# Speech Processing

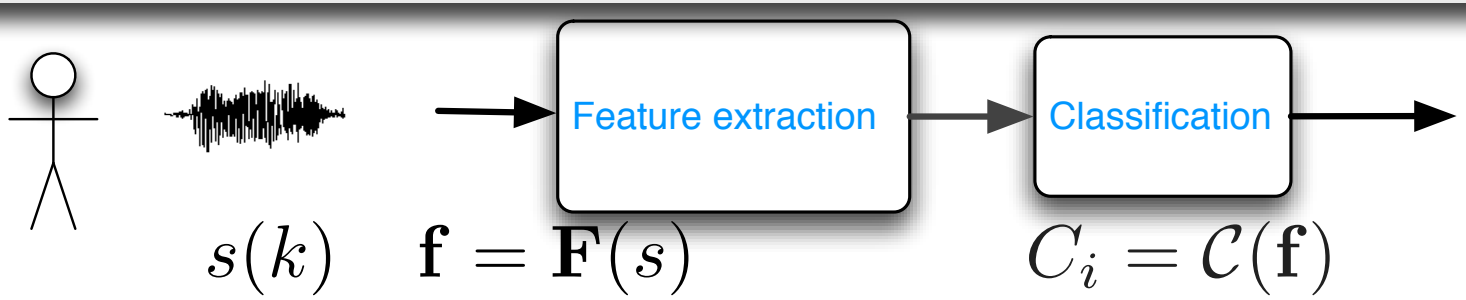
Mohamed CHETOUANI
Professeur des Universités
Institut des Systèmes Intelligents et de Robotique (ISIR)
Sorbonne Université

mohamed.chetouani@sorbonne-universite.fr




# Speech processing in Social Interactions

Problem statement: From speech samples to social information: Identity, Emotion, Intention, Pathology...




LPC, MFCC, TRAPs, NPC...

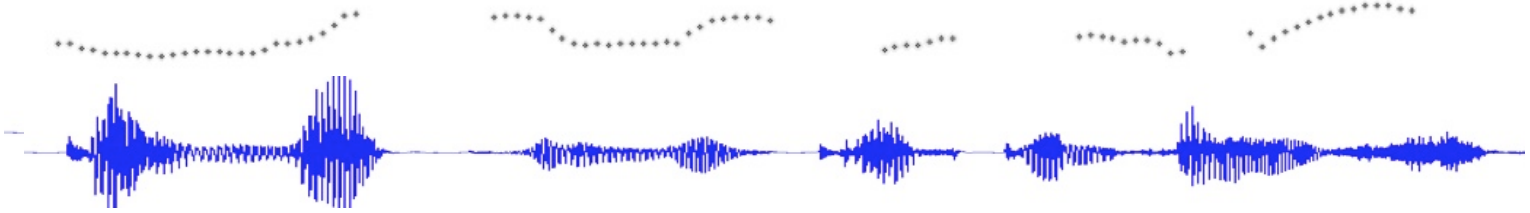
k-ppv, GMM, HMM, SVM...

Idea: Characterization of speech signals by feature extraction, time-scale and classification methods

Units?

Voiced, unvoiced, vowels...

Subjectivity?

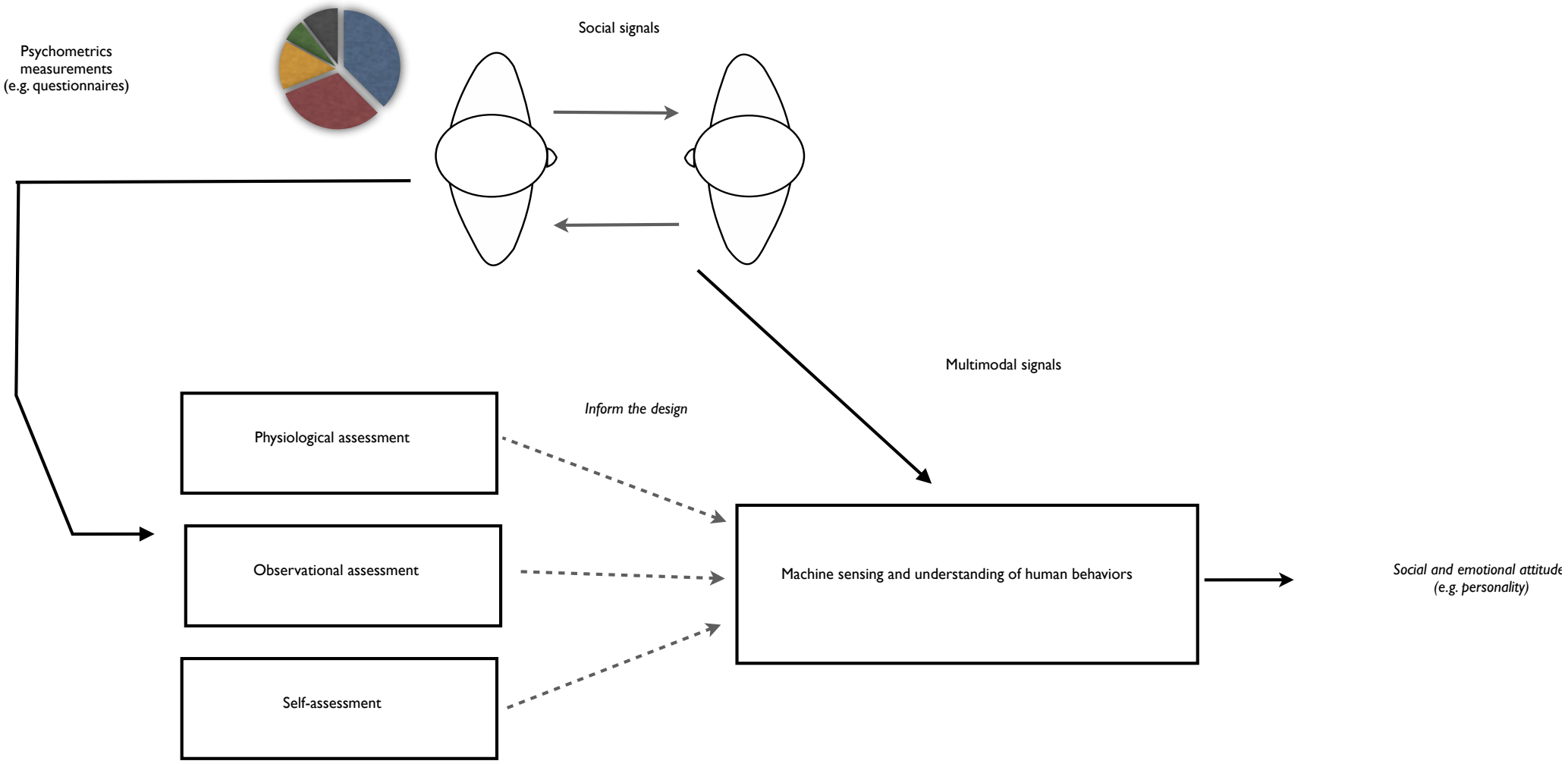


Statistical nature of speech signals?

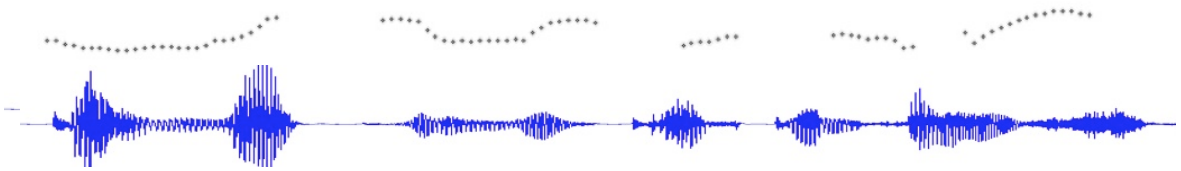
Gaussian vs Non-gaussian, Stationarity...

Manual vs Automatic annotations

# Affective Computing



# Speech Affective Computing Methodology



(1) Extraction of Low-Level Descriptors (LLD):

- Pitch (fundamental frequency)
- Energy
- Spectrum
- …

(2) Application of functionals

- Statistics
- Duration
- Linear prediction
- …

(3) Classification

- Speech turn
- Utterance turn

# Speech Affective Computing

## Lessons learnt from the First Computational Paralinguistics Challenge

Table 6
ComParE acoustic feature set: 65 provided **low-level descriptors** (LLD).

| **4 Energy Related LLD** | Group |
|---|---|
| Sum of Auditory Spectrum (Loudness) | Prosodic |
| Sum of RASTA-Style Filtered Auditory Spectrum | Prosodic |
| RMS Energy, Zero-Crossing Rate | Prosodic |
| **55 Spectral LLD** | **Group** |
| RASTA-Style Auditory Spectrum, Bands 1−26 (0−8 kHz) | Spectral |
| MFCC 1−14 | Cepstral |
| Spectral Energy 250−650 Hz, 1 k−4 kHz | Spectral |
| Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90 | Spectral |
| Spectral Flux, Centroid, Entropy, Slope, Harmonicity | Spectral |
| Spectral Psychoacoustic Sharpness | Spectral |
| Spectral Variance, Skewness, Kurtosis | Spectral |
| **6 Voicing Related LLD** | **Group** |
| $F_0$ (SHS & Viterbi Smoothing) | Prosodic |
| Probability of Voicing | Sound Quality |
| Log. HNR, Jitter (Local, Delta), Shimmer (Local) | Sound Quality |

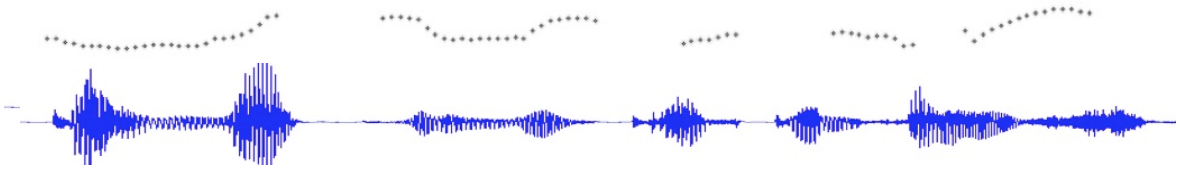
Björn Schuller, Felix Weninger, Yue Zhang, Fabien Ringeval, Anton Batliner, Stefan Steidl, Florian Eyben, Erik Marchi, Alessandro Vinciarelli, Klaus Scherer, **Mohamed Chetouani,** Marcello Mortillaro, Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2019)

Table 7
ComParE acoustic feature set: **functionals** applied to LLDs as defined in Table 6.

| |
|---|
| **Mean Values** |
| Arithmetic Mean$^{A\triangleq,B}$, Arithmetic Mean of Positive Values$^{A^\Delta,B}$ |
| Root-Quadratic Mean, Flatness |
| **Moments:** Standard Deviation, Skewness, Kurtosis |
| **Temporal Centroid**$^{A\triangleq,B}$ |
| **Percentiles** |
| Quartiles 1−3, Inter-Quartile Ranges 1−2, 2−3, 1−3 |
| 1%-tile, 99%-tile, Range 1−99% |
| **Extrema** |
| Relative Position of Maximum and Minimum, Full Range (Maximum−Minimum) |
| **Peaks and Valleys**$^{A}$ |
| Mean of Peak Amplitudes, |
| Difference of Mean of Peak Amplitudes to Arithmetic Mean |
| Mean of Peak Amplitudes Relative to Arithmetic Mean |
| Peak to Peak Distances: Mean and Standard Deviation |
| Peak Range Relative to Arithmetic Mean |
| Range of Peak Amplitude Values |
| Range of Valley Amplitude Values Relative to Arithmetic Mean |
| Valley-Peak (Rising) Slopes: Mean and Standard Deviation |
| Peak-Valley (Falling) Slopes: Mean and Standard Deviation |
| **Up-Level Times:** 25%, 50%, 75%, 90% |
| **Rise and Curvature Time** |
| Relative Time in which Signal is Rising |
| Relative Time in which Signal has Left Curvative |
| **Segment Lengths**$^{A}$ |
| Mean, Standard Deviation, Minimum, Maximum |
| **Regression**$^{A\triangleq,B}$ |
| Linear Regression: Slope, Offset, Quadratic Error |
| Quadratic Regression: Coefficients *a* and *b*, Offset *c*, Quadratic Error |
| **Linear Prediction** |
| LP Analysis Gain (Amplitude Error), LP Coefficients 1−5 |

$^{A}$Functionals applied only to energy related and spectral LLDs (group A)
$^{B}$Functionals applied only to voicing related LLDs (group B)
$^{\Delta}$Functionals applied only to $\Delta$LLDs
$\triangleq$ Functionals **not** applied to $\Delta$LLDs

# Speech Affective Computing Methodology



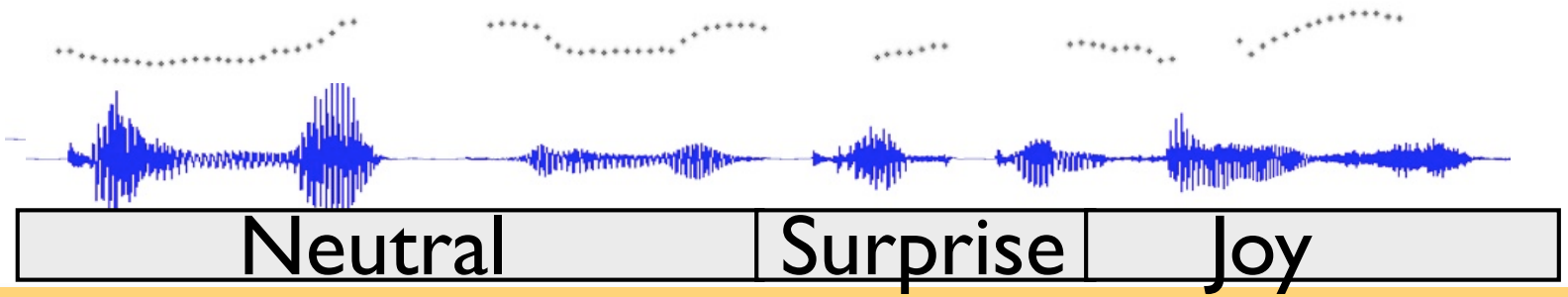
**INTERSPEECH 2013 configuration:**
6373 features

Group A: 4 energy related LLDs + 55 spectral LLDs
Group B: 6 voicing related LLDs

54 functionals applied to Group A + 46 applied to Delta LLDs -> 5900 features
39 functionals applied to Group B and Delta LLDs -> 468 features
5 temporal static descriptors for voiced segments: mean length, standard deviation of the segment length, minimum and maximum of voiced segments, ratio of non-zero F0

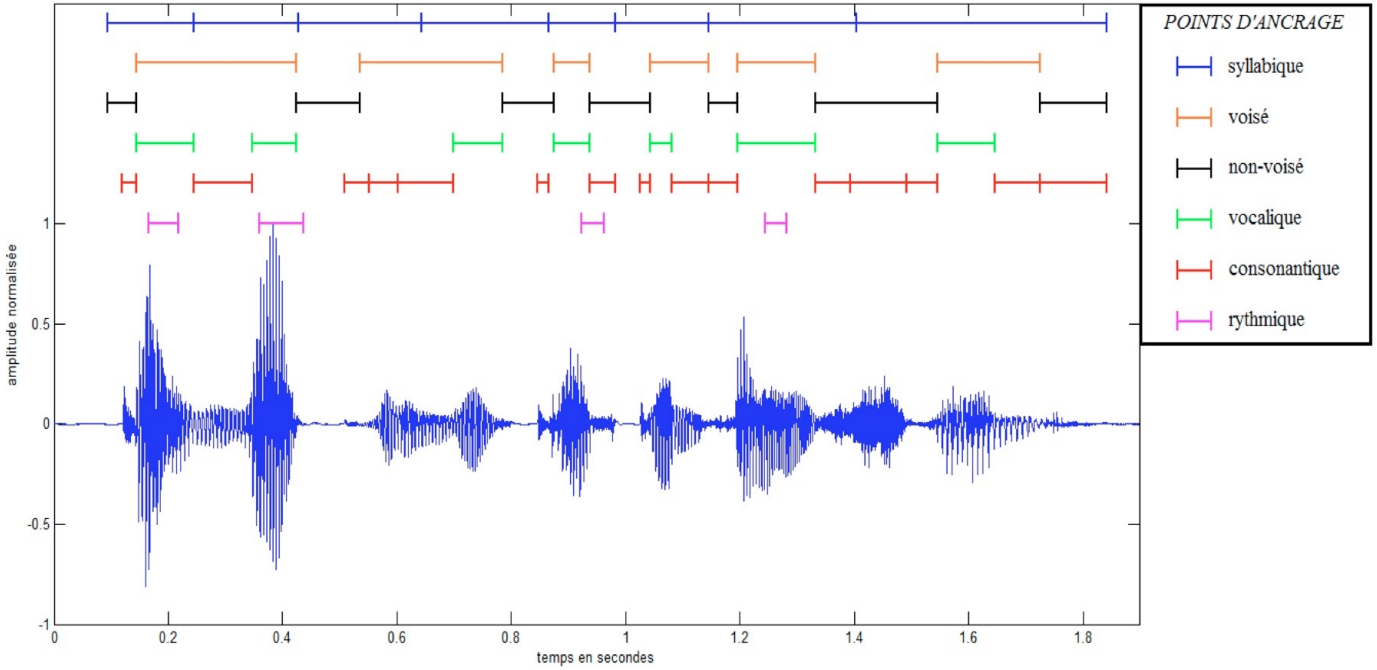
**Total:** 5900 + 468 + 5 = **6373 features**

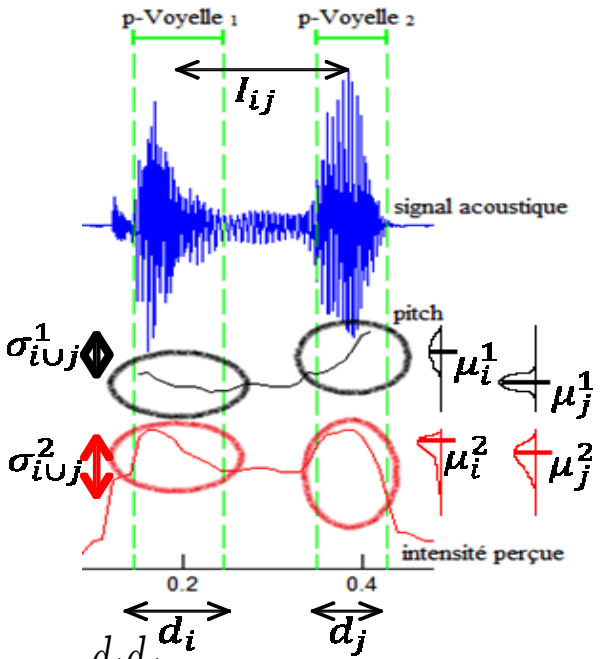
# Speech Affective Computing




Units for emotional speech processing:

- Acoustical, Prosodic and Rhythmic prominence
- Dynamics of units: Huang-Hilbert transform, Hotelling distance...



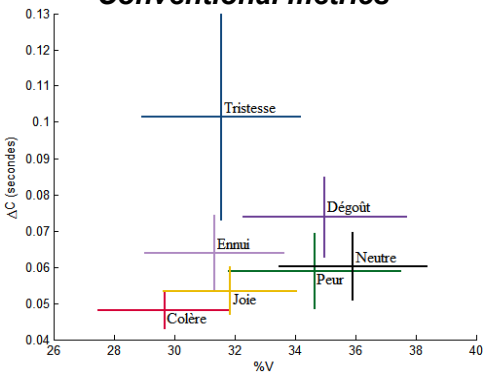

$$HD_{ij} = \frac{d_i d_j}{d_i + d_j}\left[(\mu_i - \mu_j)^T \Sigma^{-1}_{i\cup j}(\mu_i - \mu_j)\right]$$

# Speech Affective Computing

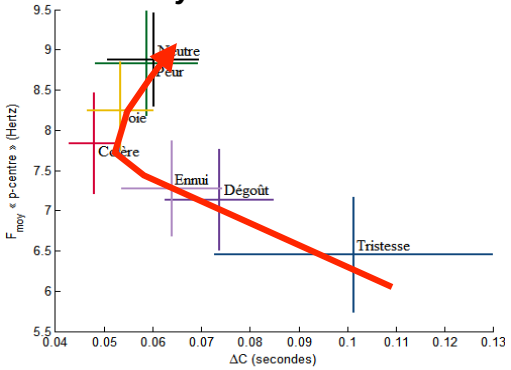
## Applications

- Acted emotion recognition
- Children assessment: Autism, Pervasive Developmental Disorders, Language impairment
- Spontaneous and atypical emotions

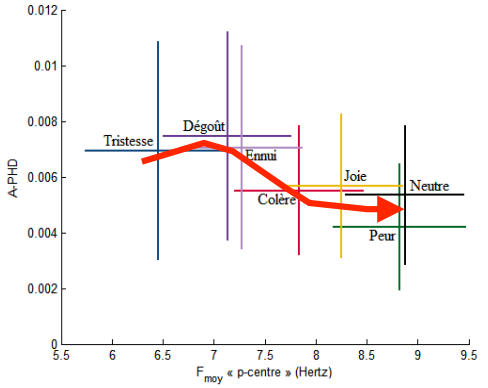
***Conventional metrics***



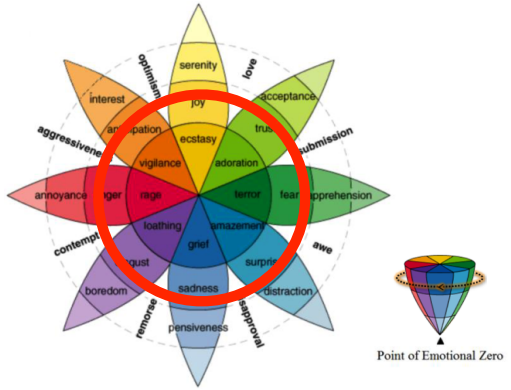

***Hybrid metrics***



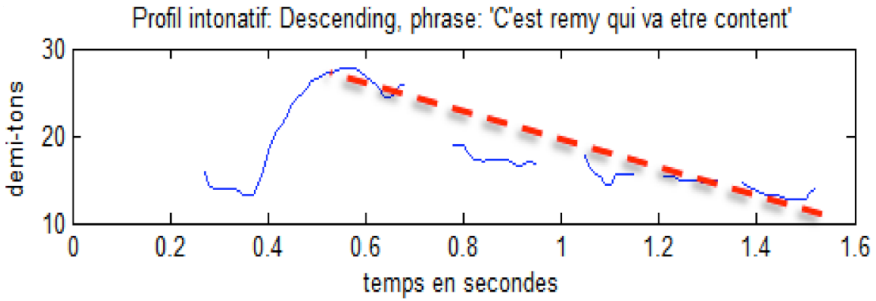

***Non-conventional metrics***

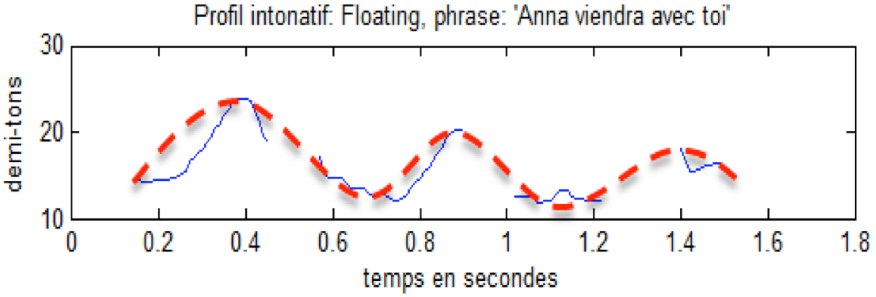



***Plutchik wheel* [Plutchik 80]**




## Automatic evaluation of prosody

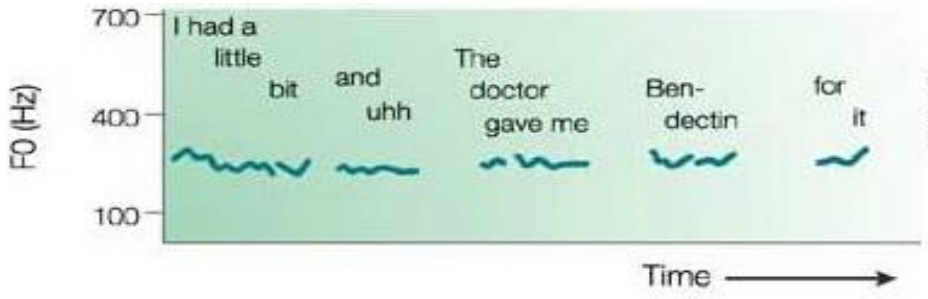


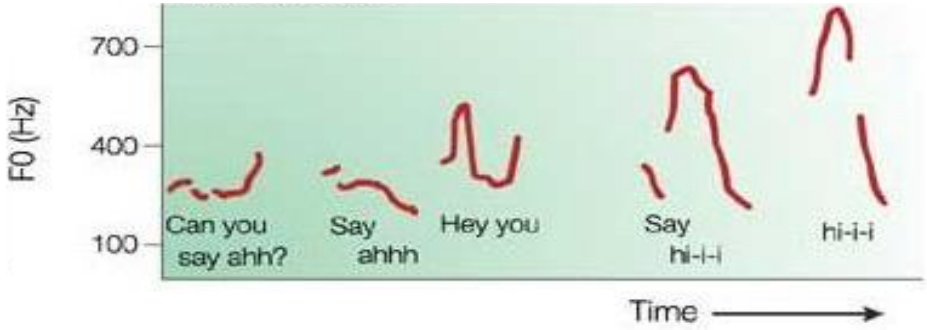

F. Ringeval et al. : Automatic intonation recognition for the prosodic assessment of language impaired children. *IEEE Trans. Audio, Speech and Language Processing* **19**(5) 1328-1342 (2011).

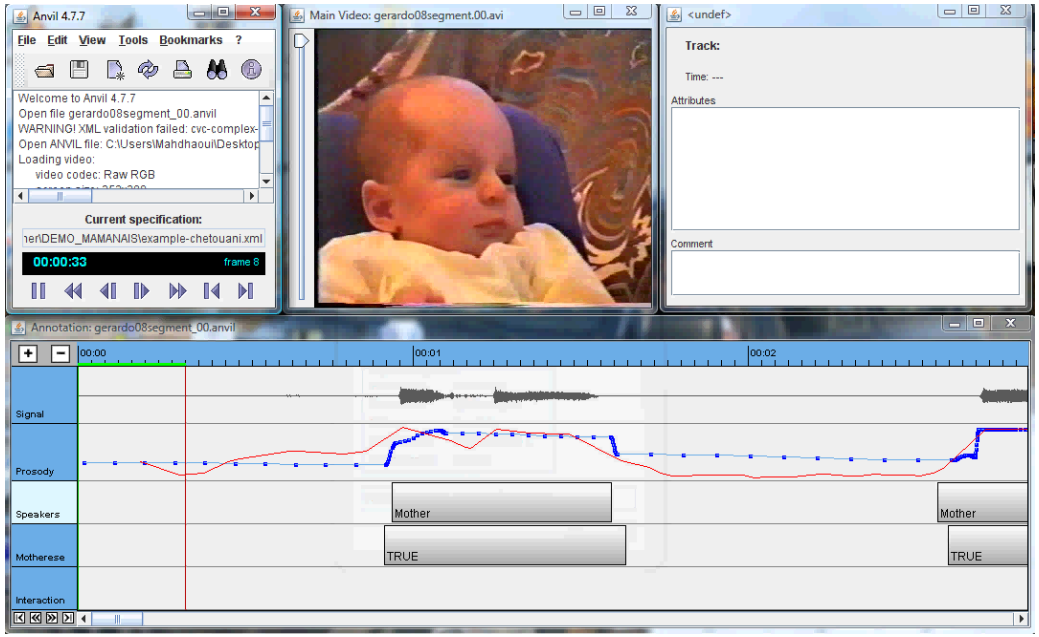
# Speech Affective Computing Methodology

**Adult-directed speech**




**Infant-directed speech**



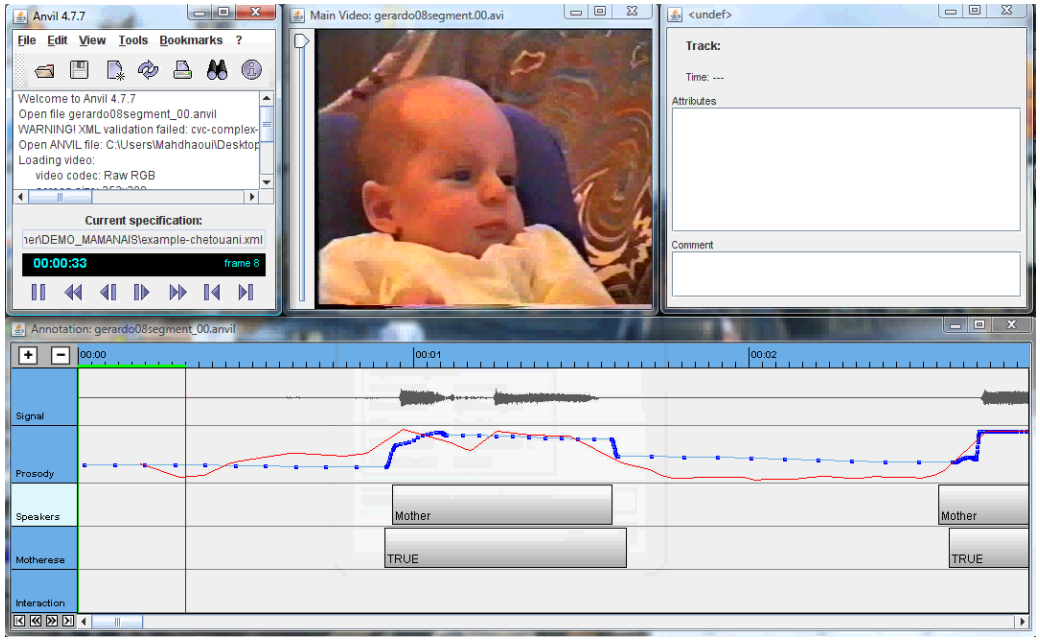



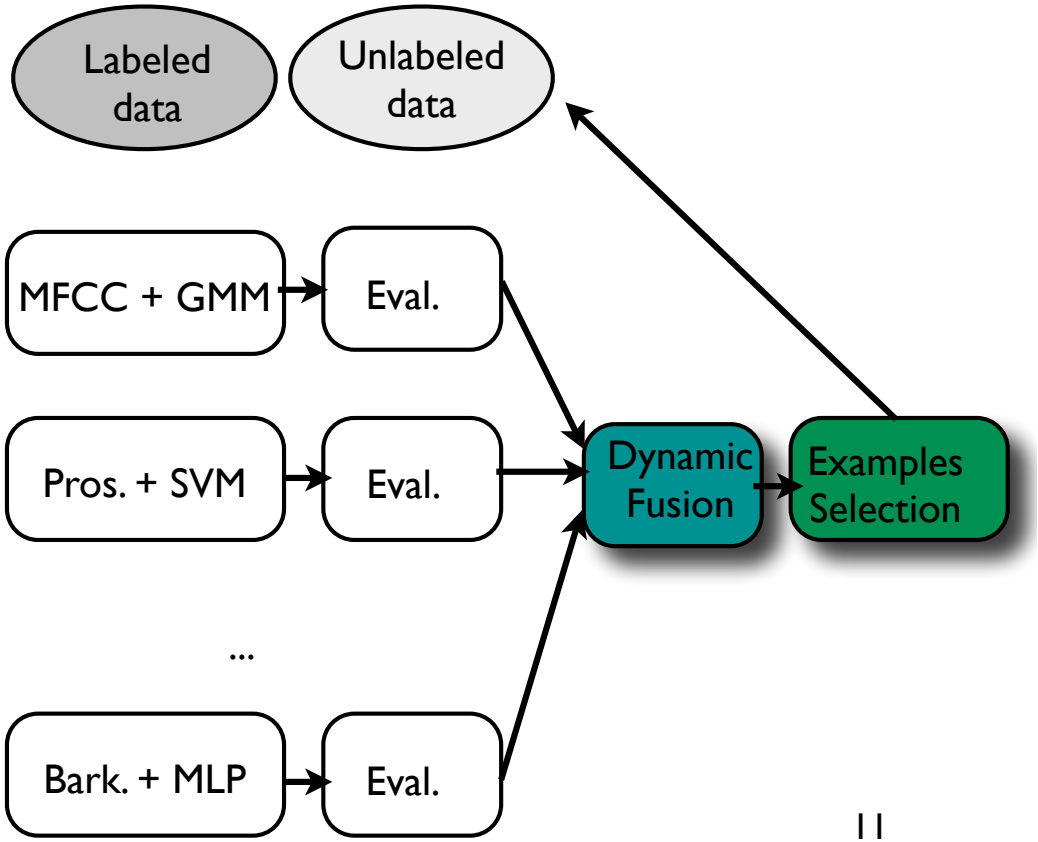

C. Saint-georges et al. : Motherese in Interaction: At the Cross-Road of Emotion and Cognition? (A Systematic Review) *Plos One*, 2013

# Speech Affective Computing
# Co-Learning approaches

Annotation and Subjectivity of social signals

- Semi-supervised learning: combining labeled and unlabeled data
- Co-training and fusion: Multi-view characterization



# Speech Affective Computing
# Co-Learning approaches

Table 3: Self-training algorithm

**Given:**
a set $L$ of Labelled examples
a set $U$ of Unlabelled examples
a number $n$ of examples to be added to $L$ in each iteraction
**Loop:**
Use $L$ to train the classifier $h$
Allow $h$ to label $U$
Let $T$ be the $n$ examples in $U$ on which $h$ makes the most confident predictions
Add $T$ to $L$
Remove $T$ from $U$
**End**

Table 4: Co-Training algorithm

**Given:**
a set $L$ of Labelled examples
a set $U$ of Unlabelled examples
**Loop:**
Use $L$ to train each classifier $h_1$
Use $L$ to train each classifier $h_2$
Allow $h_1$ to label $p_1$ positive and $n_1$ negative examples from $U$
Allow $h_2$ to label $p_2$ positive and $n_2$ negative examples from $U$
Add these self-labelled examples to $L$
Remove these self-labelled examples from $U$
**End**

A. Mahdhaoui and M. Chetouani : Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis. *Speech Communication* **53**(9) 1149-1161 (2011).

# Speech Affective Computing
# Co-Learning approaches

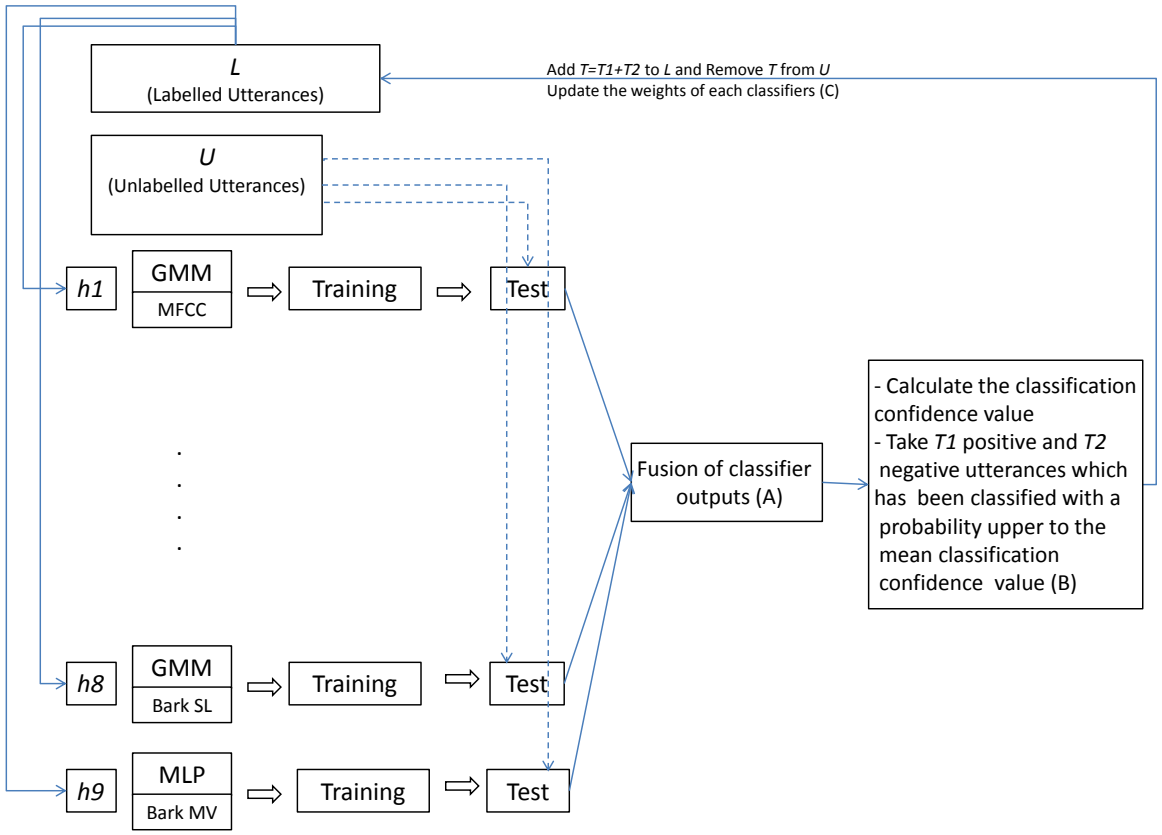



Figure 5: Structure of the proposed Co-training algorithm

A. Mahdhaoui and M. Chetouani : Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis. *Speech Communication* **53**(9) 1149-1161 (2011).

# Speech Affective Computing Co-Learning approaches

Table 5: The proposed Co-Training algorithm

**Given:**

a set $L$ of $m$ Labelled examples $\{(l_1^1, ..., l_1^v, y_1), ..., (l_m^1, ..., l_m^v, y_m)\}$ with labels $y_i = \{1, 2\}$

a set $U$ of $n$ Unlabelled examples $\{(x_1^1, ..., x_1^v), ..., (x_n^1, ..., x_n^v)\}$

$v$ = number of view (classifier)

**Initialization:**

$\omega_k$ (weights of classifier)= $1/v$ for all the view

**While U not empty**

**A. Classify all the example of the test database:**

Do for $k = 1, 2, ..., v$

1. Use $L$ to train each classifier $h_k$
2. Classify all examples of $U$ by each $h_k$
3. Calculate the probability of classification for each example $x_i$ from $U$, $p(C_j|x_i) = \sum_{k=1}^{v} \omega_k \times h_k(C_j|x_i^k)$
4. $Labels(x_i) = argmax(p(C_j|x_i))$

End for

**B. Update the training ($L$) and test ($U$) databases:**

$U_j = \{z_1, ..., z_{nj}\}$ the ensemble of example classified $C_j$

Do for $i = 1, 2, ..., nj$

$p(C_j|z_i) = \frac{\sum_{k=1}^{v} \omega_k \times h_k(C_j|z_i^k)}{\sum_{k=1}^{v} \omega_k}$

End for

$margin_j = \frac{\sum_{1}^{n_j} p(C_j|z_i)}{n_j}$

Take $T_j$ from $U_j$ the examples which has classified on $C_j$ with a probability upper to $margin_j$.

$T = \sum T_j$

Add $T$ to $L$ and remove it from $U$

**C. Update weights:** $\omega_k = \frac{\sum_{i=1}^{size(T)} h_k(z_i^k)}{\sum_{k=1}^{v} \sum_{i=1}^{size(T)} h_k(z_i^k)}$

**End While**

A. Mahdhaoui and M. Chetouani : Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis. *Speech Communication* **53**(9) 1149-1161 (2011).

# Speech Affective Computing
# Co-Learning approaches

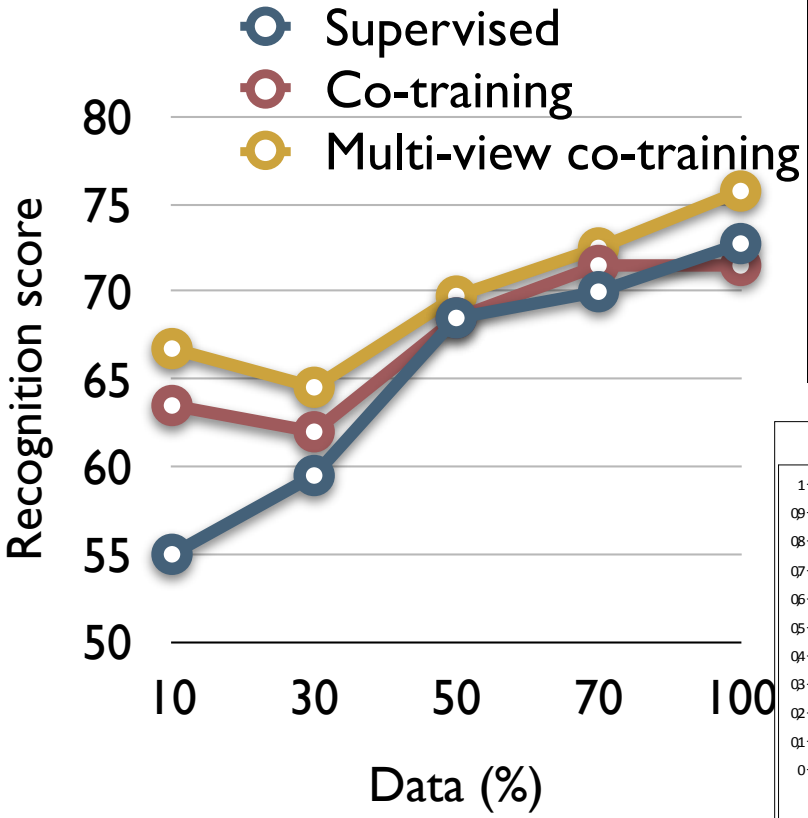
Table 8: Classification accuracy with different numbers of annotations

| Number of annotations | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Proposed Co-training method | 66.75 | 65.25 | 63.5 | 67 | 69.75 | 72.25 | 72.5 | 71.75 | 74 | 75.75 |
| *Co-training* standard (using h1 and h4) | 63.5 | 62.5 | 62 | 64.5 | 68.5 | 69.75 | 71.25 | 69.5 | 71 | 71.5 |
| *Co-training* standard (using all the classifiers h1-h9) | 57 | 58.5 | 58.5 | 61 | 64 | 67 | 67.25 | 68 | 69 | 68.5 |
| *Self-training (using h1: MFCC-GMM)* | 52 | 50 | 50 | 54 | 55 | 62.5 | 61 | 65 | 69 | 70.25 |
| *Self-training (using h2: prosody-GMM)* | 54 | 52.5 | 53 | 52 | 53.5 | 58 | 59 | 62 | 64.5 | 67.75 |
| Supervised method: MFCC-GMM (best configuration) | 55 | 59.25 | 59.5 | 61.5 | 68.5 | 71 | 70 | 69.75 | 71.5 | 72.75 |

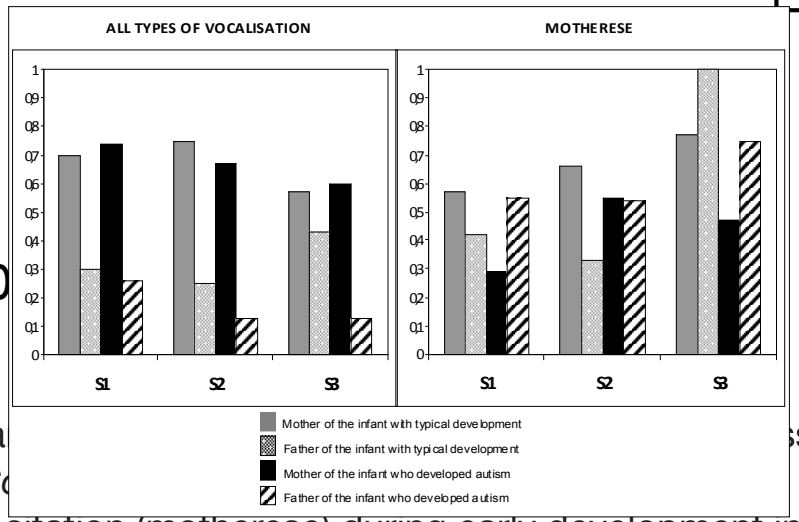
A. Mahdhaoui and M. Chetouani : Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis. *Speech Communication* **53**(9) 1149-1161 (2011).

# Speech Affective Computing
# Co-Learning approaches

- Motherese detection in Family Home Movies
- Early signs of Autism Spectrum Disorders




Motherese detection




A. Mahdhaoui and M. Chetouani : Supervised a ssification for parent-infant interaction analysis. *Speech Co*

R. Cassel et al.: Course of maternal prosodic incitation (motherese) during early development in autism: an exploratory study. *Interaction Studies. Vol 14 Pages 480-496 2014.*

# Lessons learnt from the First Computational Paralinguistics Challenge



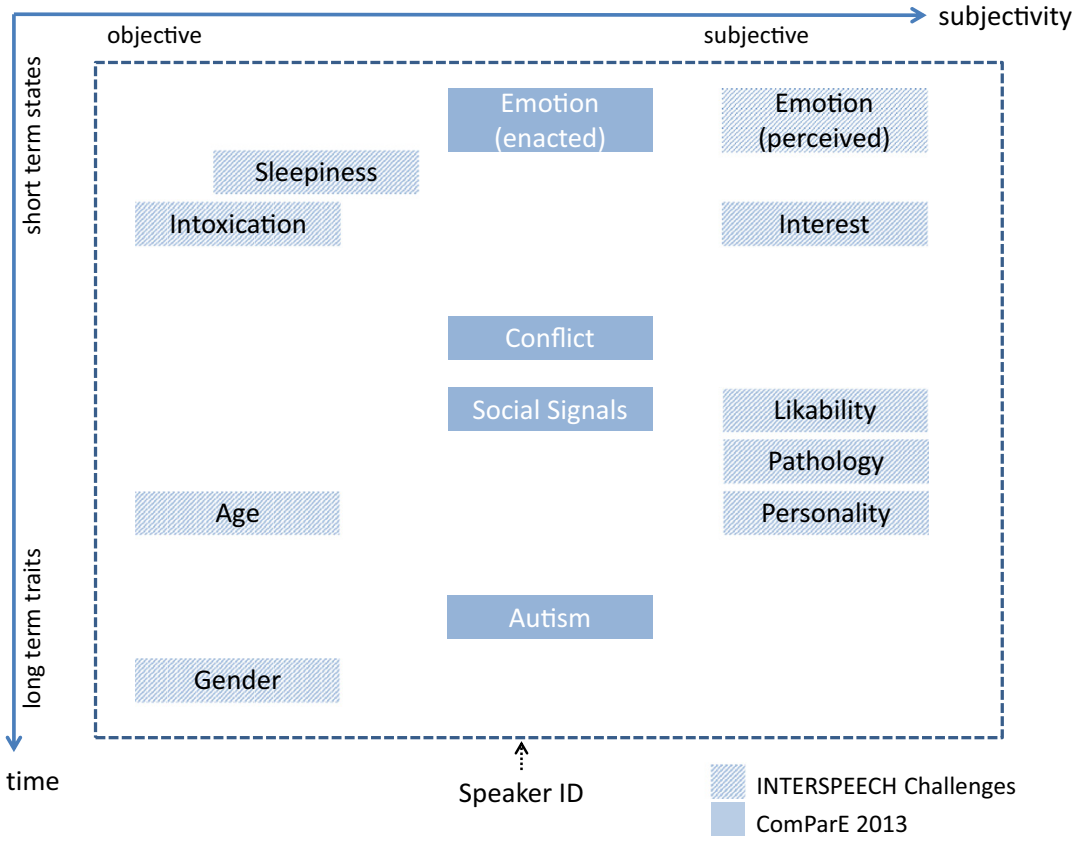



Fig. 1. Speaker characteristics investigated in the INTERSPEECH Challenges 2009–2012 and the First Computational Paralinguistics Challenge (ComParE) 2013.

Björn Schuller, Felix Weninger, Yue Zhang, Fabien Ringeval, Anton Batliner, Stefan Steidl, Florian Eyben, Erik Marchi, Alessandro Vinciarelli, Klaus Scherer, **Mohamed Chetouani,** Marcello Mortillaro, Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2019)

# Lessons learnt from the First Computational Paralinguistics Challenge

Table 4
Partitioning of the GEMEP database into train, dev(elopment), and test set for 12-way classification by emotion category, and binary classification by pos(itive)/neg(ative) arousal (A) and valence (V).

| # | Train | Dev | Test | A | V | Σ |
|---|---|---|---|---|---|---|
| Admiration$^{+}$ | 20 | 2 | 8 | pos | pos | 30 |
| Amusement | 40 | 20 | 30 | pos | pos | 90 |
| Anxiety | 40 | 20 | 30 | neg | neg | 90 |
| Cold anger | 42 | 12 | 36 | neg | neg | 90 |
| Contempt$^{+}$ | 20 | 6 | 4 | neg | neg | 30 |
| Despair | 40 | 20 | 30 | pos | neg | 90 |
| Disgust$^{+}$ | 20 | 2 | 8 | −* | −* | 30 |
| Elation | 40 | 12 | 38 | pos | pos | 90 |
| Hot anger | 40 | 20 | 30 | pos | neg | 90 |
| Interest | 40 | 20 | 30 | neg | pos | 90 |
| Panic fear | 40 | 12 | 38 | pos | neg | 90 |
| Pleasure | 40 | 20 | 30 | neg | pos | 90 |
| Pride | 40 | 12 | 38 | pos | pos | 90 |
| Relief | 40 | 12 | 38 | neg | pos | 90 |
| Sadness | 40 | 12 | 38 | neg | neg | 90 |
| Shame$^{+}$ | 20 | 2 | 8 | pos | neg | 30 |
| Surprise$^{+}$ | 20 | 6 | 4 | −* | −* | 30 |
| Tenderness$^{+}$ | 20 | 6 | 4 | neg | pos | 30 |
| Σ | 602 | 216 | 442 | | | 1260 |

$^{+}$ Mapped to 'other' and excluded from evaluation in 12-class task.
* Mapped to 'undefined' and excluded from evaluation in binary tasks.

Björn Schuller, Felix Weninger, Yue Zhang, Fabien Ringeval, Anton Batliner, Stefan Steidl, Florian Eyben, Erik Marchi, Alessandro Vinciarelli, Klaus Scherer, **Mohamed Chetouani,** Marcello Mortillaro, Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2019)

# Deep Learning for Human Affect Recognition: Insights and New Developments

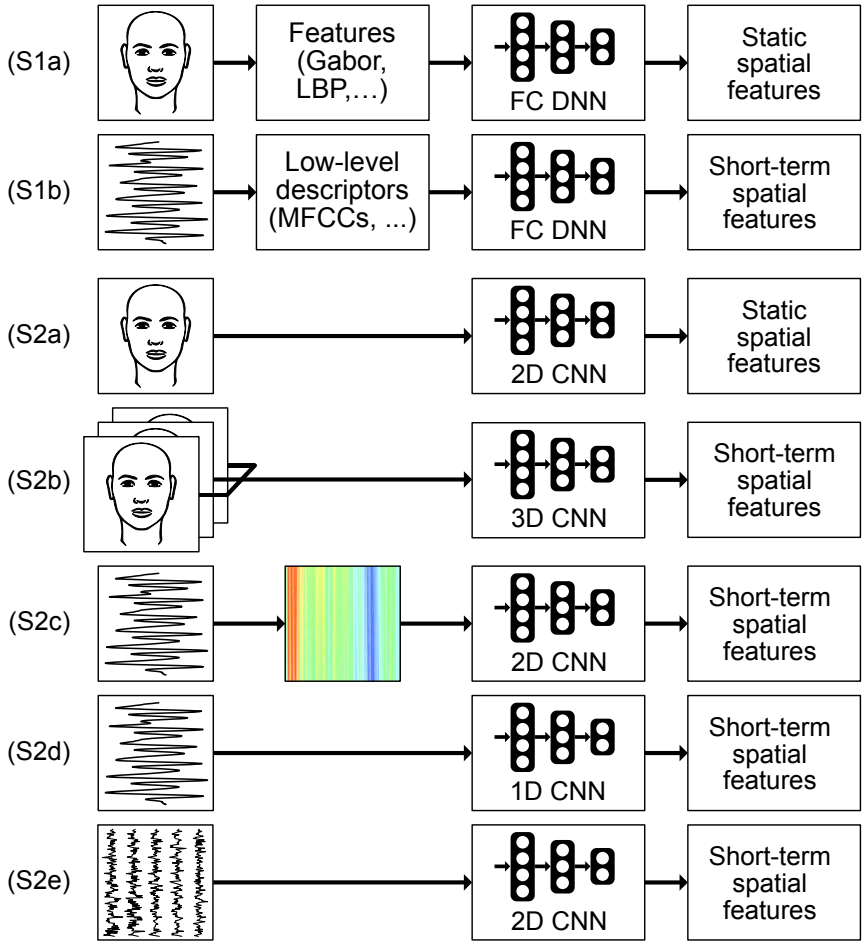



Fig. 2. Applications of deep learning for spatial feature learning with fully-connected DNNs (S1a–S1b) and CNNs (S2a–S2e).

P. V. Rouast, M. Adam and R. Chiong, "Deep Learning for Human Affect Recognition: Insights and New Developments," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2018.2890471.

# Deep Learning for Human Affect Recognition: Insights and New Developments

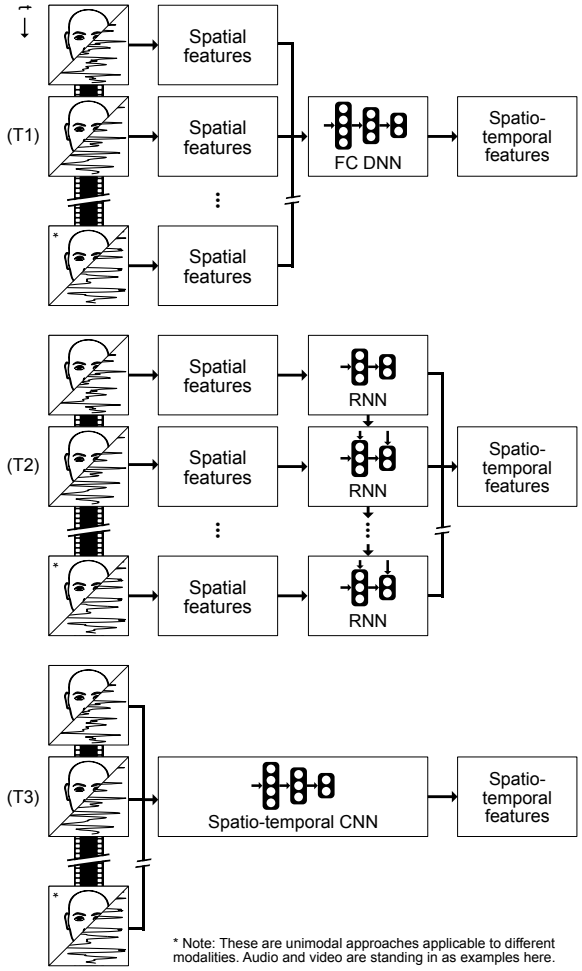



Fig. 5. Applications of deep learning for temporal feature learning with fully-connected DNNs (T1), RNNs (T2), and CNNs (T3).

P. V. Rouast, M. Adam and R. Chiong, "Deep Learning for Human Affect Recognition: Insights and New Developments," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2018.2890471.

# Deep Learning for Human Affect Recognition: Insights and New Developments

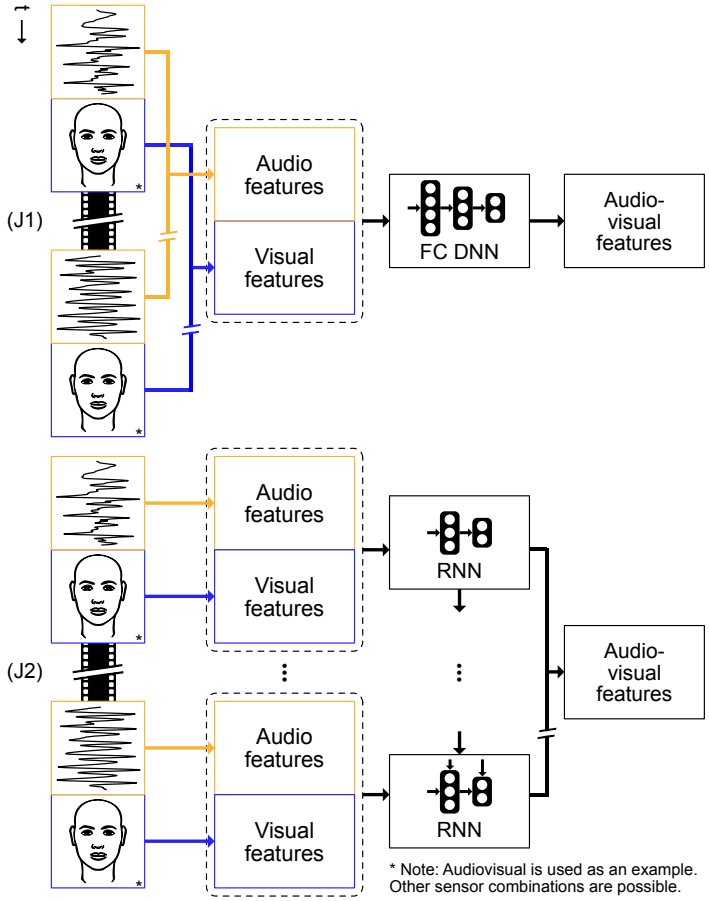
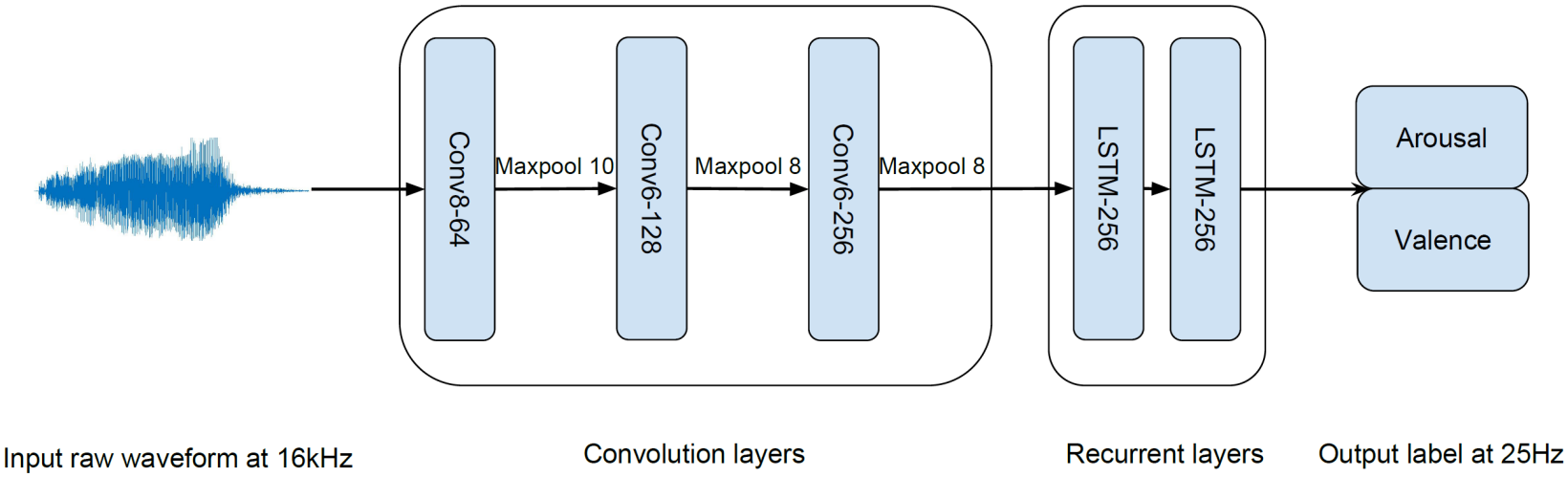



Fig. 6. Applications of deep learning for joint multimodal feature learnin with fully-connected fusion DNNs (J1) and fusion RNNs (J2).

P. V. Rouast, M. Adam and R. Chiong, "Deep Learning for Human Affect Recognition: Insights and New Developments," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2018.2890471.

# End-to-end approach

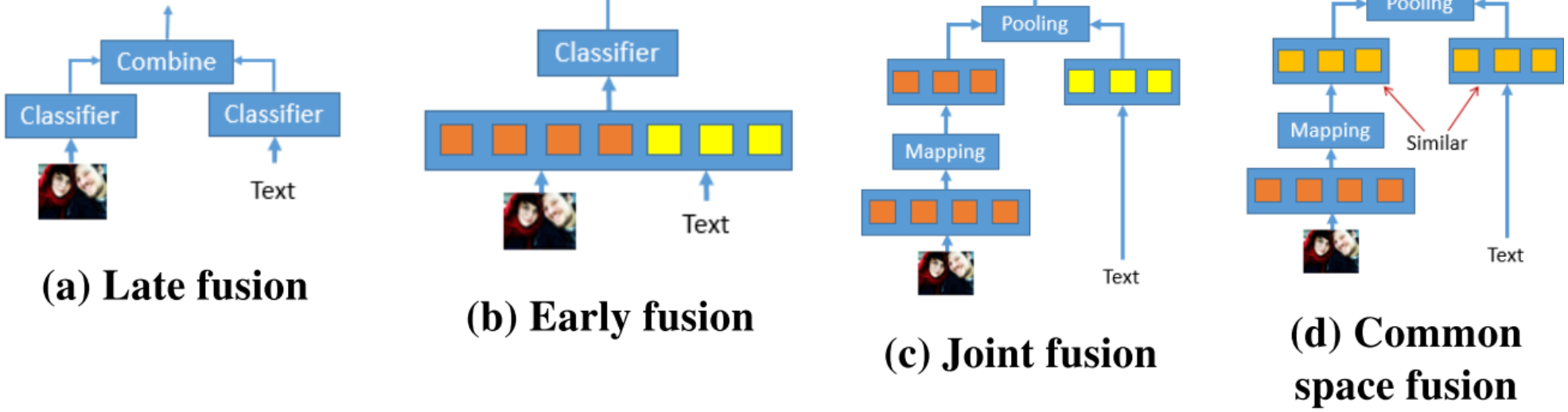



## Convolution operation:

$$(f \star h)(i,j) = \sum_{k=-T}^{T} \sum_{m=-T}^{T} f(k,m) \cdot h(i-k, j-m) \quad (1)$$

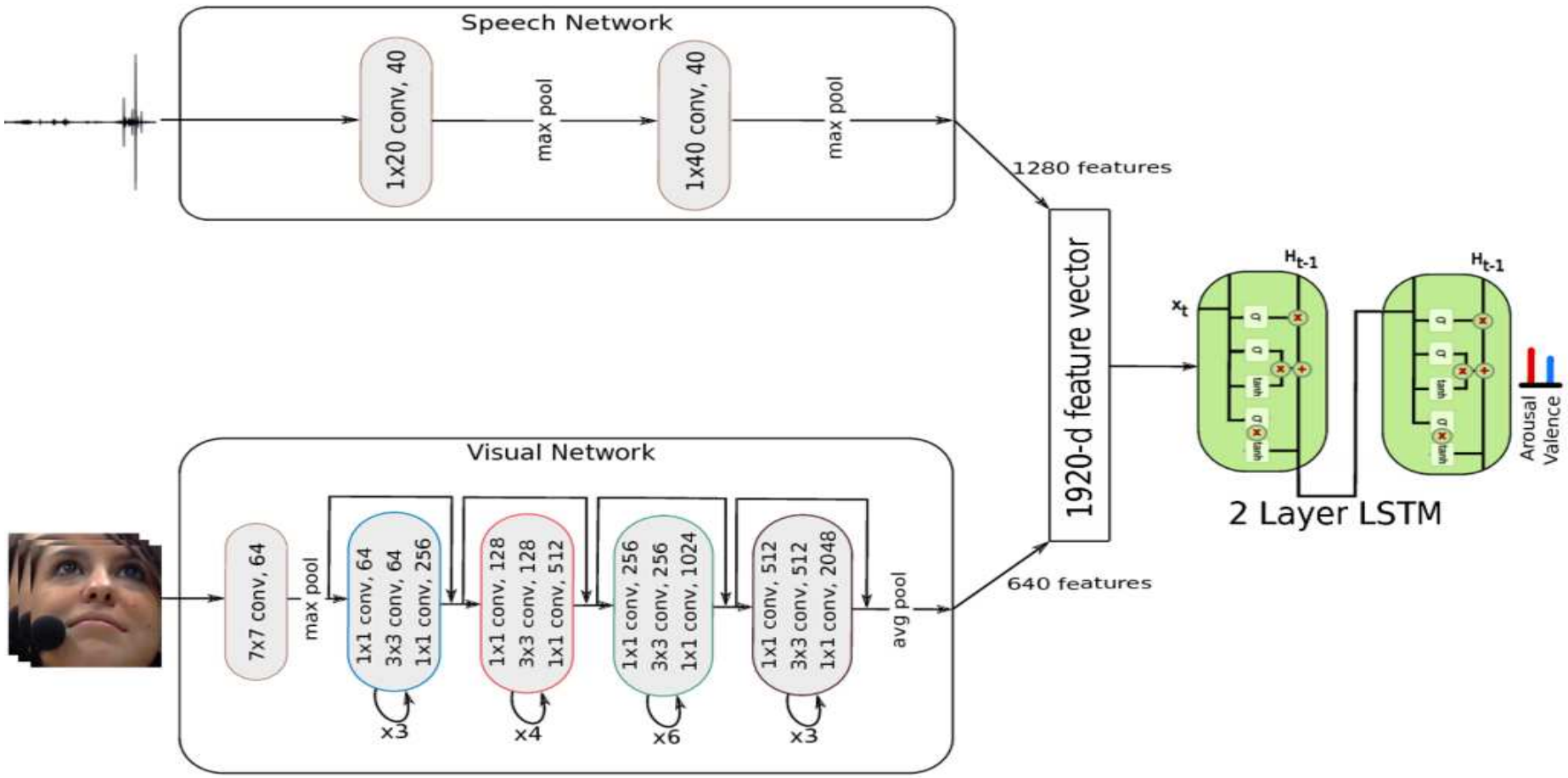
P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017, doi: 10.1109/JSTSP.2017.2764438.

# End-to-End Multimodal Emotion Recognition



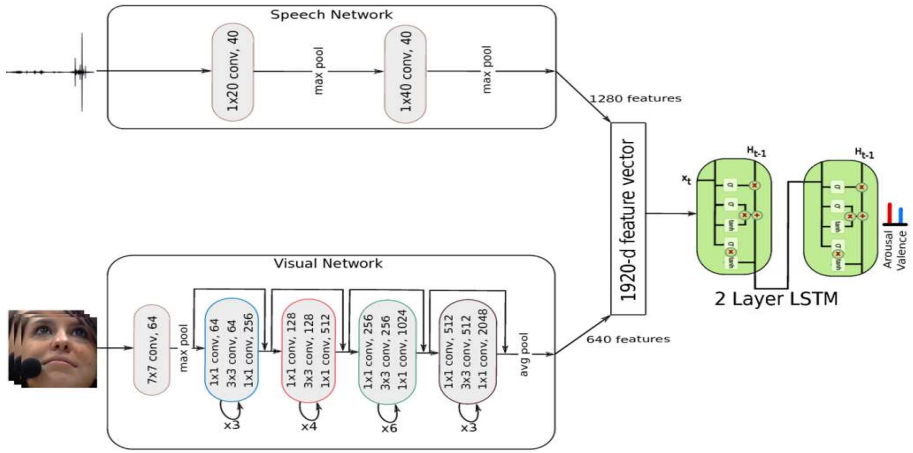

P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017, doi: 10.1109/JSTSP.2017.2764438.

# End-to-End Multimodal Emotion Recognition




P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017, doi: 10.1109/JSTSP.2017.2764438.

# End-to-End Multimodal Emotion Recognition



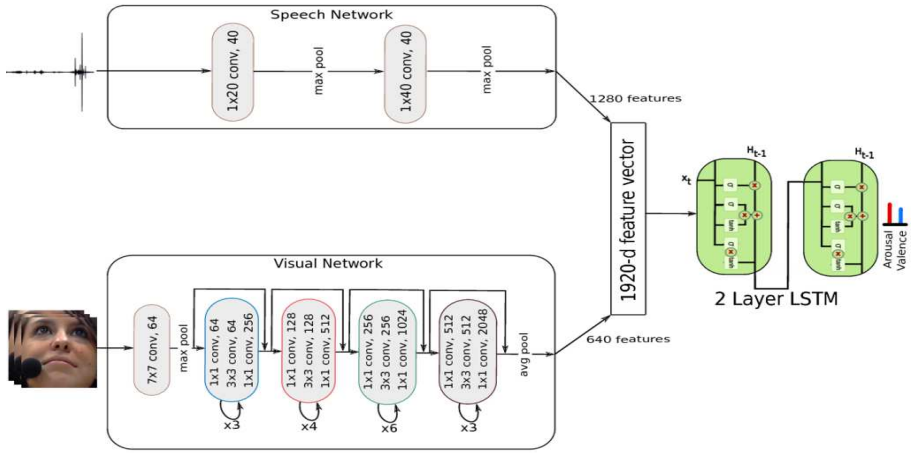

ResNet

**Speech Network:**

- Raw waveform (6s)
- Temporal Convolution (F=20): space time finite impulse filters with 5ms window in order to extract fine-scale spectral information
- Pooling across time: Half-wave rectifier (analogous to the cochlear transduction step in the human ear) and then downsampled to 8kHz (pool size = 2)
- Temporal convolution (F=40): Extract more long-term characteristics of the speech signal
- Max pooling across channels
- Dropout

P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017, doi: 10.1109/JSTSP.2017.2764438.

# End-to-End Multimodal Emotion Recognition




| Sequence length | Arousal | Valence |
|---|---|---|
| *Visual network* | | |
| 75 | .293 | .276 |
| 150 | .363 | .488 |
| 300 | .193 | .496 |
| *Speech network* | | |
| 75 | .727 | .345 |
| 150 | .744 | .369 |
| 300 | .685 | .130 |

Table II: Results (in terms of $\rho_c$) on arousal and valence after 60 epochs when varying sequence length for speech and visual networks.

**Objective function:**
Instead of MSE, correlation

$$\mathcal{L}_c = 1 - \rho_c = 1 - \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$
$$= 1 - 2\sigma_{xy}^2 \psi^{-1} \quad (3)$$

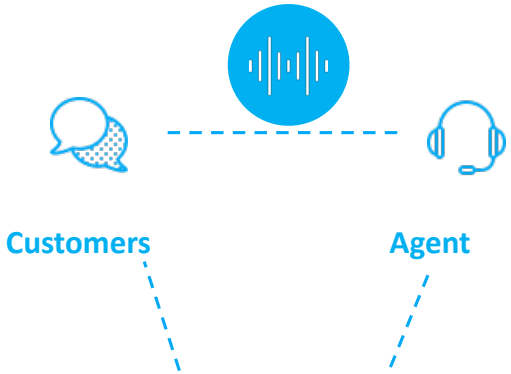
where $\psi = \sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2$ and $\mu_x = \mathbb{E}(\mathbf{x})$, $\mu_y = \mathbb{E}(\mathbf{y}), \sigma_x^2 = \text{var}(\mathbf{x}), \sigma_y^2 = \text{var}(y)$ and $\sigma_{xy}^2 = \text{cov}(\mathbf{x}, \mathbf{y})$ Thus, to minimise $\mathcal{L}_c$ (or maximise $\rho_c$), we backpropagate the gradient of the last layer weights with respect to $\mathcal{L}_c$,

$$\frac{\partial \mathcal{L}_c}{\partial \mathbf{x}} \propto 2\frac{\sigma_{xy}^2(\mathbf{x} - \mu_y)}{\psi^2} + \frac{\mu_y - \mathbf{y}}{\psi}, \quad (4)$$

where all vector operations are done element-wise.

P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017, doi: 10.1109/JSTSP.2017.2764438.

# Customer Satisfaction & Emotion




A typical 5-point Likert scale:

- Completely Dissatisfied
- Somewhat Dissatisfied
- Neutral
- Somewhat Satisfied
- Completely Satisfied

- **H1 Customers' emotions and CSAT response**
  - **H1a**: Customers expressing positive emotions respond more to CSAT questionnaires;
  - **H1b**: Customers expressing negative emotions respond less to CSAT questionnaires;
  - **H1c**: Customers expressing anger respond less to CSAT questionnaires;
- **H2 Customers' emotions and self-reported satisfaction**
  - **H2a**: Customers expressing positive emotions report higher satisfaction;
  - **H2b**: Customers expressing negative emotions report lower satisfaction;
  - **H2c**: Customers expressing anger report lower satisfaction;
- **H3 Customers' emotional profiles and CSAT response rate**
  - **H3a**: Customers manifesting upward positive valence dynamics (more positive emotions towards the end of the call) exhibit a higher CSAT response rate compared to flat or negative dynamics;
  - **H3b**: Customers manifesting downward negative valence dynamics (fewer negative emotions towards the end of the call) exhibit a higher CSAT response rate compared to flat or positive dynamics;
  - **H3c**: Customers manifesting downward anger dynamics (fewer anger events towards the end of the call) exhibit a higher CSAT response rate;
- **H4 Customers' emotional profiles and self-reported satisfaction**
  - **H4a**: Customers manifesting upward positive valence dynamics report higher satisfaction;
  - **H4b**: Customers manifesting downward negative valence dynamics report higher satisfaction;
  - **H4c**: Customers manifesting downward anger dynamics report higher satisfaction;

Eric Bolo, Muhammad Samoul, Nicolas Seichepine, Mohamed Chetouani. Quietly Angry, Loudly Happy: Self-Reported Customer Satisfaction Vs. Automatically Detected Emotion In Contact Center Calls. Interaction Studies, 2023, 24 (1), pp.168-192. 10.1075/is.22038.bol. hal-04109350

# Customer Satisfaction & Emotion

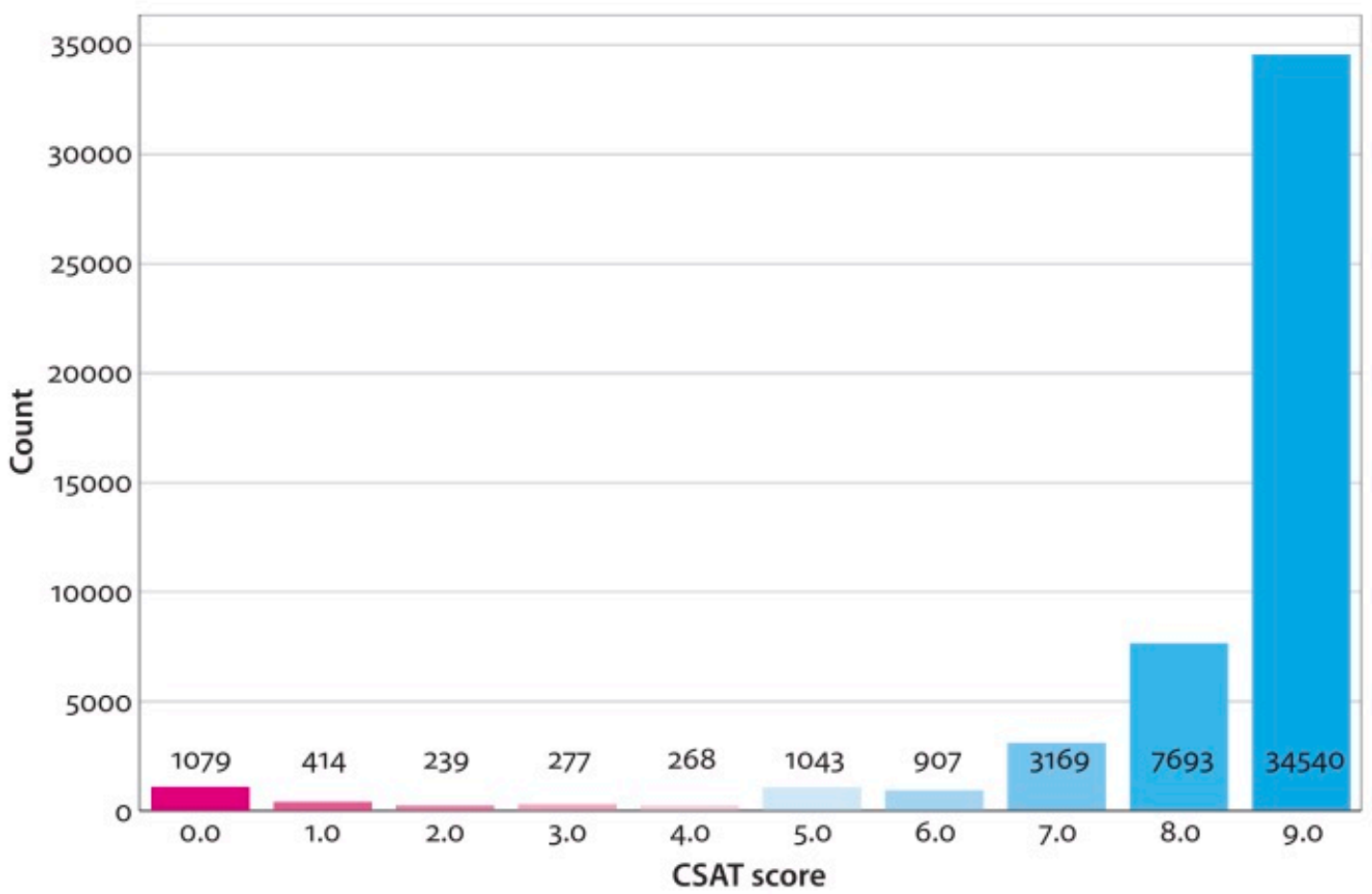



**Figure 1.** CSAT scoring distribution

Eric Bolo, Muhammad Samoul, Nicolas Seichepine, Mohamed Chetouani. Quietly Angry, Loudly Happy: Self-Reported Customer Satisfaction Vs. Automatically Detected Emotion In Contact Center Calls. Interaction Studies, 2023, 24 (1), pp.168-192. 10.1075/is.22038.bol. hal-04109350

# Customer Satisfaction & Emotion

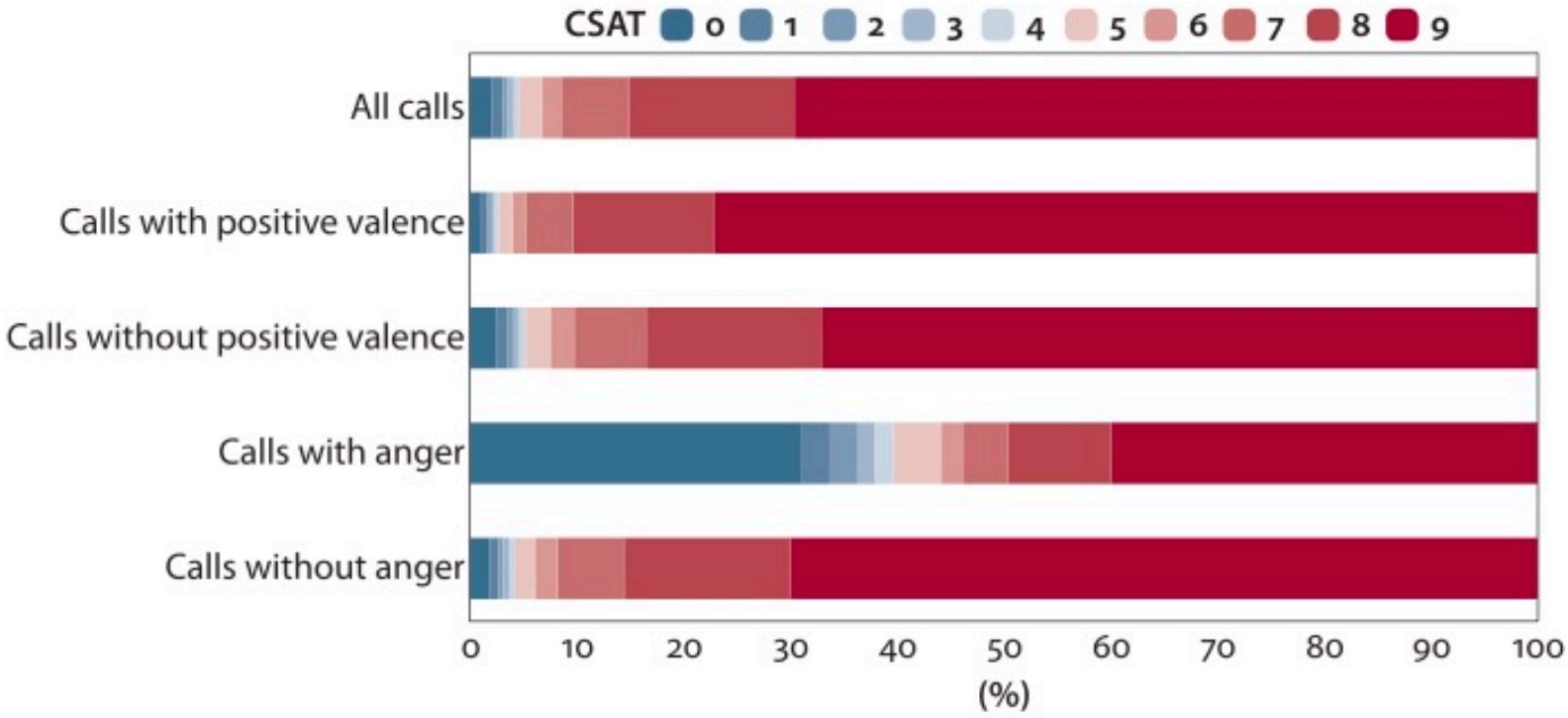



**Figure 3.** CSAT scoring distribution with respect to emotional analysis. Number of calls are reported in Table 6

Eric Bolo, Muhammad Samoul, Nicolas Seichepine, Mohamed Chetouani. Quietly Angry, Loudly Happy: Self-Reported Customer Satisfaction Vs. Automatically Detected Emotion In Contact Center Calls. Interaction Studies, 2023, 24 (1), pp.168-192. 10.1075/is.22038.bol. hal-04109350

# Customer Satisfaction & Emotion

$$\text{CSAT response rate} = \frac{\text{Nbr of Calls with CSAT scoring}}{\text{Total Nbr of Calls}}$$

$$\Delta = \frac{N_{end} - N_{beg}}{N_{beg} + N_{mid} + N_{end}}$$

$$\text{aggH-CSAT} = \frac{\sum_{i=6}^{9} CSATscore(i)}{\sum_{i=1}^{9} CSATscore(i)}$$

Table 4: CSAT response rate and emotion in respect to the emotion analysis. * indicates a significant difference wit the baseline (p<0.05)

| # of calls | With CSAT scoring | Without CSAT Scoring | CSAT response rate |
|---|---|---|---|
| All calls | 49 629 | 111 001 | 0.29 |
| Calls with positive valence | 13 491 | 27 212 | **0.33*** |
| Calls with negative valence | 31 722 | 67 907 | **0.32*** |
| Calls with anger | 892 | 3 586 | **0.20*** |

Eric Bolo, Muhammad Samoul, Nicolas Seichepine, Mohamed Chetouani. Quietly Angry, Loudly Happy: Self-Reported Customer Satisfaction Vs. Automatically Detected Emotion In Contact Center Calls. Interaction Studies, 2023, 24 (1), pp.168-192. 10.1075/is.22038.bol. hal-04109350

# Customer Satisfaction & Emotion

Table 5: Proportion of detected emotion in calls with and without CSAT scoring.

| Proportion in | With CSAT scoring ($\pi_1$) | Without CSAT Scoring ($\pi_2$) | $\pi_1$ vs. $\pi_2$ |
|---|---|---|---|
| Calls with positive valence | 0.27 | 0.24 | $\mathbf{z = 11.36, p < 0.05}$ |
| Calls with negative valence | 0.63 | 0.61 | $\mathbf{z = 10.46, p < 0.05}$ |
| Calls with anger | 0.01 | 0.03 | $\mathbf{z = -16.12, p < 0.05}$ |

Table 6: Number of calls per CSAT scoring in respect of affective computing analysis (see also figure 3) and the Aggregate High Customer Satisfaction score (aggH-CSAT) in respect to the emotion analysis. * indicates a significant difference with the baseline (p<0.05)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | aggH-CSAT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| All calls | 1079 | 414 | 239 | 277 | 268 | 1043 | 907 | 3169 | 7693 | 34540 | 0.93 |
| Calls with positive valence | 150 | 94 | 44 | 43 | 53 | 182 | 165 | 602 | 1814 | 10344 | **0.96*** |
| Calls without positive valence | 929 | 320 | 195 | 234 | 215 | 861 | 742 | 2567 | 5879 | 24196 | **0.92*** |
| Calls with negative valence | 883 | 299 | 180 | 184 | 193 | 747 | 556 | 2012 | 4819 | 21849 | **0.92*** |
| Calls without negative valence | 196 | 115 | 59 | 93 | 75 | 296 | 351 | 1157 | 2874 | 12691 | **0.95*** |
| Calls with anger | 184 | 17 | 17 | 11 | 10 | 39 | 26 | 51 | 115 | 422 | **0.69*** |
| Calls without anger | 895 | 397 | 222 | 266 | 258 | 1004 | 881 | 3118 | 7578 | 34118 | **0.94*** |

Eric Bolo, Muhammad Samoul, Nicolas Seichepine, Mohamed Chetouani. Quietly Angry, Loudly Happy: Self-Reported Customer Satisfaction Vs. Automatically Detected Emotion In Contact Center Calls. Interaction Studies, 2023, 24 (1), pp.168-192. 10.1075/is.22038.bol. hal-04109350

# Customer Satisfaction & Emotion

Table 7: CSAT response rate with respect to the customers' emotional profiles. * indicates a significant difference with the baseline ($p<0.05$)

| # of calls | With CSAT scoring | Without CSAT Scoring | CSAT response rate |
|---|---|---|---|
| All calls | 49 629 | 111 001 | 0.29 |
| Calls with upward positive valence dynamics | 11 896 | 24 577 | **0.32*** |
| Calls with downward positive valence dynamics | 302 | 521 | **0.36*** |
| Calls with upward negative valence dynamics | 15 489 | 40 017 | **0.27*** |
| Calls with downward negative valence dynamics | 5 854 | 9 621 | **0.37*** |
| Calls with upward anger dynamics | 437 | 2 155 | **0.16*** |
| Calls with downward anger dynamics | 82 | 230 | **0.38*** |

Eric Bolo, Muhammad Samoul, Nicolas Seichepine, Mohamed Chetouani. Quietly Angry, Loudly Happy: Self-Reported Customer Satisfaction Vs. Automatically Detected Emotion In Contact Center Calls. Interaction Studies, 2023, 24 (1), pp.168-192. 10.1075/is.22038.bol. hal-04109350

# Summary

- Multiple features
- Deep learning architectures for speech processing
- Interplay between speech and machine learning approaches

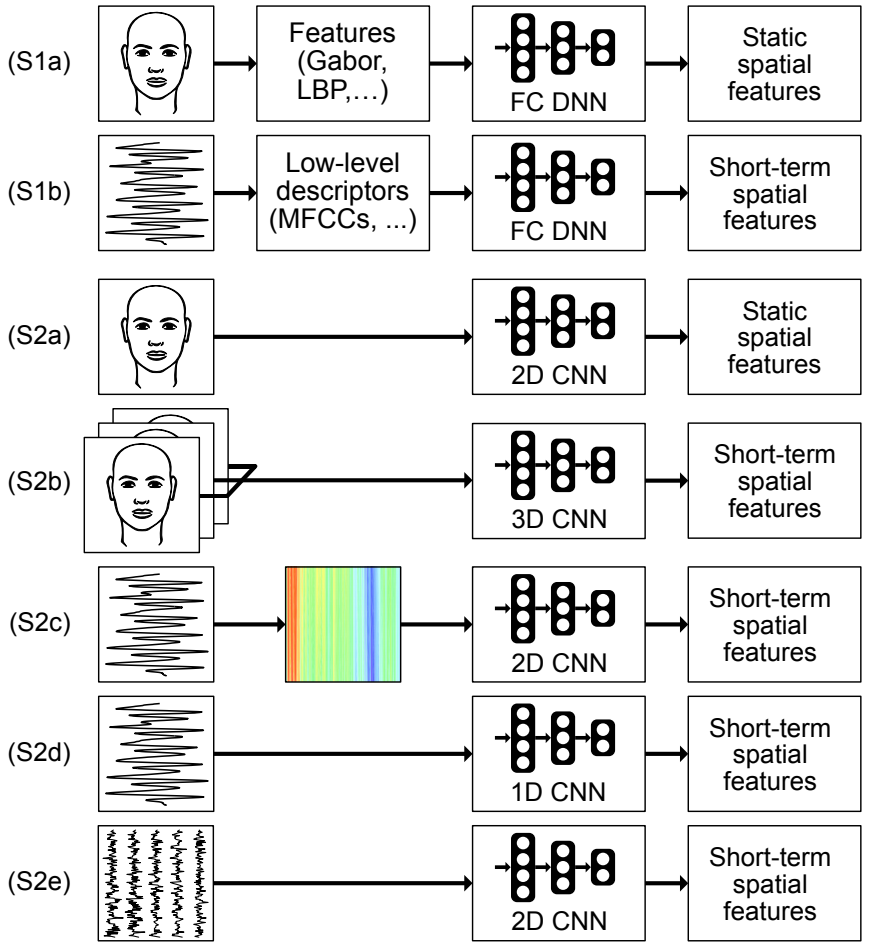



Fig. 2. Applications of deep learning for spatial feature learning with fully-connected DNNs (S1a–S1b) and CNNs (S2a–S2e).

# Next steps

## Practicals

Intention recognition from speech signals

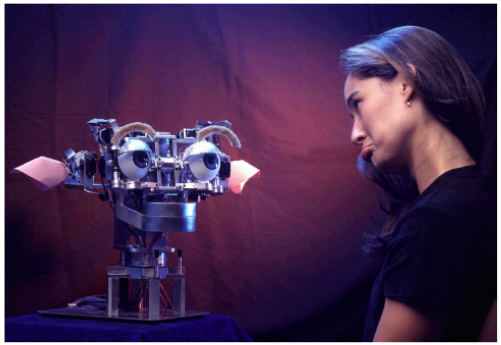


**Fig. 2.** Kismet is an expressive robotic creature designed for natural social interaction with people. See text.

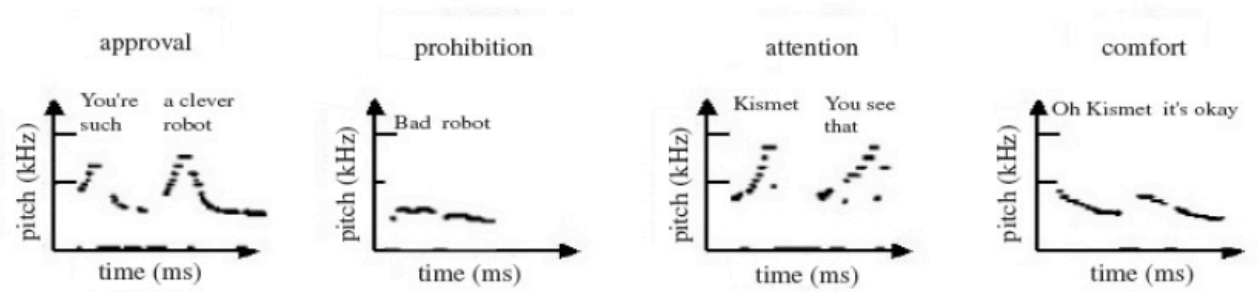



Fig. 2. Fernald's prototypical prosodic contours shown in robot directed speech for approval, attentional bid, prohibition, and soothing.

# Next steps

## Practicals

Intention recognition from speech signals

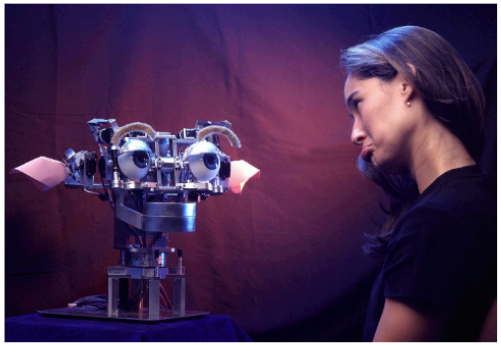


**Fig. 2.** Kismet is an expressive robotic creature designed for natural social interaction with people. See text.

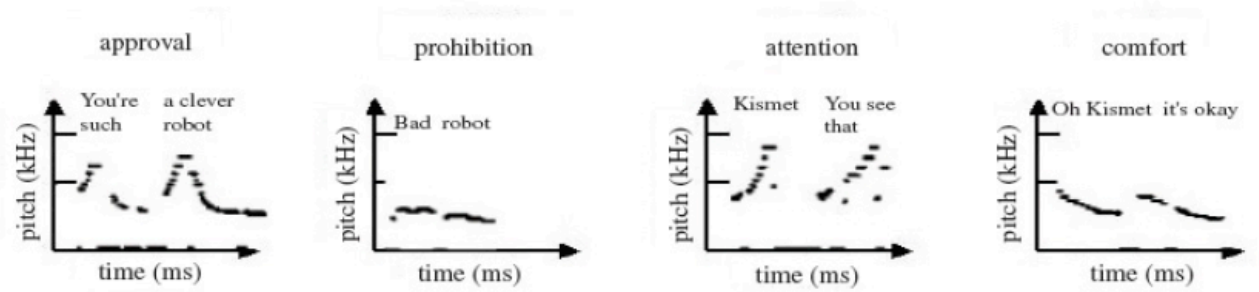



Fig. 2. Fernald's prototypical prosodic contours shown in robot directed speech for approval, attentional bid, prohibition, and soothing.

# Thank you for your attention

Questions?