Affective Computing

Acoustic Feature Extraction

Emilia Parada-Cabaleiro
emilia.parada-cabaleiro@jku.at

Institute of Computational Perception

JOHANNES KEPLER
UNIVERSITY LINZ

Institute of
Computational
Perception

Which is the time-line?

05.05.2022:	**Acoustic Feature Extraction (S3 047 + ZOOM)**
12.05.2022:	SER + Release Assignment 3 (S3 047 + ZOOM)
19.05.2022:	NO LECTURE
26.05.2022:	PUBLIC HOLIDAY
02.06.2022:	Discussion/Deadline Assignment 3 (ZOOM)
09.06.2022:	Recap. before Exam (S3 047 + ZOOM)
16.06.2022:	PUBLIC HOLIDAY
23.06.2022:	Exam (ZOOM)

FEATURE EXTRACTION – overview

- **OpenSMILE & Feature sets**
- **LLDs & Functionals**
- **OpenXBOW**
- **ComParE – Computational Paralinguistics challengE**

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia,* pp. 1459-1462.

Schmitt, M. & Schuller, B. (2017). openXBOW - Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit, *The Journal of Machine Learning Research*, 18(96): 1-5.

FEATURE EXTRACTION: OpenSMILE

- **OpenS** (Open-Source)
- **MILE** (Media Interpretation by Large-space Extraction)

  OpenSMILE :) is a toolkit for real-time feature extraction of (mainly) audio sources (speech and music) used for signal processing and machine learning applications. It is written in C++ but we will use the wrapper available for Python.

  DOCUMENTATION: https://audeering.github.io/opensmile/about.html
  INSTALLATION: https://audeering.github.io/opensmile-python/

  pip install opensmile

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia,* pp. 1459-1462.

FEATURE EXTRACTION: OpenSMILE

Audio Features

- Feature extractor for audio signal processing

  OpenSMILE

  https://www.audeering.com/opensmile/

- Feature sets for SER

  ComParE (Computational Paralinguistics challengE)

  6373 features by applying functionals to 65 LLDs+Deltas

  eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set) 88 functionals from 25 LLDs

FEATURE EXTRACTION: LLDs & Functionals

Audio Features

- Low Level Descriptors (LLDs)

  Short-term acoustic properties of the vocal signal extracted over time, e.g. every 10ms

- Statistical Functionals

  Mean, standard deviation, coefficient of variance … extracted for the whole sample (used for classification)

FEATURE EXTRACTION: LLDs & Functionals

- **Low Level Descriptors (LLDs)**

  Acoustic parameters extracted over time (overlapping frames) with a specified *hop size* and *frame length*

  e.g., F0 extracted each 10ms over a frame length of 60ms

  For each input (audio file), a matrix, where the rows are the time; the columns are the LLDs, is generated.

  LLDs can be used to feed dynamic ML models, e.g., Recurrent Neural Networks (RNNs).

FEATURE EXTRACTION: LLDs & Functionals

Low Level Descriptors (LLDs)

name	F0final_sma	voicingFinalUnclipped_sma	jitterLocal_sma	jitterDDP_sma
'0a94RFIXCVVsrqpP'	128.996	0.7768089	0	0
'0a94RFIXCVVsrqpP'	133.8499	0.7621633	0.03255364	0
'0a94RFIXCVVsrqpP'	137.1216	0.7414457	0.07903271	0.1182196
'0a94RFIXCVVsrqpP'	109.6528	0.7489837	0.216016	0.2073679
'0a94RFIXCVVsrqpP'	82.34254	0.7553298	0.2263948	0.3289911
'0a94RFIXCVVsrqpP'	56.57745	0.7670681	0.2535968	0.2821849
'0a94RFIXCVVsrqpP'	56.31411	0.7670981	0	0.2304451
'0a94RFIXCVVsrqpP'	56.11495	0.7649633	0.0006468152	0
'0a94RFIXCVVsrqpP'	55.8317	0.754817	0	0.0006468152
'0a94RFIXCVVsrqpP'	57.05762	0.7404714	0.004828005	0
'0a94RFIXCVVsrqpP'	58.2954	0.731442	0	0.004828005

FEATURE EXTRACTION: LLDs & Functionals

- **Functionals**

  Statistical operations performed for a specific LLD

  e.g., F0_mean, i.e., the mean of the LLD F0

  For each input (audio file), a vector, where each element is a functional of an LLD, is generated.

  Functionals can be used to feed static ML models, e.g., Support Vector Machines (SVM).

FEATURE EXTRACTION: LLDs & Functionals

Functionals

name	audspec_lengthL1norm_sma_range	audspec_lengthL1norm_sma_
'0a94RFIXCVVsrqpP'	5.035054	(
'0a35hcJVI4T8XzQi'	5.952072	(
'0a3daseMeMctSIYn'	3.774093	(

- **Other feature types:**
  - **Bag of Audio Words (openXBOW)**
  - **I-vectors (kaldi)**

FEATURE EXTRACTION: Feature sets

- **ComParE**

  6 373 acoustic features divided into four sub-sets:

  - Mel-Frequency Cepstral Coefficients (MFCC)
  - Spectral features (spectral slope ...)
  - Prosodic features (energy, loudness, pitch ...)
  - Micro-prosodic features (jitter, shimmer ...)

  Statistical functionals from 65 LLDs and the Delta Coefficients

Schuller, Björn, et al. (2013) "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism." Proc. of Interspeech 2013, Lyon, France, ISCA, pp. 148-152.

FEATURE EXTRACTION: Feature sets

- **eGeMAPS**

  88 acoustic features divided into three parameter groups:

  - Frequency related features (F0, jitter, formants ...)
  - Energy/amplitude related features (shimmer, loudness...)
  - Spectral features (spectral slope ...)

  Statistical functionals from 25 LLDs

Eyben, Florian, et al. (2015) "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing." IEEE transactions on affective computing 7(2), pp. 190-202.

LLDs

name	F0final_sma	voicingFinalUnclipped_sma	jitterLocal_sma	jitterDDP_sma
'0a94RFIXCVVsrqpP'	128.996	0.7768089	0	0
'0a94RFIXCVVsrqpP'	133.8499	0.7621633	0.03255364	0
'0a94RFIXCVVsrqpP'	137.1216	0.7414457	0.07903271	0.1182196
'0a94RFIXCVVsrqpP'	109.6528	0.7489837	0.216016	0.2073679
'0a94RFIXCVVsrqpP'	82.34254	0.7553298	0.2263948	0.3289911
'0a94RFIXCVVsrqpP'	56.57745	0.7670681	0.2535968	0.2821849
'0a94RFIXCVVsrqpP'	56.31411	0.7670981	0	0.2304451
'0a94RFIXCVVsrqpP'	56.11405	0.7640622	0.0006468153	0

Functionals

name	audspec_lengthL1norm_sma_range	audspec_lengthL1norm_sma_
'0a94RFIXCVVsrqpP'	5.035054	0
'0a35hcJVI4T8XzQi'	5.952072	0
'0a3daseMeMctSIYn'	3.774093	0

FEATURE EXTRACTION: OpenXBOW Bag of (Audio) Words

The Passau Open-Source Crossmodal Bag-of-Words Toolkit

openXBOW generates a bag-of-words representation from a sequence of numeric and/or textual features, e.g., **acoustic LLDs**, visual features, and transcriptions of natural speech.

- **Clone the following GitHub repository:**

  https://github.com/openXBOW/openXBOW

Maximilian Schmitt and Björn Schuller: "openXBOW - Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit", The Journal of Machine Learning Research, Volume 18, No. 96, pp. 1-5, October 2017.

FEATURE EXTRACTION: OpenXBOW Bag of (Audio) Words

From LLDs to Functionals

name	F0final_sma
'0a94RFIXCVVsrqpP'	128.996
'0a94RFIXCVVsrqpP'	133.8499
'0a94RFIXCVVsrqpP'	137.1216
'0a94RFIXCVVsrqpP'	109.6528
'0a94RFIXCVVsrqpP'	82.34254
'0a94RFIXCVVsrqpP'	56.57745
'0a94RFIXCVVsrqpP'	56.31411
'0a94RFIXCVVsrqpP'	56.11495
'0a94RFIXCVVsrqpP'	55.8317
'0a94RFIXCVVsrqpP'	57.05762
'0a94RFIXCVVsrqpP'	58.2954

mean

name	F0final_sma_mean
'0a94RFIXCVVsrqpP'	**84,741279**
'0a35hcJVI4T8XzQi'	**????**
'0a3daseMeMctSIYn'	**????**

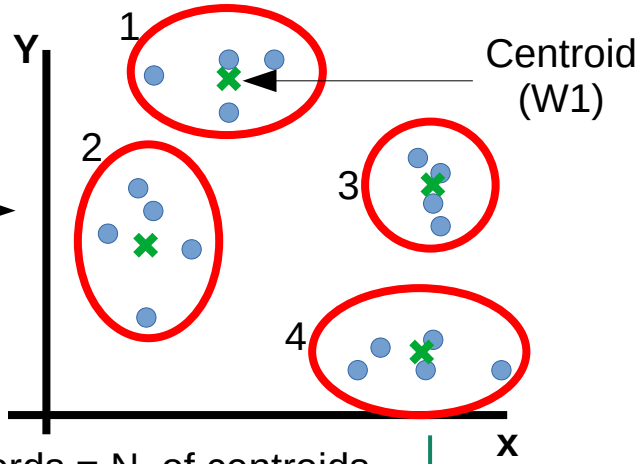
JOHANNES KEPLER UNIVERSITY LINZ

FEATURE EXTRACTION: OpenXBOW Bag of (Audio) Words

From LLDs to Bag of Audio Words (openXBOW)

F0final_sma	voicingFinalUnclipped_sma	jitterLocal_sma	jitterDDP_sma
128.996	0.7768089	0	0
133.8499	0.7621633	0.03255364	0
137.1216	0.7414457	0.07903271	0.1182196
109.6528	0.7489837	0.216016	0.2073679
82.34254	0.7553298	0.2263948	0.3289911
56.57745	0.7670681	0.2535968	0.2821849
56.31411	0.7670981	0	0.2304451
56.11495	0.7649633	0.0006468152	0
55.8317	0.754817	0	0.0006468152
57.05762	0.7404714	0.004828005	0
58.2954	0.731442	0	0.004828005

LLDs from a variety of samples



N. of words = N. of centroids

W1	W2	W3	W4	W5	W6	W7	Wn
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_n
Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_n

Codebook

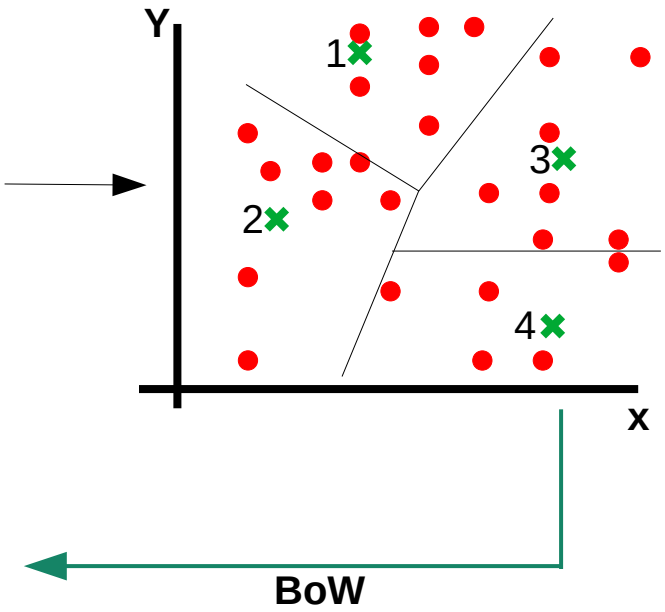
N. of dimensions = N. of LLDs

JOHANNES KEPLER UNIVERSITY LINZ

FEATURE EXTRACTION: OpenXBOW Bag of (Audio) Words

From LLDs to Bag of Audio Words (openXBOW)

name	F0final_sma	voicingFinalUnclipped_sma	jitterLocal_sma	jitterDDP_sma
'0a94RFIXCVVsrqpP'	128.996	0.7768089	0	0
'0a94RFIXCVVsrqpP'	133.8499	0.7621633	0.03255364	0
'0a94RFIXCVVsrqpP'	137.1216	0.7414457	0.07903271	0.1182196
'0a94RFIXCVVsrqpP'	109.6528	0.7489837	0.216016	0.2073679
'0a94RFIXCVVsrqpP'	82.34254	0.7553298	0.2263948	0.3289911
'0a94RFIXCVVsrqpP'	56.57745	0.7670681	0.2535968	0.2821849
'0a94RFIXCVVsrqpP'	56.31411	0.7670981	0	0.2304451
'0a94RFIXCVVsrqpP'	56.11495	0.7649633	0.0006468152	0
'0a94RFIXCVVsrqpP'	55.8317	0.754817	0	0.0006468152
'0a94RFIXCVVsrqpP'	57.05762	0.7404714	0.004828005	0
'0a94RFIXCVVsrqpP'	58.2954	0.731442	0	0.004828005



Name	W1	W2	W3	W4	W5	W6	Wn
0a9R...	6	8	7	5	...	...	...
???	???	???	???	???	???	???	???

N. of dimensions for each vector = N. of words in the codebook

JOHANNES KEPLER UNIVERSITY LINZ

FEATURE EXTRACTION: ComParE

ComParE

The Interspeech Computational Paralinguistics ChallengE is an open Challenge in the field of Computational Paralinguistics.

- **Tasks: To recognize states and traits from vocal signals.**
  - Primates species classification
  - Stress detection
  - Baby crying recognition
  - Austrian dialects identification
- **When: Every year at INTERSPEECH 2009-2021. This year ACM-MM.**
  - http://www.compare.openaudio.eu/

FEATURE EXTRACTION: ComParE

ComParE Baseline

Available at the ISCA (International Speech Communication Association) archive:

https://www.isca-speech.org/archive/index.html

- **2021:**
  https://www.isca-speech.org/archive/pdfs/interspeech_2021/schuller21_interspeech.pdf
- **2020:**
  - https://www.isca-speech.org/archive/pdfs/interspeech_2020/schuller20_interspeech.pdf
- **2019:**
  - https://www.isca-speech.org/archive/pdfs/interspeech_2019/schuller19_interspeech.pdf

In-class learning: **L7.ipynb**

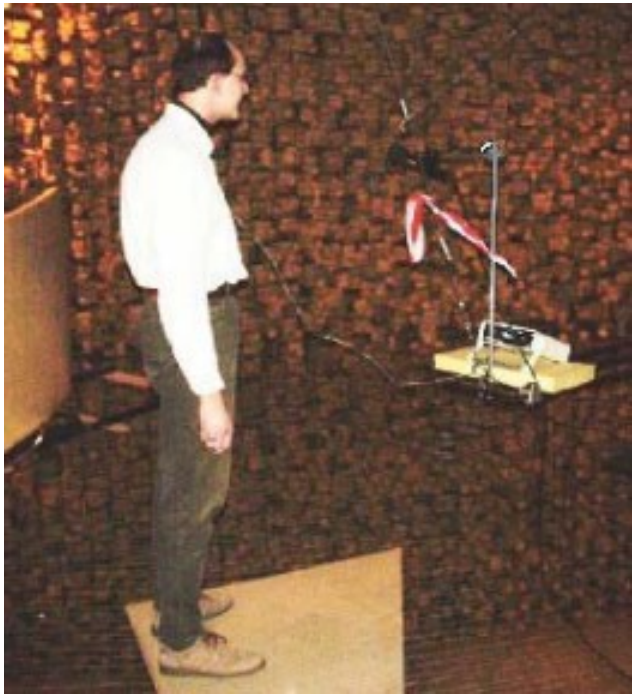
Berlin Database of Emotional Speech (EmoDB)

German emotional speech produced by actors according to the categorical model

http://emodb.bilderbar.info/docu/

Download it here:

http://emodb.bilderbar.info/download/



Burkhardt, F., et al. (2005). A database of German emotional speech. Proceedings of Interspeech.

OpenSMILE

The Munich open-Source Media Interpretation by Large feature-space Extraction

OpenSMILE :) is a toolkit for real-time feature extraction of audio sources (speech and music) used for signal processing and machine learning applications. It is written in C++ but we will use the wrapper available for Python.

DOCUMENTATION: https://audeering.github.io/opensmile/about.html
INSTALLATION: https://audeering.github.io/opensmile-python/

pip install opensmile

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia,* pp. 1459-1462.

OpenXBOW

The Passau Open-Source Crossmodal Bag-of-Words Toolkit

openXBOW generates a bag-of-words representation from a sequence of numeric and/or textual features, e.g., **acoustic LLDs**, visual features, and transcriptions of natural speech.

- **Clone the following github repository:**

  https://github.com/openXBOW/openXBOW

Maximilian Schmitt and Björn Schuller: "openXBOW - Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit", The Journal of Machine Learning Research, Volume 18, No. 96, pp. 1-5, October 2017.