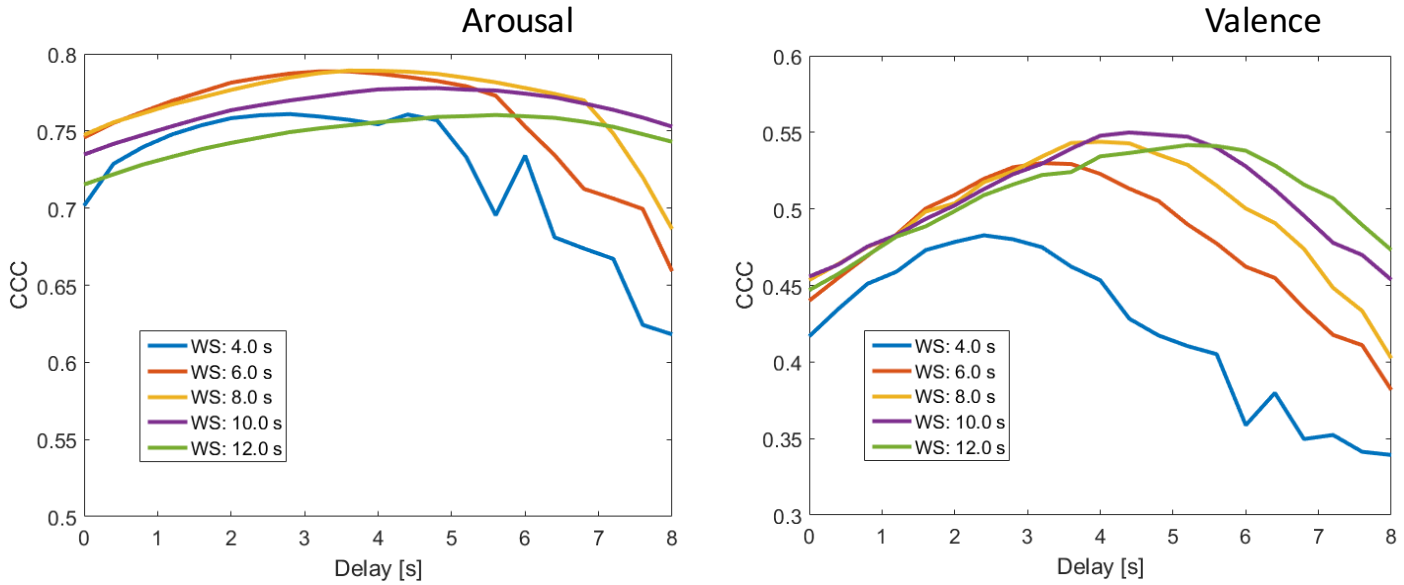
openWORD

$$TF'(w) = \lg(TF(w) + 1)$$

Maximilian Schmitt, Fabien Ringeval, and Björn Schuller. At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech. In *Proc. INTERSPEECH 2016*, San Fransisco (CA), USA, September 2016.

Experiments: Bag of audio words

- Results: learning time-delay



Experiments: Bag of audio words

- Results: comparison with state-of-the-art

Model	Ref.	CCC			
		Arousal		Valence	
		Valid	Test	Valid	Test
BLSTM-RNN	[26]	**.800**		.398	
CNN (end-to-end)	[4]	.741	.686	.325	.261
Proposed (BoAW)	Table 2	.793	**.753**	**.550**	.430
Proposed (early fusion)	Table 4	.799	.738	.521	**.465**

Conclusion

- **Affective computing is a “hot” topic**
- **Vast application fields, robotics, behaviomedics, media, …**
- **Shifted from small and acted to large and spontaneous**
- **Context modelling with LSTM helps, CNN+LSTM performs well**
- **Simple but well defined systems can perform better than end-to-end**

=> not (yet) the end of signal processing!

Perspectives

- **Toward BIG DATA**
  - **ER community => 10 hours of data**
  - **ASR community => 10.000 hours of data**

  - **Crawl the web – vlogs, vimeo, youtube, …**
  - **Use novelty detection to identify non-neutral**
  - **Use active learning to automatically label easiest**

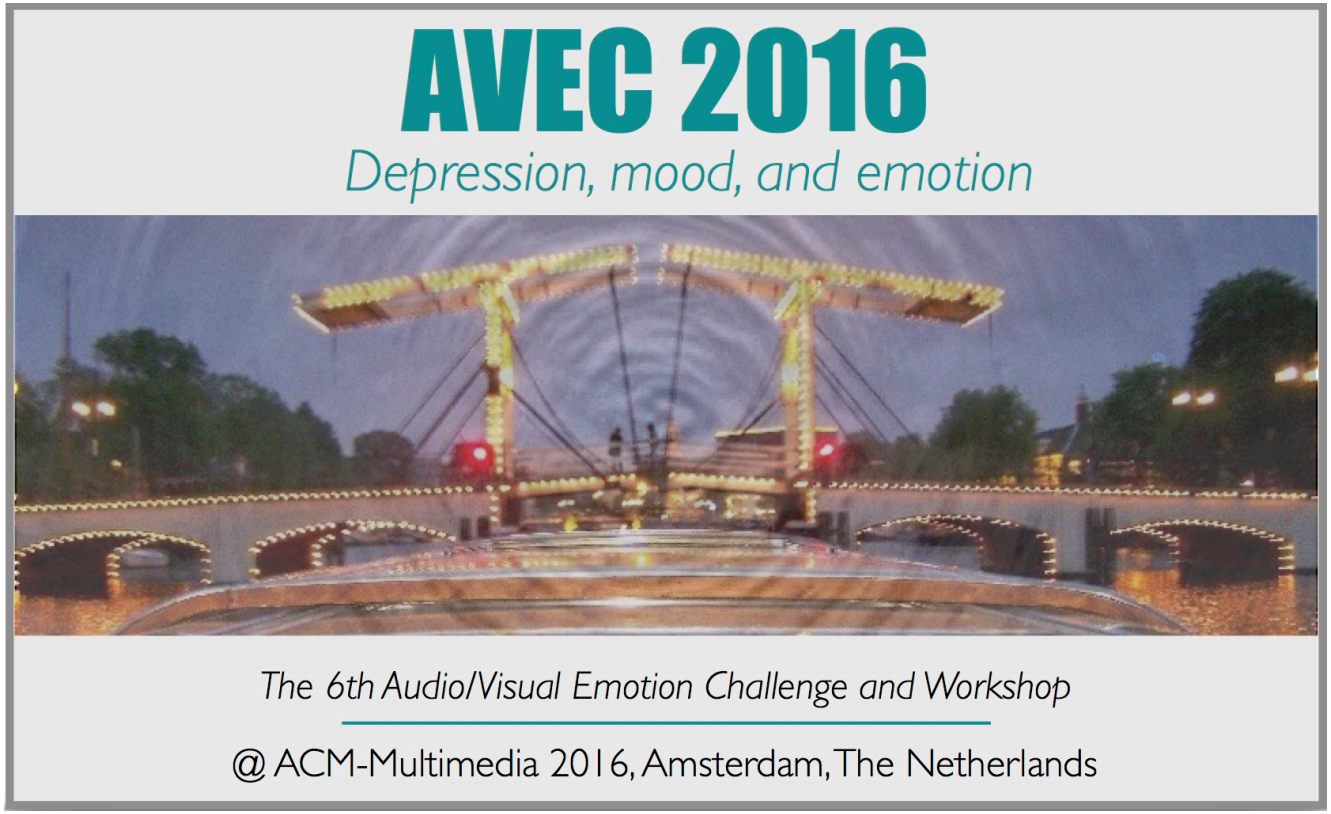
Perspectives

- **Toward real-life data**
  - **Not “properly” recorded => issue**
  - **But that’s real life!!!**
  - **Wearable sensors, smartphones, webcam**
  - **Face diversity of individuals (transfer learning)**

Perspectives

- **Toward multimodality**
  - **Datasets are not all with ECG, EDA, FAU, …**
  - **Use pseudo-multimodality to cross fertilise the missing modalities**
  - **Model context with available sensors (e.g., GPS, movements, logs, calls)**

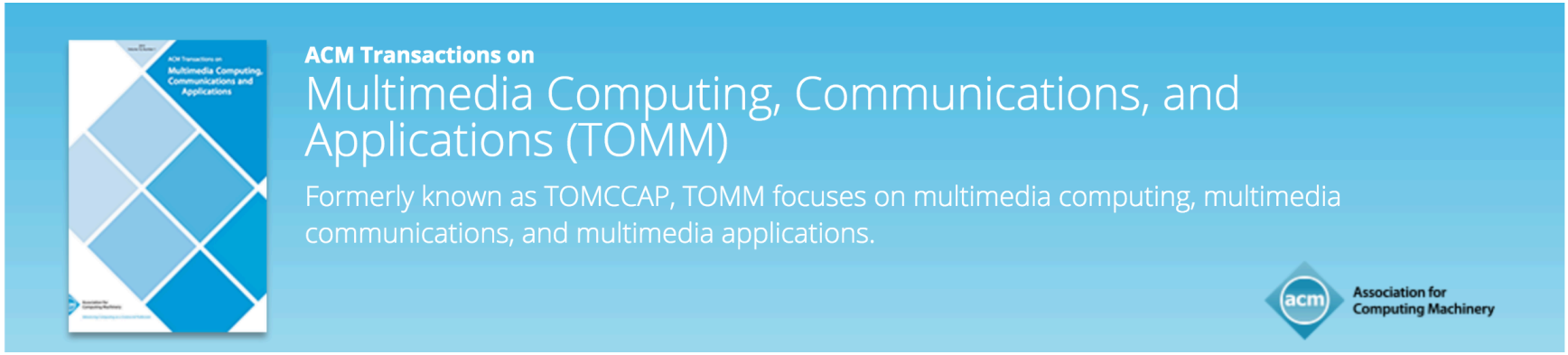
Commercial break – AVEC'16 16th October



Commercial break – ACM TOMM 31st October

Special Section on Multimedia Computing and Applications of Socio-Affective Behaviors in the Wild

Fabien Ringeval, Björn Schüller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Maja Pantic



Thank you!

Imperial College London

UPMC
SORBONNE UNIVERSITÉS

UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Università di Roma
Tor Vergata

UNIVERSITÄT PASSAU

Technische Universität München

UNIVERSITÉ Grenoble Alpes

LIG