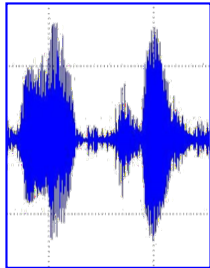




Institut
Mines-Télécom



Social signal processing and speech emotion recognition

Chloé Clavel



Analyze, recognize





Introduction

Affective computing and human-agent interaction

« *Affective Computing* » what is at stake?

■ Born in 1997

■ Book

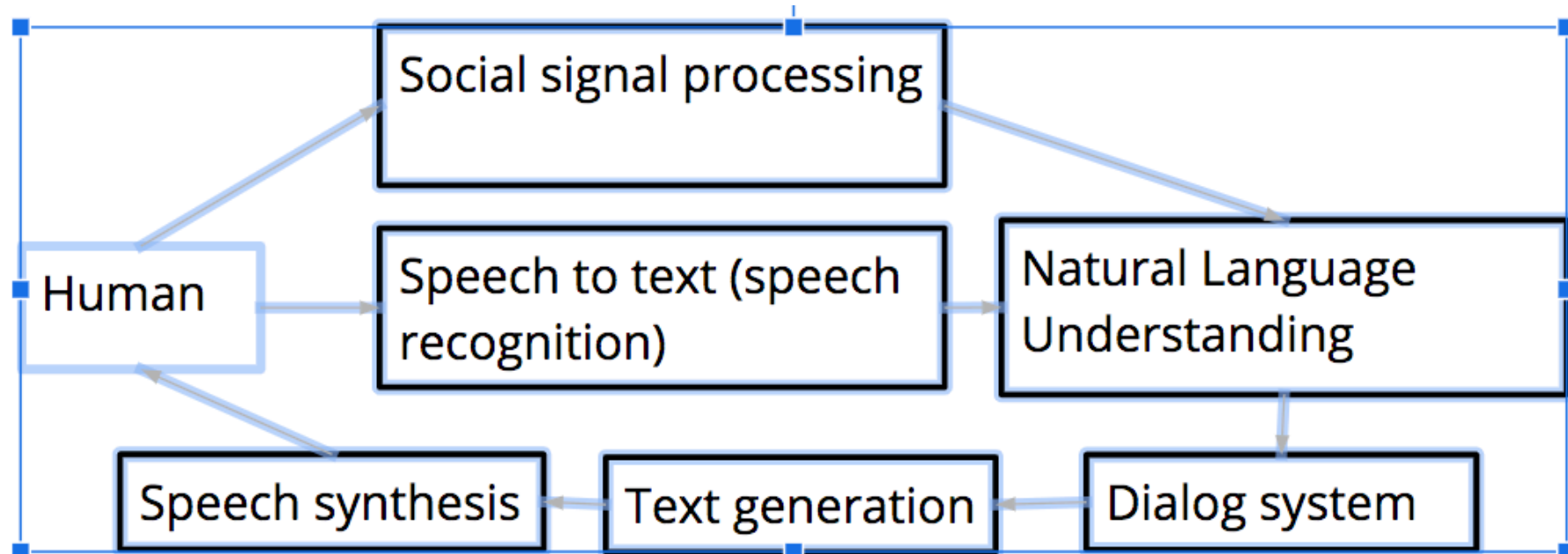
- « *Affective Computing* », R. Picard, 1997, MIT Press

■ Challenge : more natural human-agent interactions by integrating socio-emotional behaviors



Image From
http://sunway.edu.my/university/research_aat/AAT/groups/Intelligent

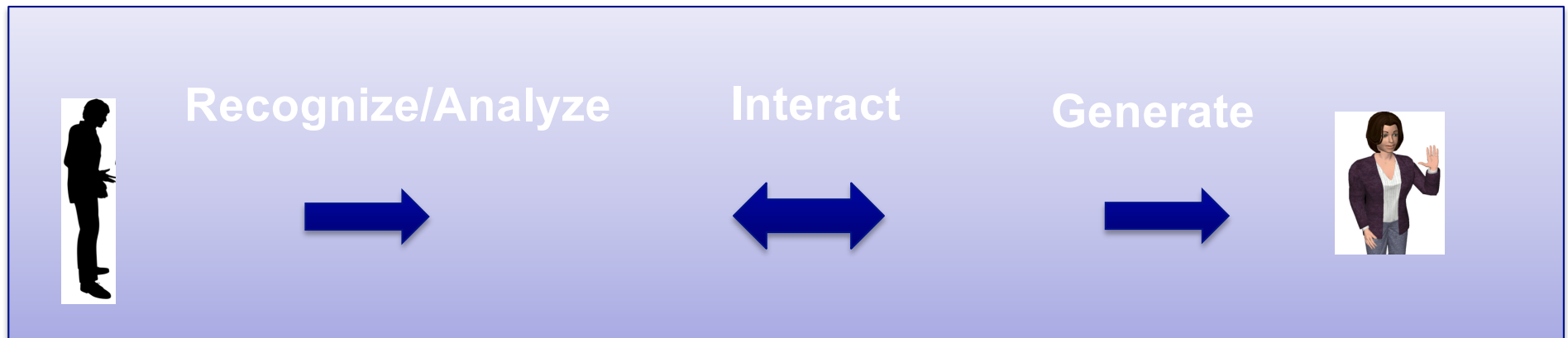
Speech-based dialog modules



Affective Computing and multimodal dialog

■ Give the computer (agent/robot) the ability to

- Recognize emotions and social behaviors
- process/interact,
- generate emotions and social behaviors





Social signal processing

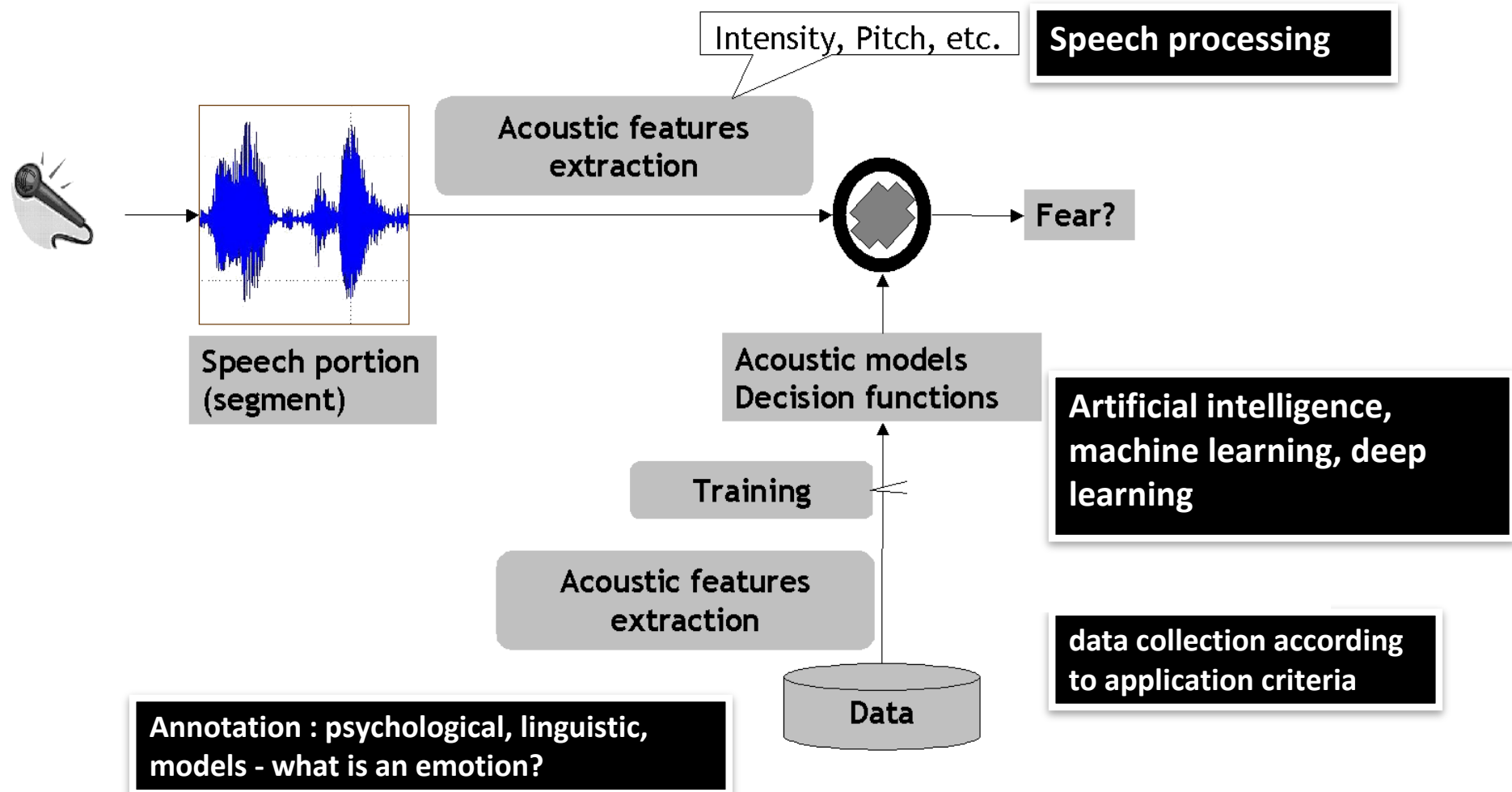
- **An evolution of Affective Computing domain**
- **Towards the analysis of a larger set of socio-emotional behaviors**
 - social stances, personality ,etc.



Socio-emotional behaviors

- Scherer's definitions [Scherer, 2005]
 - Emotion: short phenomenon, physiological reaction, appraisal of a major event (stimulus)
 - Mood: diffuse non-caused low-intensity long-duration change in subjective feeling
 - Interpersonal stances: affective stance toward another person in a specific interaction
 - Attitudes: enduring, affectively colored beliefs, dispositions towards objects or persons
 - Personality traits: stable personality dispositions and typical behavior tendencies
- PRACTICE : link the following terms to the most relevant phenomenon
 - liking, gloomy, contemptuous, jealous, sad

How does it work?

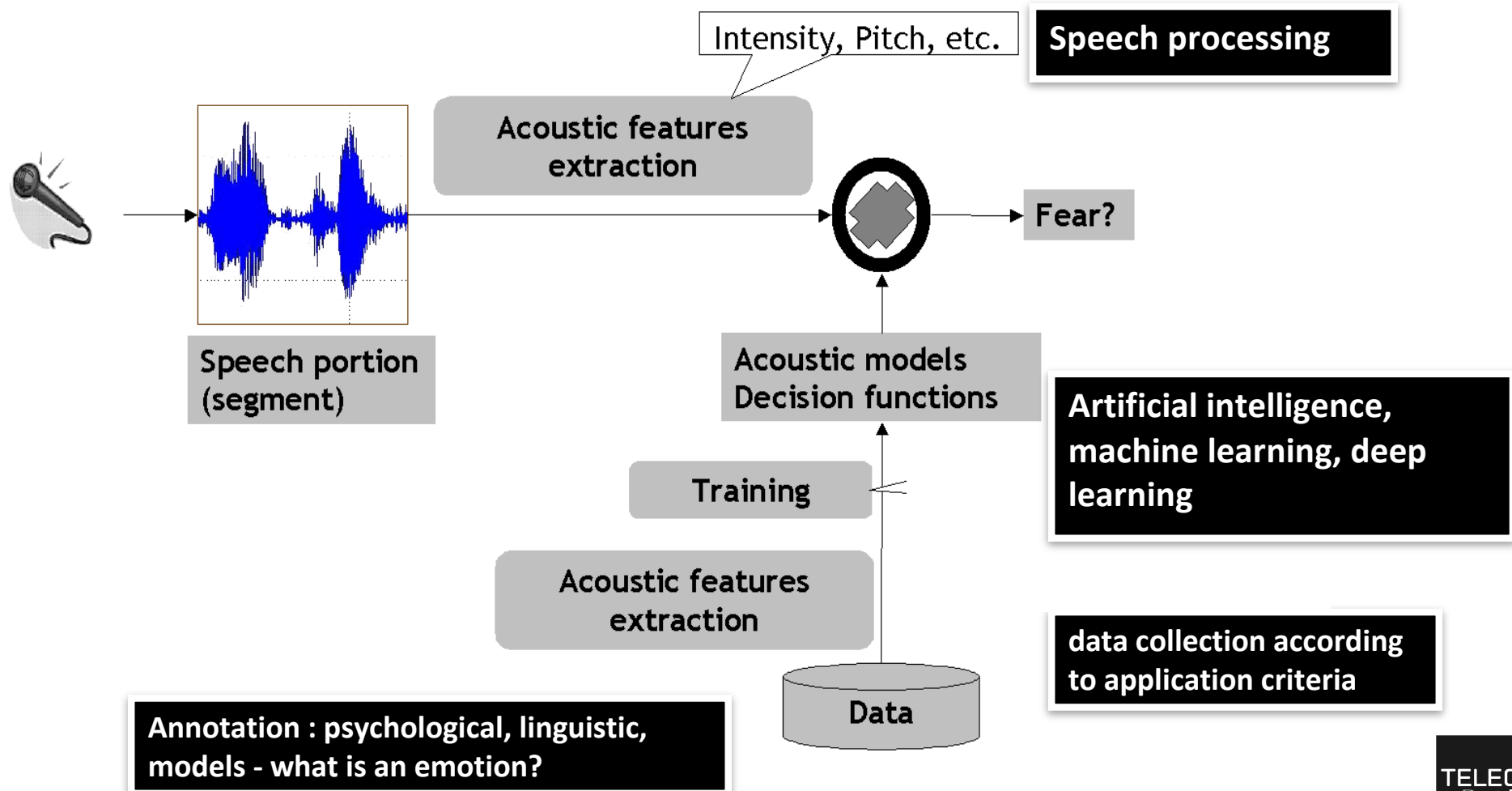


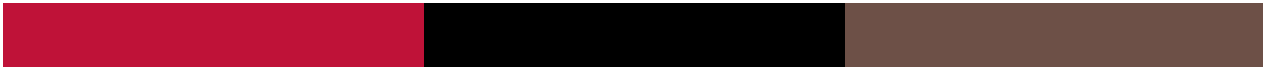
Scientific context – evolution of a technology

- 1996: First audio-based system
 - developed on acted database collected in laboratory [Dellaert et Polzin, 1996]
- 1997:
 - Identification of potential applications
 - definition of Affective Computing research domain [Picard, 1997]
- Now :
 - Deep learning approaches for multimodal recognition, End-to-end approaches [Trigeorgis et al. 2016], [Ghosh et al. 2016]
 - Industrial solutions
 - In call centers : ex. <http://www.nice.com/speech-analytics>

Steps/Outline of the lecture

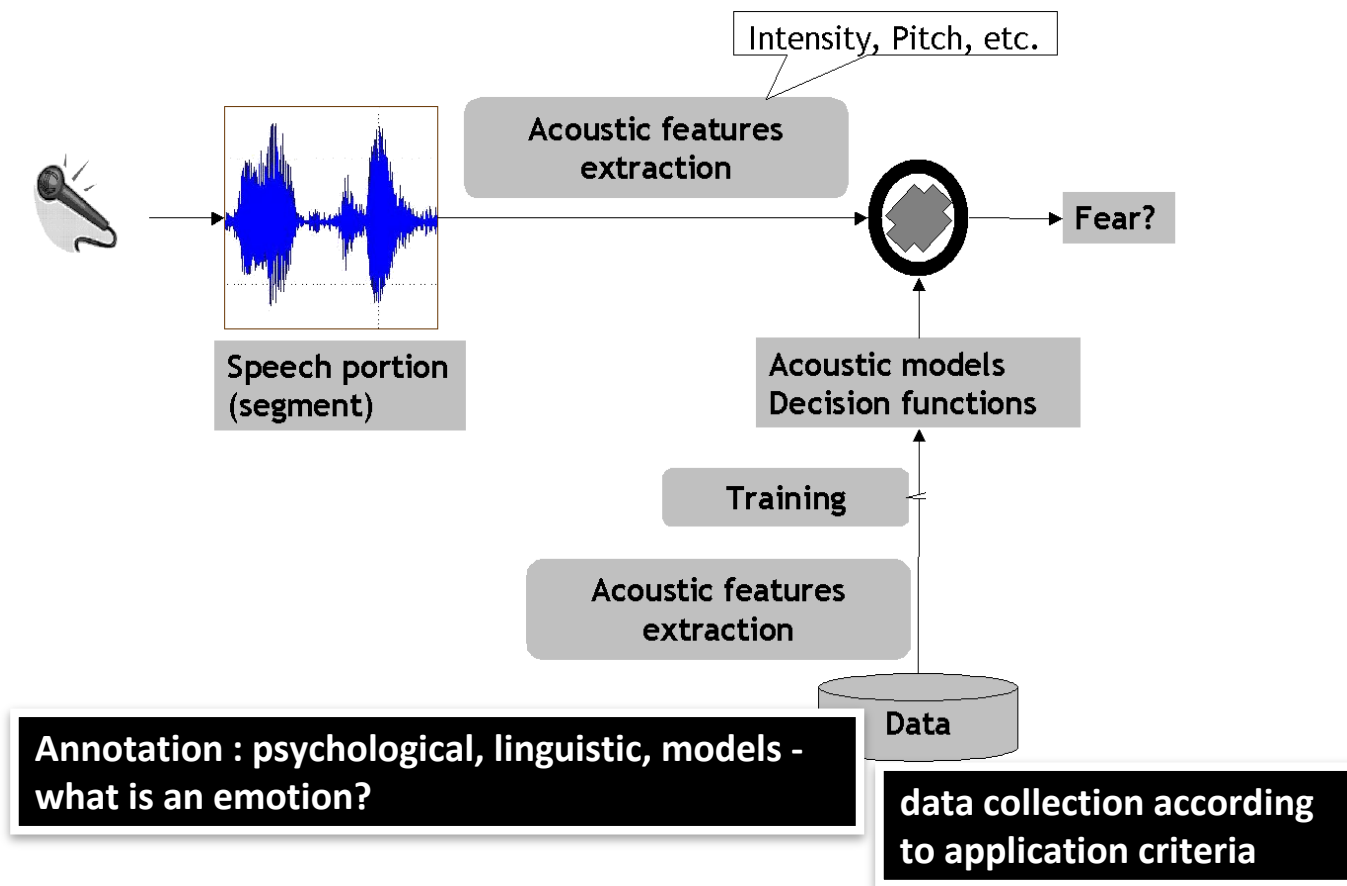
1. Collection and annotation of an emotional database
2. Acoustic representation of the data
3. Machine/deep learning





Collection and annotation of an emotional corpus

Acquisition et annotation of an emotional corpus





Challenges for emotional corpora

■ Quality of the learned models

Depends on :

- the sufficient quantity of the data
- the emotional content illustrated in the data
- the quality of the emotion labels

Emotional content in data: difficulties

- Complex phenomena:
 - Unpredictable: for a same event, variants of emotional reactions according to the personalities:

Their First Murder, 1941 © Weegee



- Mixed emotions: The study of the spontaneous voice shows that emotions are often mixed or hidden in natural interactions (ex: fear and anger)



How to evaluate the quality of the emotional content in data?

■ Criteria:

- How close is the emotional content
 - to the targeted application ?
 - to real-life contexts?
- How large is the variety of emotional manifestations ?



+/- close to real-life context

1/ acted corpora :

Emotions played by actors with or without context

Sometimes not realistic manifestations

- : far from real-life context

+ : control the type emotions illustrated in the corpus



+/- close to real-life context

2/ elicited corpora: the speaker is put in situations likely to trigger emotions

— Wizard of Oz protocol :

- the speaker believes that she interacts with the computer while a human is actually behind the (mis)functioning of the computer

+: closer to real-life context

+ : control the type emotions illustrated in the corpus



+/- close to real-life context

3/ « real-life » or spontaneous corpora

- Corpus recorded in daily interactions (e.g. call-centers)
- sometimes difficult, ex: fear

++ : close to real-life context

- : control the type emotions illustrated in the corpus



Annotation of emotional content

How to provide robust and reliable emotional annotations for automatic emotion recognition systems?

■ **Difficulty:**

- Subjective phenomena
 - We don't have the same perception of an emotion expressed by our interlocutor

■ **Recommendation:**

- Ask several annotators
- Provide them with an annotation guide
- Measure their agreement



Annotation guide

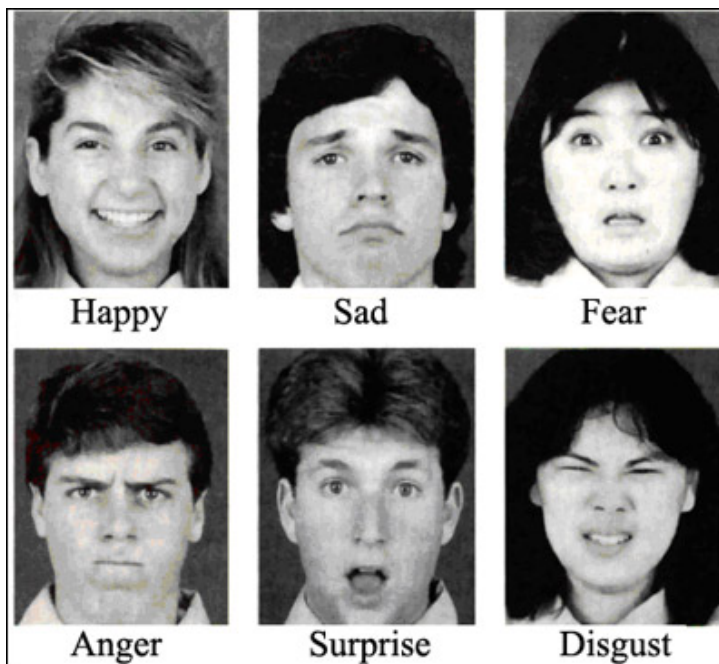
- emotion categories or dimensions
- context to annotate
- continuous vs. segmental annotations
- units of annotation

Choice of emotion categories or dimensions

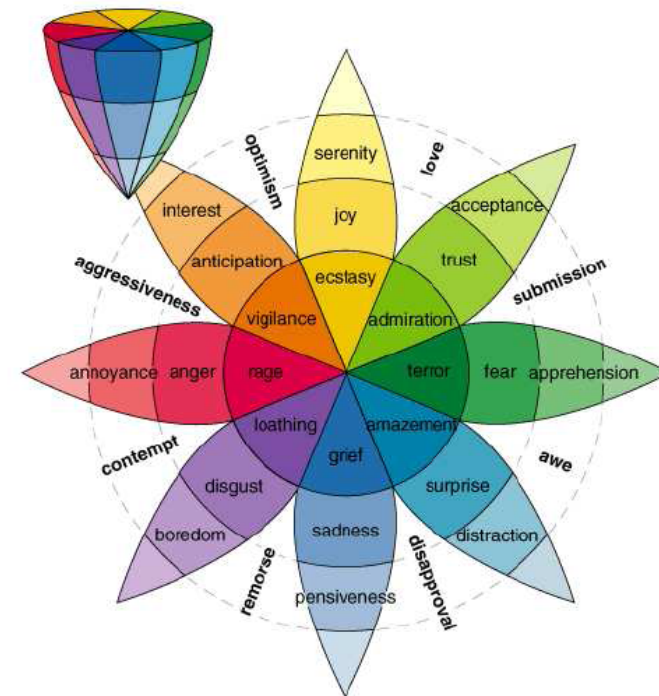
■ From psychological models

— Categories : denomination of emotions by lexical items

- Ex: the Big-six (fear, anger, sadness, joy, disgust, surprise) [Ekman, 1999]
- More and finer emotional categories (ex Plutchik wheel [Plutchik, 1984])



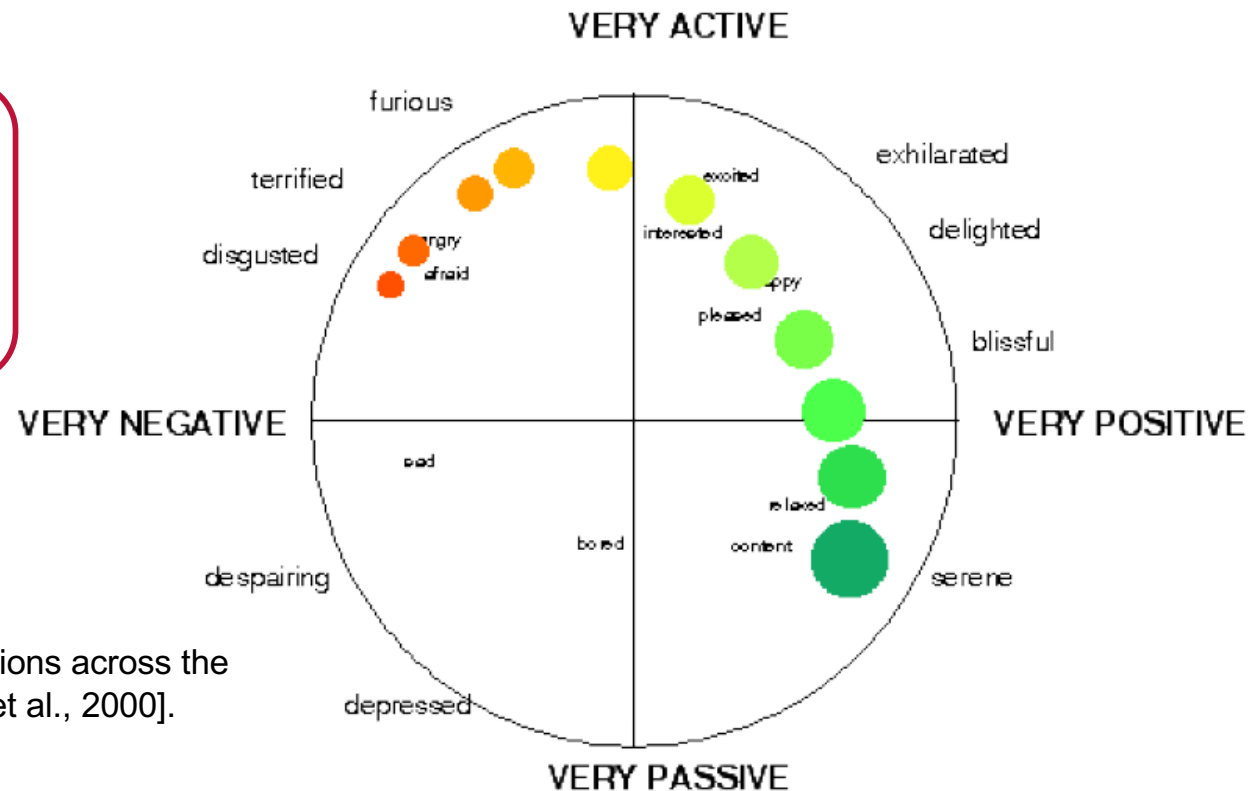
© Ekman



Choice of emotion categories or dimensions

- Abstract Dimensions to represent emotions [Osgood et Mai, 1975]
 - Valence ou evaluation : positive/negative axis (anger vs. Joy)
 - Activation
 - Control/Dominance : speaker's effort to control her emotion

Continuous vs. discrete annotation of emotion dimensions



Annotation of emotional variations across the time using Feeltrace [Cowie et al., 2000].



Annotation of the emotional context

■ Emotional context:

- Situation
 - ex: topic of the conversation, events triggering the emotions, etc.
- Speakers (gender, name, age)
- Other modalities (verbal, facial cues, etc.)
 - Transcription of verbal (text) and paralinguistic contents (laughs, breathings)
 - norm : LDC (Linguistic Data Consortium)
- Sound environment:
 - Speech Quality/Intelligibility



Choice of annotation unit

- Word, chunk, sentence,
- Interpausal unit
- speaker turn,
- conversation,
- etc.



Emotion annotation Standards

■ EARL Emotion Annotation and Representation Language

- defined in W3C World Wide Web Consortium
<https://www.w3.org/TR/emotionml/>



Annotation tools

- Dimensional annotation of emotions:
 - Feeltrace
- Multimodal annotation:
 - Elan
- Audio annotation
 - Praat
 - Transcriber



How to ensure the reliability of subjective annotations?

- **Opinion/Emotion phenomenon = subjective phenomenon**
 - Annotated by multiple annotators
 - Assess the degree of reliability of the annotations

Measure the reliability of annotations

■ Measures

- Cohen's kappa [Carletta, 1996]:
 - agreement corrected for what it would be under the mere fact of chance

$$\kappa = \frac{\bar{p}_o - \bar{p}_e}{1 - \bar{p}_e}$$

- P_o is the proportion of agreement observed and P_e the probability that the annotators agree by chance

Mesure de la fiabilité des annotations

PRACTICE

$$\kappa = \frac{\bar{p}_o - \bar{p}_e}{1 - \bar{p}_e}$$

■ Kappa values ?

- When annotators agree as much as chance
- When the annotators agree totally

■ Exercice

- 50 audio sequences annotated by 2 people (Ann1 / Ann2) in 2 categories positive / negative
- Calculating kappa between the two annotators

Ann1\Ann2	Positive	Negative
Positive	20	5
Negative	10	15

Measure the reliability of annotations $\kappa = \frac{\bar{p}_o - \bar{p}_e}{1 - \bar{p}_e}$

■ $P_o = (20+15)/50 = 0,7$

■ **Calculate P_e :**

- Ann1 uses positive label 50% of the time
- Ann2 uses positive label 60% of the time
- Probability that Ann1 and Ann2 use the positive label:
 $0.5 \times 0.6 = 0.3$
- Probability that Ann1 and Ann2 use the negative label :
 $0.5 \times 0.4 = 0.2$
- Probability to agree by chance : $0.2 + 0.3 = 0.5$

■ **Kappa computation:**

- $\text{Kappa} = 0.2 / 0.5 = 0.4$

Measure the reliability of annotations

- Moderate agreement = standard for emotions [Landis et Koch, 1977]

Accord	Kappa
Excellent	$\geq 0,81$
Bon	0,80-0,61
Modéré	0,60-0,41
Médiocre	0,40-0,21
Mauvais	0,20-0
Très mauvais	< 0

TAB. 4.4 – Degré d'accord en fonction des valeurs de Kappa

- Other measure: Cronbach's Alpha [Cronbach, 1951] for dimensions



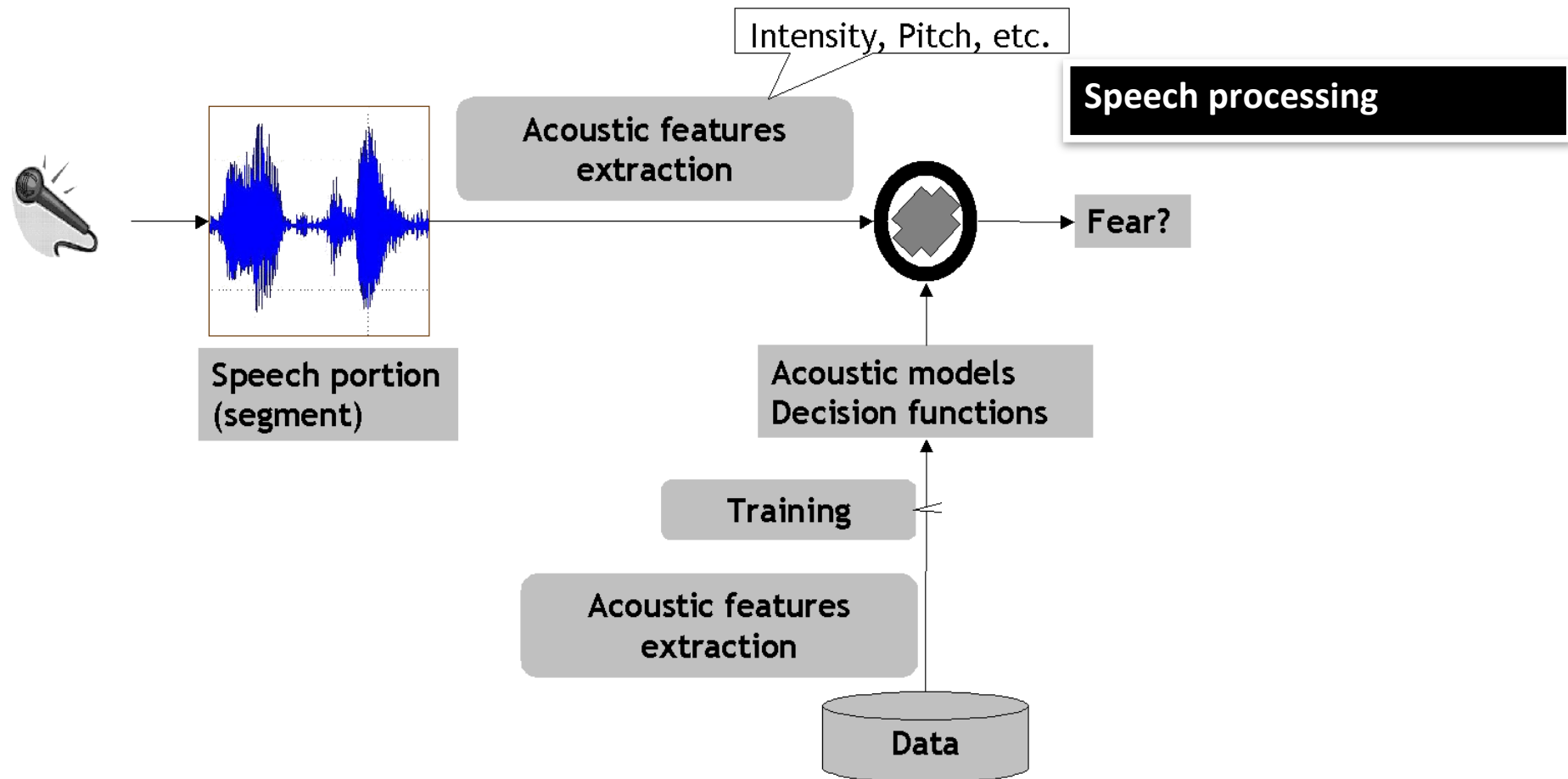
Take home message

- Different types of emotional corpora (real-life, Woz or acted)
- Annotating emotion : a difficult task
- Build annotation schema and measure annotators' agreement



Extraction of acoustic features of emotions

Extraction of acoustic features



Reminder (lecture speech recognition)

■ Processus of speech production

1. Generation of ventilatory energy
2. Vibration of vocal chords (voiced sounds) and or apparition of explosive and friction noises
3. Realisation of an articulatory mechanism at the level of supra-glottic cavities (vocal tract and nasal cavity)

=> Source-filter model [Fant, 1960]



Speech physical model and emotion signal

- In the case of emotions, model :
 - the modifications of acoustic signals
 - triggered by physiological modifications of the glottis provoked by emotions

Emotions

- ⇒ Physiological modifications of the glottis
- ⇒ Modifications of the speech production
- ⇒ Modifications of the acoustic signal



Speech physical model and emotion signal

■ Appraisal theory [Scherer, 2003]

- Emotion results from:
 - Various evaluations of an external stimulus.
 - Physiological modification provoked by this evaluation

Reaction to an unpleasant event
⇒ Tension of the vocal tract
⇒ more energy in high frequencies



Appraisal theory of Scherer

■ Physiological modifications => signal modifications

- Fear : pulse augmentation, dryness of the mouth,
=> Stronger and higher voice, high speech rate
- Boredom, sadness: decrease of cardiac rythm
=> Lower and deeper voice, slower speech rate



Acoustic features of emotions

■ Types of features :

- Prosodic (fundamental frequency, intensity, speech rate, see previous lecture)
- Voice quality
- Spectral and cepstral

Voice quality features

■ Jitter

- Deviation of the fundamental frequency
- Formula (T_n period at time n)

$$Jitter = \frac{\sum_{n=2}^{N-1} |2T_n - T_{n-1} - T_{n+1}|}{\sum_{i=1}^N T_n}$$

- Characterize creaky voice

~Vibrato

Voice quality features

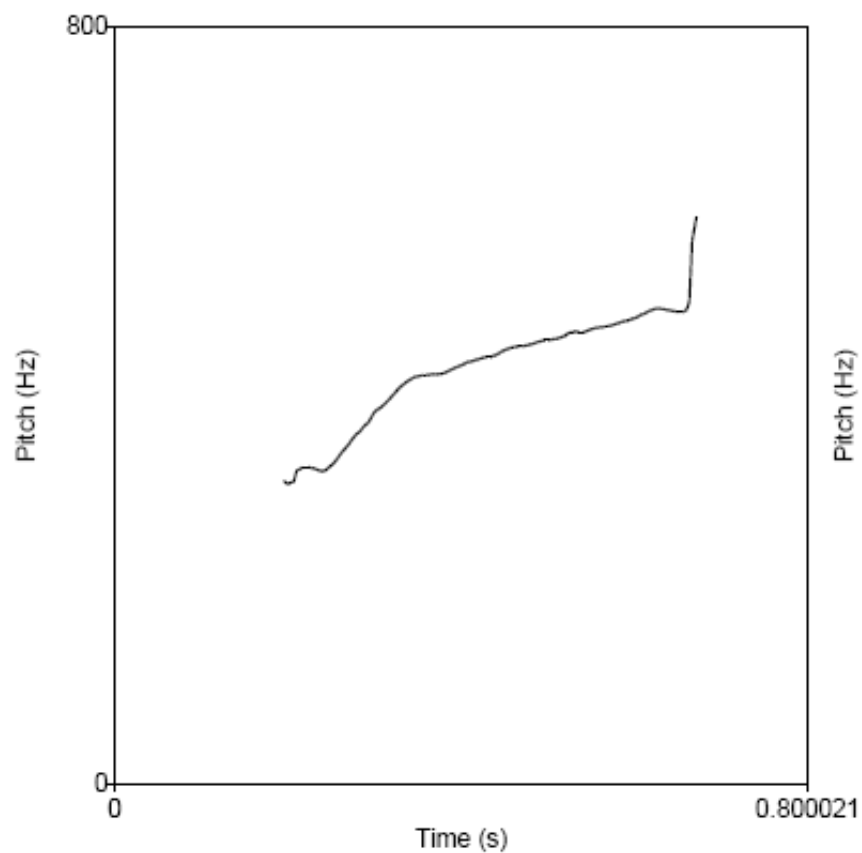
■ Shimmer

- Amplitude modulation A_n during period T_n

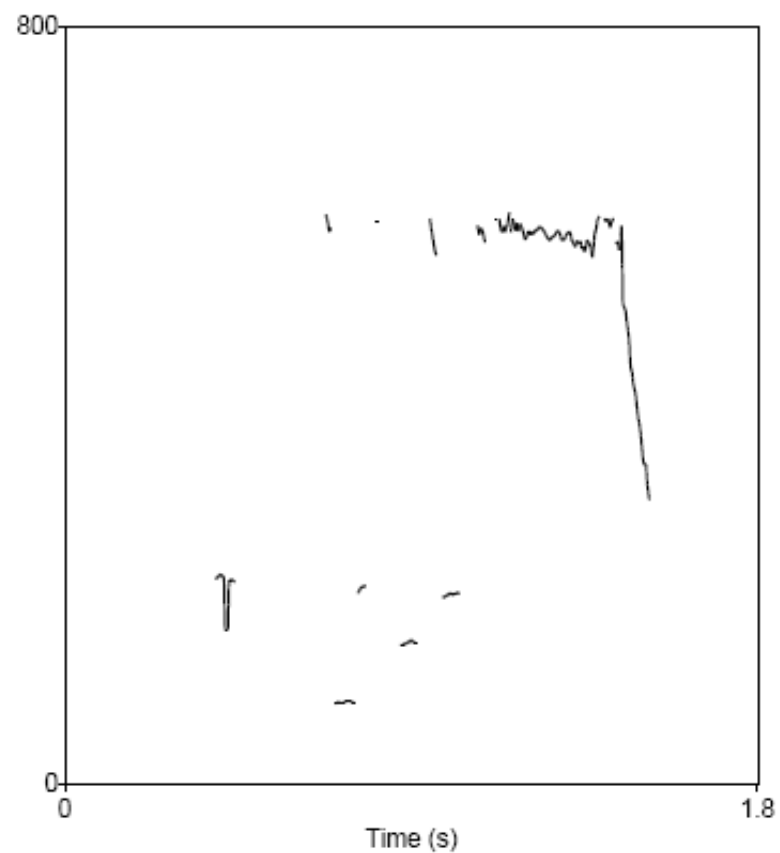
$$Shimmer = \frac{\sum_{n=2}^{N-1} |2A_n - A_{n-1} - A_{n+1}|}{\sum_{i=1}^N A_n}$$

~Tremolo

Fundamental frequency and jitter



"Josh?"



"Jooosh!"





Voice quality features

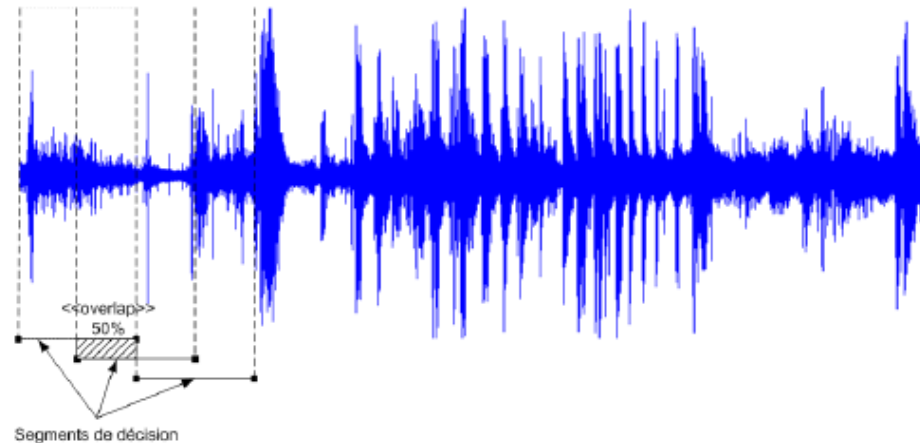
■ HNR : harmonic to noise ratio

- Indicates the level of breathing in the voice
[Yegnanarayana et al., 1998]

Acoustic features of emotions

■ Temporal unit of the analysis

- Windows 20ms-40ms with overlap
- Statistics (mean, variance, and standard deviation) on syllabs, words, sentences and speaker turns





Acoustic features of emotions

■ Recent approaches:

- Learning representations through auto-encodeurs
 - Ghosh, S., E. Laksana, L.-P. Morency et S. Scherer. 2016a, «Learning Representations of Affect from Speech», Iclr 2016
 - Freitag, M., S. Amiriparian, S. Pugachevskiy, N. Cummins et B. Schuller. 2017, «auDeep : Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks»,



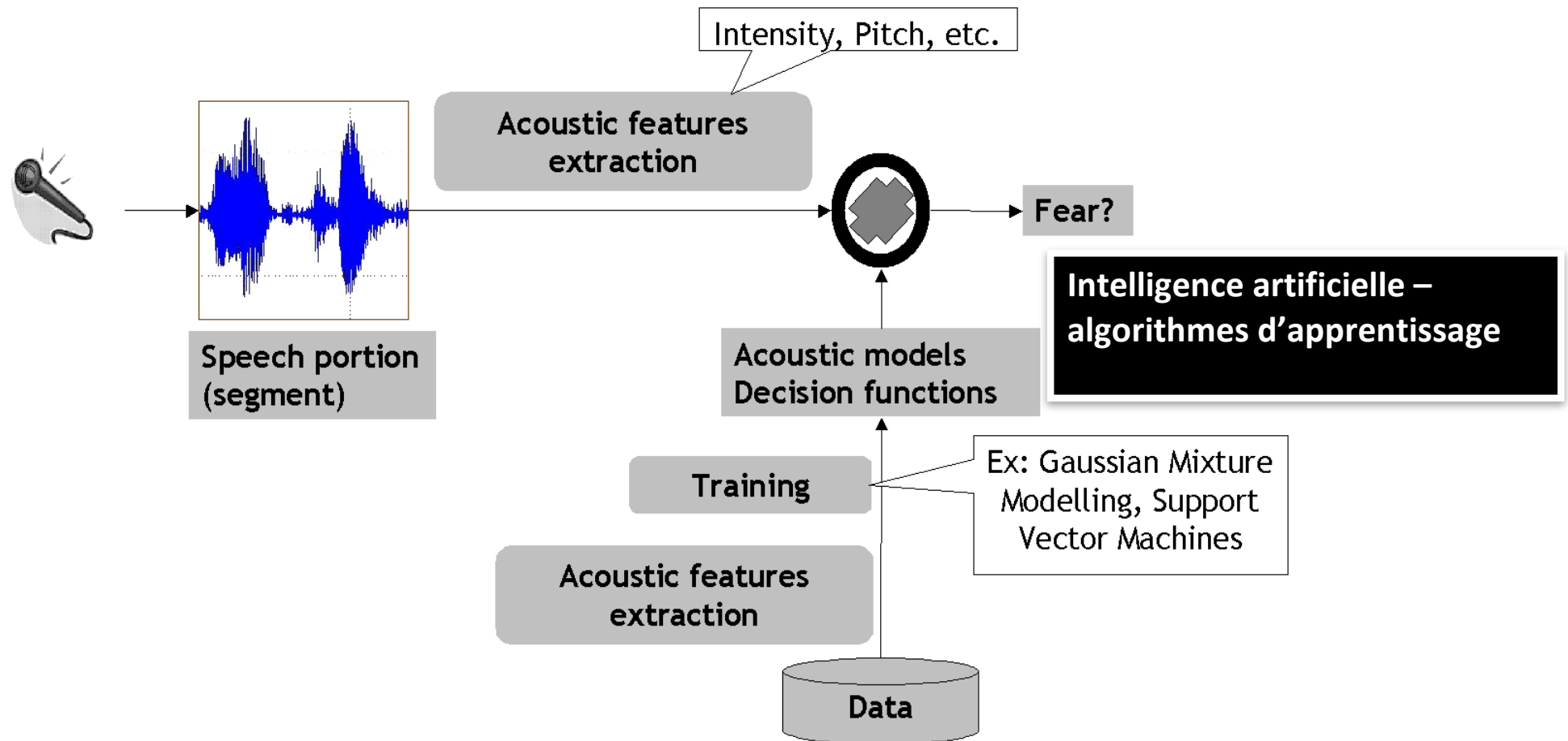
Tools for acoustic features extraction

- **Praat** www.praat.org
- **Matlab, toolbox voice processing**
<http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>
- **Opensmile**
- **COVAREP (matlab)**
- **librairie python pour l'audio**
<https://librosa.github.io/librosa/>



Machine/Learning/Deep Learning

Machine learning





Normalization and feature space reduction

■ Preliminary steps

■ Motivation :

- Normalize feature range between 0 and 1
 - (from one feature to another)
 - From one speaker to another (ex: fundamental frequency)
 - From one recording to another



Normalization and feature space reduction

■ Technics for the normalization by gender/speaker/phonemes

- min-max
- sigma-mu



Feature space reduction

■ Motivation

- Reduce complexity
- Avoid a performance decrease



Feature space reduction

■ Classical Technics [Duda, 1973]

- Projection of the data representation space onto a smaller dimensional space
 - Principal component analysis,
 - discriminant analysis,
 - representation learning



Examples of classification methods

■ Used for speech-based emotion recognition

- Logistic regression
- Gaussian Mixtures Models
- Recent approaches:
 - End-to-end machine learning : from signal to emotion (without using any descriptors)
 - Trigeorgis, G., F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller et S. Zafeiriou. 2016, «Adieu features ? End-to-end speech emotion recognition using a deep convolutional recurrent network»
 - Mix CNN (to learn local structure in speech signal) with RNN (LSTM, to learn temporal dependencies)



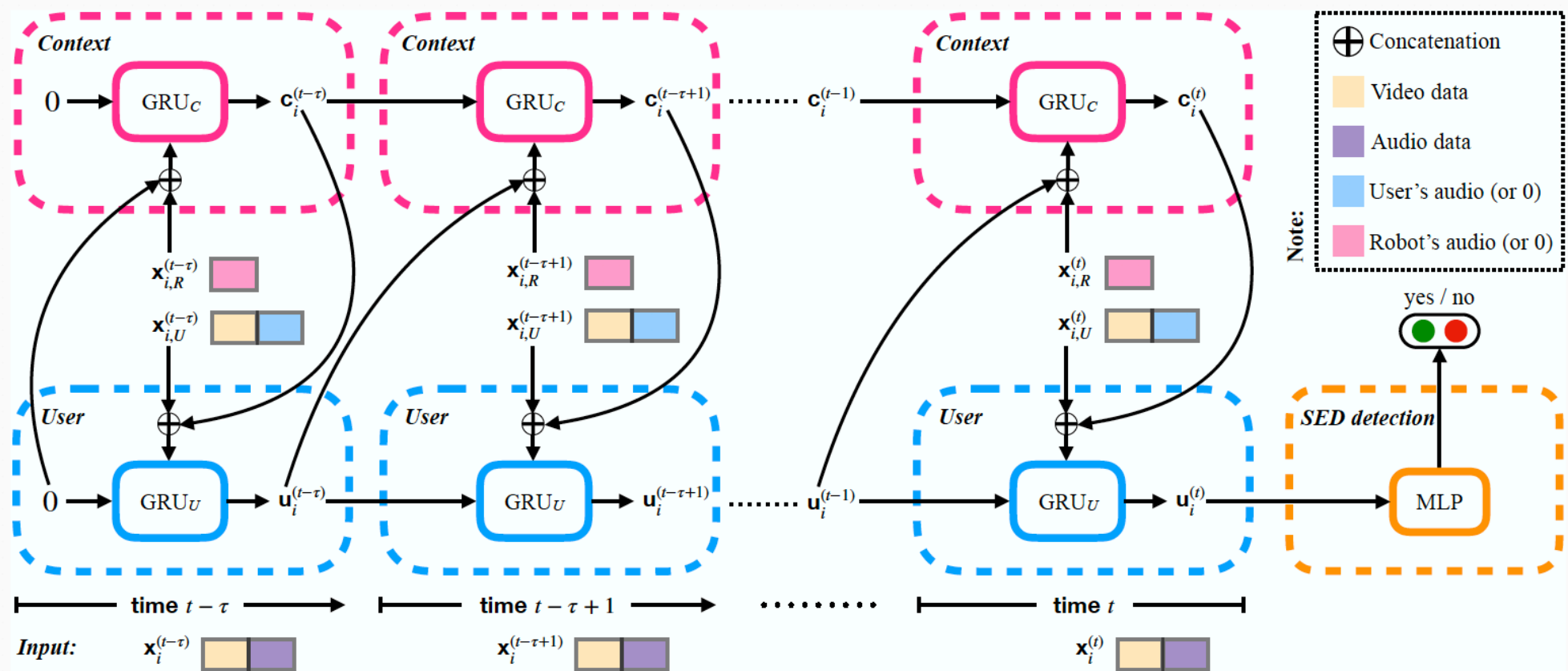
Recurrent neural networks for emotion recognition

■ Recurrent networks:

- Good to model the sequential nature of audio
- Can be extended to model the conversation sequentiality
 - Hazarika, Devamanyu, et al. "Conversational memory network for emotion recognition in dyadic dialogue videos." *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. Vol. 2018. NIH Public Access, 2018.
 - gated recurrent units to model past utterances of each speaker into memories.
 - memories are then merged using attention-based hops to capture inter-speaker dependencies.

Recurrent approaches to integrate interaction context in human-robot interaction

- Use of the robot data as contextual information to help assess user engagement in a human-robot interaction
 - Atamna, A. and Clavel, C.,. HRI-RNN: A User-Robot Dynamics-Oriented RNN for Engagement Decrease Detection. In INTERSPEECH 2020





Subjectivity on labels in machine learning models

- **Rizos, G. and Schuller, B.W., 2020, June. Average Jane, Where Art Thou?—Recent Avenues in Efficient Machine Learning Under Subjectivity Uncertainty. In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 42-55). Springer, Cham.**



Emotion transfer learning

- <https://www.sciencedirect.com/science/article/pii/S1566253520303018>



Performances and evaluation

■ Performances depend on :

- the test corpus (actées vs. "real-life", diversity of data)
- The number of classes (max 4-5 classes processed)
- Emotional classes considered (fear/anger more subtle than fear/neutral)

■ Human vs. system performance confrontation



Performances and evaluation

■ Example:

- 94% on data recorded in the laboratory [Schuller et al, 2004].
- 55% on call center data for 5 classes [Devillers, 2007].
- 80% for two valence classes (negative/neutral) [Devillers, 2007].
- First comparison effort (HUMAN, AIBO data, [Batliner07])
- 1st Emotion Challenge Interspeech 2009
- Now : regular AVEC challenge



To go further

■ Conferences:

- ACII Affective Computing and Intelligent Interaction
- Interspeech
- ICASSP

References: speech emotion recognition

- [Batliner et al., 2006] Combining efforts for improving automatic classification of emotional user states. A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, et V. Aharonson. Dans Proc. Of Proc. IS-LTC, Ljubljana, 2006.
- [Breazeal, 2002] *Recognizing affective intent in robot directed speech*, Breazeal, C. & Aryananda, L. Autonomous Robots , 2002
- [Calliope, 1997] *La parole et son traitement automatique*. Calliope, Dunod, 1997.
- [Dellaert, 1996] *Recognizing Emotion in Speech*, Dellaert, F., Polzin, T. & Waibel, A., Proc. of ICSLP 1996
- [Fant, 1960] *Acoustic theory of speech production*, Gunnar Fant, Mouton, The Hague.
- [Plutchik, 1984] *A General Psychoevolutionary Theory*, R. Plutchik, Erlbaum, Hillsdale, NJ, Scherer, K.R. and Ekman, P. Eds, 1984
- [Scherer, 2003] *Vocal communication of emotion : a review of research paradigms*, K. Scherer. Speech Communication, 40(1-2) :227–256, 2003.

References: speech emotion recognition

- [Varadarajan 2006] *UT-SCOPE - A corpus for Speech under Cognitive/Physical task Stress and Emotion*, Varadarajan, V., Hansen, J. & Ayako, I., Proc. of LREC Workshop on Corpora for Research on Emotion and Affect
- [Schuller et al., 2004] *Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture*. B. Schuller, G. Rigoll, et M. Lang. Dans Proc. of ICASSP, Montreal, 2004.
- [Trigeorgis et al. 2016] Trigeorgis, G., F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller et S. Zafeiriou. 2016, «Adieu features ? End-to-end speech emotion recognition using a deep convolutional recurrent networks»
<https://www.dropbox.com/s/zjmbranh68g9sng/AdieuFeatures.pdf?dl=0>
- [Ghosh et al. 2016] Ghosh, S., E. Laksana, L.-P. Morency et S. Scherer. 2016a, «Learning Representations of Affect from Speech», Iclr 2016
<https://arxiv.org/pdf/1511.04747.pdf>



References: multimodal emotion recognition

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabas Póczos. Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities. 12 2018.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. 2019.