Speech, Language, and Movement Processing to Model Parkinson's Disease

Juan Rafael Orozco-Arroyave (Rafa)

GITA Lab
School of Engineering, University of Antioquia
Medellín, Colombia

gita.udea.edu.co
rafael.orozco@udea.edu.co

UNIVERSIDAD DE ANTIOQUIA
1 8 0 3

GITA

Invited Talk at:
Conversational AI Reading Group
Quebec AI Institute (Mila)
Montreal, Canada
27.11.2025

Agenda

1. Introduction
2. Classical approaches
3. Speech and movement
4. Transitions in facial expressions
5. Speech and language analysis
6. Summary and outlook

Agenda

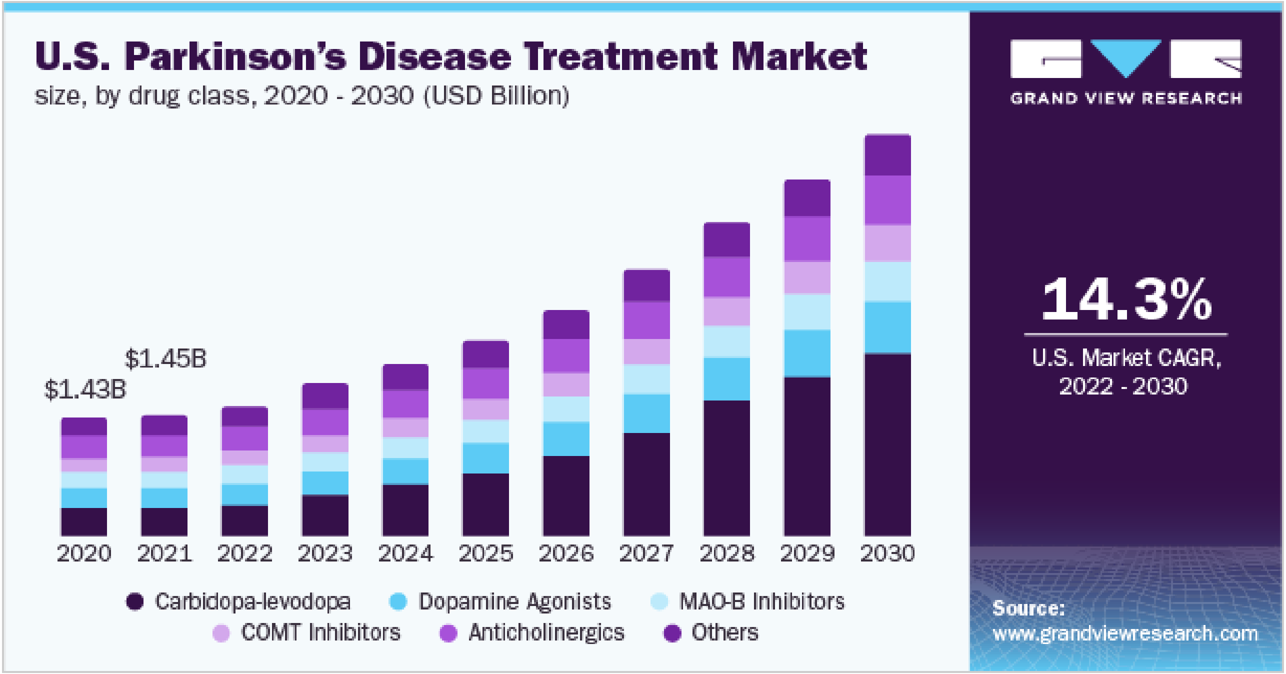
1. Introduction
2. Classical approaches
3. Speech and movement
4. Transitions in facial expressions
5. Speech and language analysis
6. Summary and outlook

Introduction

It affects +6M people

Source:
https://www.silverbook.org/fact/new-cases-of-parkinsons-disease-each-year/

Introduction (cont.)



Source: https://www.grandviewresearch.com/industry-analysis/parkinsons-disease-treatment-market

Introduction (cont.)

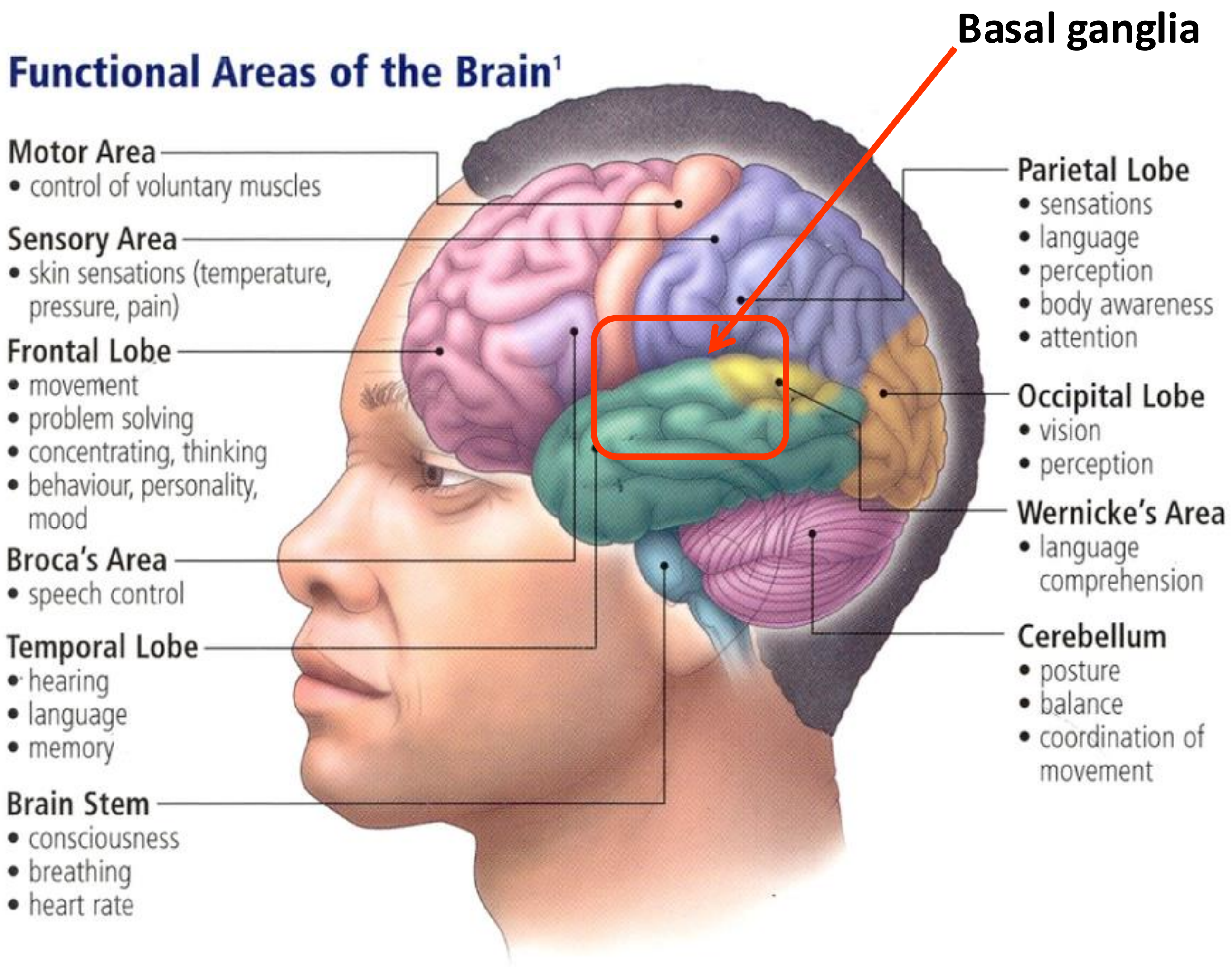


Figure retrieved from:
https://quizlet.com/313112455/functional-areas-of-the-brain-diagram/

Introduction (cont.)

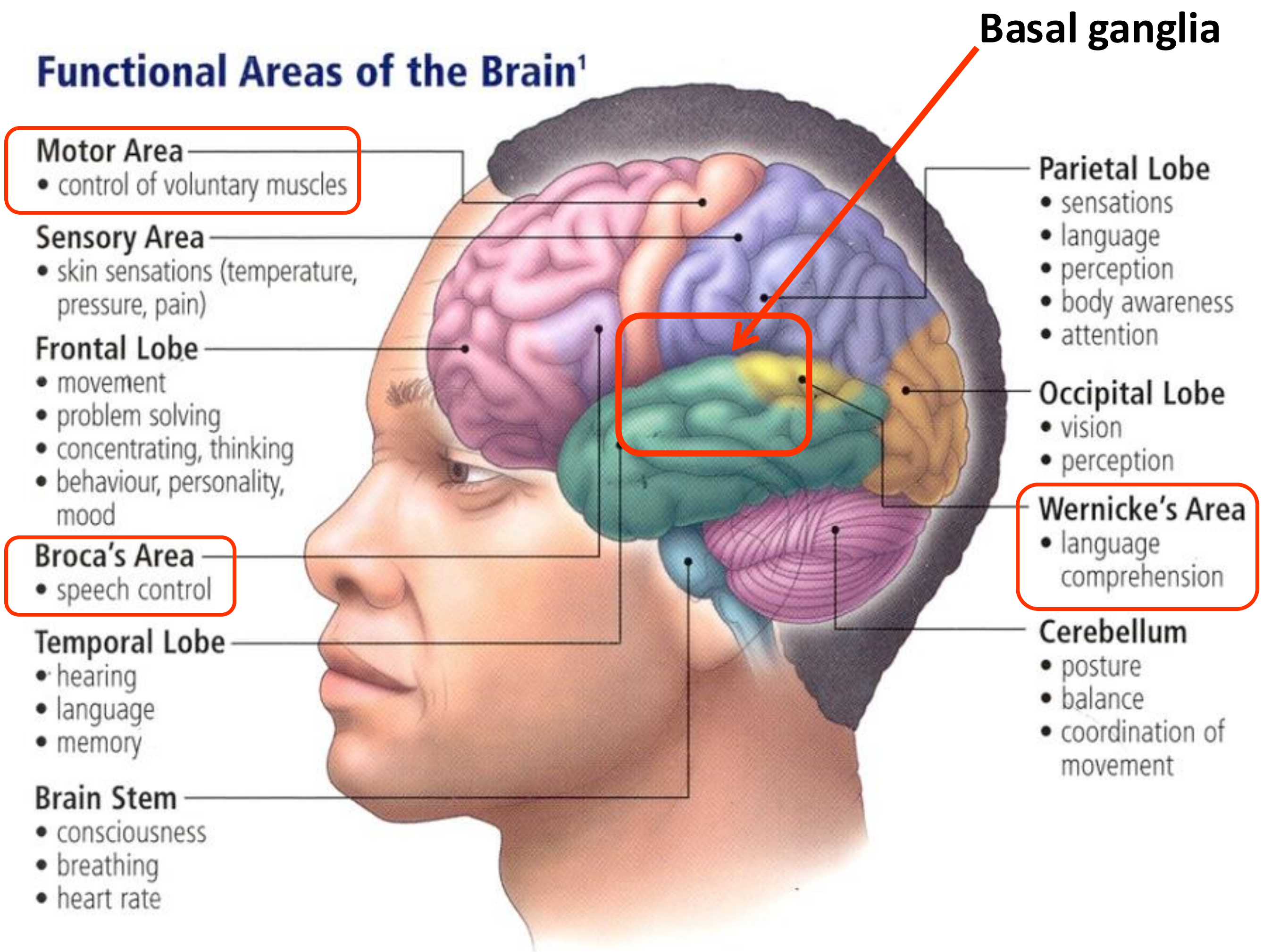


Figure retrieved from:
https://quizlet.com/313112455/functional-areas-of-the-brain-diagram/

Introduction (cont.)

- Typical diagnosis is expensive and time-consuming

MEMORIA

Lea la lista de palabras, el paciente debe repetirlas. Haga dos intentos. Recuérdeselas cinco minutos mas tarde.

	ROSTRO	SEDA	IGLESIA	CLAVEL	ROJO	Sin puntos
1er intento						
2do intento						

ATENCIÓN

Lea la serie de números(1 número /sg)
El paciente debe repetirla [ ]2 1 8 5 4
El paciente debe repetirla a la inversa [ ]7 4 2 ____/2

Lea la serie de letras. El paciente debe dar un golpecito con la mano cada vez que se diga la letra A.
No se asignan puntos si tiene ≥ 2 errores.
[ ]F B A C M N A A J K L B A F A K D E A A A J A M O F A A B ____/1

Restar de 7 en 7 empezando desde 100
[ ]93 [ ]86 [ ]79 [ ]72 [ ]65 ____/3
4 o 5 sustracciones correctas: **3 puntos**, 2 o 3 correctas: **2 puntos**, 1 correcta: **1 punto**, 0 correctas: **0 puntos**.

LENGUAJE

Repetir: El gato se esconde bajo el sofá cuando los perros entran en la sala. [ ]
Espero que él le entregue el mensaje una vez que ella se lo pida. [ ] ____/2

Fluidez del lenguaje. Decir el mayor número posible de palabras que comiencen por la letra "P" en 1 min.
[ ]________ (N ≥ 11 palabras) ____/1

Neuropsychological tests require expert knowledge

Clinical imaging requires highly sophisticated machinery

Not all patients/societies can afford such costs!

Introduction (cont.)

New approaches

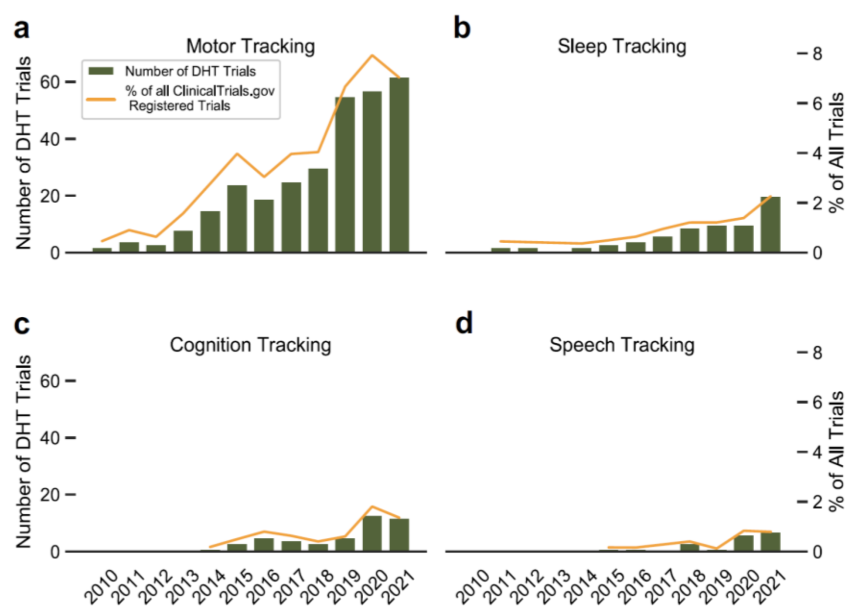
npj | digital medicine   www.nature.com/npjdigitalmed

BRIEF COMMUNICATION   OPEN

Check for updates

Evidence from ClinicalTrials.gov on the growth of Digital Health Technologies in neurology trials

Lars Masanneck[1,2], Pauline Gieseler[2], William J. Gordon[3,4,5], Sven G. Meuth[1] and Ariel D. Stern[2,6,7]



Introduction (cont.)

Which skills are affected by PD?

Voice	Speech	Language	Emotions
Gait	Handwriting	Face Expressions	Depression

Introduction (cont.)

SPA GER

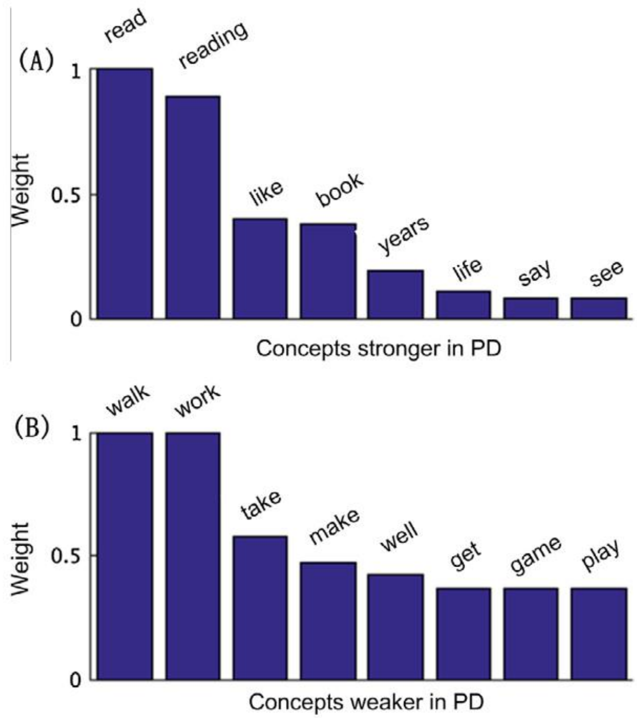
Voice	Speech	Language	Emotions
Gait	Handwriting	Face Expressions	Depression

Introduction (cont.)

- «When I was young, I liked to dance Tango but not now, I am not in the mood to dance anymore»
- «I feel bad today»
- «I have lost the appetite today»

Voice	Speech	Language	Emotions
Gait	Handwriting	Face Expressions	Depression

Introduction (cont.)



<u>Semantics</u>

PD patients use less motor verbs than HC subjects

A.M. García et al., «How language flows when movements don't: An automated analysis of spontaneous discourse in Parkinson's disease» Brain & Language, 162: 19-28, 2016.

Voice	Speech	Language	Emotions
Gait	Handwriting	Face Expressions	Depression

Introduction (cont.)

Healthy

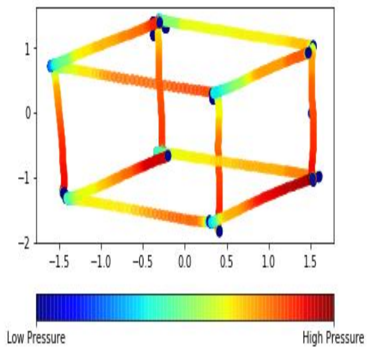
Parkinson's

Pictures used with explicit permission from the subjects

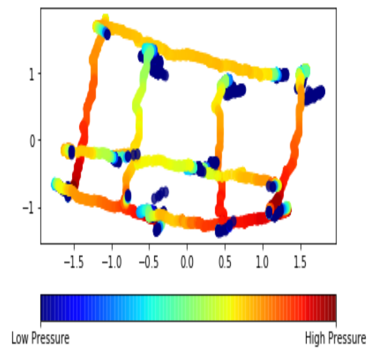
Voice	Speech	Language	Emotions
Gait	Handwriting	Face Expressions	Depression

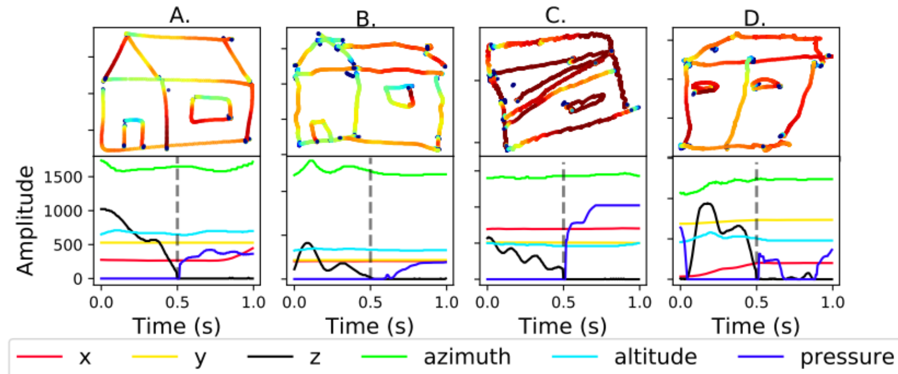
Introduction (cont.)

Healthy



Parkinson





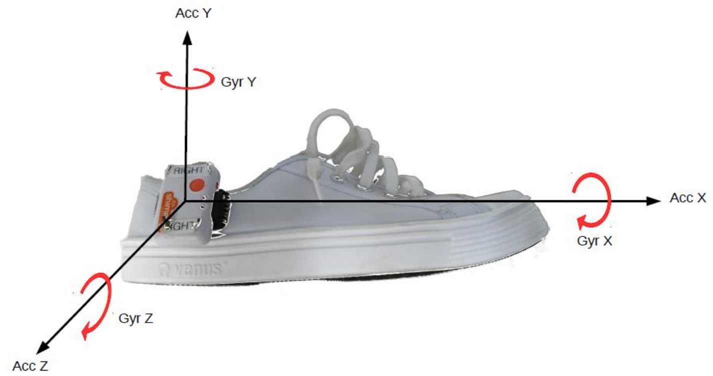
Pressure, in-air movements, tremor, micrographia

P. Drotar et al., «Analysis of in-air movement in handwriting: A novel marker for Parkinson's disease» Computer Programs and Methods in Biomedicine, 117(3): 405-411, 2014.

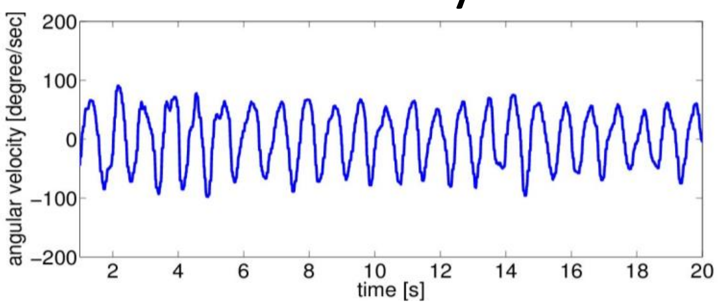
JC Vásquez-Correa et al., «Multimodal assessment of Parkinson’s disease: a deep learning approach» IEEE Journal of Biomedical and Health Informatics, 23(4): 21-36, 2019.

Voice	Speech	Language	Emotions
Gait	Handwriting	Face Expressions	Depression

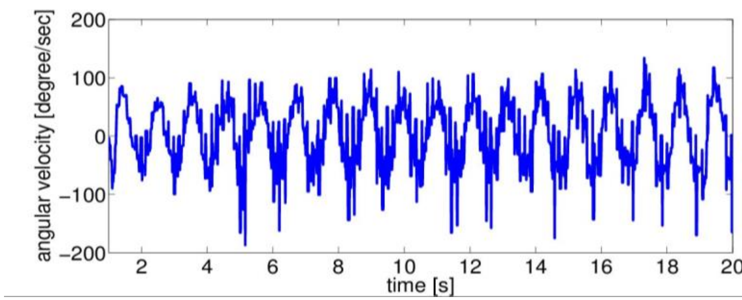
Introduction (cont.)



Healthy



Parkinson



Voice	Speech	Language	Emotions
Gait	Handwriting	Face Expressions	Depression

Introduction (cont.)

Evaluation of the neurological state

- Movement Disorder Society – Unified Parkinson's Disease Rating Scale (MDS-UPDRS)
  - Part I (13 items) non-motor experiences of daily living [0-52]
  - Part II (13 items): motor experiences of daily living [0-2]
  - Part III (33 items): motor examination [0-132]
  - Part IV (6 items): motor complications [0-24]

  - ➢ 14 items for upper limbs (UL) -> handwriting
  - ➢ 14 items for lower limbs (LL) -> gait, heel-toe, etc.
  - ➢ Only one item for speech
  - ➢ Only one item for depression
  - ➢ Speech & Language evaluation is not performed by an expert phoniatrician
- Hoehn & Yahr scale (H&Y)
  - One item with 8 possible values between 0 and 5.

Introduction (cont.)

Evaluation of the neurological state

- Movement Disorder Society – Unified Parkinson's Disease Rating Scale (MDS-UPDRS)
  - Part I (13 items) non-motor experiences of daily living [0-52]
  - Part II (13 items): motor experiences of daily living [0-2]
  - **Part III (33 items): motor examination [0-132]**
  - Part IV (6 items): motor complications [0-24]

  - 14 items for upper limbs (UL) -> handwriting
  - 14 items for lower limbs (LL) -> gait, heel-toe, etc.
  - Only one item for speech
  - Only one item for depression
  - Speech & Language evaluation is not performed by an expert phoniatrician
- Hoehn & Yahr scale (H&Y)
  - One item with 8 possible values between 0 and 5.

Agenda

1. Introduction
2. Classical approaches
3. Speech and movement
4. Transitions in facial expressions
5. Speech and language analysis
6. Summary and outlook

2. Classical approaches

- Speech
  -- How a person speaks -> Dysarthric speech

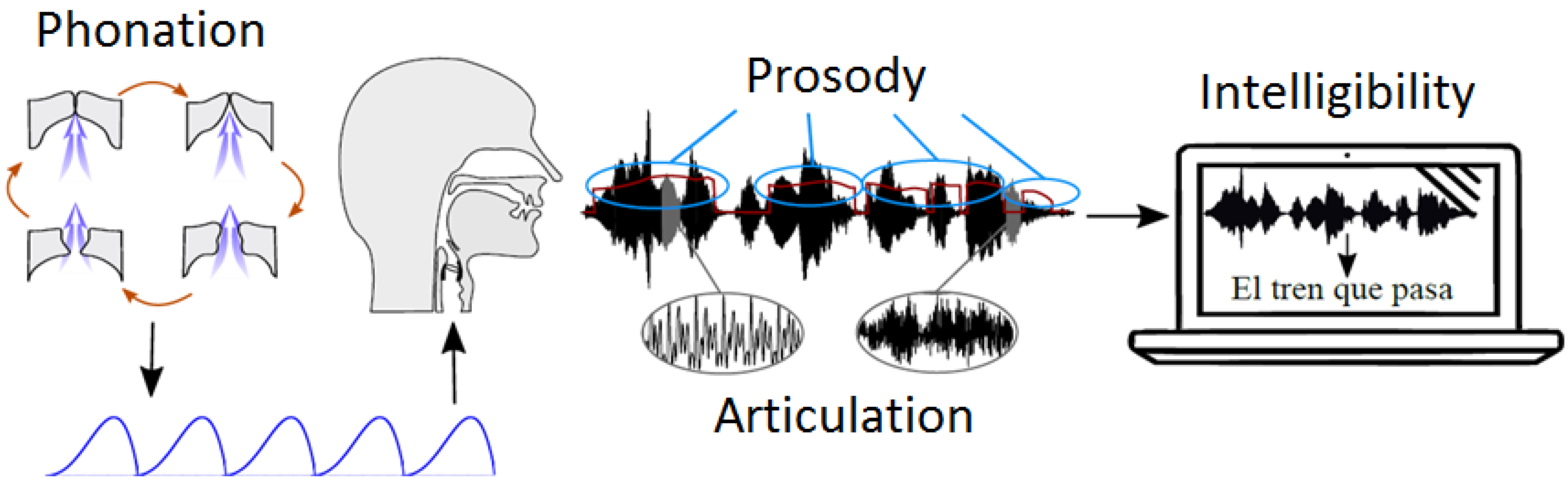
- Handwriting
  -- Micrographia -- Tremor -- Rigidity

- Gait
  -- Bradykinesia -- Postural instability -- Freezing of gait (FoG)

- Facial expression
  -- Hypomimia

Speech modeling

It can be studied considering different aspects/dimensions of speech



JR. Orozco-Arroyave et al., «NeuroSpeech: an open-source software for Parkinson's speech analysis» Digital Signal Processing, 77: 207-221, 2018.

Toolkit to extract features: DisVoice < https://disvoice.readthedocs.io/en/latest/ >

Phonation

- Process to produce the excitation signal: take air from the lungs and make the vocal folds vibrate.
  - Harmonicity, periodicity, and regularity.

Articulation

- Movement of articulators: tongue, lips, jaw, velum.
  - Position, time, and energy.

Prosody

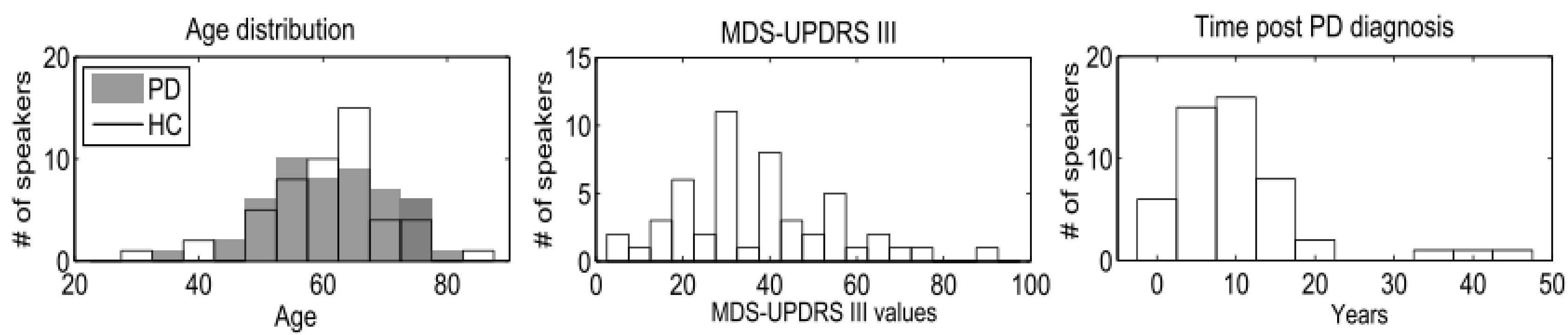
- Intonation and time to produce natural speech.
  - Control of the tone, pauses, speech rate and time.

Intelligibility

- Is the person understandable?
  - Number of words/syllables/phonemes correctly recognized.

Summary of experiments and results with speech signals

- Phonation, articulation, prosody and intelligibility
- 100 speakers (50 with PD and 50 HC; 25 male in each group)



- K-fold cross-validation
- Support Vector Machine (SVM)
- Support Vector Regressor (SVR)

	Phonation (vowels)	Phonation (words)	Articulation (sent+monol)	Prosody (sent+monol)	Intelligibility (DDK tasks)
Accuracy (%)	86 ± 4	76 ± 4	83 ± 3	76 ± 5	60 ± 3
AUC	0.87	0.78	0.85	0.80	0.59

- Vowels seem to be a good choice, so why to evaluate other speech tasks? R/ because sustained vowels are not a natural way of communication, and we need non intrusive evaluations.

	Phonation (vowels)	Phonation (words)	Articulation (sent+monol)	Prosody (sent+monol)	Intelligibility (DDK tasks)
Accuracy (%)	86 ± 4	76 ± 4	83 ± 3	76 ± 5	60 ± 3
AUC	0.87	0.78	0.85	0.80	0.59

- Vowels seem to be a good choice, so why to evaluate other speech tasks? R/ because sustained vowels are not a natural way of communication, and we need non intrusive evaluations.

- Intelligibility needs to be studied in more detail, and remember: **it highly depends on your ASR system**

	Phonation (vowels)	Phonation (words)	Articulation (sent+monol)	Prosody (sent+monol)	Intelligibility (DDK tasks)
Accuracy (%)	86 ± 4	76 ± 4	83 ± 3	76 ± 5	60 ± 3
AUC	0.87	0.78	0.85	0.80	0.59

- Vowels seem to be a good choice, so why to evaluate other speech tasks? R/ because sustained vowels are not a natural way of communication, and we need non intrusive evaluations.

- Intelligibility needs to be studied in more detail, but remember: **it highly depends on your ASR system**

- Articulation seems to give good results, robust, stable … let's see how generalizable are these results to other languages

Articulation in sentences and monologues only

	Spanish	German*	Czech**
Accuracy (%)	81	79	95
AUC	0.82	0.78	0.94

*German data: 88 with PD and 88 HC

** Czech data: 20 with PD and 16 HC

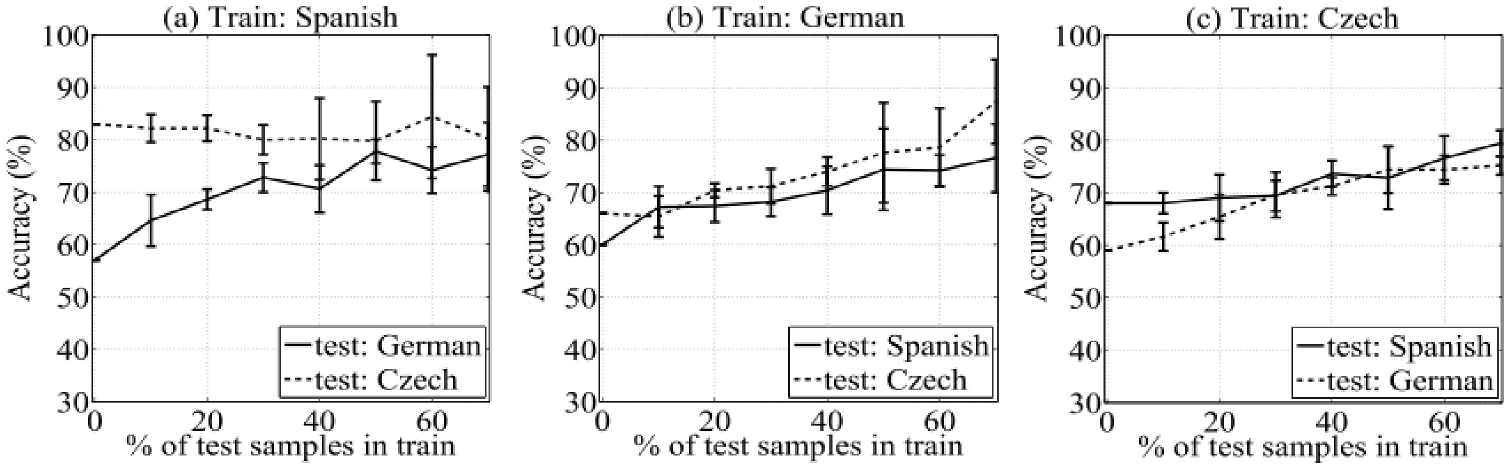
What about crossing the languages?

Articulation in sentences and monologues only

	Spanish	German*	Czech**
Accuracy (%)	81	79	95
AUC	0.82	0.78	0.94

*German data: 88 with PD and 88 HC

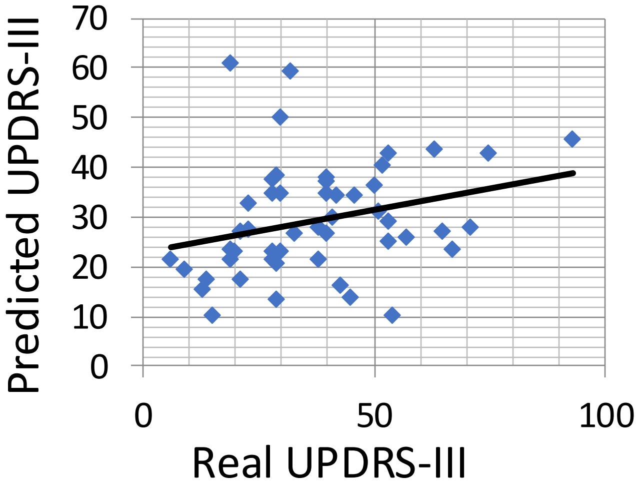
** Czech data: 20 with PD and 16 HC



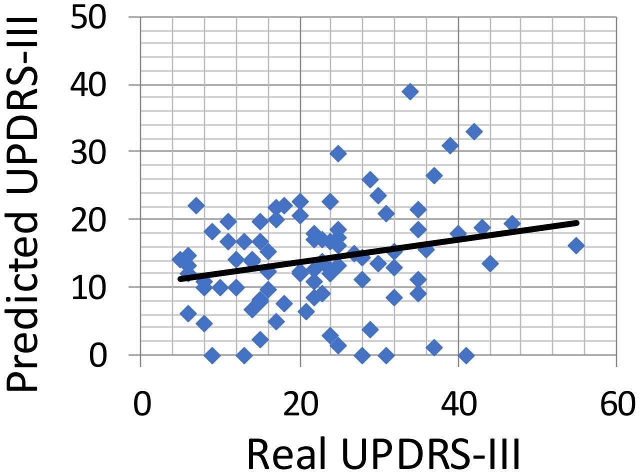
Monologues only

Evaluation of the neurological state

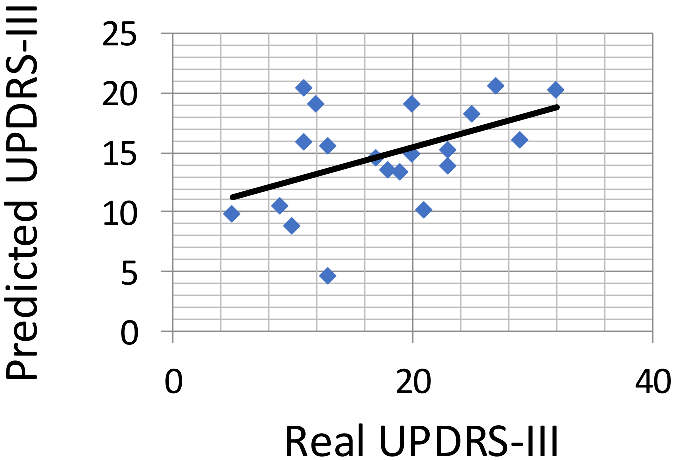
Spanish: $\rho = 0.36; p < 0.0001$



German: $\rho = 0.22; p < 0.0001$



Czech: $\rho = 0.45; p < 0.0001$



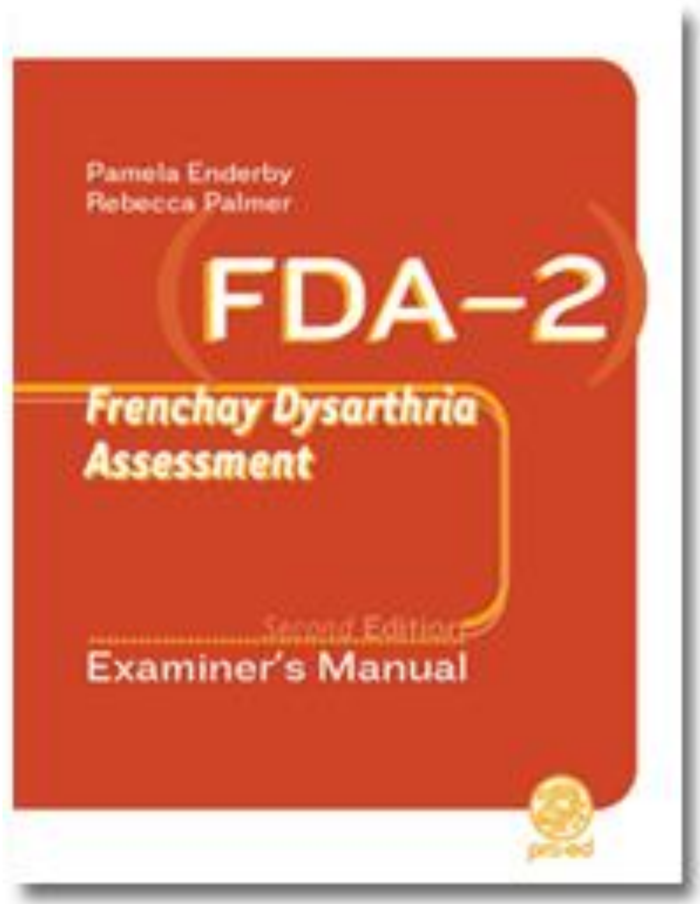
ρ : Spearman's correlation coefficient

But the MDS-UPDRS-III is a general scale to evaluate many different motor aspects.

If only speech signals are available, a dedicated scale is required!

Evaluation of the degree of dysarthria: Parkinson's Disease

Frenchay Dysarthria Assessment (FDA)



Authors: Pam Enderby & Rebecca Palmer, 2008

Requires the patient to visit the expert at the clinic because it includes swallowing tasks

Modified FDA (m-FDA)

- 13 items
- Range per item: 0 ... 4
- Total range: 0 ... 52

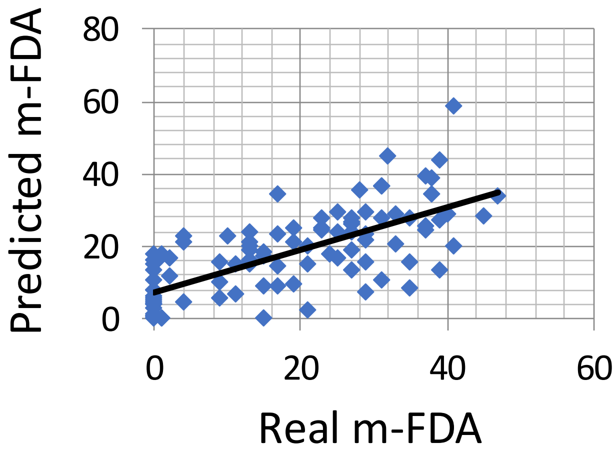
Factor	Speech Task
1- Respiration	Sustained vowels and monologue
2- Lips	Monologue and /pa-ta-ka/
3- Palate/velum	Monologue and /pa-ta-ka/
4- Larynx	Monologue and read text
5- Tongue	Monologue and /pa-ta-ka/
6- Intelligibility	Monologue and read text

JC. Vásquez-Correa et al., «Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease» Journal of Communication Disorders, 76: 21-36, 2018.

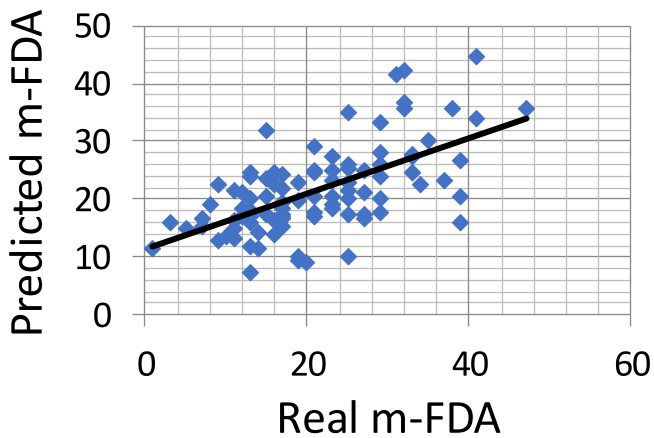
- Three experts phoniatricians agreed on the first 10 evaluations (we had PC-GITA with 100 subjects)
- The other 90 evaluations (40 patients and 50 healthy controls) were individually performed per each phoniatrician
- Inter-rater reliability: 0.75
- Spearman's correlation (ρ) between **articulation features** and expert evaluations

	Expert 1	Expert 2	Expert 3	Median
Monologue	0.43	0.39	0.28	0.35
Read text	0.58	0.42	0.47	0.52
/pa-ta-ka/	0.72	0.62	0.62	0.67

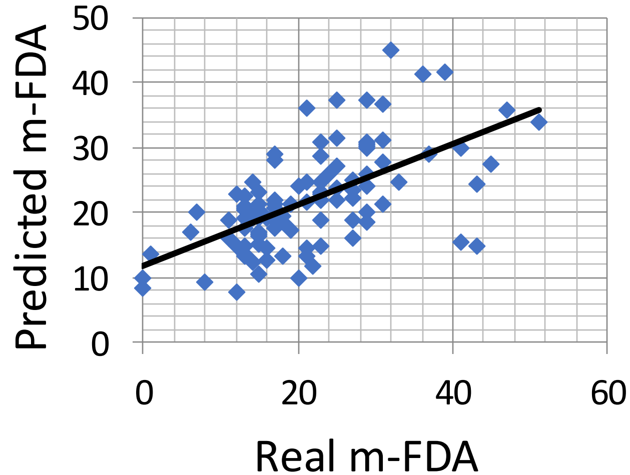
Expert 1: $\rho = 0.72; p < 0.0001$



Expert 2: $\rho = 0.62; p < 0.0001$

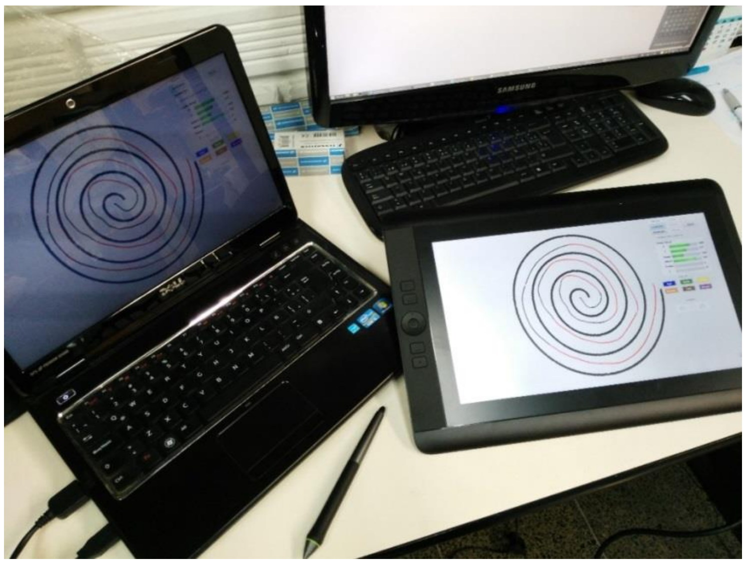


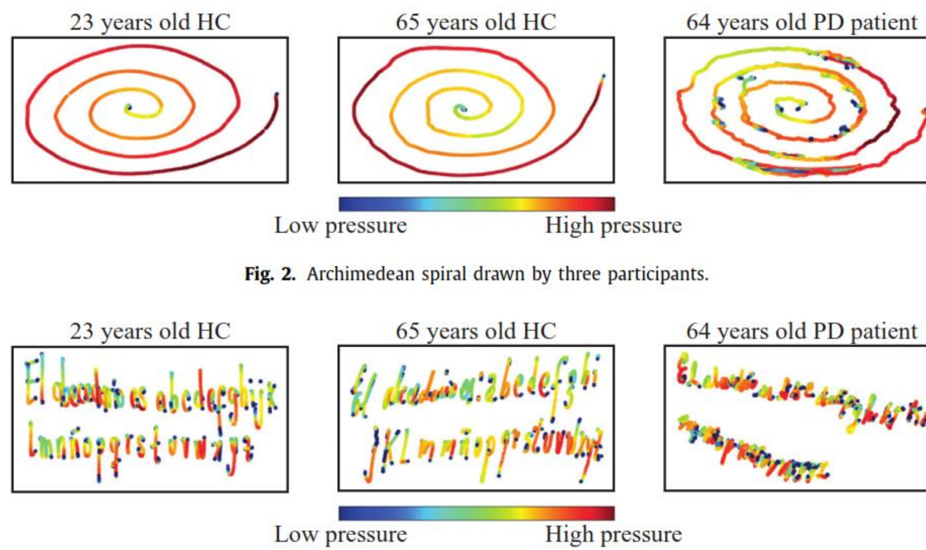
Expert 3: $\rho = 0.62; p < 0.0001$



Handwriting modeling

Dataset: 39 Patients with PD, 39 elderly HC (eHC), and 40 young HC (yHC)



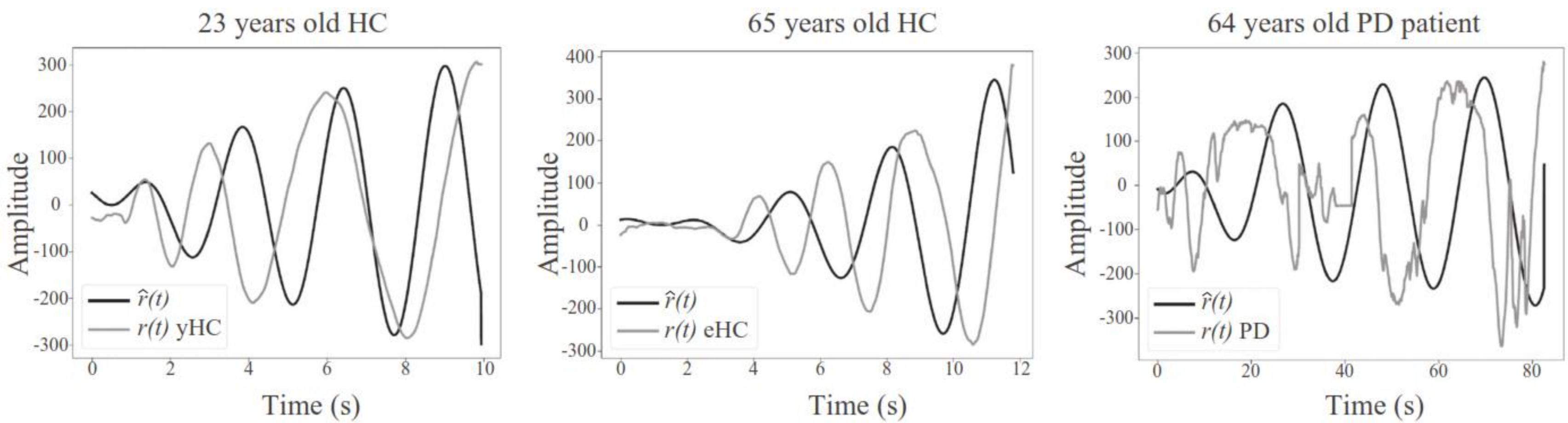


C.D. Ríos-Urrego et al., «Analysis and evaluation of handwriting in patients with Parkinson's disease using kinematic, geometrical, and nonlinear features» Computer Methods and Programs in Biomedicine, 173: 43-52, 2019.

Standard pipeline with an SVM as classifier

- Kinematics: position, speed, acceleration, pressure, distance from the pen to the table's surface, etc.

- Geometrical: spiral's trajectory is modeled as an amplitude-modulated signal: $r(t) = (at^3 + bt^2 + ct + d)\sin(2\pi f t)$

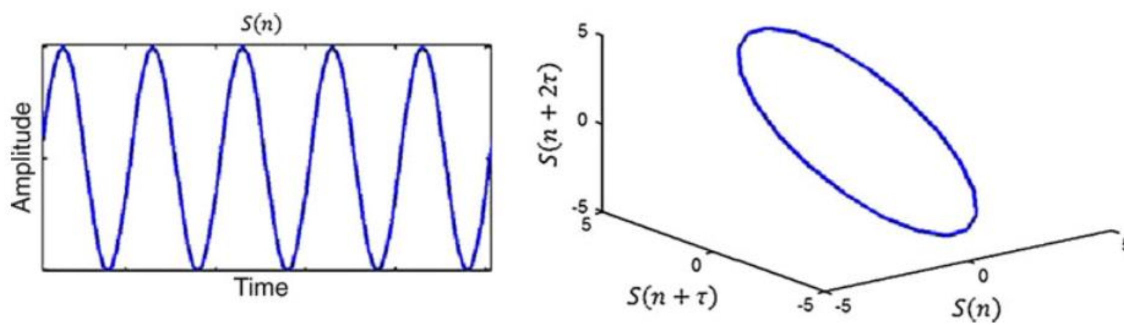


Standard pipeline with an SVM as classifier

- Kinematics: position, speed, acceleration, pressure, distance from the pen to the table's surface, etc.

- Geometrical: spiral's trajectory is modeled as an amplitude-modulated signal: $r(t) = (at^3 + bt^2 + ct + d)\sin(2\pi f t)$

- Non-linear: complexity and entropy measures computed upon embedded attractors

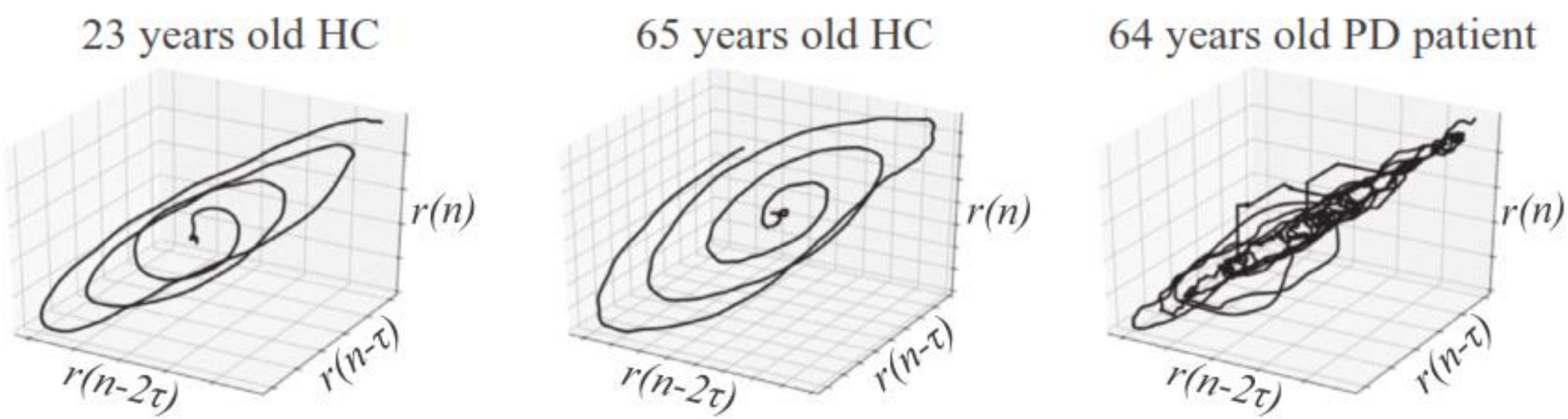


Standard pipeline with an SVM as classifier

- Kinematics: position, speed, acceleration, pressure, distance from the pen to the table's surface, etc.

- Geometrical: spiral's trajectory is modeled as an amplitude-modulated signal: $r(t) = (at^3 + bt^2 + ct + d)\sin(2\pi f t)$

- Non-linear: complexity and entropy measures computed upon embedded attractors



	PD vs. eHC	**PD vs. yHC**
Accuracy	89 %	94 %

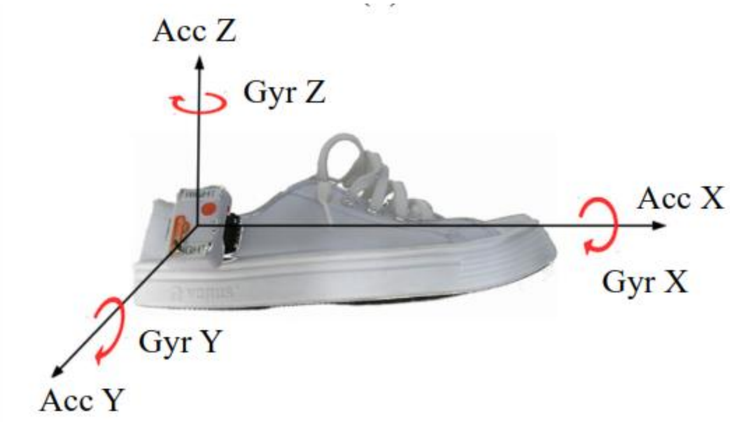
Good results, but …

generalizable?

easy to administer?

Main drawback: several databases worldwide collected with different settings and acquisition protocols.

Gait modeling



eGaIT system

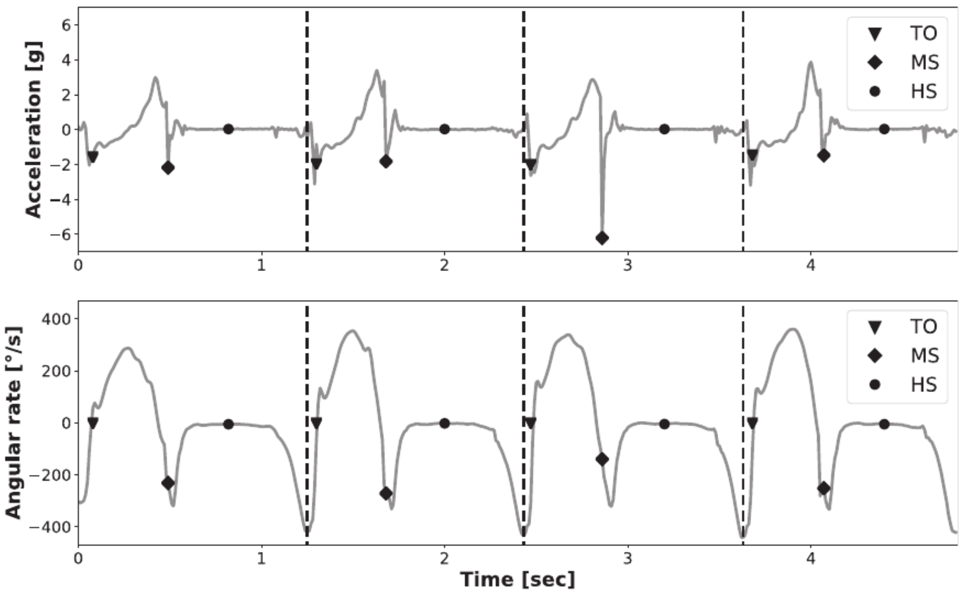
ASTRUM IT

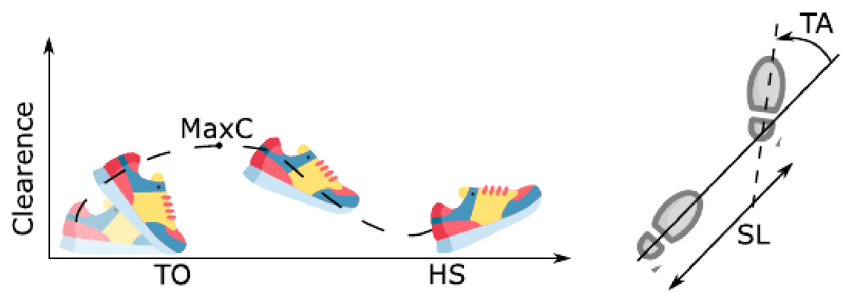
FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Task

4x10m walk

Dataset: 45 PD, 45 elderly HC (eHC), and 44 young HC (yHC)





TA: Turning angle; SL: stride length.

TO: Toe Off; MD: Midstance; HS: Heel strike.

H. Carvajal-Castaño, J.D. Lemos-Duque, and J.R. Orozco-Arroyave, "Effective detection of abnormal gait patterns in Parkinson's disease patients using kinematics, nonlinear, and stability gait features", Human movement science, 81: 1-33, 2022.

Dataset: 45 PD, 45 elderly HC (eHC), and 44 young HC (yHC)

Kinematics

Stride time
Swing time
Stance time
Stride length (SL)
Stride velocity
Turning angle

Nonlinear Dynamics

Lempel-Ziv complexity
Entropy
Hurst Exponent

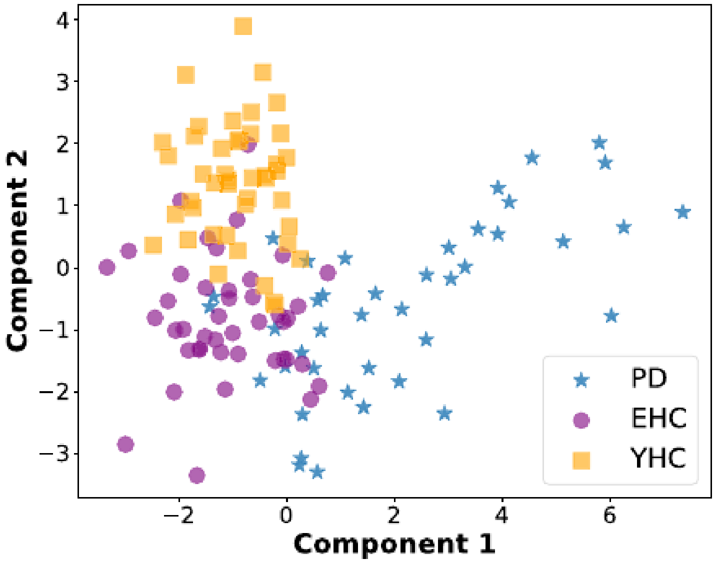
Stability (S)

Temporal S ~ jitter
Amplitude S ~ shimmer
LogEn

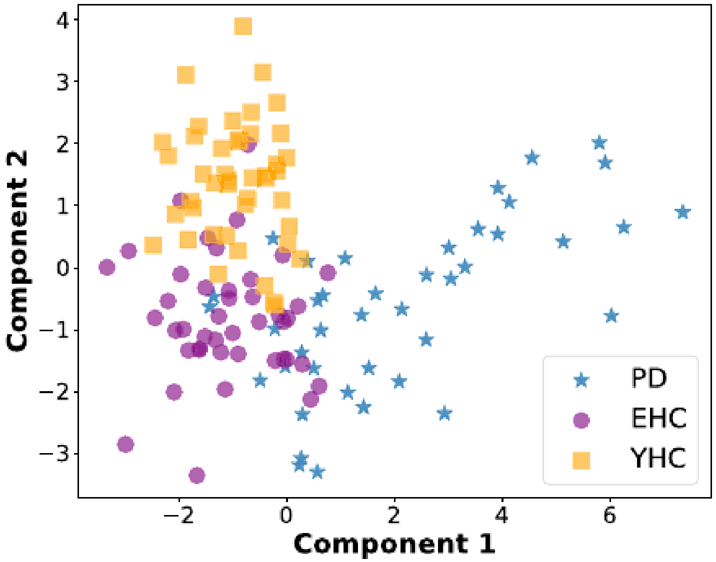
	PD vs. eHC	PD vs. yHC
Accuracy (4X10m)	81 %	88 %

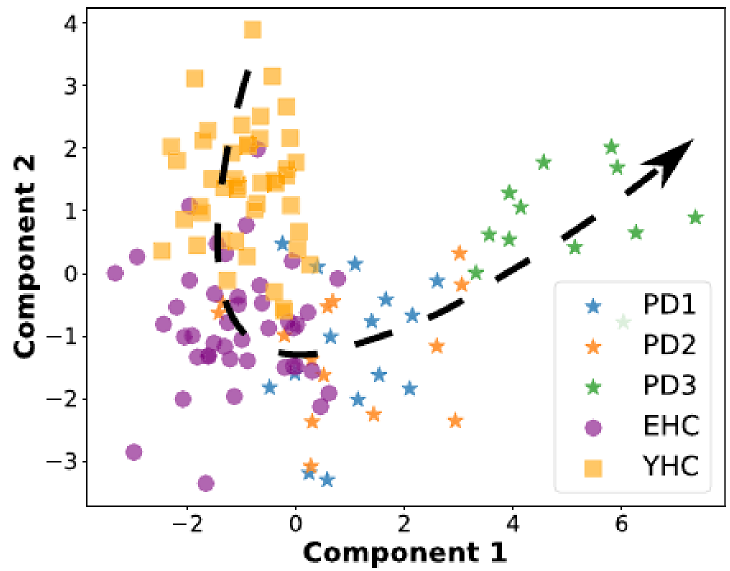
H. Carvajal-Castaño, J.D. Lemos-Duque, and J.R. Orozco-Arroyave, "Effective detection of abnormal gait patterns in Parkinson's disease patients using kinematics, nonlinear, and stability gait features", Human movement science, 81: 1-33, 2022.

Dataset: 45 PD, 45 elderly HC (eHC), and 44 young HC (yHC)



Dataset: 45 PD, 45 elderly HC (eHC), and 44 young HC (yHC)





PD1: initial stage

PD2: intermediate stage

PD3: advanced stage

Agenda

1. Introduction
2. Classical approaches
3. Speech and movement
4. Transitions in facial expressions
5. Speech and language analysis
6. Summary and outlook

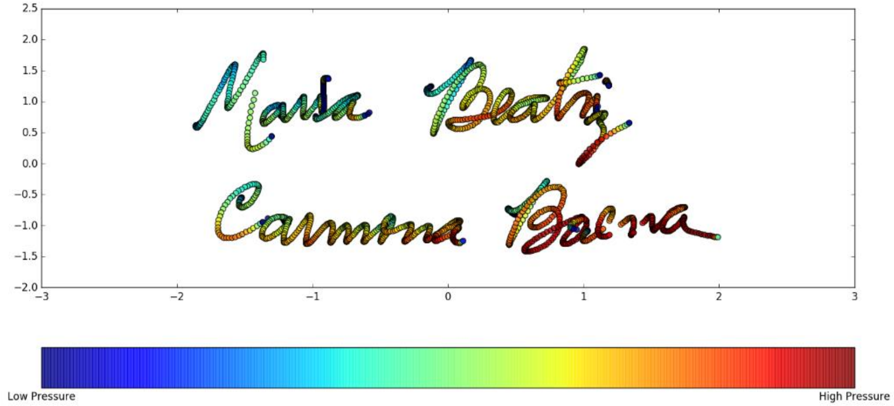
3. Speech and movement

Results with speech seem reasonable and relatively good in performance.

Why should we incorporate other biosignals/modalities?

3. Speech and movement

- MDS-UPDRS-III: 43
- Normal gait
- Normal handwriting
- Normal speech
- Left arm out-of-control

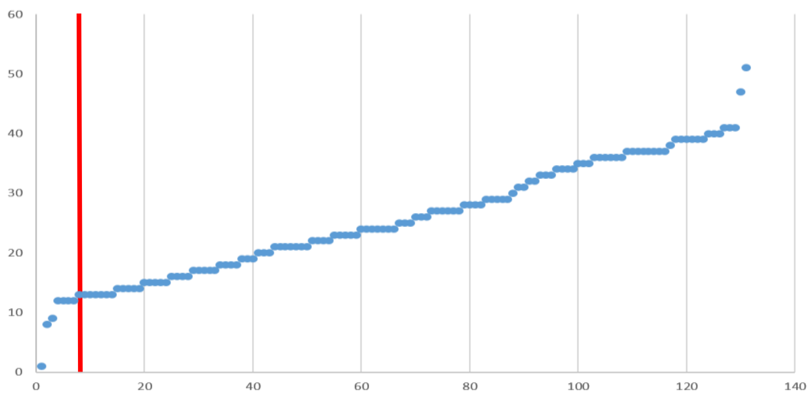


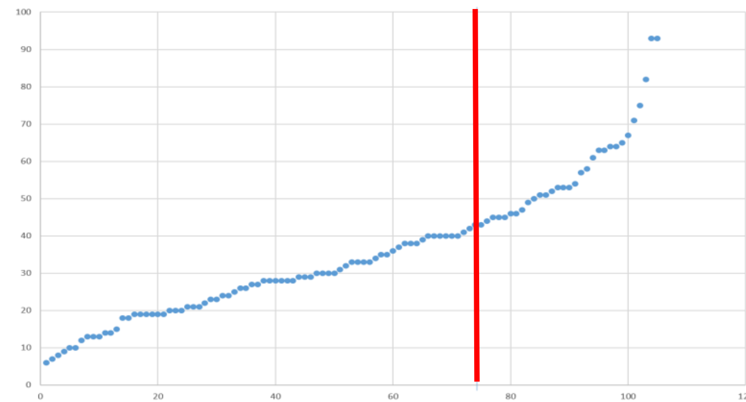
Not all patients reflect symptoms in the same biosignals

Dysarthria (m-FDA)

Same patient

MDS- UPDRS III





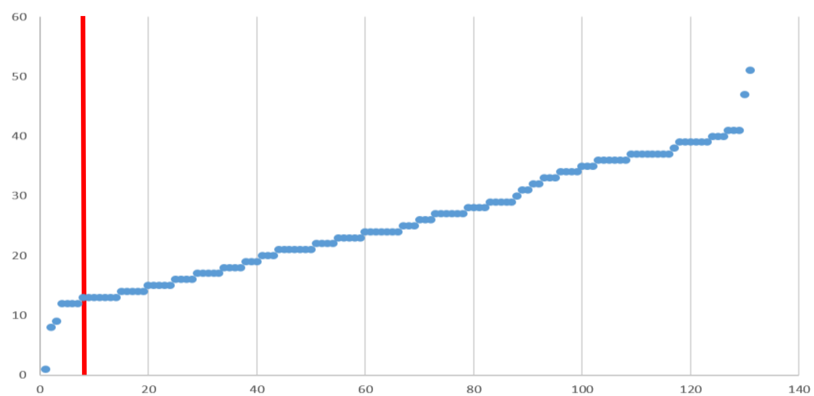
Patients

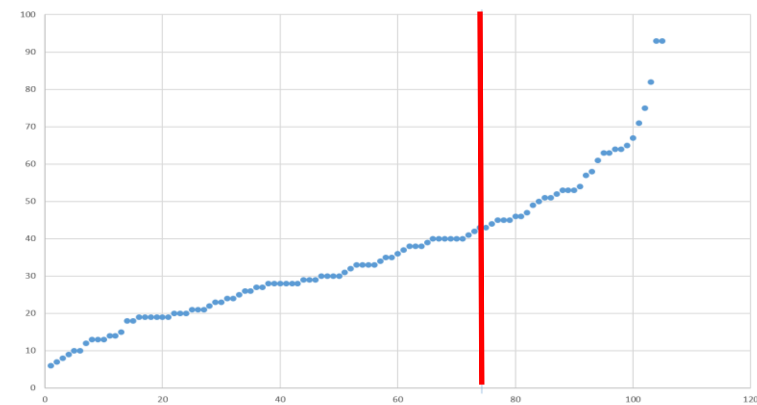
Not all patients reflect symptoms in the same biosignals

Dysarthria (m-FDA)

Same patient

MDS- UPDRS III





Patients

Not all patients are affected in all modalities

Multimodal evaluation is required!

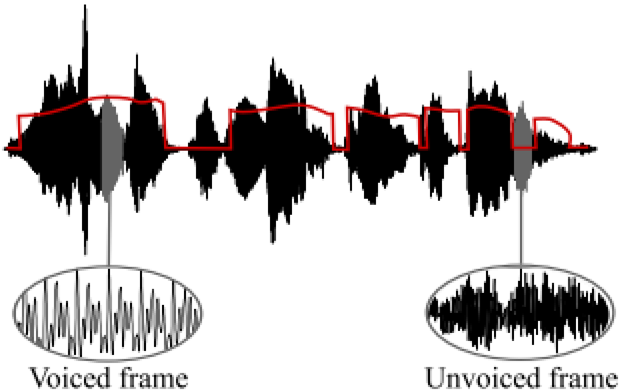
We wanted to propose an interpretable approach

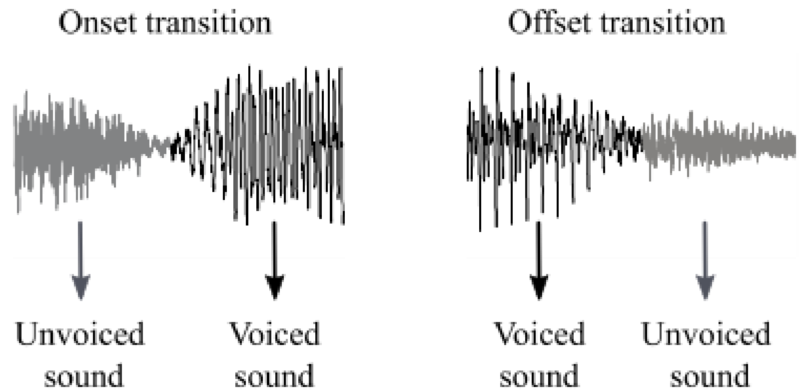
To model difficulties of PD patients to start/stop movements

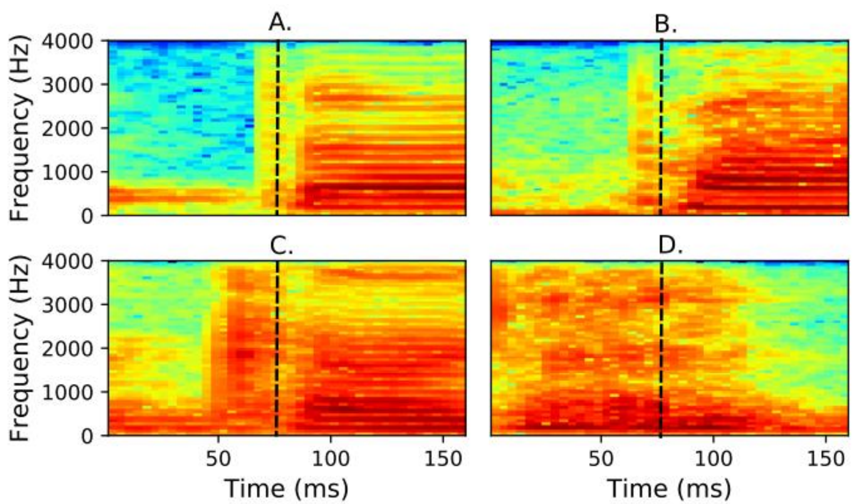
- Transitions in speech
- Transitions in gait
- Transitions in handwriting
- Transitions in facial expressions production

Why to look at the transitions?

- Transitions in speech





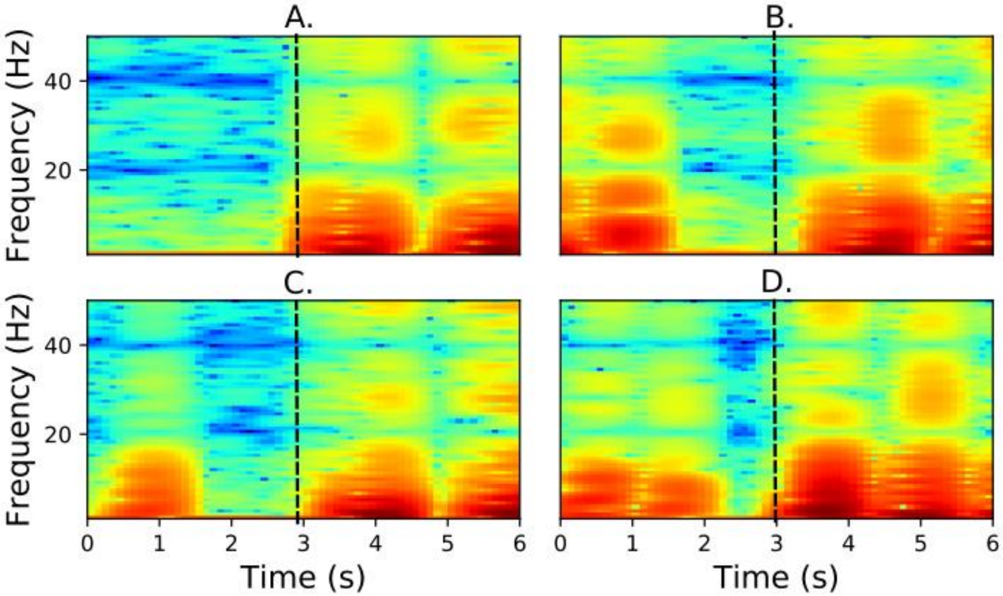


A. HC subject
B. PD in initial state
C. PD in intermediate state
D. PD in severe state

- The spectrograms of the transitions were used as input to the CNN, i.e., 1 channel.

- Each transition contains 80ms to each side (chunks of 160ms)

- Transitions in gait



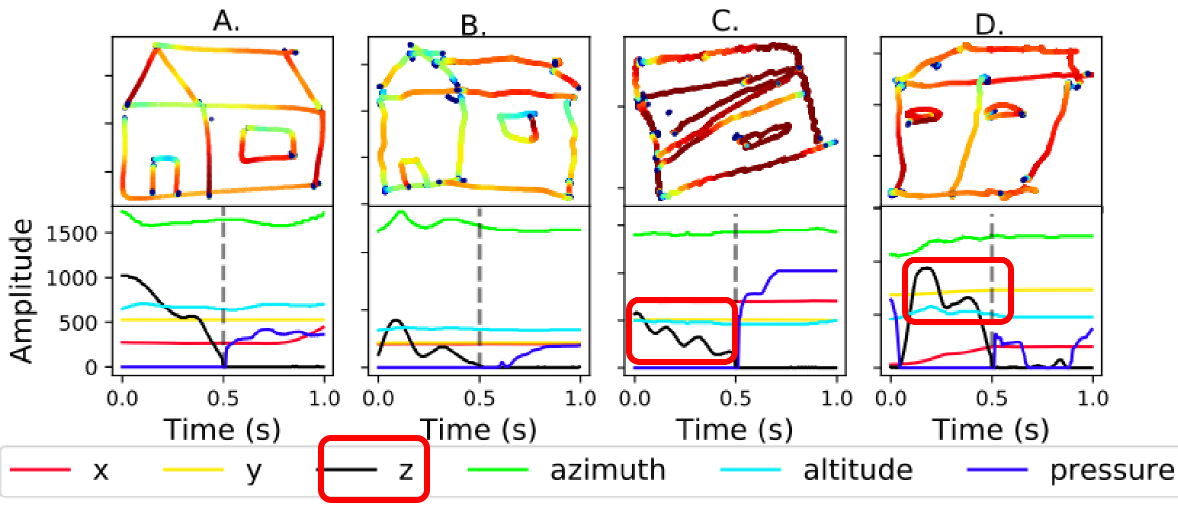
A. HC subject
B. PD in initial state
C. PD in intermediate state
D. PD in severe state

- Six signals per foot: 3D accelerometer and 3D gyroscope

- The spectrograms of each signal in the gait onset and offset are used as input to the CNN, i.e., 12 channels.

- Each transition contains 3s to each side -> 6s per chunk

- Transitions in handwriting



A. HC subject
B. PD in initial state
C. PD in intermediate state
D. PD in severe state

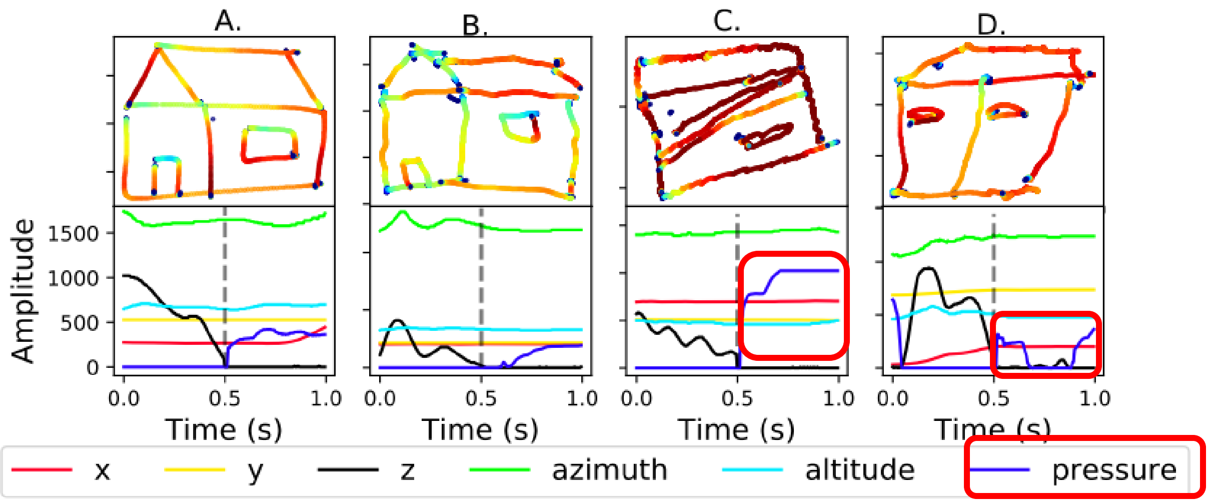
- Transition appears when the patient takes-off the tablet's surface after drawing a stroke (**offset**) and when the patient starts a stroke (**onset**)

- 8 signals are collected: x-position, y-position, z-position, pressure, azimuth angle, altitude angle, on-surface trajectory, and angle of the trajectory

- The raw version of the signals and their derivatives are used as input to the CNN, i.e., 16 channels

- Each transition contains 200ms to each side -> 400ms per chunk

- Transitions in handwriting



A. HC subject
B. PD in initial state
C. PD in intermediate state
D. PD in severe state

- Transition appears when the patient takes-off the tablet's surface after drawing a stroke (**offset**) and when the patient starts a stroke (**onset**)

- 8 signals are collected: x-position, y-position, z-position, pressure, azimuth angle, altitude angle, on-surface trajectory, and angle of the trajectory

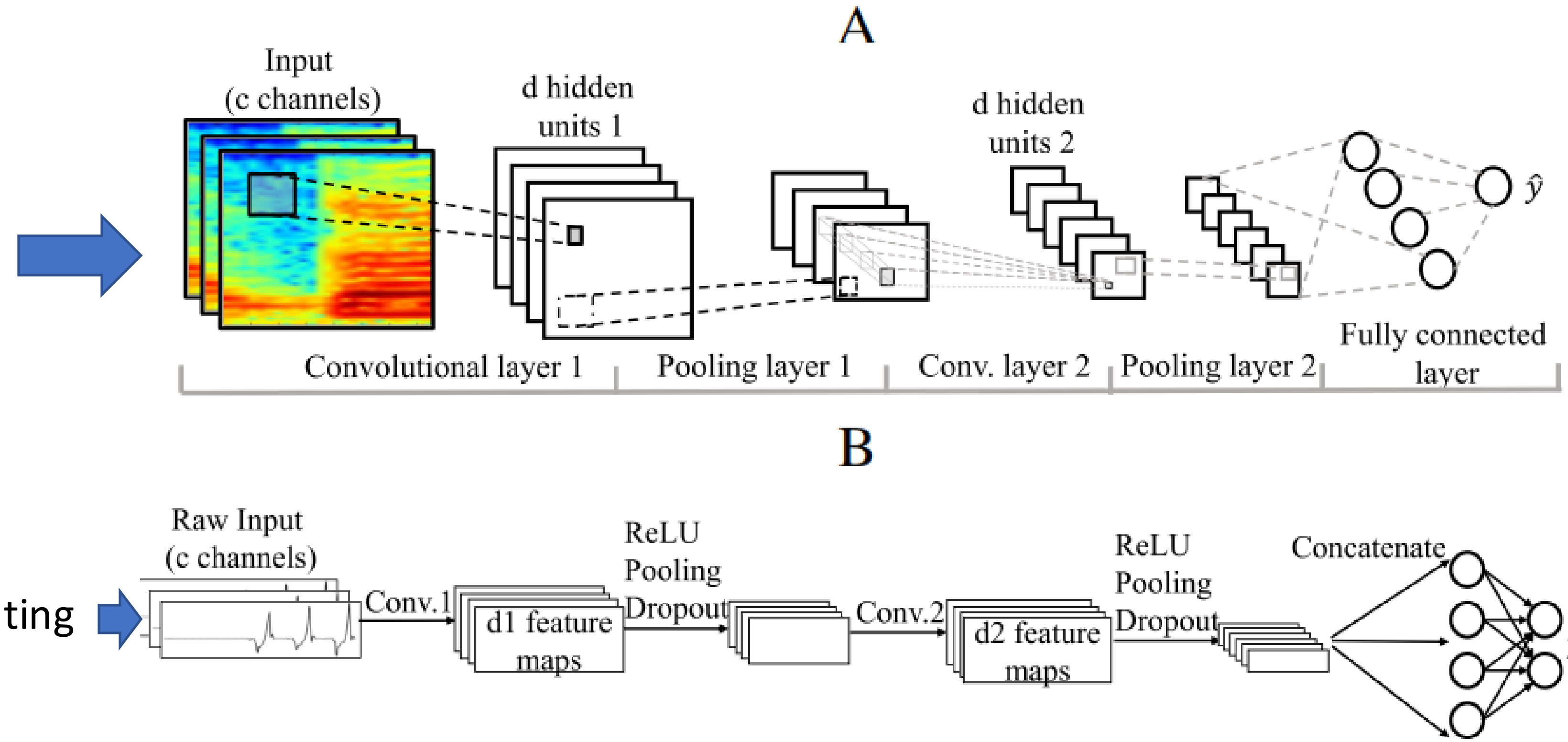
- The raw version of the signals and their derivatives are used as input to the CNN, i.e., 16 channels

- Each transition contains 200ms to each side -> 400ms per chunk

Multimodal architecture

Speech

Gait



Handwriting

- Stochastic Gradient Descent
- Loss function: cross-entropy
- Activation function: Rectifier Linear Unit (ReLU)
- Dropout in training to avoid over-fitting

JC Vásquez-Correa et al., «Multimodal assessment of Parkinson's disease: a deep learning approach» IEEE Journal of Biomedical and Health Informatics, 23(4): 21-36, 2019.

Results obtained with the multimodal architecture

Bio-signal	Acc. Test	Acc. Dev.	AUC	N.
Speech baseline	74.5±1.7	77.0±2.4	0.841	
Speech onset	92.3±12.3	99.4±0.7	0.963	140055
Speech offset	83.5±6.6	99.1±0.7	0.925	135389
Gait baseline	63.0±8.9	66.0±3.1	0.725	
Gait onset	80.3±10.3	83.3±8.9	0.878	326977
Gait offset	78.8±16.0	87.8±5.1	0.901	1231016
Handwriting baseline	67.1±4.2	67.7±1.7	0.725	
Handwriting onset	60.4±3.5	95.7±4.0	0.634	142517
Handwriting offset	66.5±5.5	98.1±1.7	0.699	255560
Fusion baseline	89.0±7.8	87.8±3.1	0.944	
Fusion onset	**97.6±2.9**	98.8±0.6	0.988	609549
Fusion offset	84.3±5.8	86.0±1.4	0.890	1621965

JC Vásquez-Correa et al., «Multimodal assessment of Parkinson's disease: a deep learning approach» IEEE Journal of Biomedical and Health Informatics, 23(4): 21-36, 2019.

Agenda

1. Introduction
2. Classical approaches
3. Speech and movement
4. Transitions in facial expressions
5. Speech and language analysis
6. Summary and outlook

4. Transitions in facial expressions (hypomimia)

Datasets

- Face recognition: VGGFace2 -> 3,31 millions of faces from 9.131 subjects
- Facial expressions - Emotions: EmotioNet -> 950.000 faces with 12 Action Units
- Parkinson: FacePark-GITA -> 30 PD patients and 24 healthy subjects.
    - **Videos** with 30fps, non-controlled environments.
    - Patients and controls are matched by age and gender.

L.F. Gómez-Gómez, A. Morales, J. Fierrez, and JR. Orozco-Arroyave «Exploring Facial Expressions and Affetive Domains for Parkinson Detection» PLoS ONE, 18(2): 1-25, 2023.

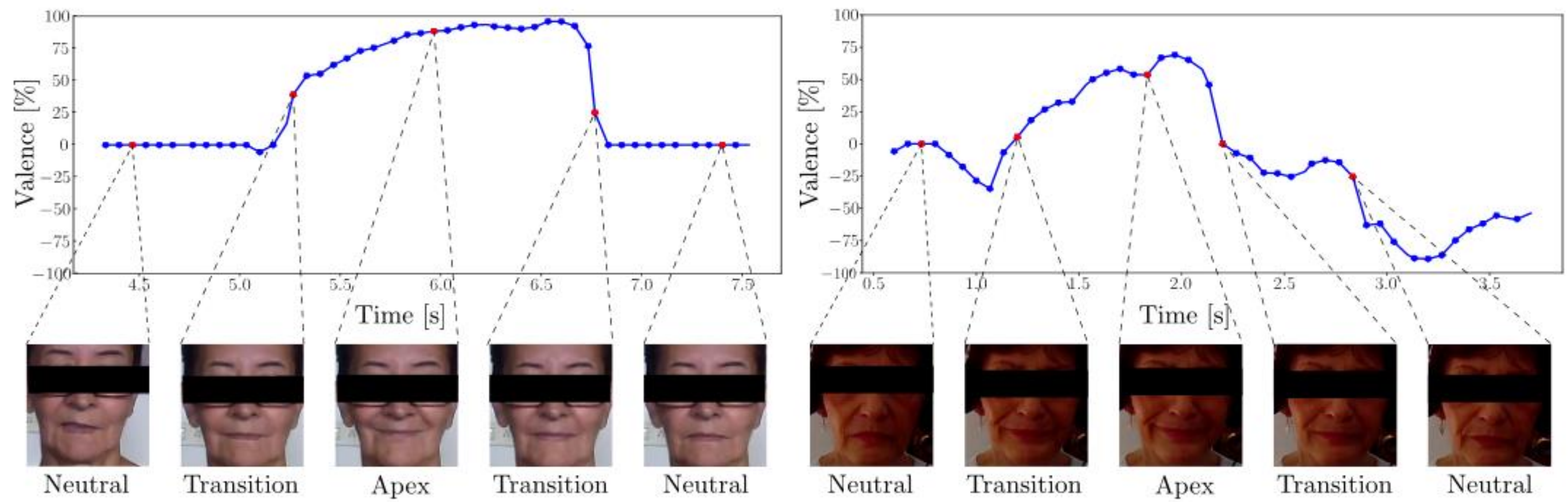


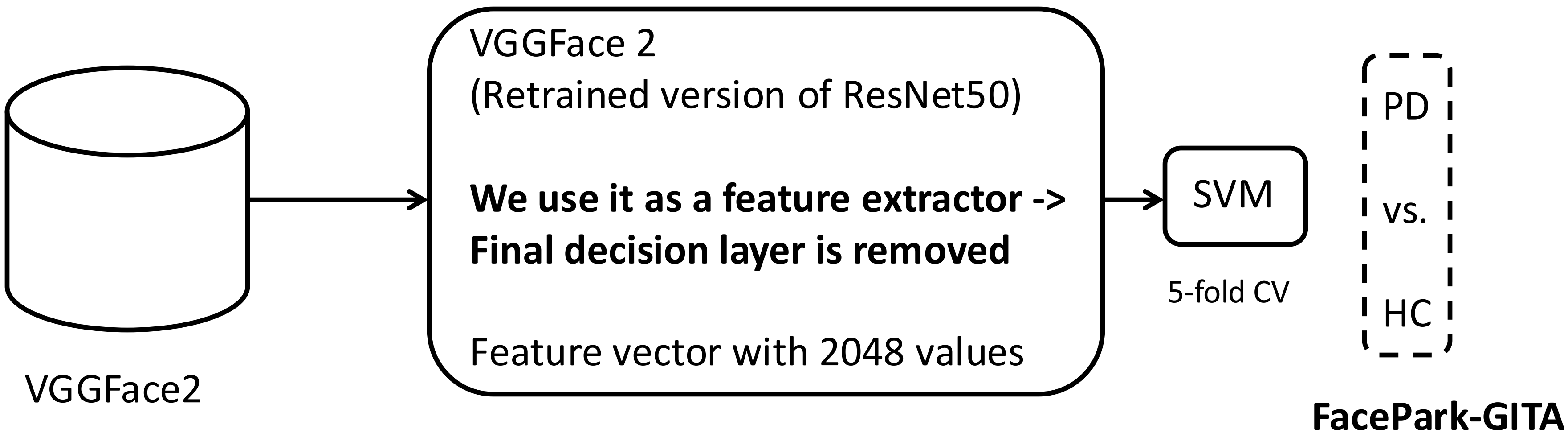
Figure 2: Emotion stages according to the evoked valence measured with the Affectiva tool. (left) Healthy woman 63 years old; (right) Woman with Parkinson's disease, 67 years old, facial expression item = 2.

In this case transitions are extracted from the Valence % using the software Affectiva*: Neutral (N), Transition (Onset), Apex (A), Transition (Offset), and Neutral (N).

* https://www.affectiva.com/

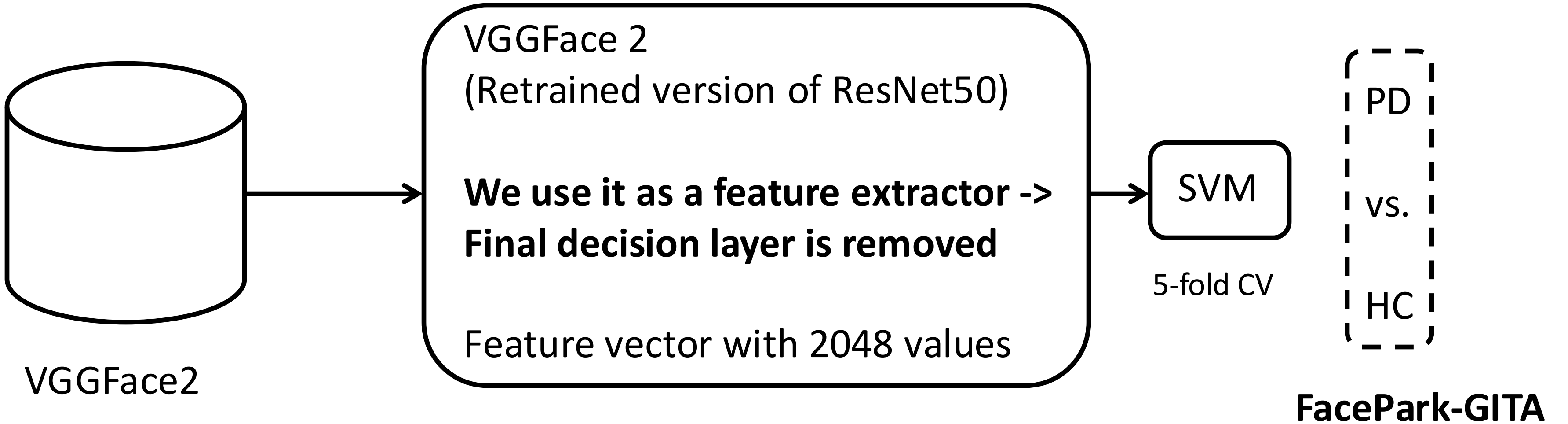
Experiment 1: face recognition level

Single images vs. sequence of images



Experiment 1: face recognition level

Single images vs. sequence of images



E.S.	Kernel*	Acc[%]	Sens[%]	Spec[%]	F1[%]
Neutral	C=1e+01; γ=1e-04	69.0 ± 10.1	74.0 ± 11.6	63.0 ± 9.7	67.8 ± 10.1
Apex	C=1e+01; γ=1e-04	70.0 ± 9.1	84.4 ± 7.9	53.3 ± 24.0	61.0 ± 18.6
Onset	C=1e+01; γ=1e-04	71.4 ± 3.2	88.6 ± 7.0	50.0 ± 9.0	63.1 ± 6.6
Offset	C=1e+01; γ=1e-04	71.6 ± 5.2	79.5 ± 3.3	61.9 ± 13.5	68.6 ± 8.2
Neutral	C=1e-03	70.8 ± 9.6	77.3 ± 10.2	63.0 ± 9.7	69.3 ± 9.7
Apex	C=1e-03	70.8 ± 9.1	83.7 ± 7.3	55.7 ± 21.6	63.8 ± 16.3
Onset	**C=1e-02**	**72.9 ± 4.2**	**88.6 ± 7.8**	**53.4 ± 7.7**	**66.1 ± 5.9**
Offset	C=1e-01	72.8 ± 4.3	81.5 ± 4.5	61.9 ± 13.5	69.2 ± 7.9

E.S.: Expression stage. First three rows: Gaussian kernel. Last three rows: Linear kernel.
*Column with optimal hyper-parameters.

With sequences of the transitions there is an **improvement of about 6%**

Sequences	Kernel*	Acc[%]	Sens[%]	Spec[%]	F1[%]
NOnA	C=1e+02; γ=1e-04	77.4 ± 8.7	89.3 ± 4.6	63.0 ± 16.1	72.9 ± 11.2
AOffN	C=1e+01; γ=1e-04	76.3 ± 8.0	86.8 ± 12.0	63.5 ± 22.4	69.2 ± 17.8
NOnAOffN	C=1e+01; γ=1e-04	77.2 ± 8.6	86.1 ± 14.8	67.2 ± 10.3	74.2 ± 8.5
NOnA	C=1e-03	78.2 ± 9.8	90.1 ± 5.2	63.8 ± 17.1	73.8 ± 12.6
AOffN	C=1e-03	77.8 ± 9.1	88.8 ± 9.4	64.2 ± 24.1	70.4 ± 20.5
NOnAOffN	**C=1e-03**	**78.4 ± 7.1**	**87.8 ± 11.4**	**67.7 ± 11.6**	**75.4 ± 7.9**

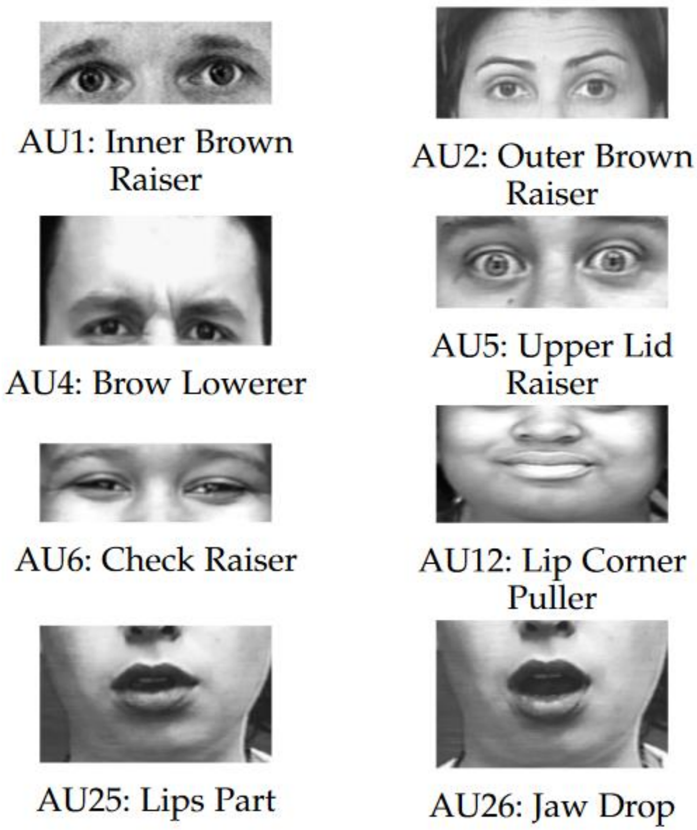
First three rows: Gaussian kernel. Last three rows: Linear kernel.

*Column with optimal hyper-parameters.

Early fusion was used to merge the information from the sequences

Experiment 2: transfer learning from the affective domain

Base model: 8 Action Units (AUs) from the **EmotioNet database**



The ResNet50-based model is retrained with selected AUs.
This model is complemented with information from EmotioNet by freezing some layers. Three TL strategies:

- Freeze 50: the remaining 50% is retrained with EmotioNet

- Freeze 75: the remaining 25% is retrained with EmotioNet

- Freeze 100 -> original FR model (baseline)

FacePark-GITA is only used to train the SVM

Experiment 2: transfer learning from the affective domain (cont.)

The model was first tested on the automatic classification of AUs

Models	**Metrics**	**AU 1**	**AU 2**	**AU 4**	**AU 5**	**AU 6**	**AU 12**	**AU 25**	**AU 26**
Baseline ($\mathbf{x}_{FR}$)	AUC	0.83	0.83	0.87	0.80	0.94	0.95	0.92	0.80
	EER [%]	24.58	23.78	21.01	27.13	12.82	12.11	15.38	27.32
Freeze 75 ($\mathbf{x}_{AF}$)	AUC	0.84	0.84	0.86	0.84	0.92	0.93	0.95	0.85
	EER [%]	21.84	20.80	19.90	21.65	14.34	10.42	8.63	22.48
Freeze 50 ($\mathbf{x}_{AF}$)	AUC	0.84	0.87	0.87	0.87	0.93	0.95	0.90	0.83
	EER [%]	20.56	19.29	18.92	19.53	13.22	10.58	10.99	24.32

Experiment 2: transfer learning from the affective domain (cont.)

Representations obtained from the freezing of layers are further used to discriminate between PD vs. HC. (**Freeze 75 model)**

Sequence	Kernel*	Acc[%]	Sens[%]	Spec[%]	F1[%]
NOnA	C=1e+01; γ=1e-04	84.2 ± 5.4	90.0 ± 8.3	77.2 ± 10.8	82.3 ± 6.3
AOffN	C=1e+02; γ=1e-04	81.6 ± 8.6	87.8 ± 7.4	73.9 ± 11.5	80.0 ± 9.5
NOnAOffN	C=1e+02; γ=1e-04	86.7 ± 8.9	91.2 ± 4.7	81.6 ± 15.5	85.5 ± 10.2
NOnA	C=1e-01	84.7 ± 5.4	89.5 ± 9.4	78.9 ± 11.3	82.9 ± 6.5
AOffN	C=1e-01	82.6 ± 9.6	87.8 ± 8.3	76.1 ± 13.3	81.2 ± 10.4
NOnAOffN	**C=1e-01**	**87.3 ± 8.0**	**90.6 ± 5.0**	**83.6 ± 13.1**	**86.6 ± 8.8**

Improvement of around 9% w.r.t. experiment 1

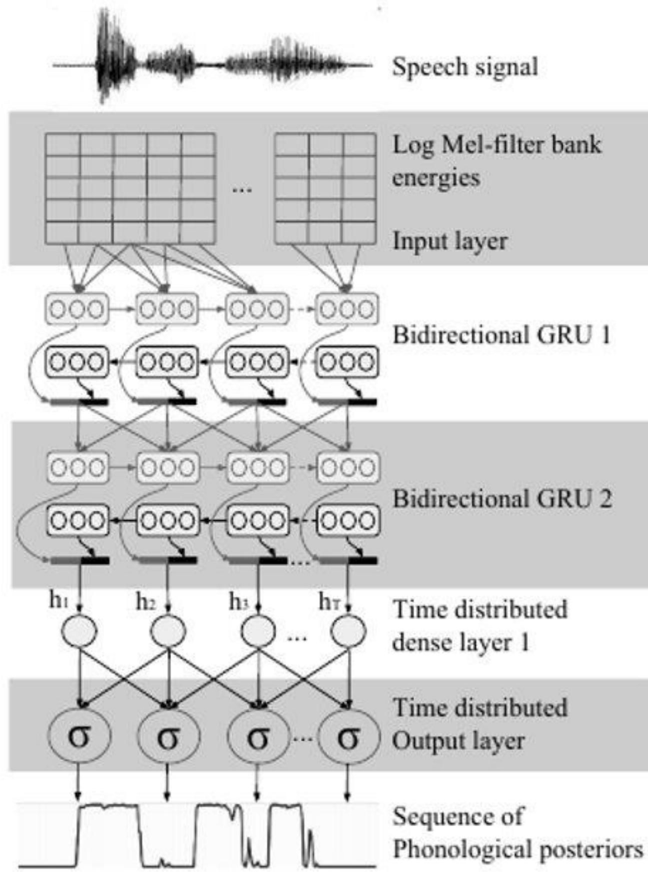
Agenda

1. Introduction
2. Classical approaches
3. Speech and movement
4. Transitions in facial expressions
5. Speech and language analysis
6. Summary and outlook

5. Speech and language analysis

Easy to interpret speech models

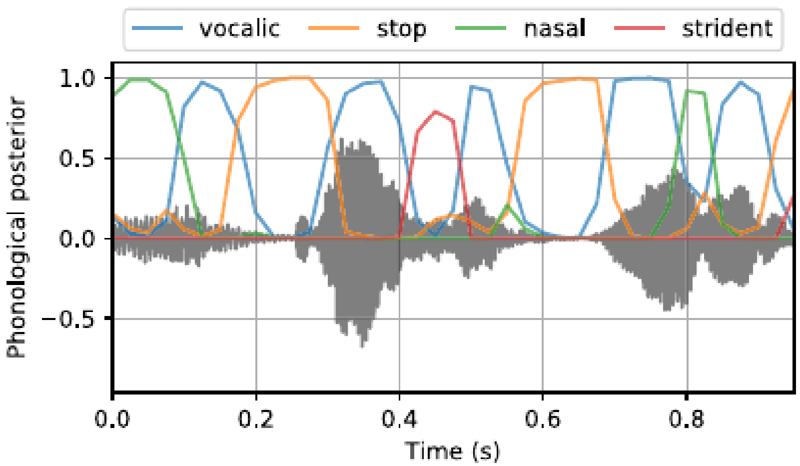
Phonemic identifiability

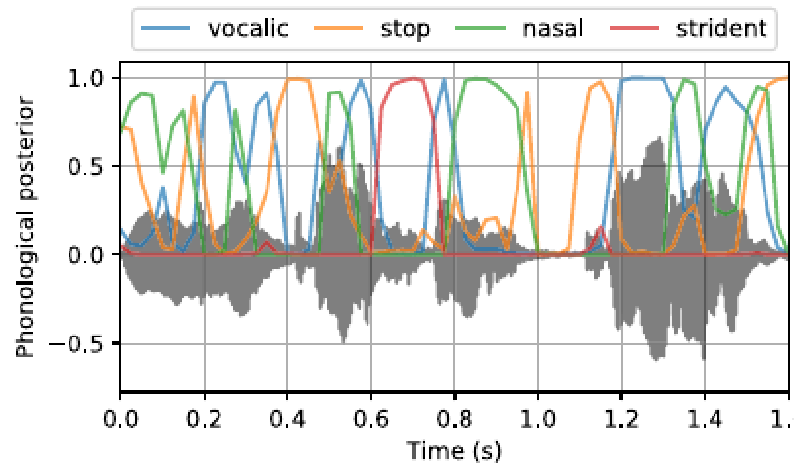


Bidirectional GRUs and time-distributed layers => sequence to sequence model

Easy to interpret speech models

Phonemic identifiability

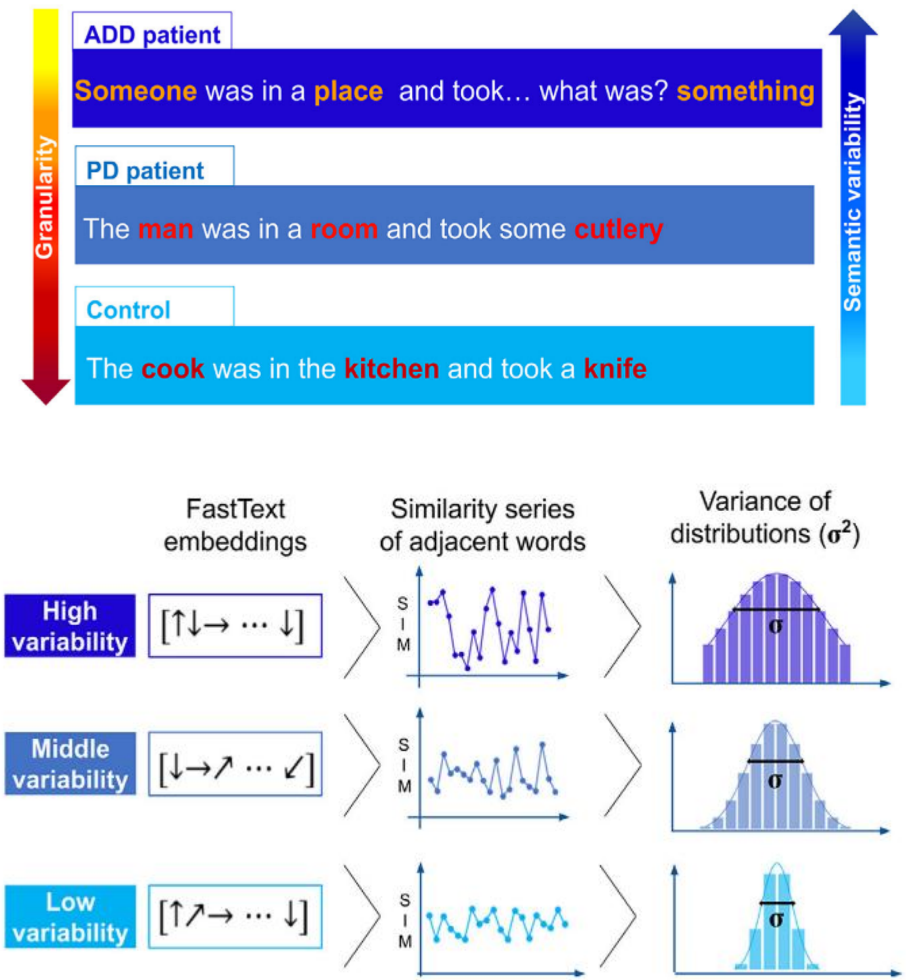




General accuracy of 80% (PD vs. HC)

With **clinical** interpretability!

Easy to interpret language models



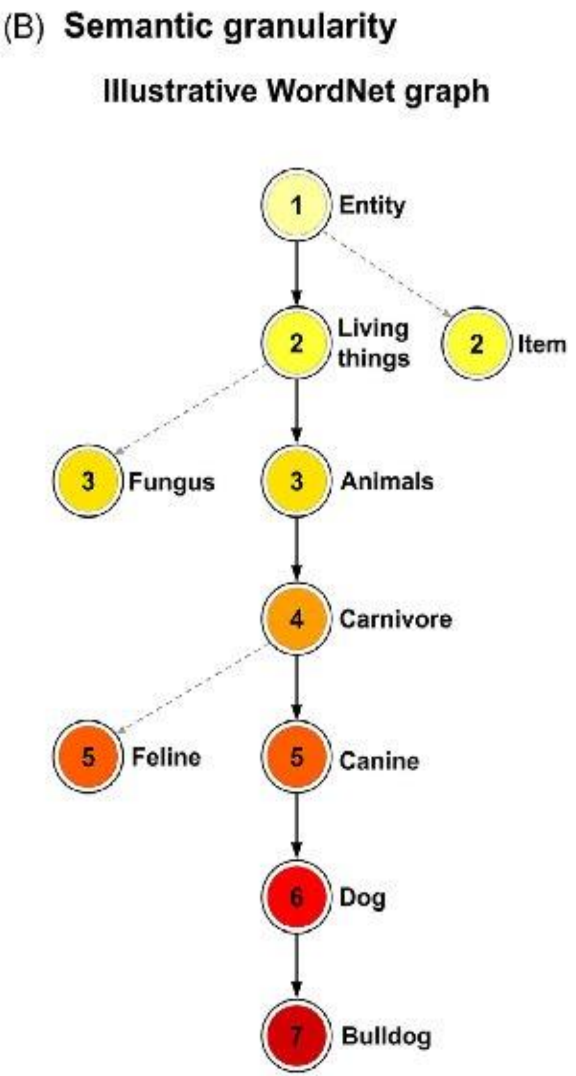


Figure adapted from:
C. Sanz et al. "Automated text-level semantic markers of Alzheimer's disease" Alzheimers Dementia (Amst), 14(1):e12276, 2022.

Different dimensions of analysis (language)

Lexico-semantics: Link of word forms with context-sensitive conceptual information.

Morphosyntax: Morphological processes (word-formation) and syntactic patterning (e.g., word sequencing and hierarchization).

Discourse-level processing: Language production with information-rich contexts and cultural expectations.

Examples of speech/language analysis

Example 1: Cognitive decline evaluation in PD patients

Cohort
40 PD (16 MCI and 24 non-MCI)
40 HC

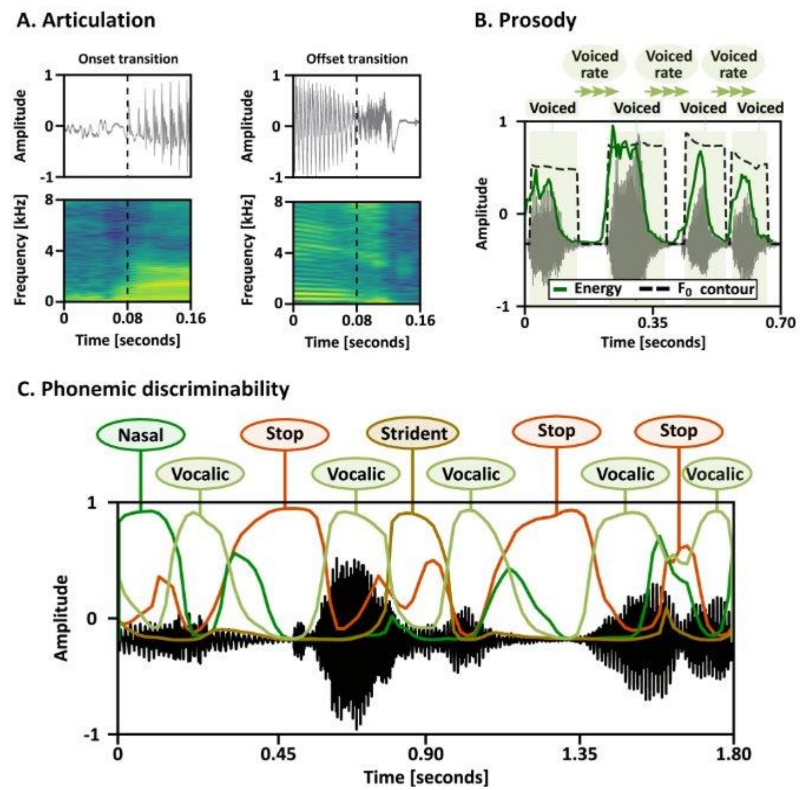
Models
Articulation
Prosody
Phonemic identifiability

MCI: Mild Cognitive Impairment

A.M. García, T. Arias-Vergara, J.C. Vásquez-Correa, E. Nöth, M. Schuster, A.E. Welch, Y. Bocanegra, A. Baena, and J.R. Orozco-Arroyave, "Cognitive determinants of dysarthria in Parkinson's disease: An automated machine learning approach", Movement Disorders, 2021, Aug 14. doi: 10.1002/mds.28751. Epub ahead of print. PMID: 34390508.

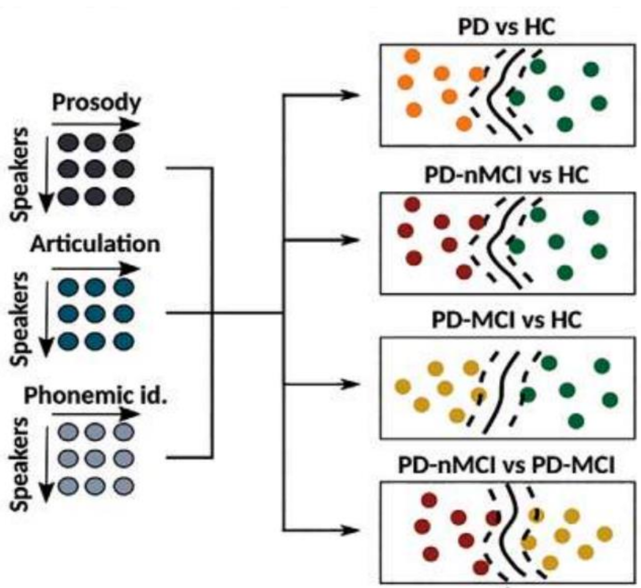
Examples of speech/language analysis

Example 1: Cognitive decline evaluation in PD patients



Examples of speech/language analysis

Example 1: Cognitive decline evaluation in PD patients

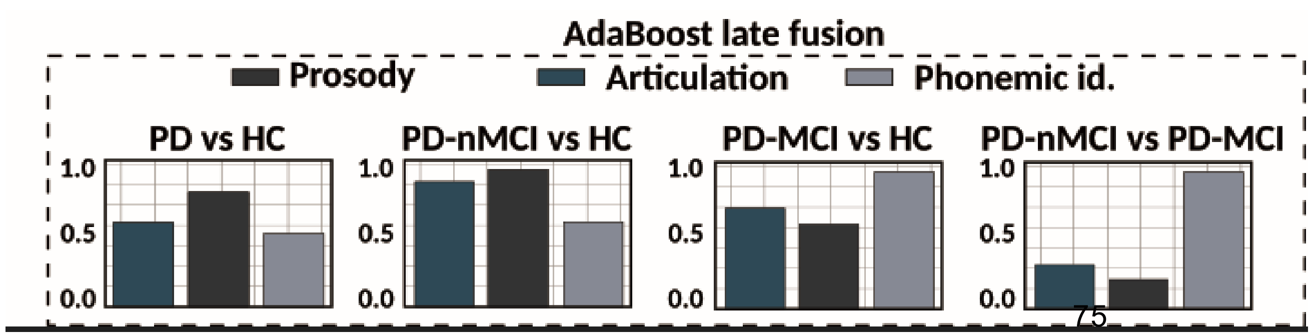


84%, articulation was the best: 75%

80%, articulation was the best: 76%

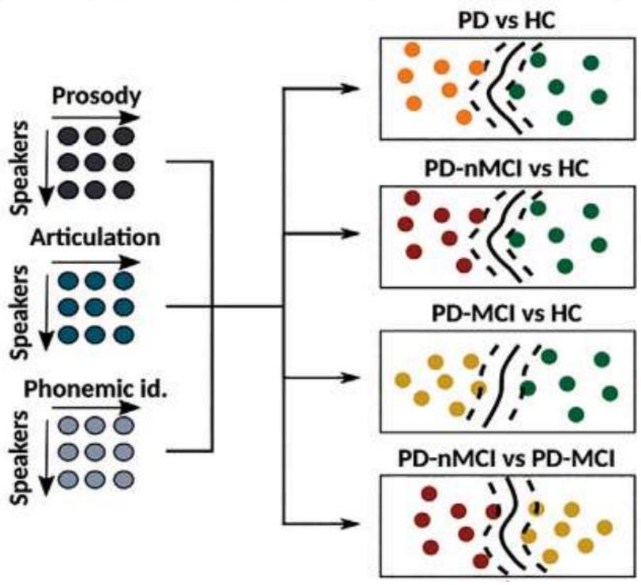
87%, phonemic was the best: 79%

72%, phonemic was the best: 72%



Examples of speech/language analysis

Example 1: Cognitive decline evaluation in PD patients

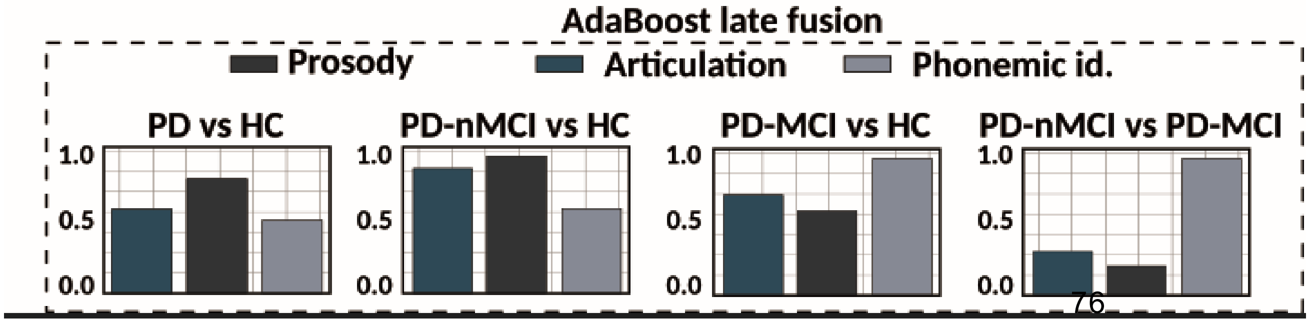


84%, articulation was the best: 75%

80%, articulation was the best: 76%

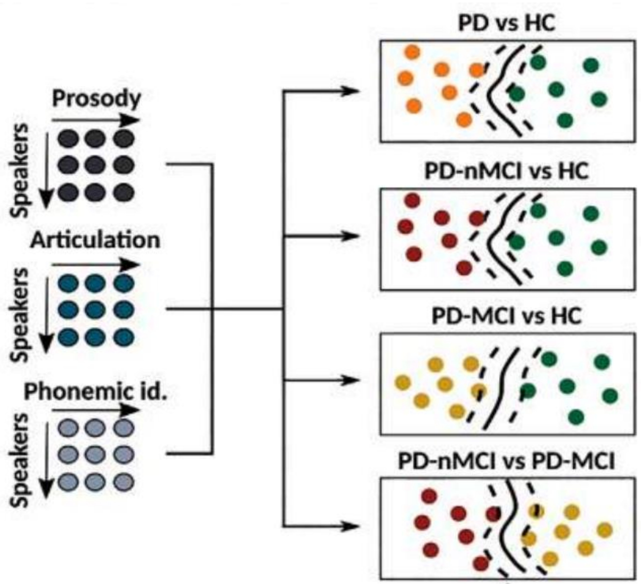
87%, phonemic was the best: 79%

72%, phonemic was the best: 72%



Examples of speech/language analysis

<u>Example 1:</u> Cognitive decline evaluation in PD patients

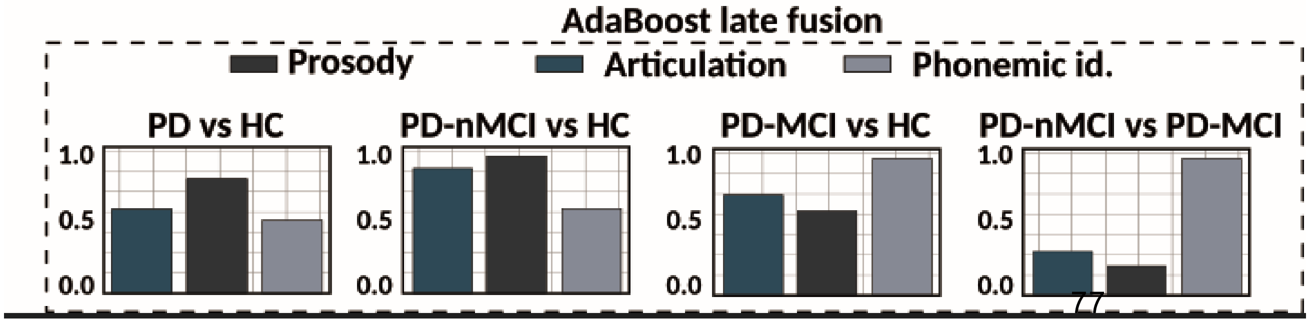


84%, articulation was the best: 75%

80%, articulation was the best: 76%

87%, phonemic was the best: 79%

72%, phonemic was the best: 72%



Examples of speech/language analysis

<u>Example 2:</u> Automated semantic analyses of action stories

Cohort
40 PD (16 MCI and 24 non-MCI)
40 HC

Two texts: Action text (AT) and non-Action text (nAT)

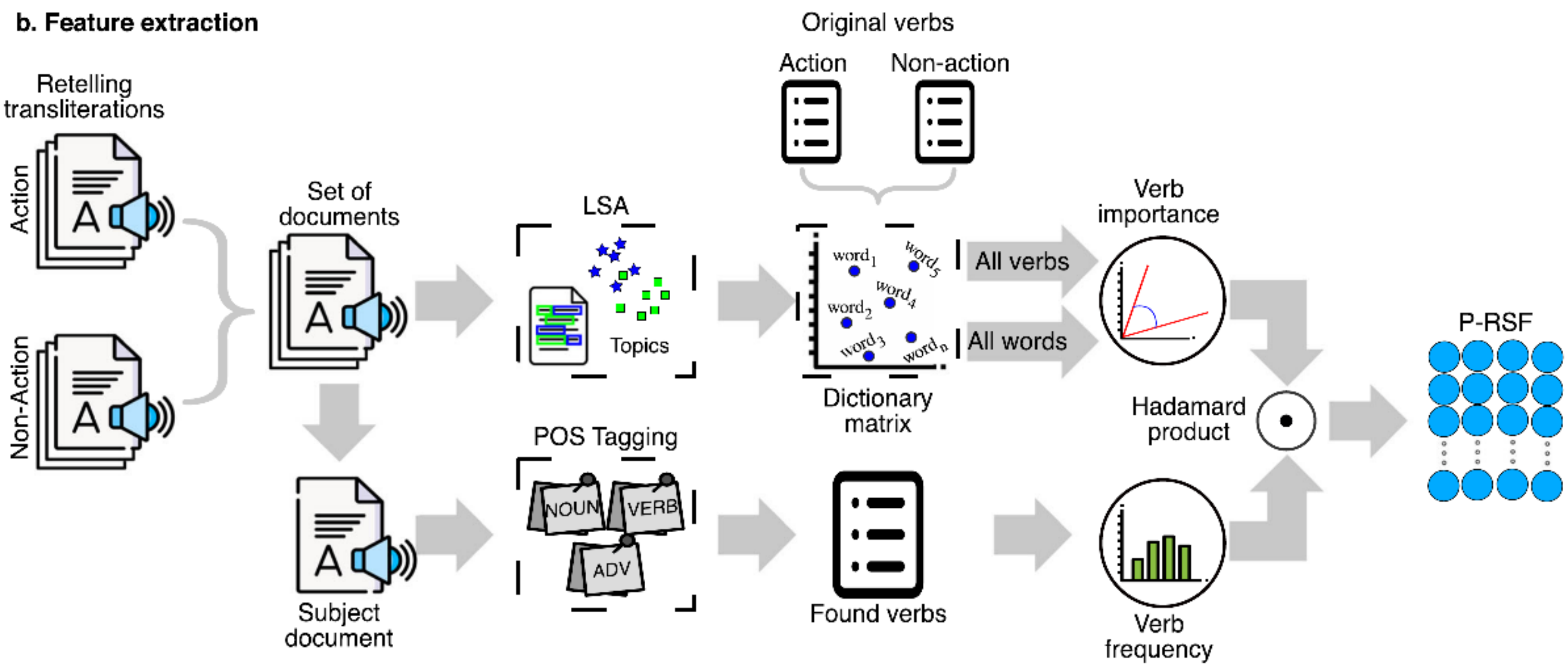
Models
Proximity-to-Reference-Semantic-Field (P-RSF)
Glove

MCI: Mild Cognitive Impairment

A.M. García, D. Escobar-Grisales, J.C. Vásquez-Correa, Y. Bocanegra, L. Moreno, J. Carmona, and J.R. Orozco-Arroyave, "Detecting Parkinson's disease and its cognitive phenotypes via automated semantic analyses of action stories", npj Parkinson's Disease, 2022, (8):163 ; https://doi.org/10.1038/s41531-022-00422-8.

Examples of speech/language analysis

Example 2: Automated semantic analyses of action stories



Examples of speech/language analysis

Example 2: Automated semantic analyses of action stories

Text	PD vs HC	PD-nMCI vs HC	PD-MCI vs HC	PD-nMCI vs PD-MCI
AT	0.80	0.93	0.90	0.82
nAT	0.60	0.55	0.80	0.53

AT: Action Text; nAT: non-Action Text; AUC values.

Examples of speech/language analysis

Example 2: Automated semantic analyses of action stories

Text	PD vs HC	PD-nMCI vs HC	PD-MCI vs HC	PD-nMCI vs PD-MCI
AT	0.80	0.93	0.90	0.82
nAT	0.60	0.55	0.80	0.53

AT: Action Text; nAT: non-Action Text; AUC values.

Examples of speech/language analysis

Example 2: Automated semantic analyses of action stories

Text	PD vs HC	PD-nMCI vs HC	PD-MCI vs HC	PD-nMCI vs PD-MCI
AT	0.80	0.93	0.90	0.82
nAT	0.60	0.55	0.80	0.53

AT: Action Text; nAT: non-Action Text; AUC values.

Agenda

1. Introduction
2. Classical approaches
3. Speech and movement
4. Transitions in facial expressions
5. Speech and language analysis
6. Summary and outlook

5. Summary and outlook

- [Movement] **Transitions (onsets & offsets)** are suitable to model abnormal behavior of different biosignals in PD patients

- [Multimodal] Allows a better understanding of PD symptoms

- [Multimodal] There is still a lot of space for other approaches

- [Face] Further research is required to find better models with smaller architectures: full dynamics (i.e., video) in the transitions using **RNNs or 3D Convolutions**

- [Face + Speech] **Synchronous fusion** of speech and facial movements are among the next steps

- [Data collection and processing] More data are required and **Federated Learning** emerges as an alternative when data sharing is not possible due to privacy reasons

Speech, Language, and Movement Processing to Model Parkinson's Disease

Juan Rafael Orozco-Arroyave (Rafa)

GITA Lab
School of Engineering, University of Antioquia
Medellín, Colombia

gita.udea.edu.co
rafael.orozco@udea.edu.co

UNIVERSIDAD DE ANTIOQUIA
1803

GITA

Invited Talk at:
Conversational AI Reading Group
Quebec AI Institute (Mila)
Montreal, Canada
27.11.2025

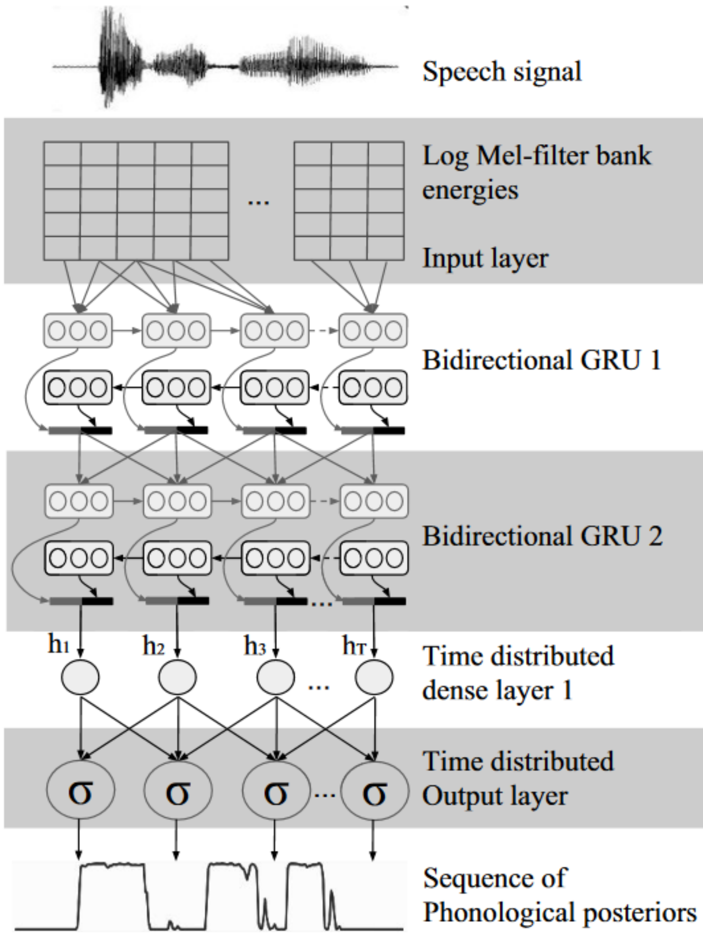
Multimodal architecture

- Speech
  - 1 channel and STFT with 128 points -> 65 frequency indexes
  - Frames of 16ms length and overlap of 4ms -> 40 frames per speech chunk
  - # inputs: 65*40*1 = 2600

- Gait
  - 12 channels and STFT with 128 points -> 65 frequency indexes
  - Frames with 200ms length an overlap of 100ms -> 60 frames per chunk
  - # inputs: 65*60*12 = 46800

- Handwriting
  - 16 channels and raw signals with a sampling rate of 180 Hz.
  - # inputs: 180*16 = 2880

Architecture of Phonet



- Chunks of speech: 500ms
- Widows of speech: 25ms with 10ms step size
- Input features: 33 log-energy of the signal according to the Mel scale

- Two bidirectional GRU layers: to model future (forward) and past (backward) states

- Time-distributed dense layer: fully connected dense layer with shared weights on each time-step -> produces an output sequence with same length as the input (sequence-to-sequence model).
  This gives a one-to-one relation between input and output sequences.

- Time-distributed output: softmax activation function to get the posteriors

JC. Vásquez-Correa, P. Klumpp, JR. Orozco-Arroyave, and E. Nöth, «*Phonet: a Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech*» Proceedings of INTERSPEECH 2019.