Affective computing on speech
–
Self-supervised representation learning of primate vocalisations : from analysis to synthesis

Jules Cauzinille

Annotation of expressive dimensions on a multimodal French corpus of political interviews

Master's internship

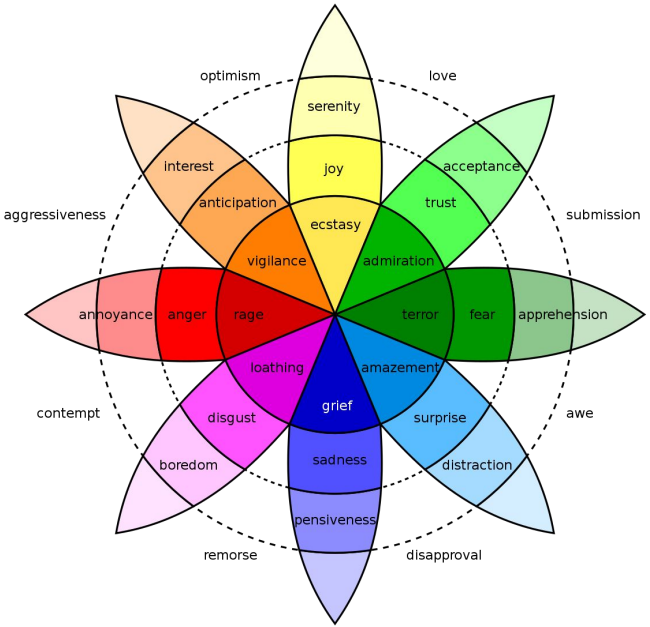
LISN
Marc Evrard, Albert Rillard

Introduction

- Automatically identify conflictual interactions in political interviews
- Analyse vocal expressivity
- 32 audiovisual interviews (7.5 h)
    - Bourdin Direct
    - Les 4 Vérités
- Multimodality :
    - Automatic transcription
    - Speech quality
    - Video

Expressivity models

- Complicated annotation task : low inter-annotator agreement
- Affective models:
  - **Categorical models**:
    Darwin (1872), Ekman (1969) 6 basic emotions, Plutchik (1980) 8 basic emotions, etc.
  - **Dimensional models**:
    Magda Arnold (1960) and James Russel (1980) circumplex model of affect



Plutchik (1980)'s 8 basic emotions

The Circumplex model

Arousal

Physiological activation, vocal excitement :
Calm (passive arousal) / Excited (active arousal)

Valence

Level of pleasure :
Negative / Positive

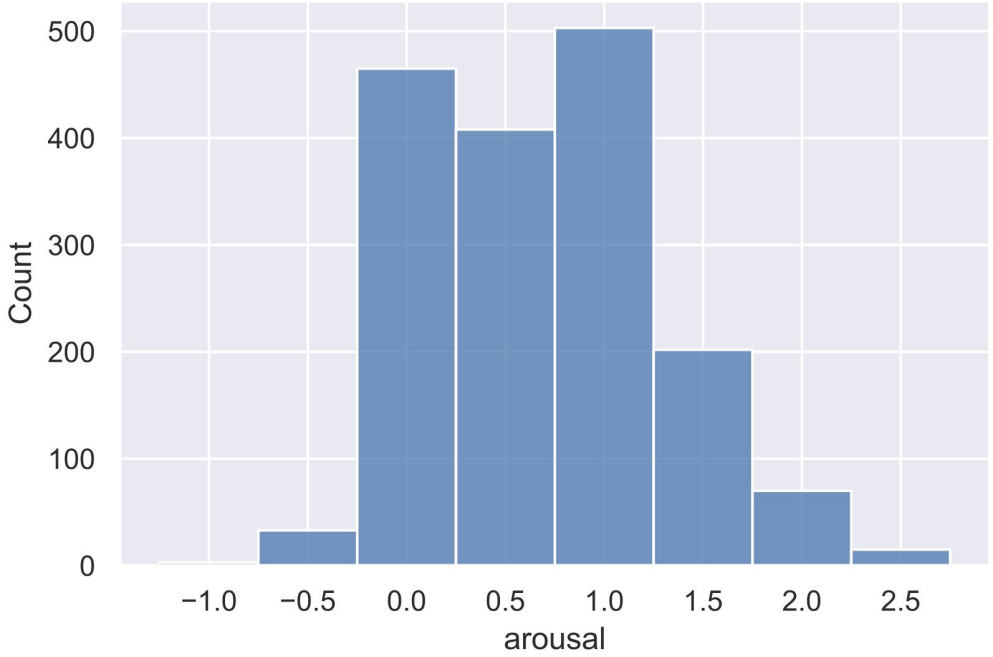
Geneva wheel of emotions (Scherer et al., 2013)

Vocal expressivity

- Mostly prosodical :
  - F0, intensity, speech rate, vocal quality
- Frequency code (Ohala, 1984)
  - Acoustic projection of physical force
- Effort code (Gussenhoven, 2004)
  - Articulatory effort
  - F0 variation

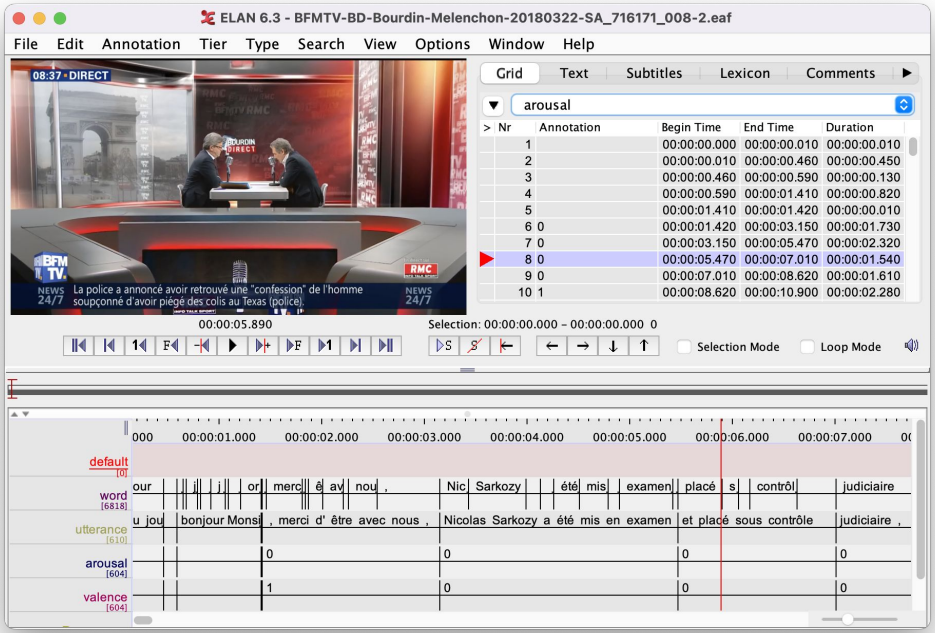
Vocal expressivity in a broadcast political context

- Cold anger display (average arousal)
- Minimal hot anger (Fonagy, 1976)
- Arousal histogram (2 annotators / 1h corpus)



The segmented approach

- 3 criteria:
  - Expressive variation
  - Semantic unit (clause)
  - 3 seconds threshold

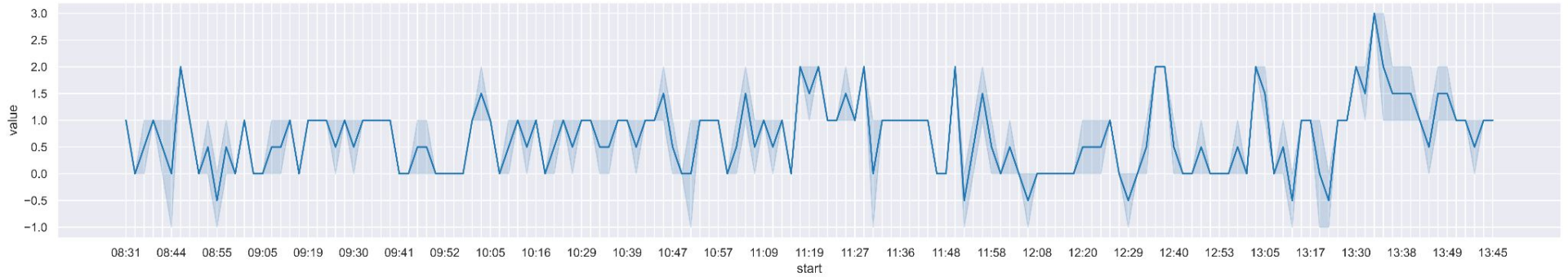


Discrete levels of arousal

- 7 levels likert scale (Joshi et al., 2015)
- Highly frequent neutral arousal
- Most bursts go higher rather than lower
- 7 levels allow for more generalization of the framework

Inter-annotator agreement

- First publication : 1h annotation (12 interviews - 5 min each)
    - 2 annotators
- Quadratic-weighted Kappa (Artstein and Poesio, 2008) gives a “moderated” agreement $\kappa_w = 0.546$



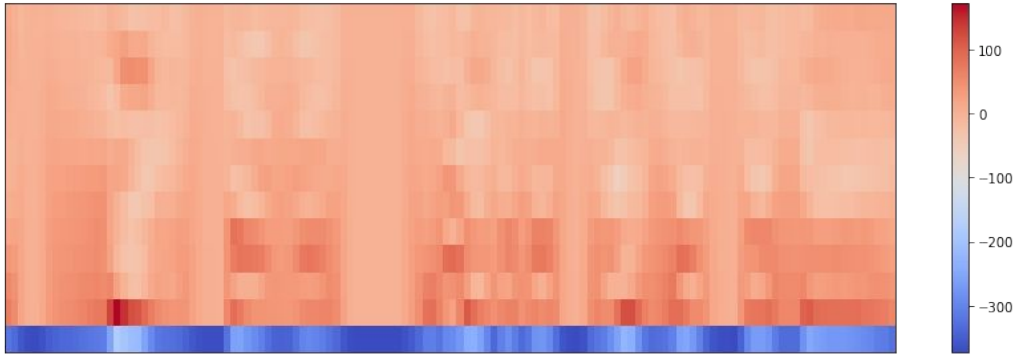
“MFCC-based model”

- Extraction
  - 25 ms frames
  - 13 MFCCs per frame
  - Trimmed and padded segments 3 s (120 frames)

- Architecture and hyperparameters inspired by Zhao et al. (2019)

  - 2 conv layers : kernel (3 × 3) with ReLU activation
  - 64 et 128 filters for each layer
  - Two max-pooling layers (2 × 2)
  - 3 fully connected linear layers
  - 30% dropout

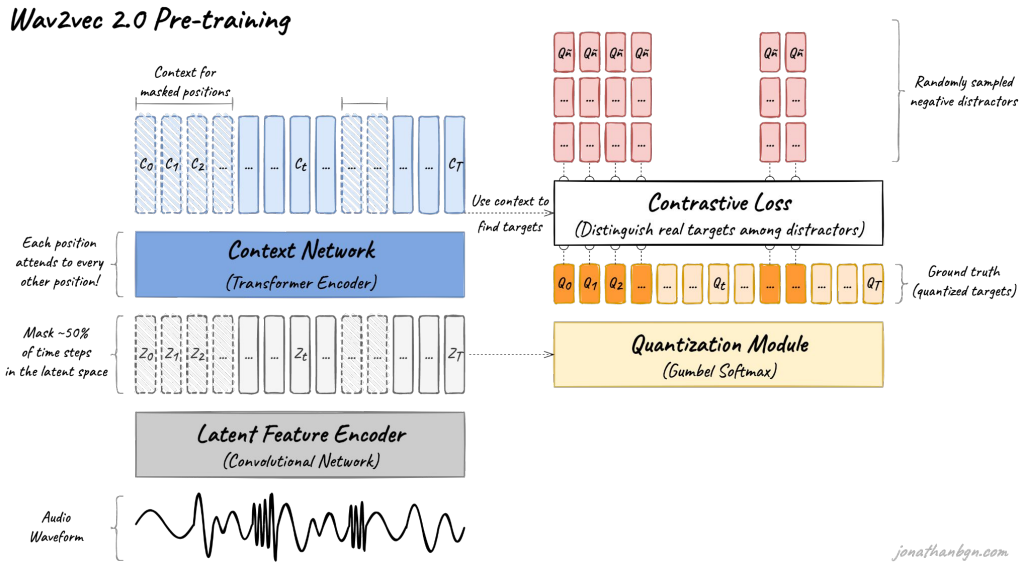
Example of MFCCs representations



“Wav2vec-based model”

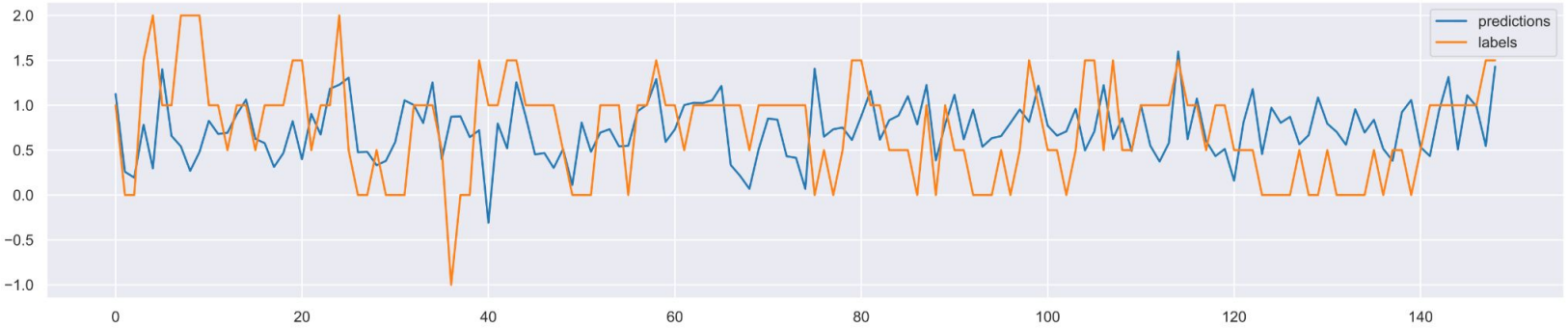
- wav2vec2.0 *facebook/wav2vec2-large-xlsr-53* feature extraction (Conneau et al., 2020, Evain et al., 2021)

- Best model: GRU
  - One hidden layer of size 128
  - Sigmoid activation
  - 10% dropout



Results

Model	RMSE	MSE	MAE
MFCC & CNN	0.555 (+/-0.064)	0.322 (+/-0.081)	0.464 (+/-0.053)
Wav2vec & GRU	0.577 (+/-0.062)	0.336 (+/-0.073)	0.461 (+/-0.051)



Self-supervised representation learning of primate vocalizations, from *analysis* to *synthesis*

P.h.D project - ILCB

Jules Cauzinille

Multidisciplinarity

1. Speech processing - Computational linguistics

2. Bioacoustics - Primatology - Origins of language

3. Signal processing - Deep learning - Self-supervised approaches

BENOÎT FAVRE

RICARD MARXER

THIERRY LEGOUT

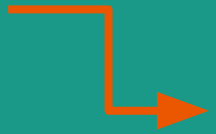
ARNAUD REY

Université Paris Cité

Context

- Success of **self-supervised** representation learning in speech processing
  - wavenet, GSLM, ZRSC

- Deep Learning in **bioacoustics:** increasing research and impressive implications
  - Stowell [2022]



- Training and probing models with animal vocalisations
  - Probing tasks
  - Synthesis

Objectives

- **Human speech bias:**
  - acoustic units
  - vocabulary size
  - overlapping and noise

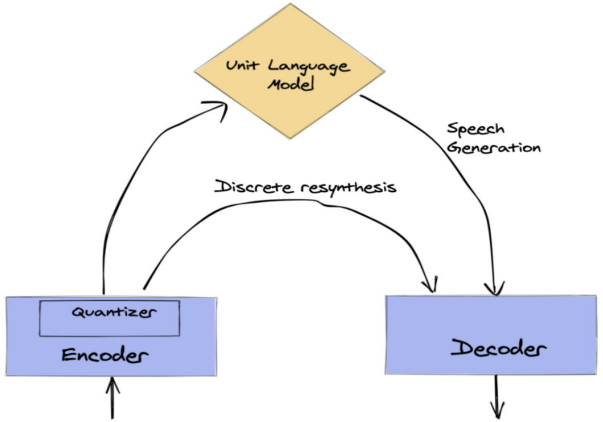
- **Environmental soundscape:**
  - separation
  - information extraction

- **Synthesis quality:**
  - domain shifts
  - experimental parameters

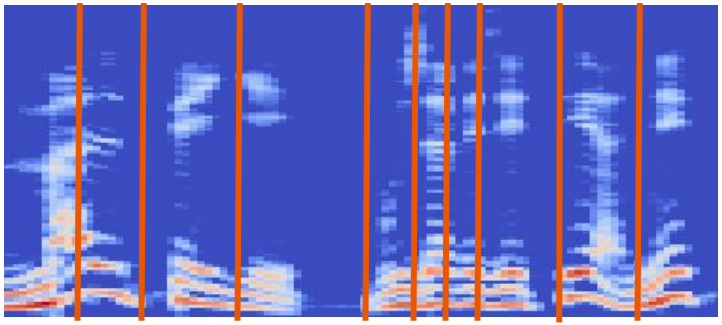
Methodology

1. Representation learning

Auto-encoders, predictive models (CPC, APC), adversarial models (GAN)



Acoustic unit discovery



Methodology

1. Representation learning
2. **Probing methods**

Bioacoustic tasks: classification in
- species
- call type
- identification (diarisation)
- physiological traits

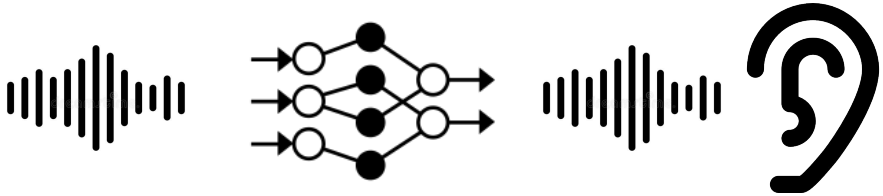
Soundscape:
- time-of-day prediction
- sound tagging…

…**Unsupervised tasks:**
- Odd-man-out

Methodology

1. Representation learning
2. Probing methods
3. **Synthesis**

Acoustic probing



Bioacoustic implications :
- Data-driven VS Physical
- Control
- Playback experiments

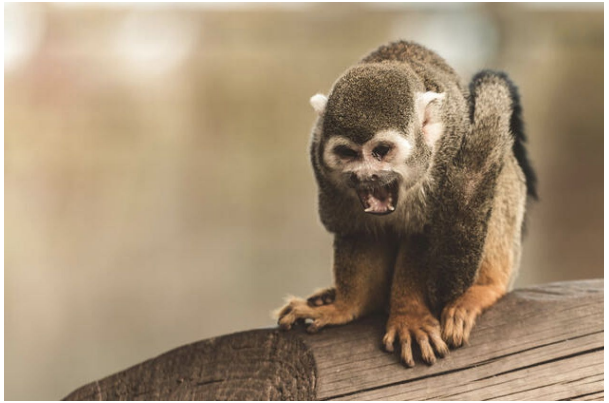
Datasets

Primates :

- Rousset (*saïmiris*, *papio papios*)
- INT (*marmosets*)
- Angela Dassow (*lar gibbons*)
- Vallée des singes *(bonobos)*

Other :

- *Human* speech (Librispeech, MSWC)
- Noise (Audioset)



Different species and vocalisation systems

Recording setup - environmental noise

Relevance for the ILCB

Understanding the model before leveraging its performances

- Probing the **black box** and making DL a truly scientific tool
- Jointly processing humans and primates to study the **origins of language**
- Contribute in making the ILCB a leading actor in the **computational modeling of language**

Supervisory team

LABORATOIRE D'INFORMATIQUE & SYSTÈMES

Benoît Favre

Multimodal speech processing
Unsupervised representation learning
Probing and explanation methods

Laboratoire Parole et Langage 50 1972 2022

Thierry Legou

Primatology
Primate bioacoustics
Animal acoustic monitoring

Ricard Marxer

Self-supervised acoustic representation learning
Bioacoustics and Deep Learning
Animal acoustic monitoring

Laboratoire de Psychologie Cognitive

Arnaud Rey

Sequence learning in non-human primates
Primate bioacoustics