INSENSION

www.insension.eu

Vocalization Recognition of People with Profound Intellectual and Multiple Disabilities (PIMD) Using Machine Learning Algorithms

Waldemar Jęśko

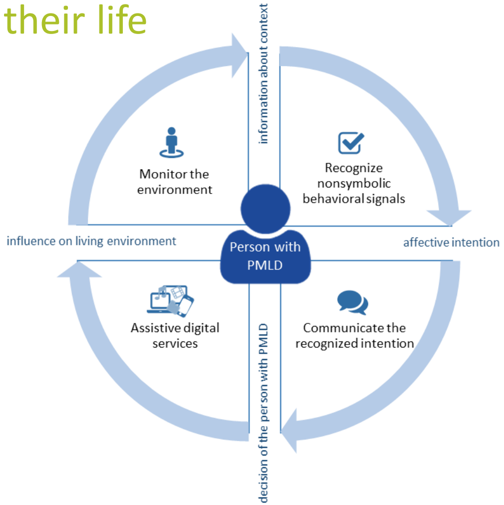
waldemarj@man.poznan.pl

Poznan Supercomputing and Networking Center, Poznan, Poland

Personalized intelligent platform enabling interaction with digital services to individuals with profound and multiple learning disabilities

ICT platform main objectives

- Improve the quality of their life
- Increase their ability to self-determination
- Enrich their life



INSENSION platform

Systems

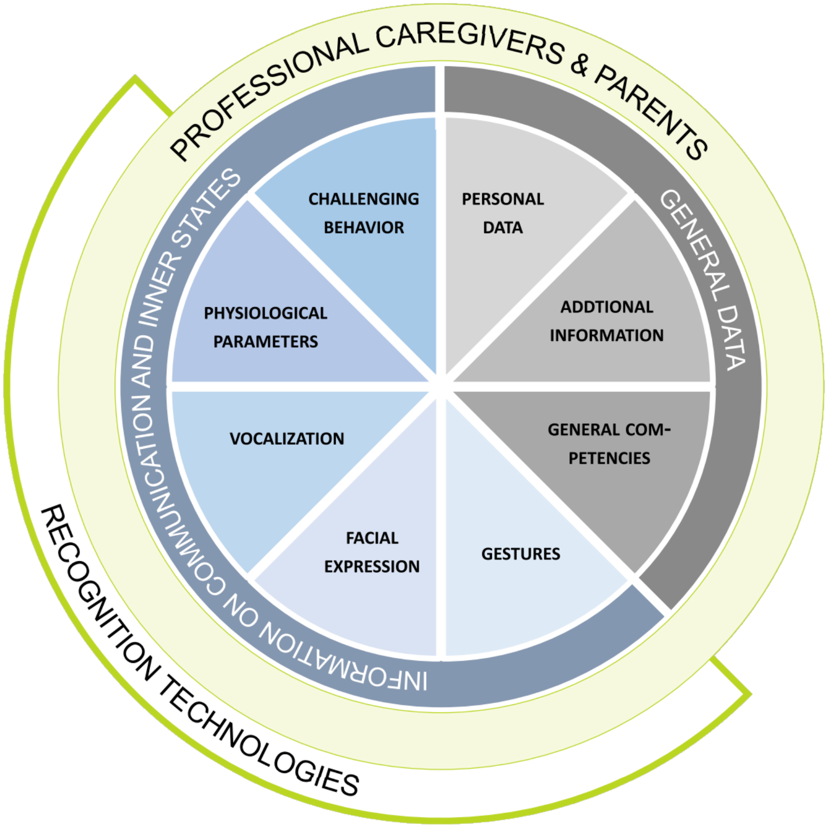
- Physiological Parameters Monitoring
- <u>Vocalization Recognition</u>
- Facial Expression Recognition
- Gesture Recognition

Make decision about


mood

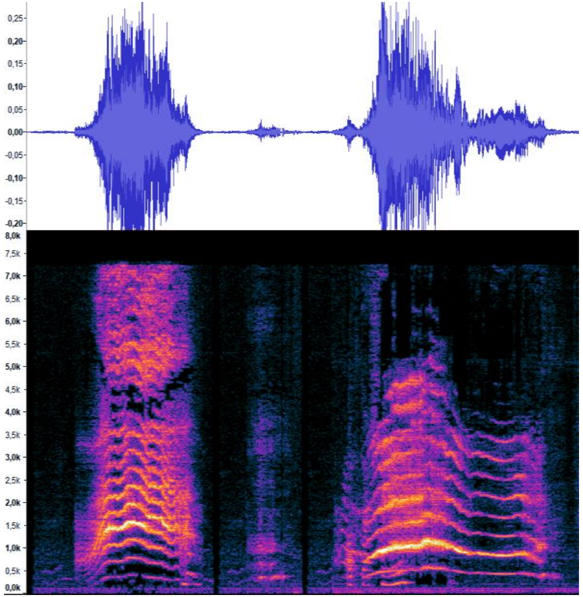

communication attempt


reason



Vocalizations characteristics

- **Vocalizations:** non-linguistic sound events (non-verbal communication)
- **Most common vocalizations types:**
  – Animal vocalizations
  – Child vocalizations
    – babbling phase (syllables ta, ma, ba, la etc.)



- Vocalizations of **people with PIMD** resemble the sounds of **vocalizing children** at the early stage of speech development

Dataset preparation

- Recording **audio sessions** of people with PIMD (2 people)
- Preparation of annotations in the **ELAN** program
- Vocalization instances:

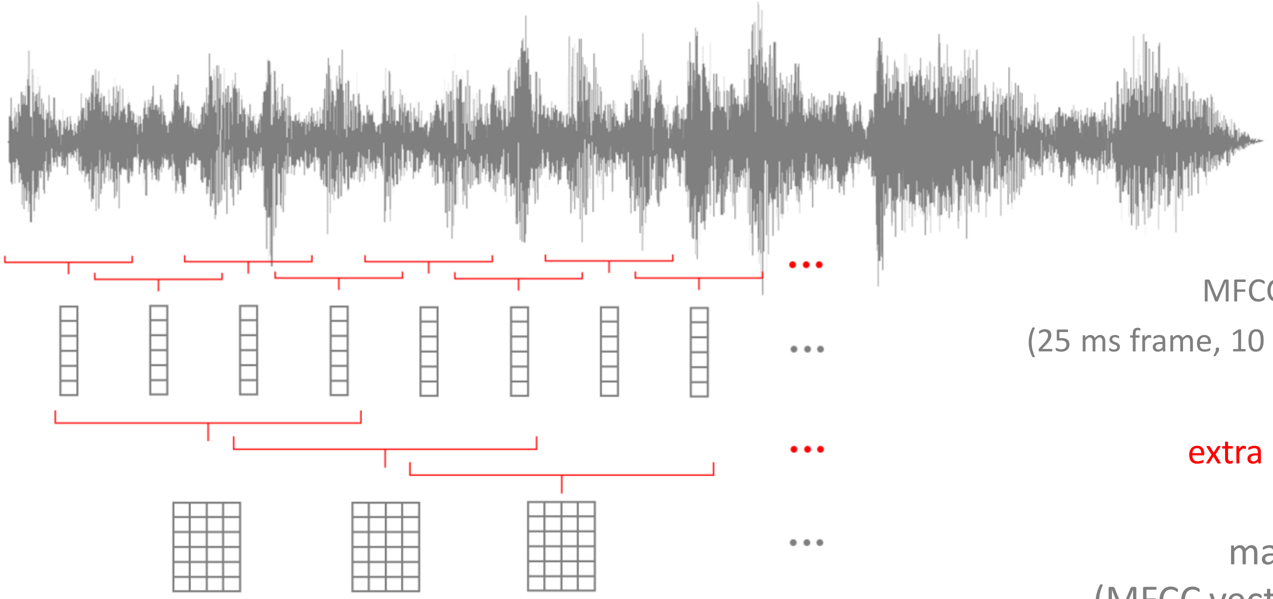
Table 1: *Number of annotated vocalization instances.*

Person	**Vocalization examples**					
Person A	eee	grunt	aaa	moan	eeh	laugh
	83	24	23	20	10	6
Person B	aaa	crying	aeaeae	cough	eee	nge
	133	51	22	11	6	4

Difficulties in obtaining recordings

- Need to record in specially **adapted rooms**
- Need for **caregivers' participation** in the process of recordings
- **Limited cooperation** from people with PIMD

Dataset preparation – pre-processing



MFCC vectors

(25 ms frame, 10 ms step, 40 features)

↓

extra buffers *

↓

matrices

(MFCC vectors over time)

* Preliminary research – testing:
  - buffer length (50 – 1000 ms),
  - buffer overlapping (0 – 90%).

Optimal conf: buffer length: 500 ms, no buffer overlapping

Dataset preparation – final dataset

Table 2: *Number of samples for tested buffer sizes – Person A.*

Buffer length [ms]	**Vocalization examples**					
	eee	grunt	aaa	moan	eeh	laugh
50	2198	749	832	901	188	136
100	1080	368	410	445	92	67
200	518	178	198	218	43	33
300	335	114	129	142	28	21
500	182	65	73	80	15	12
1000	65	25	31	34	3	4

- buffer: 500 ms
- no overlapping
- 4-fold augmentation

Table 3: *Number of examples in the final dataset – 500 ms buffer length, no overlapping, 4-fold data augm.*

Person	**Vocalization examples**					
Person A	eee	grunt	aaa	moan	eeh	laugh
	728	260	292	320	60	48
Person B	aaa	crying	aeaeae	cough	eee	nge
	1928	2884	248	56	60	20

Research methodology

- **Input:** audio signal
- 5-fold cross-validation
- data divided into folds at the input audio files level
- metrics:
  - accuracy (acc)
  - precision (prec)
  - recall (rec)
  - f1-score (f1)

Models classes:

- Person A (7):

  *eee, grunt, aaa, moan, eeh, laugh, bck**
- Person B (7):

  *aaa, crying, aeaeae, cough, eee, nge, bck**

* background class (*bck*): the remaining parts of the recordings (not being vocalizations) labeled as background

Research stages

1 Baseline

Dummy methods

- uniform
- stratified
- most frequent

< 30%

2 Basic ML methods

k-NN

NB

DT

RDF

50 / 75 %

3 Neural Networks

MLP

LSTM

LSTM
+dense

78 / 89 %

Baseline

1

Baseline

Dummy methods

- uniform
- stratified
- most frequent

Table 4: *Baseline results.*

Person	Dummy method	Metrics [%]			
		acc	prec	rec	f1
Person A	uniform	**18.1**	**30.2**	**18.1**	**21.4**
	stratified	11.2	17.4	11.2	12.8
	most frequent	10.3	1.10	10.3	1.90
Person B	uniform	15.0	32.1	15.0	**20.1**
	stratified	14.9	**33.0**	14.9	19.9
	most frequent	**33.9**	11.5	**33.9**	17.2

< 30%

Basic ML methods

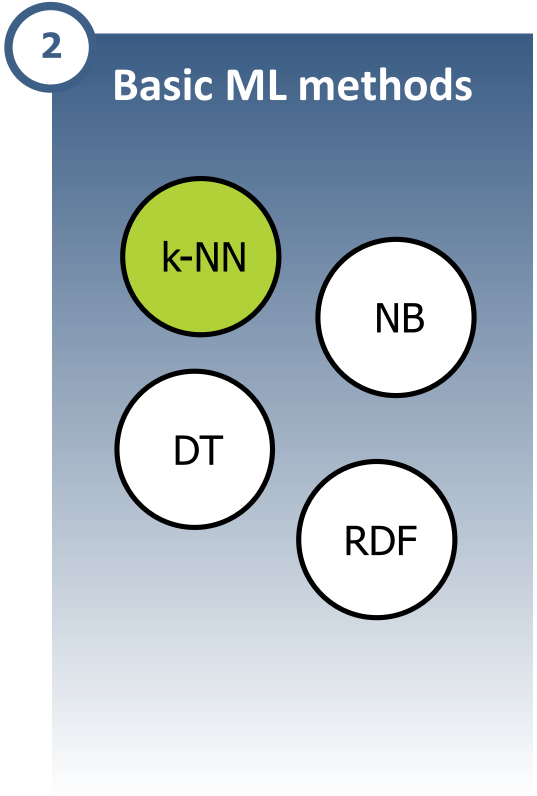
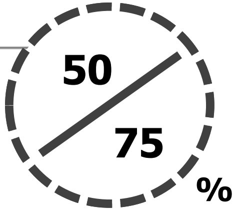


Table 5: *Basic ML methods results.*

Person	Metrics	Basic ML algorithms			
		k-NN	NB	DT	RDF
Person A	acc	**50.9**	49.1	42.2	31.0
	prec	**50.9**	47.7	40.6	25.3
	rec	**50.9**	49.1	42.2	31.0
	f1	46.4	**47.4**	34.5	18.7
Person B	acc	**74.5**	66.9	58.1	58.3
	prec	76.5	**79.8**	61.0	69.4
	rec	**74.5**	66.9	58.1	58.3
	f1	**73.0**	70.9	54.2	53.9



Neural Networks

3 Neural Networks

- MLP
- LSTM
- LSTM +dense

Table 6: *Neural networks results.*

Person	Reg. method	NN architectures MLP	LSTM	LSTM +dense
Person A	-	**70.5**	69.1	67.4
	dropout	**76.3**	71.7	66.7
	weight reg.	70.9	**72.5**	68.2
	batch norm.	**77.3**	67.6	67.3
	many methods	**78.1**	73.8	76.1
Person B	-	85.2	**86.0**	85.2
	dropout	**86.9**	86.1	84.9
	weight reg.	**84.5**	84.4	83.9
	batch norm.	**86.6**	86.5	83.3
	many methods	**88.8**	87.3	86.0

78 / 89 %

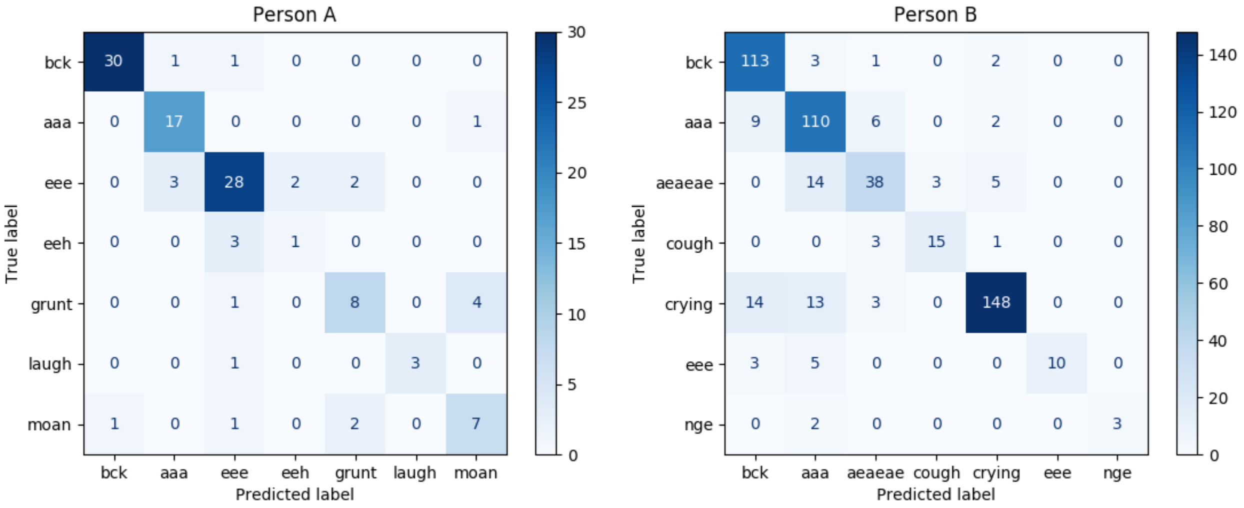
Neural Networks

3

Neural Networks

- MLP
- LSTM
- LSTM +dense

Example: CMs for one of the test folds (the best models)



Results

- The **best results** obtained: **80%** (Person A) and **90%** (Person B):
  - feedforward NN: **MLP**
  - **1-layer** with **64 neurons**
- The **best regularization** method:
  - combination of dropout and batch normalization
- **Sufficient** for enhancing communication with PIMD people
- Future work:
  - obtain more data
  - extract features with LPC, PLP, RASTA
  - classify with BLSTM, hybrid CNN-LSTM

Results – main findings

- The best results were obtained for **neural networks** (similar results for MLP and LSTM)
  - promising next step is to examine more NN architectures and feature extraction methods on a **larger vocalizations database**
- Optimal structure of tested NN:
  - **1 layer** with **64 neurons**
  - at the input and output of the layer: combination of dropout and batch norm.

  - **Conclusions**:
    - in this research problem, more extensive (deeper) networks will not be applicable
    - low network complexity (and use of regularization methods) largely prevented overfitting the training data

INSENSION

Waldemar Jęśko

waldemarj@man.poznan.pl

Poznan Supercomputing and Networking Center

www.insension.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780819.

PSNC · Pädagogische Hochschule HEIDELBERG University of Education · Jožef Stefan Institute · CTIC centro tecnológico · stowarzyszenie na tak · HARPO