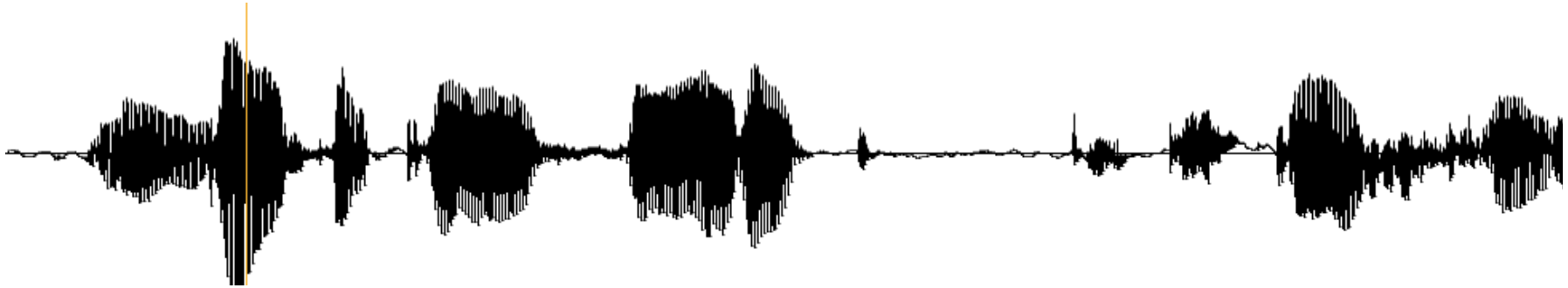
Deep Learning 101 – Speech

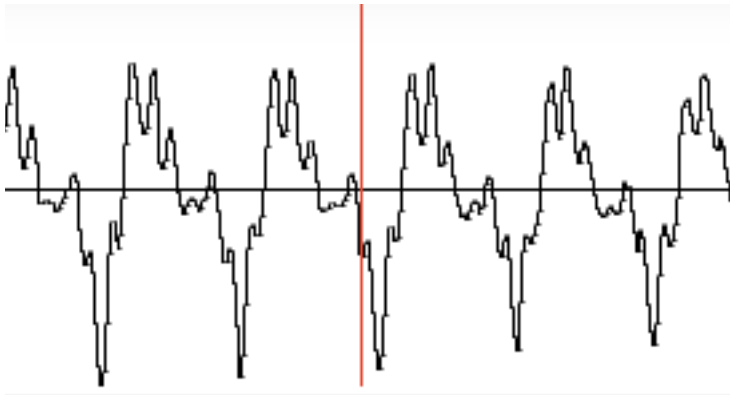
Alberto Abad ■ Catarina Botelho

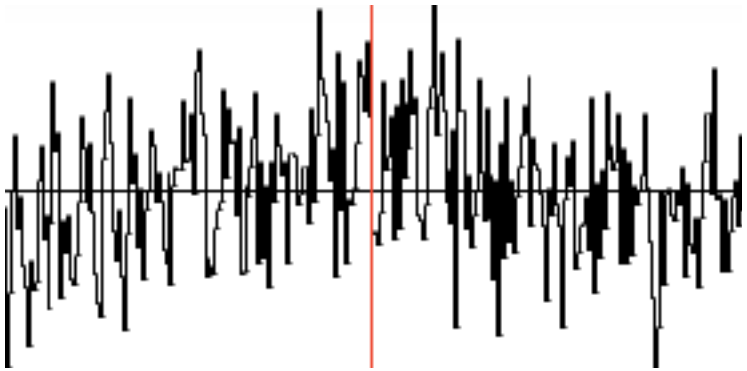
February 2024 – Deep Learning Sessions Portugal

Introduction

Speech signal in the time domain

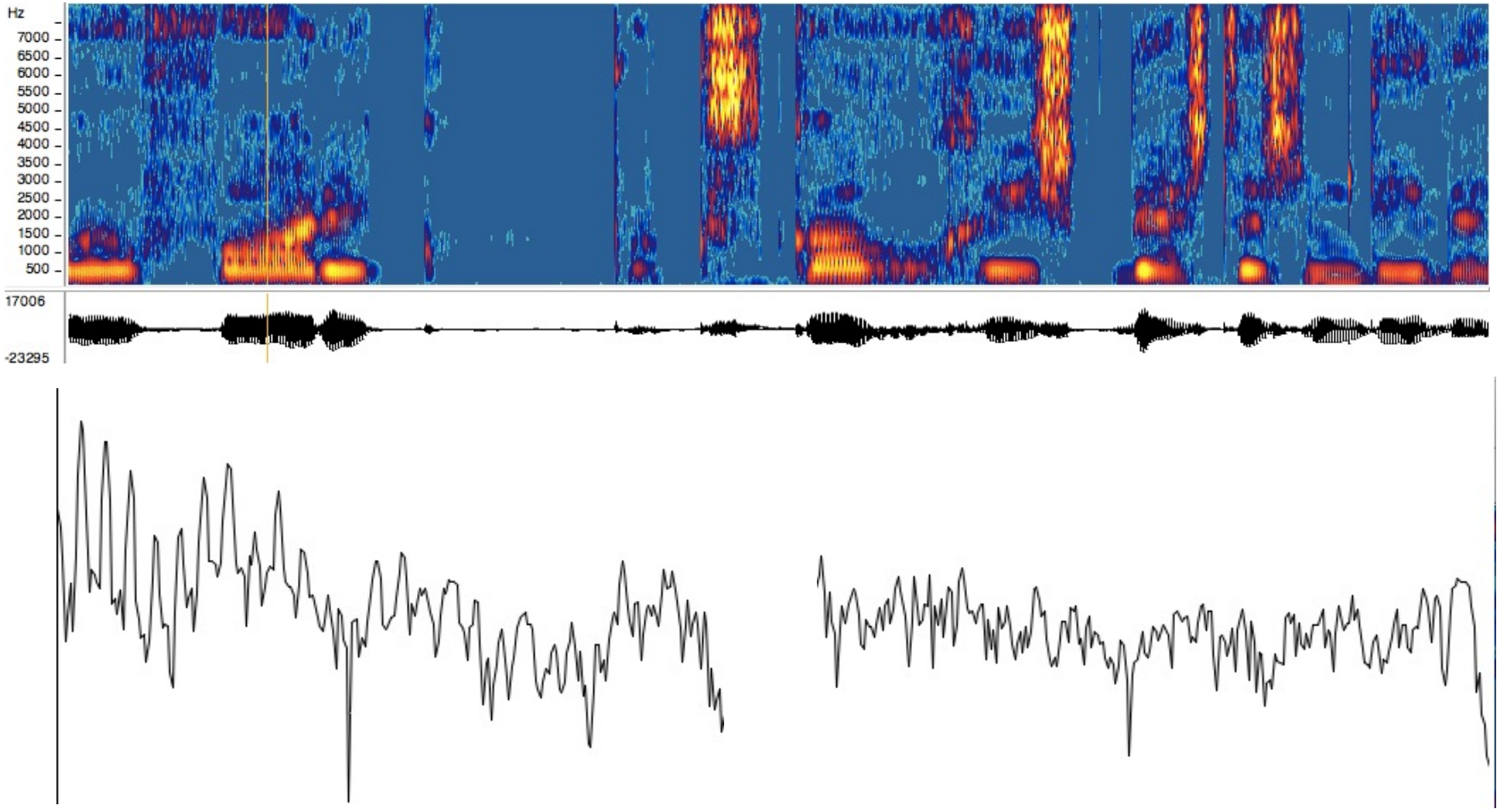






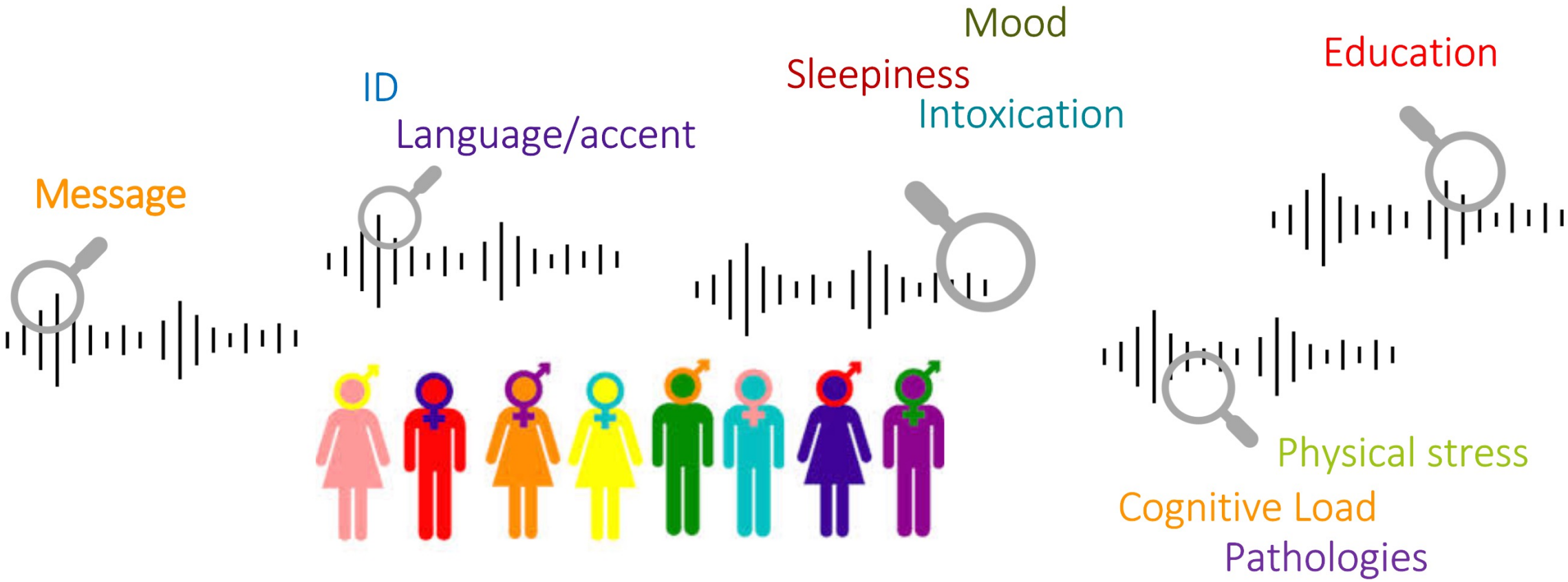
Introduction

Speech time/frequency representation



Introduction

Speech as a carrier of rich information



Introduction

Speech machine learning

Wouldn't it be dreamy if only we could extract all these information automatically!?!? Then, we could imitate human behavior (IA); or even augment capabilities (data mining); or...

YES, we can!!! (more or less)

→ **Speech + Machine Learning**

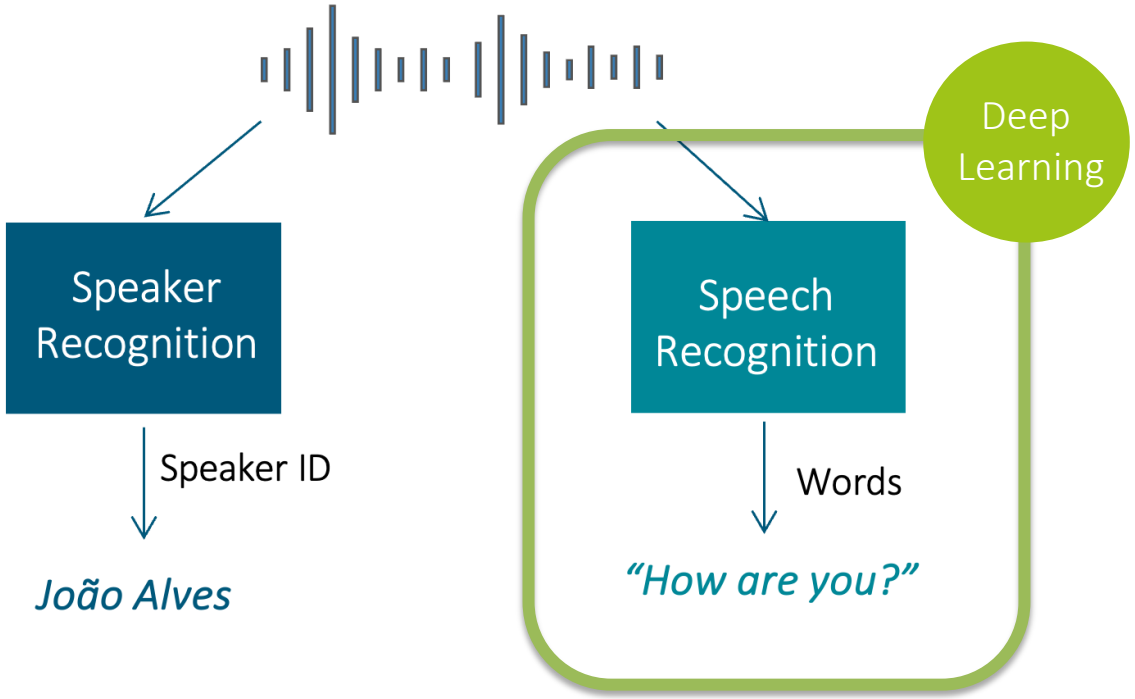
Introduction

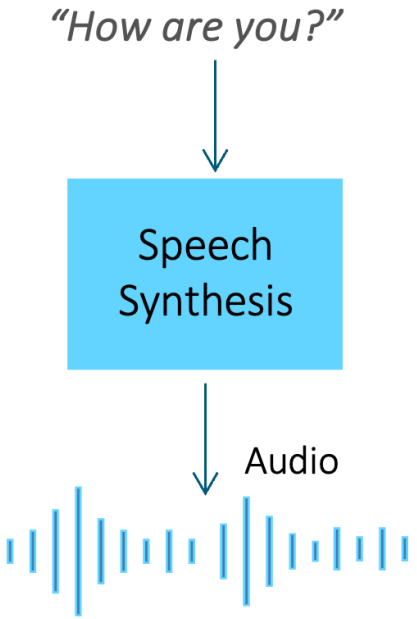
Challenges of speech ML

1. **Speech/audio variability** Same class samples can type extremely different forms
    - Source variation: speaker, gender, accent, state, volume, etc.
    - Channel variation: microphone, acoustic environment, noise, reverberation, etc.
    - Other: Intrinsic nature of the classes, etc.
2. **Data** collection and annotation challenges:
    - Scarcity of data; domain-dependence; privacy issues; inconsistency and subjectivity of annotation
3. **From the machine learning** perspective, speech is a quite remarkable problem due to the nature of the data input and class label outputs:
    - About the input:
        - Continuous signal with no discretized/finite tokens
        - Very different length of the input wrt. output → Segmentation/alignment problem
        - Elasticity of the temporal dimension
        - Discriminative cues often distributed over a reasonably long temporal span
    - Output may consist of a sequence of class labels

Introduction

Common speech tasks





Sequence -> one

Sequence -> sequence

Sequence -> sequence

Speech processing: Speech coding, Speech enhancement, Audio segmentation, Text-to-speech synthesis, Automatic speech recognition, Speaker and language identification; voice conversion; etc.

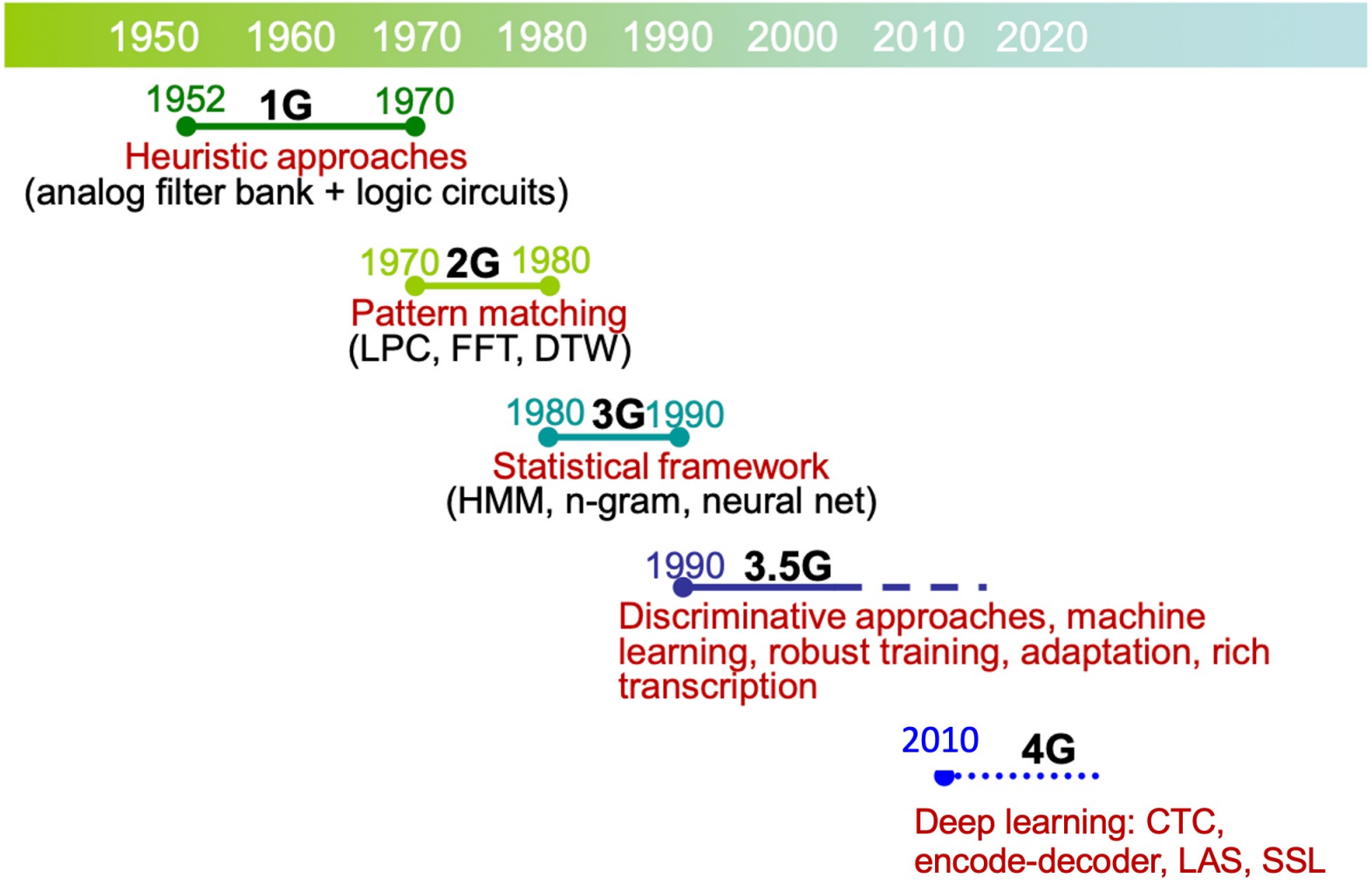
Outline

- Introduction to speech machine learning
- Automatic speech recognition overview
  - 1st Deep Learning success story in ASR
  - 2nd Deep Learning success story in ASR
- Automatic Disease Detection from Speech
  - Pipeline
  - Dealing with data scarcity
  - Complimentary modalities
  - Other challenges
- Where to get started: tools and libraries

Automatic speech recognition overview

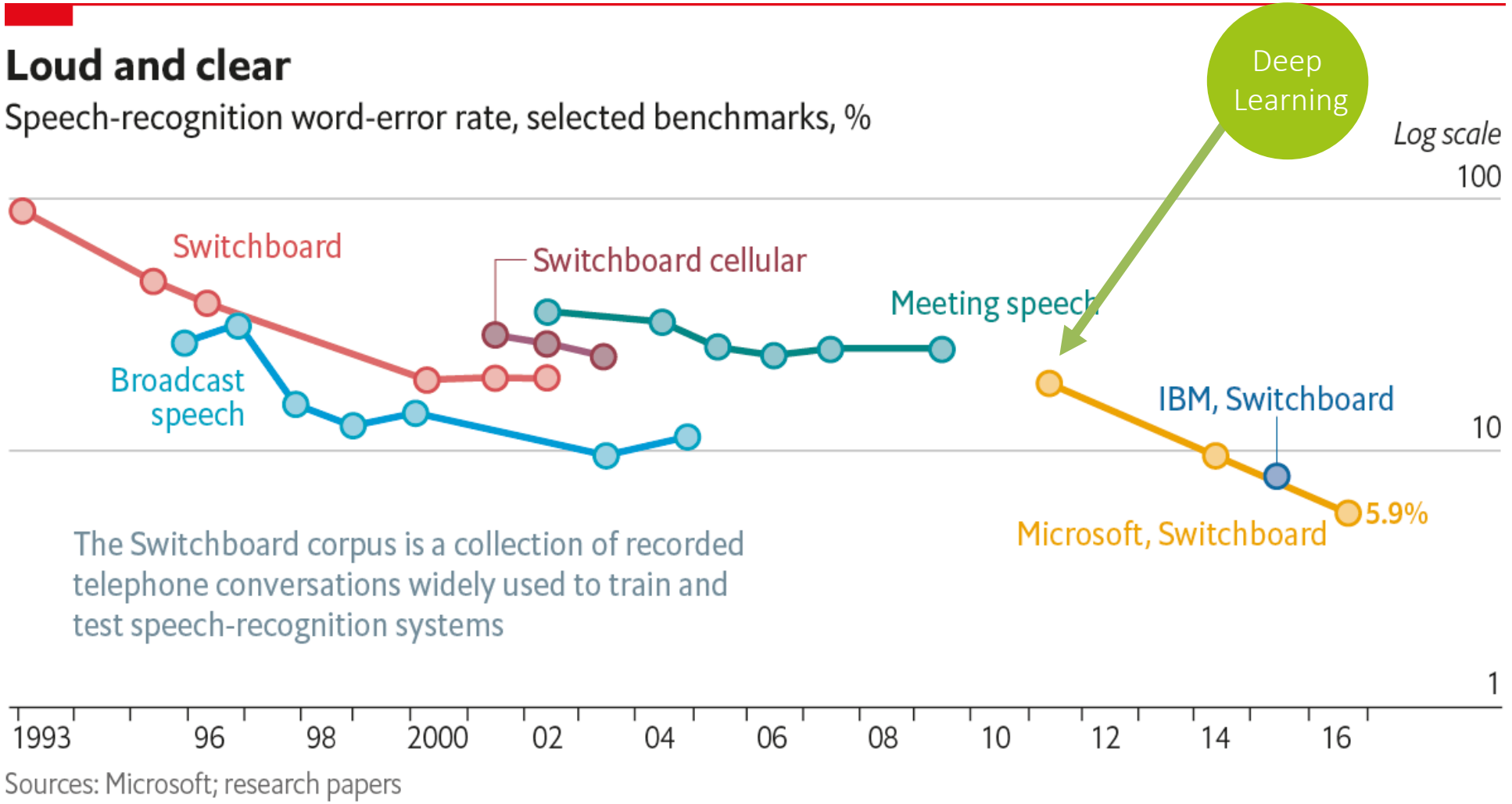
Automatic speech recognition overview

Generations of ASR technology



Automatic speech recognition overview

ASR Benchmark history



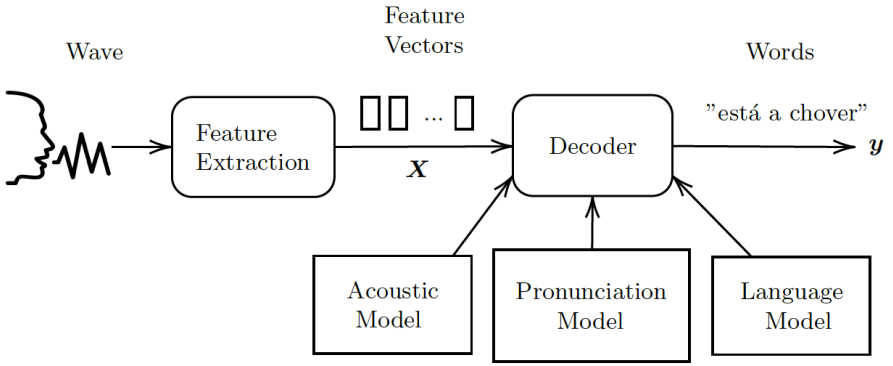
Automatic speech recognition overview

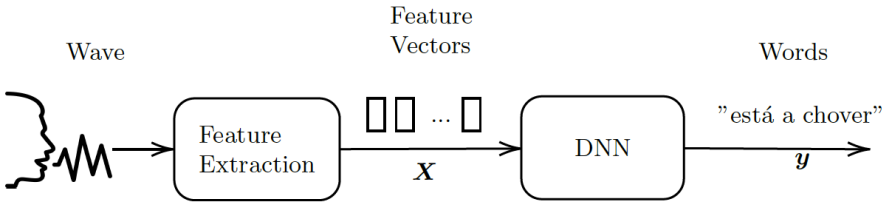
Main current ASR approaches

- Two main current (supervised) ASR trends:

1. **Hierarchical modelling of speech (3G)**
   - Speech modelling problem is structured in sub-problems
   - This is the conventional approach until ~2012
   - Today still very relevant in certain tasks/conditions

2. **End2end (4G)**
   - Direct mapping from acoustics to words/characters
   - Different flavours from 2012 (CTC, encoder-decoder, etc.)
   - State of the art (in very large data)





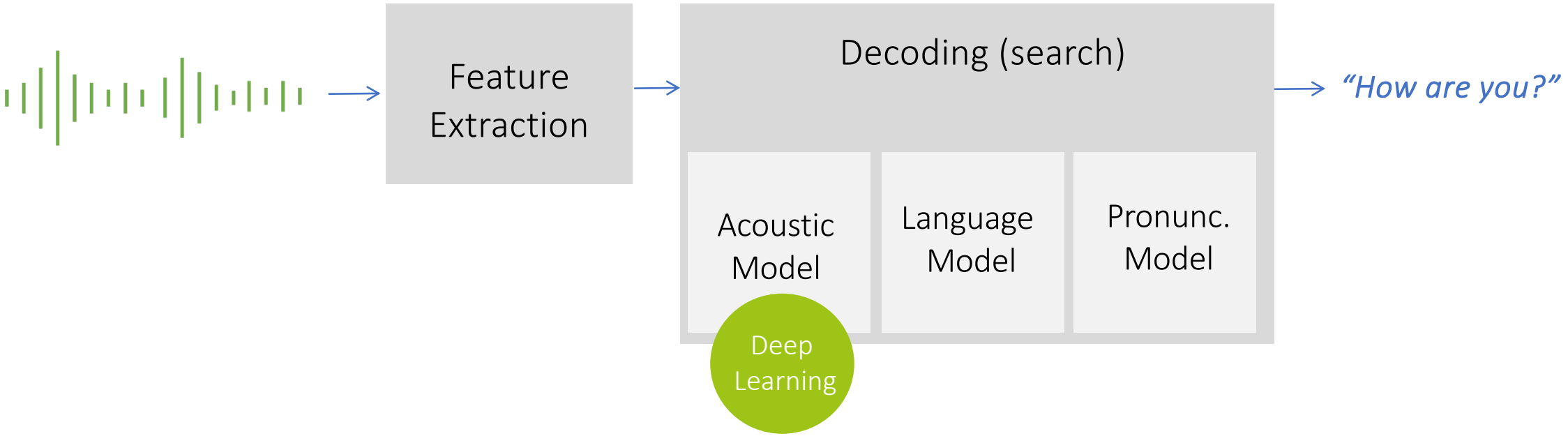
1st Deep Learning success in ASR

Hierarchical modeling of speech

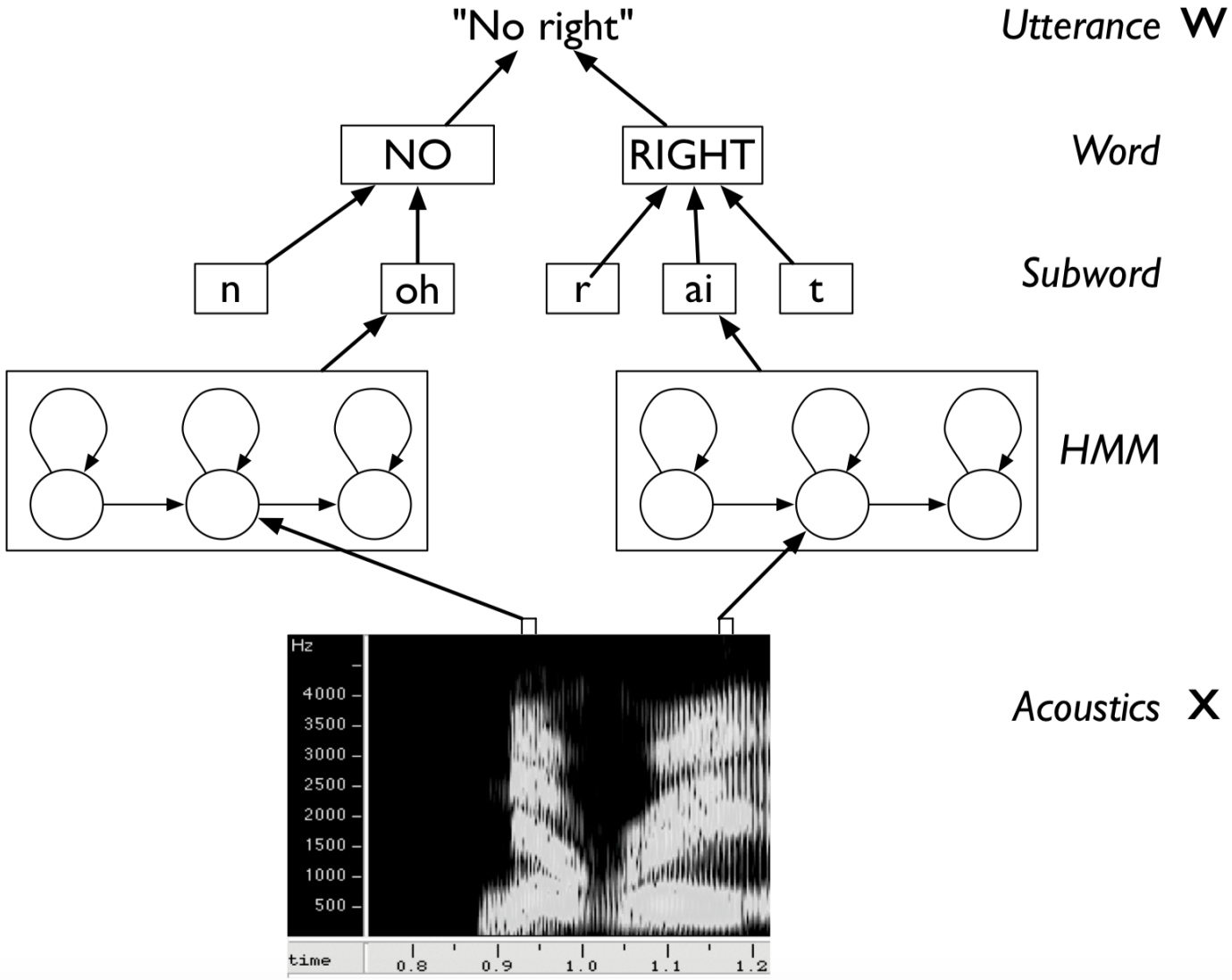
→ Automatic Speech Recognition (ASR) → *“How are you?”*

Deep Learning in 3G ASR

Hierarchical modeling of speech



Hierarchical modeling of speech



The connectionist approach (early 1990s)

- The dominant statistical model for AM until 2012 was the **Gaussian Mixture Model** (HMM/GMM).

- However, there have been attempts since early 1990 to use MLPs as acoustic models:
  - Hybrid **monophone** MLP-based systems were superior than GMM-based counterparts (but worse than **triphone**)

System	Nov92	si_dt_s6	si_dt_05.odd
HMM/GMM wint	8.11	10.39	12.40
HMM/GMM xword	6.86	9.52	10.48
ANN/HMM	9.73	13.13	14.37

Table 1: *WER results of HMM/GMM systems after [9] (word internal and cross-word) and of the baseline ANN/HMM system.*

Morgan and Bourlard (1995). Continuous speech recognition: Introduction to the hybrid HMM/connectionist approach, *IEEE Signal Processing Mag.*, 12(3):24-42

A. Abad and J. Neto (2008), Incorporating acoustical modelling of phone transitions in an hybrid ANN/HMM speech recognizer, *In* INTERSPEECH, 2008

Deep Learning in 3G ASR

What is different after 2012?

- DNNs proposed as AM for ASR:
  - **Deeper** networks, typical NNs AMs with 3-7 hidden layers:
    - This is partially possible due to different advances in ML, including, regularization strategies, hidden unit non-linearity (ReLU vs tanh vs sigmoid), architectural choices
  - **Wider** networks
    - Context-dependent HMM states or senones
  - **Computer/GPUs**
    - Permitted an increasing experiments
    - Scaling up data and parameters

G. E. Dahl, et al (2012), Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition, in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30-42, Jan. 2012.

Hinton, et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *Signal Processing Magazine, IEEE.*

Deep Learning in 3G ASR

(Early) examples of CD-HMM-DNN in LVCSR

[TABLE 3] A COMPARISON OF THE PERCENTAGE WERs USING DNN-HMMs AND GMM-HMMs ON FIVE DIFFERENT LARGE VOCABULARY TASKS.

TASK	HOURS OF TRAINING DATA	DNN-HMM	GMM-HMM WITH SAME DATA	GMM-HMM WITH MORE DATA
SWITCHBOARD (TEST SET 1)	309	18.5	27.4	18.6 (2,000 H)
SWITCHBOARD (TEST SET 2)	309	16.1	23.6	17.1 (2,000 H)
ENGLISH BROADCAST NEWS	50	17.5	18.8	
BING VOICE SEARCH (SENTENCE ERROR RATES)	24	30.4	36.2	
GOOGLE VOICE INPUT	5,870	12.3		16.0 (>> 5,870 H)
YOUTUBE	1,400	47.6	52.3	

(Hinton et al (2012))

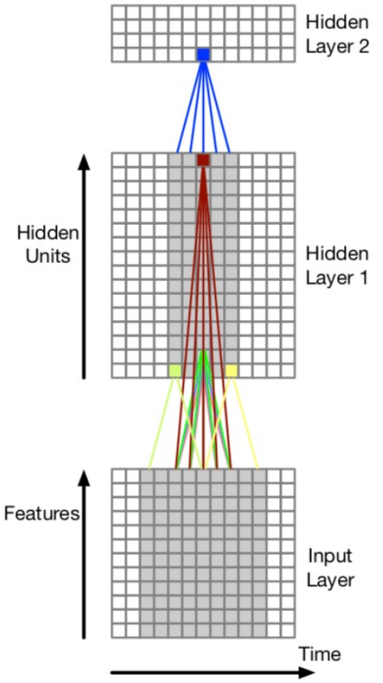
Hinton, et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine.*

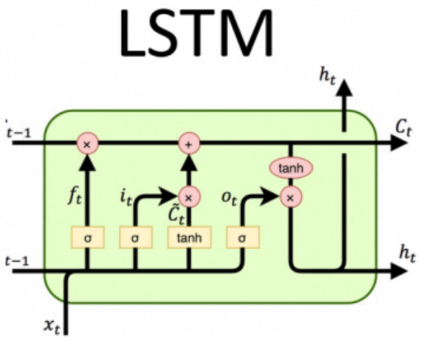
Deep Learning in 3G ASR

Since 2012 all sorts of DNN-based AM

Network Architecture	Test Set WER/% Switchboard	CallHome
GMM (ML)	21.2	36.4
GMM (BMMI)	18.6	33.0
DNN (7x2048) / CE	14.2	25.7
DNN (7x2048) / MMI	12.9	24.6
TDNN (6x1024) / CE	12.5	
TDNN (6x576) / LF-MMI	9.2	17.3
LSTM (4x1024)	8.0	14.3
LSTM (6x1024)	7.7	14.0
LSTM-6 + feat fusion	7.2	12.7

GMM and DNN results - Vesely et al (2013); TDNN-CE results - Peddinti et al (2015); TDNN/LF-MMI results - Povey et al (2016); LSTM results - Saon et al (2017)





Vesely, et al. (2013) Sequence-discriminative training of deep neural networks, *in Proc. Interspeech, 2013.*

Peddinti, et al (2015) A time delay neural network architecture for efficient modeling of long temporal contexts, *in Proc. Interspeech, 2015.*

Povey, et al. (2016) Purely sequence-trained neural networks for ASR based on lattice-free MMI, *in Proc. Interspeech, 2016.*

Saon, et al. (2017) English Conversational Telephone Speech Recognition by Humans and Machines, *in Proc. Interspeech, 2017.*

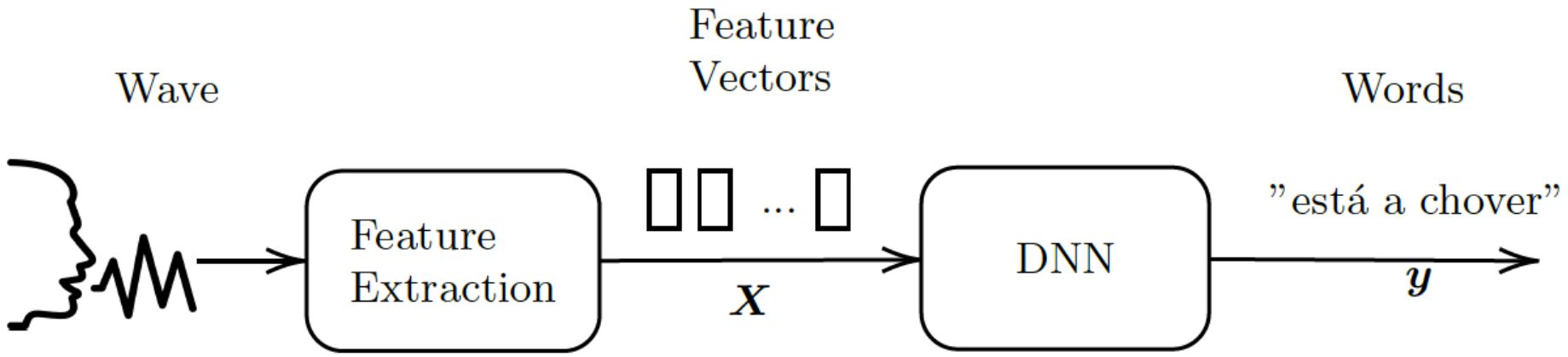
HMM-based systems problems

- It requires independent **expert knowledge** and is time-consuming to create acoustic, pronunciation and language model modules, which are then trained separately.
- The composition of the different model sources (sometimes based on finite-state transducers) typically results in huge space search graphs and quite **complex decoding strategies**.
- In practice, to develop a state-of-the-art HMM/DNN system, it is r**equired to first train an HMM/GMM** model to obtain phonetic alignments and the tied-state HMM structure.

2nd Deep Learning success in ASR

End-to-end

- A system that **directly** maps a sequence of input acoustic features into a sequence of graphemes or words



Is end-to-end possible?

- Main issue with acoustic model:
  - The input is of **variable** length.
  - The input is **much larger** than the output.
  - **We don't know how the inputs audio features align with the output characters!!**

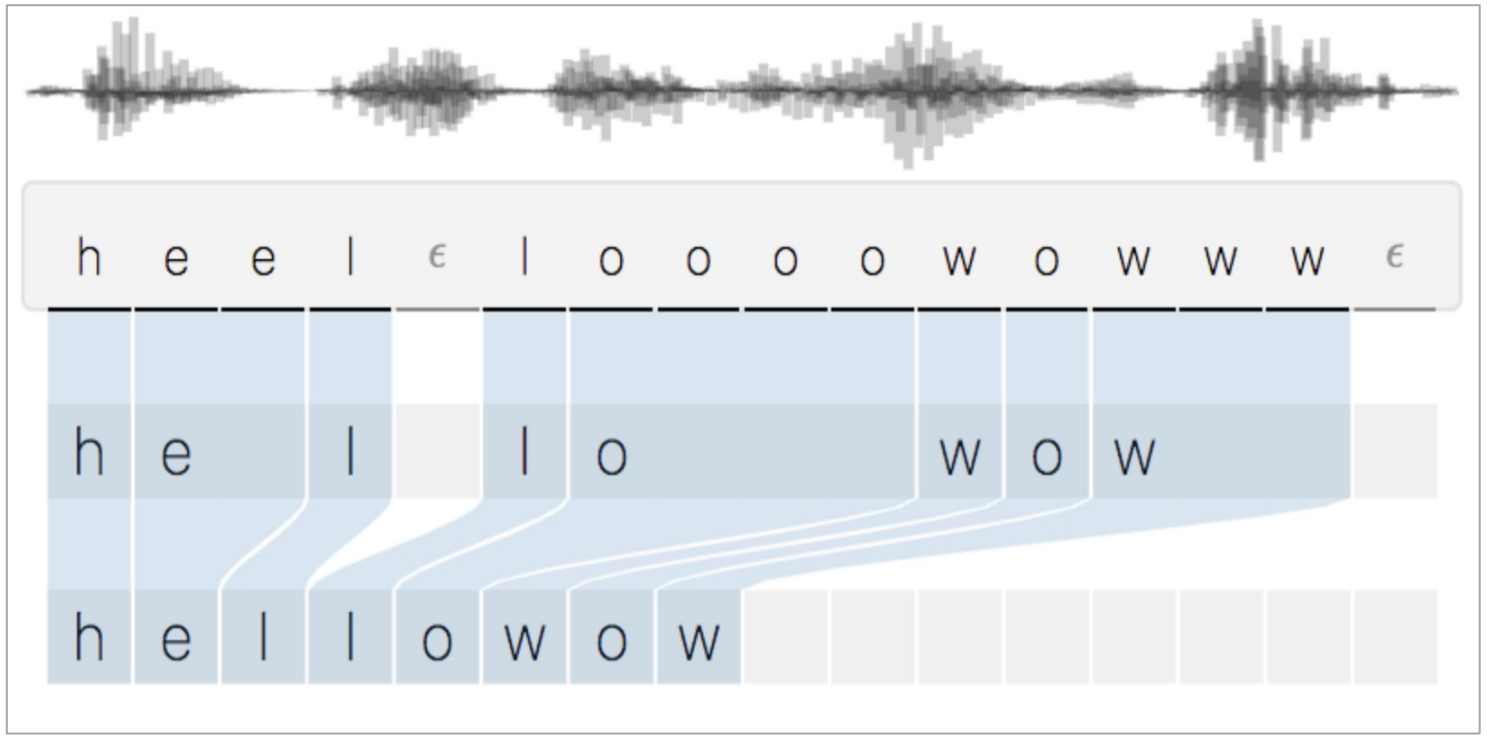


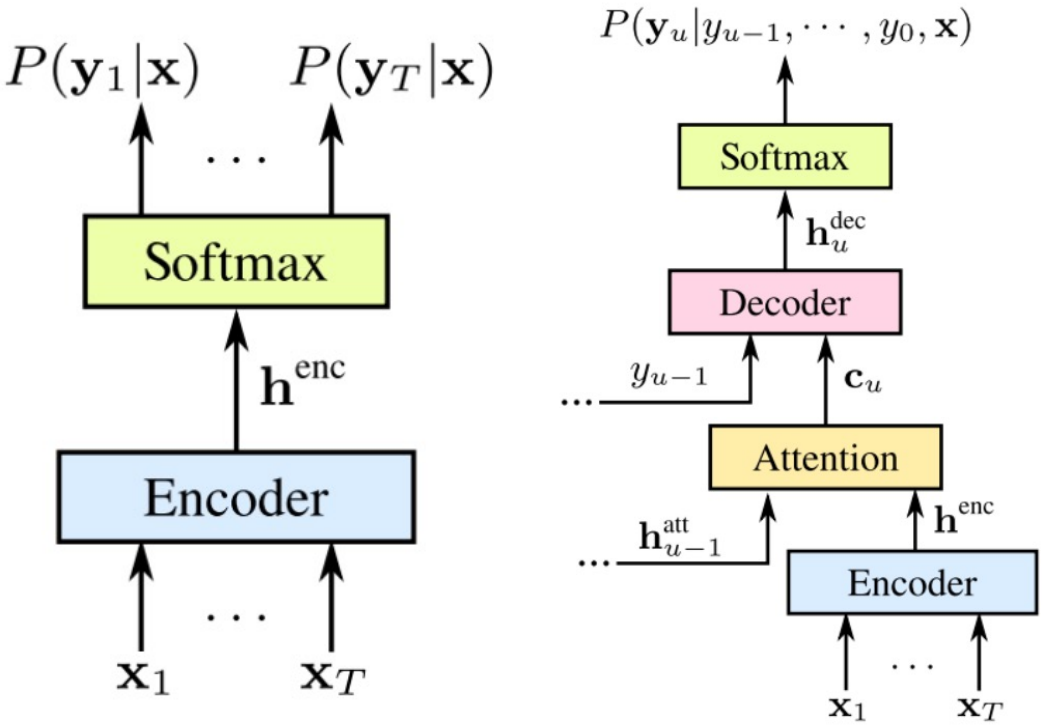
Image from https://distill.pub/2017/ctc/

Deep Learning in 4G ASR

Is end-to-end possible?

- Two main architectures to solve the alignment problem:
  - CTC [Graves et al., 2006]
  - seq2seq-attention

    [Chan et al., 2015] [Chorowski et al., 2015]



Graves, et al. (2006) Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *in Proc. of ICML, 2006.*

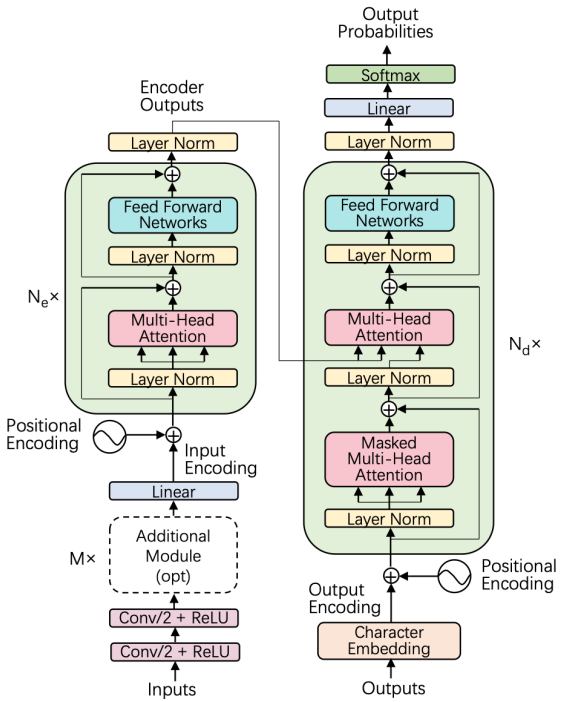
Chan, et al (2015) Listen, Attend and Spell, *arXiv preprint arXiv:1508.01211, 2015.*

Chorowski, et al. (2016) Attention-Based Models for Speech Recognition, *Advances in Neural Information Processing Systems 28 (NIPS 2015).*

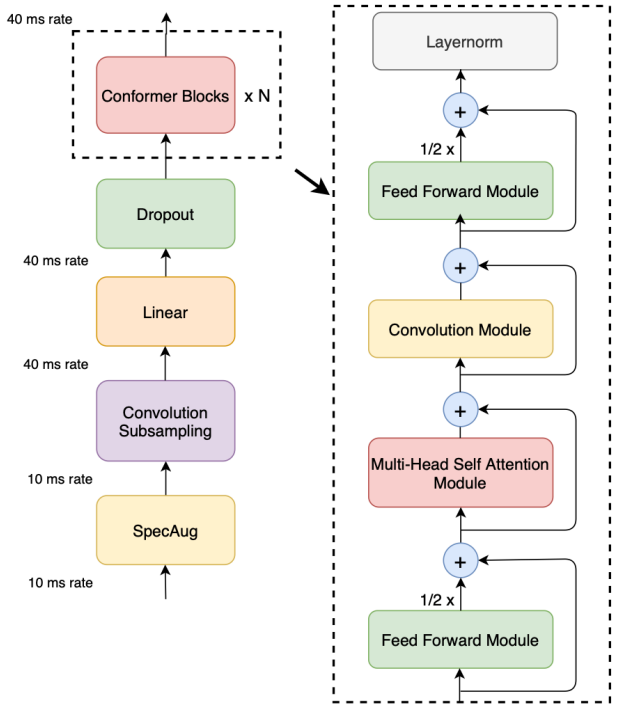
Deep Learning in 4G ASR

Transformers for ASR: Supervised approaches

Speech-Transformer



Conformer



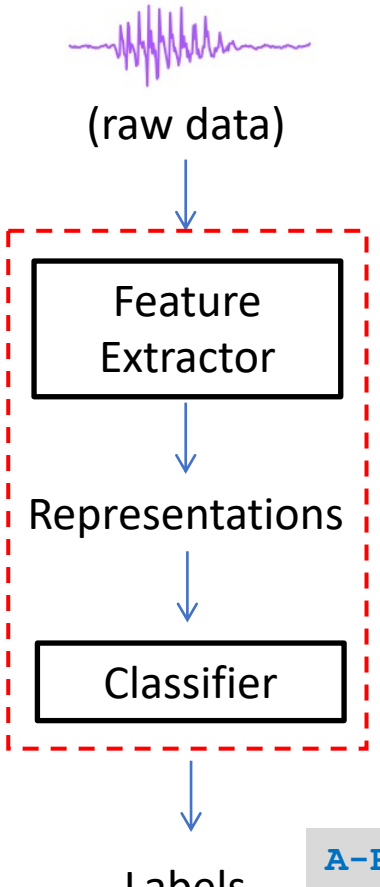
Vaswani et at. (2016) Attention is all you need, *Advances in neural information processing systems, 2017*

Dong, et al. (2018) Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition, *in Proc. ICASSP, 2018.*

Gilati, et al. (2020) Conformer: Convolution-augmented Transformer for Speech Recognition, *arXiv preprint arXiv:2005.08100 (2020).*

Deep Learning in 4G ASR

Transformers for ASR: Self-supervised approaches



- Analogous to **GPT in text,** transformer encoder-based self-supervised speech models trained on thousands of hours of unlabelled data.
- Some well-known models include
  - **wav2vec2**
  - **wavLM**
  - **HuBERT**
- This pre-trained models are used:
  - As feature extractors
  - As upstream models to fine-tuned to specific downstream tasks

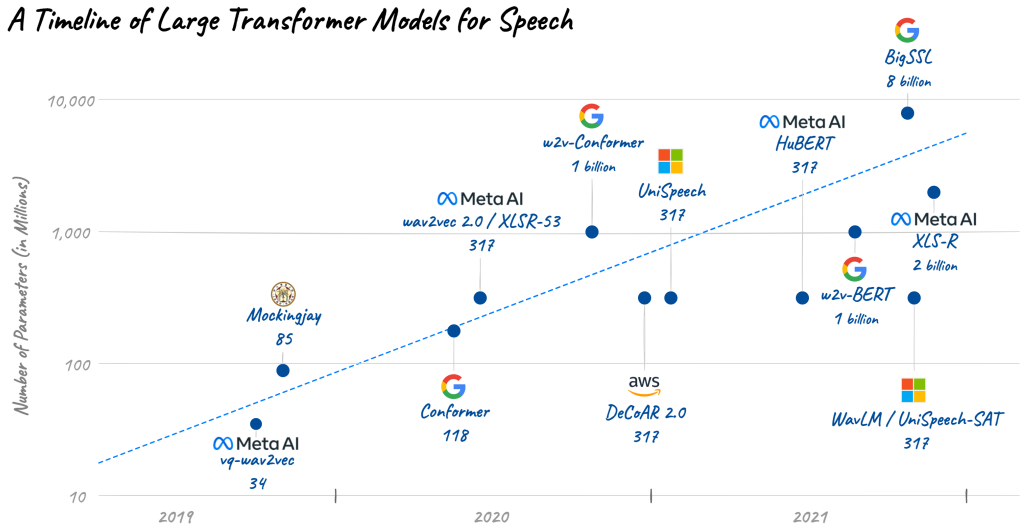


Image from https://jonathanbgn.com/2021/12/31/timeline-transformers-speech.html

A-R Mohamed et al. (2022) Self-supervised speech representation learning: A review. *IEEE JSTSP Special Issue on Self-Supervised Learning for Speech and Audio*

Hsu, Wei-Ning, et al. (2021) Hubert: Self-supervised speech representation learning by masked prediction of hidden units, *IEEE/ACM TASLP*

E2E ASR performance (Librispeech)

http://www.openslr.org/12/
https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean

work with key technologies	year	model	encoder	test-clean/other WER
Deep Speech 2: more labeled data, curriculum learning [325]	2016	CTC	bi-RNN	5.3/13.2
policy learning, joint training [326]	2018	CTC	CNN+bi-GRU	5.4/14.7
Shallow fusion, BPE, and pre-training [33]	2018	AED	BLSTM	3.8/12.8
ESPRESSO recipe: lookahead word LM, EOS thresholding [258]	2019	AED	CNN+BLSTM	2.8/8.7
SpecAugment [278]	2019	AED	CNN+BLSTM	2.5/5.8
ESPnet recipe: SpecAugment, dropout [86]	2019	AED	Transformer	2.6/5.7
Semantic mask [327]	2019	AED	Transformer	2.1/4.9
Transformer-T, SpecAugment [89]	2020	RNN-T	Transformer	2.0/4.6
Conformer-T, SpecAugment [97]	2020	RNN-T	Conformer	1.9/3.9
wav2vec 2.0: SSL with unlabeled data, DataAugment [28]	2020	CTC	Transformer	1.8/3.3
internal LM prior correction, EOS modeling[328]	2021	RNN-T	BLSTM	2.2/5.6
w2v-BERT: SSL with unlabeled data, SpecAugment [329]	2021	RNN-T	Conformer	1.4/2.5

Li, J. (2022). Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, *11*(1).

Deep Learning in 4G ASR

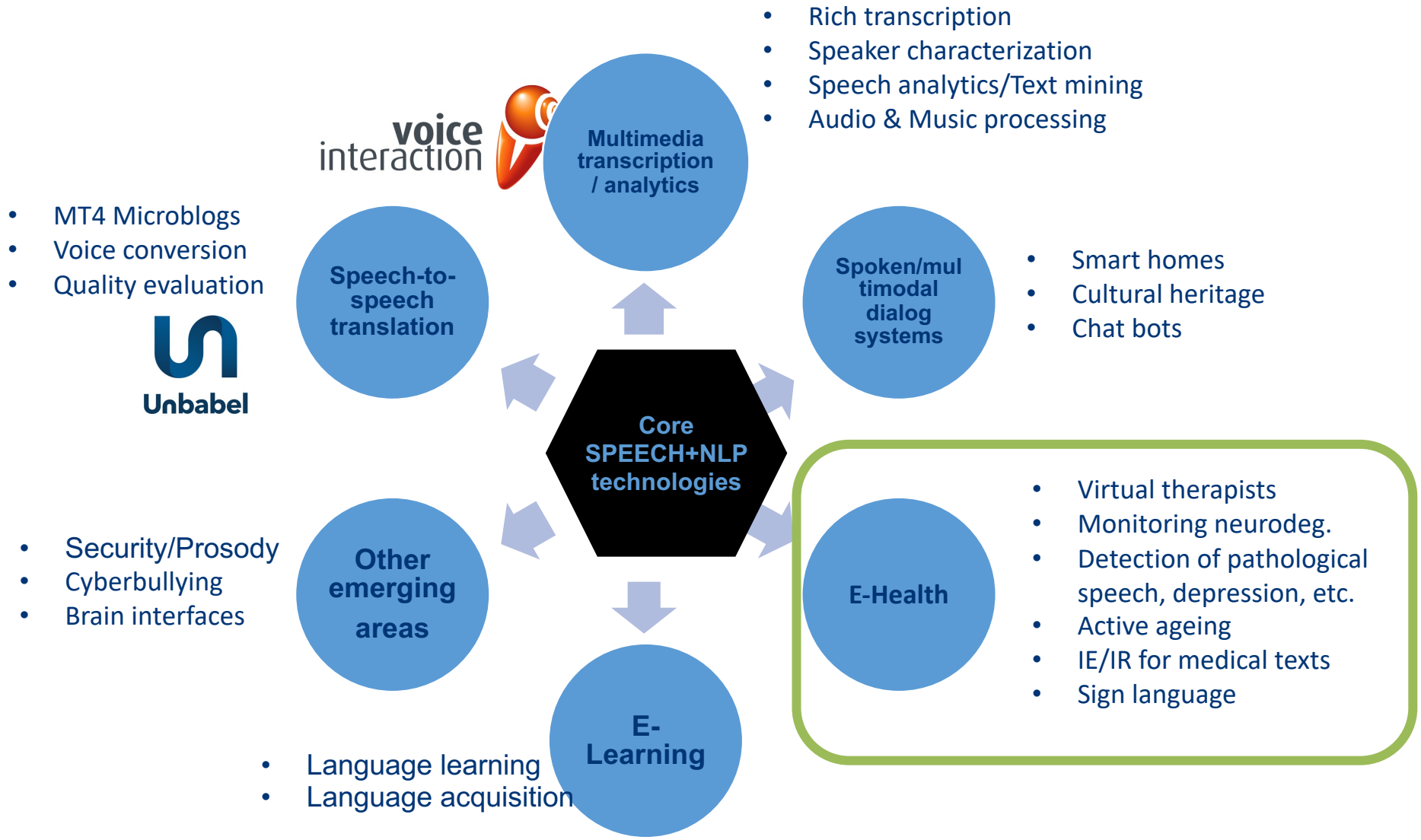
E2E ASR performance in more challenging tasks

dataset	token	error	Kaldi (s5)	ESPnet RNN (ours)	ESPnet Transformer (ours)
AISHELL	char	CER	N/A / 7.4	6.8 / 8.0	**6.0 / 6.7**
AURORA4	char	WER	(*) 3.6 / 7.7 / 10.0 / 22.3	3.5 / 6.4 / 5.1 / 12.3	**3.3 / 6.0 / 4.5 / 10.6**
CSJ	char	CER	(*) 7.5 / 6.3 / 6.9	6.6 / 4.8 / 5.0	**5.7 / 4.1 / 4.5**
CHiME4	char	WER	**6.8 / 5.6 / 12.1 / 11.4**	9.5 / 8.9 / 18.3 / 16.6	9.6 / 8.2 / 15.7 / 14.5
CHiME5	char	WER	**47.9 / 81.3**	59.3 / 88.1	60.2 / 87.1
Fisher-CALLHOME Spanish	char	WER	N/A	27.9 / 27.8 / 25.4 / 47.2 / 47.9	**27.0 / 26.3 / 24.4 / 45.3 / 46.2**
HKUST	char	CER	23.7	27.4	**23.5**
JSUT	char	CER	N/A	20.6	**18.7**
LibriSpeech	BPE	WER	3.9 / 10.4 / 4.3 / 10.8	3.1 / 9.9 / 3.3 / 10.8	**2.2 / 5.6 / 2.6 / 5.7**
REVERB	char	WER	18.2 / 19.9	24.1 / 27.2	**15.5 / 19.0**
SWITCHBOARD	BPE	WER	**18.1 / 8.8**	28.5 / 15.6	**18.1** / 9.0
TED-LIUM2	BPE	WER	**9.0** / 9.0	11.2 / 11.0	9.3 / **8.1**
TED-LIUM3	BPE	WER	**6.2 / 6.8**	14.3 / 15.0	9.7 / 8.0
VoxForge	char	CER	N/A	12.9 / 12.6	**9.4 / 9.1**
WSJ	char	WER	**4.3 / 2.3**	7.0 / 4.7	6.8 / 4.4

* CHIME4 - 4.4/2.5/7.4/4.1 (2022, Whisper medium finetuned ESPnet)
* SWITCHBOARD - 13.4/7.3 (2022, e-branchformer ESPnet)
* TED-LIUM2 - 7.3/7.1 (2022, e-branchformer ESPnet)
* TED-LIUM3 - 7.3/6.4 (2022, e-branchformer ESPnet)
* WSJ – 3.1/1.8 (2021, conformer + s3prlfrontend Hubert ESPnet)

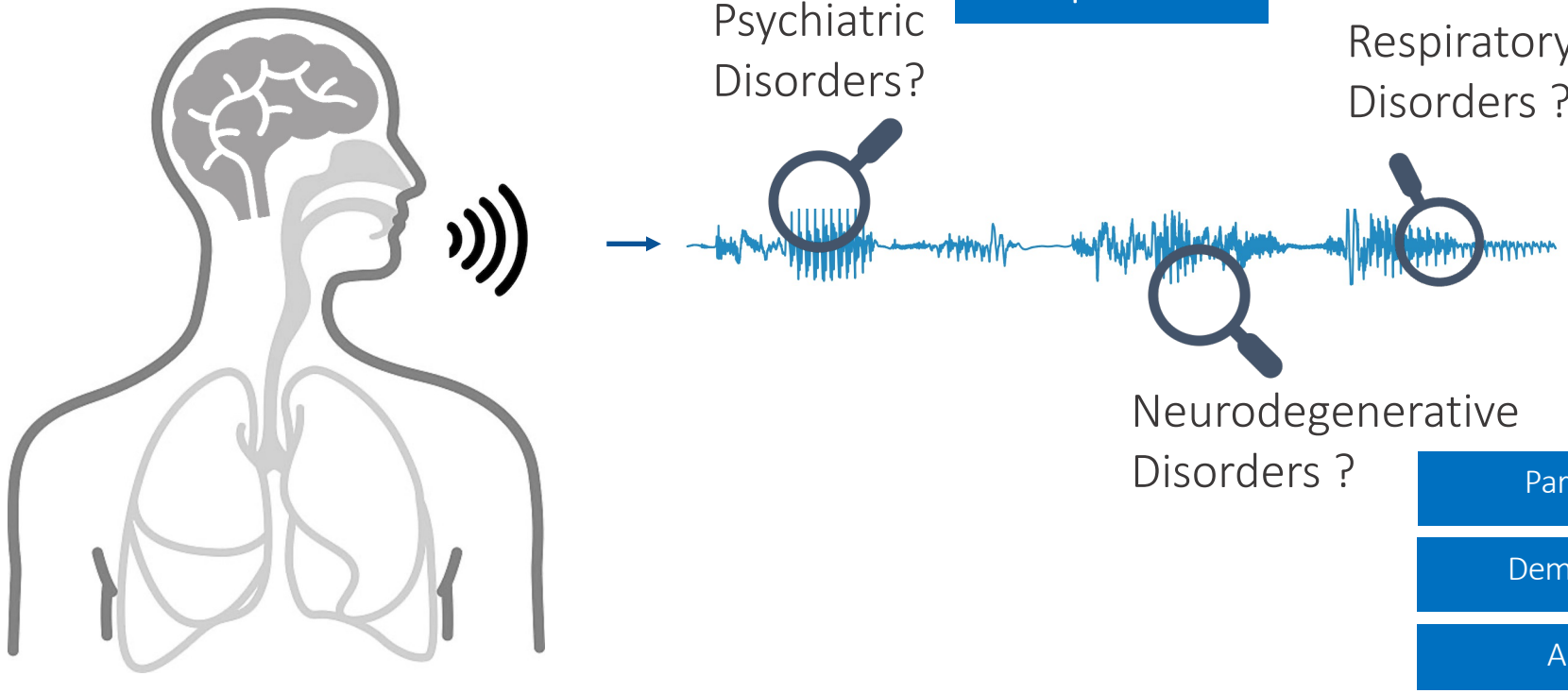
Karita, S. et al. (2019) A comparative study on transformer vs rnn in speech applications. *In Proc. IEEE ASRU*

Core technology and application areas @ HLT.INESC-ID



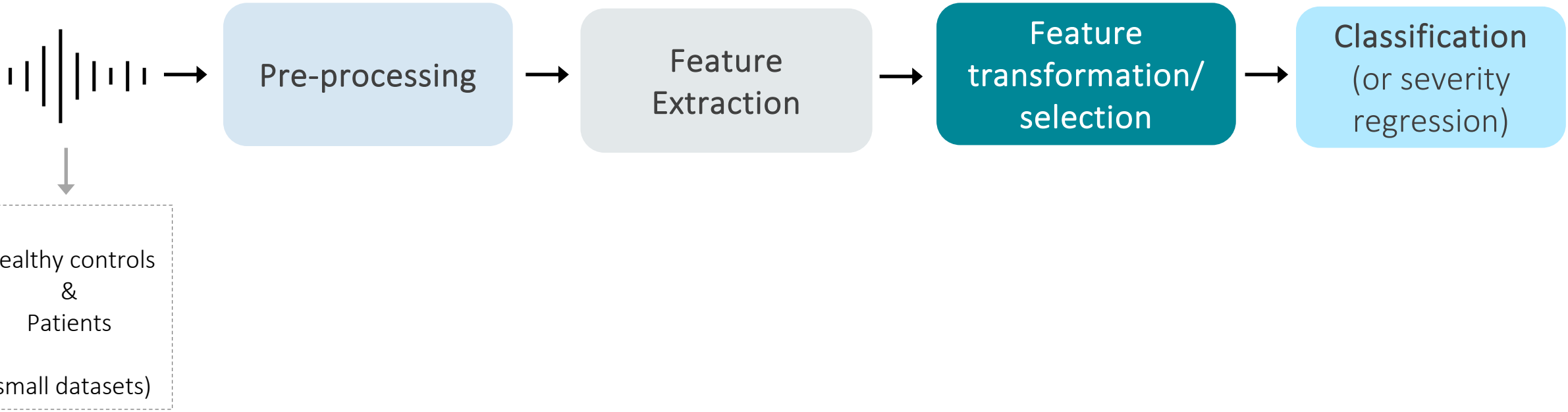
Automatic Disease Detection from Speech

Automatic Disease Detection from speech



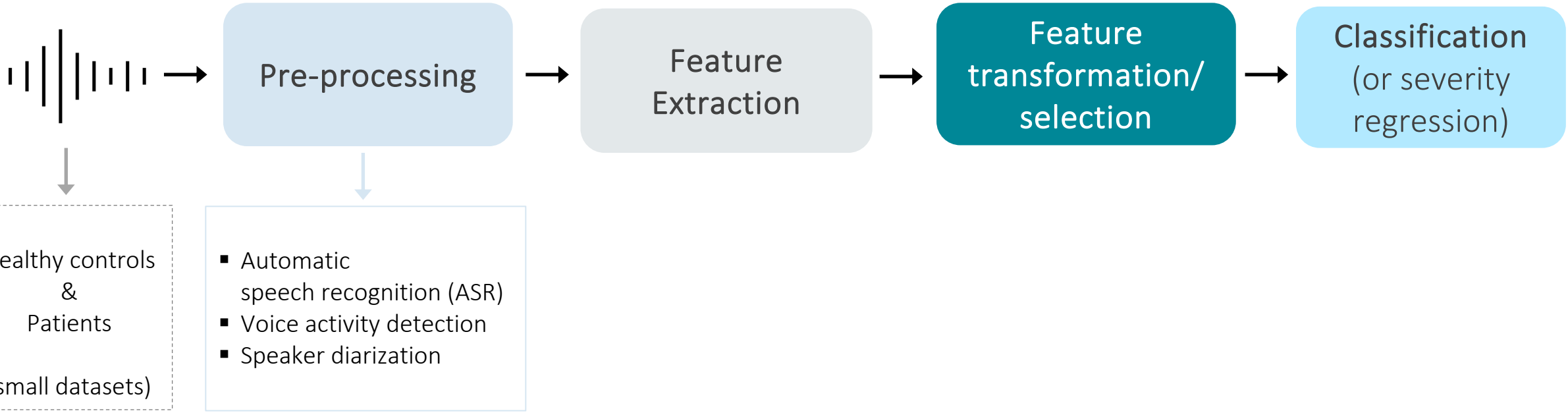
Automatic Disease Detection from speech

Pipeline



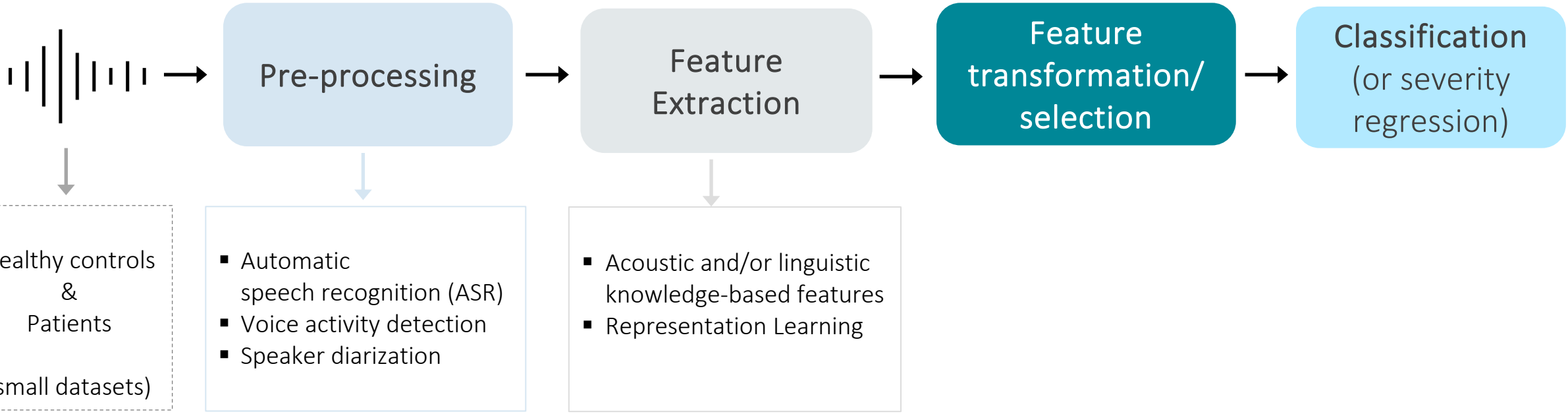
Automatic Disease Detection from speech

Pipeline



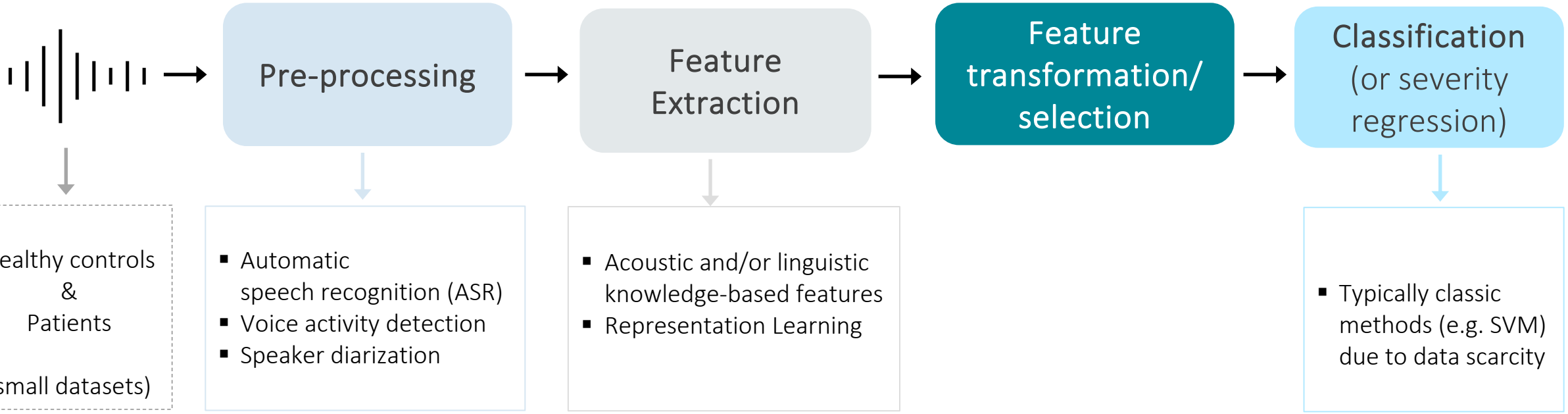
Automatic Disease Detection from speech

Pipeline



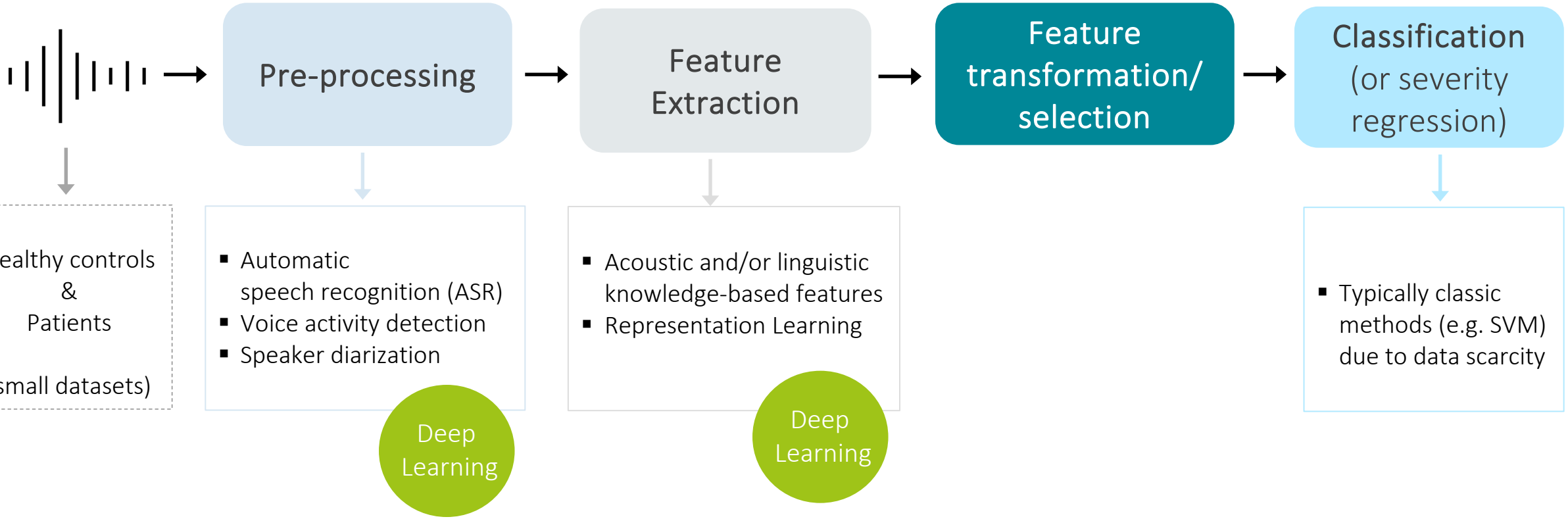
Automatic Disease Detection from speech

Pipeline



Automatic Disease Detection from speech

Pipeline



Automatic Disease Detection from speech

Pipeline

Pre-processing

Deep Learning

- Automatic Speech Recognition
- Voice activity detection
  - Example: https://huggingface.co/speechbrain/vad-crdnn-libriparty
- Diarization
  - Example: https://github.com/pyannote/pyannote-audio

Automatic Disease Detection from speech

Pipeline

Feature Extraction

Deep Learning

Knowledge-based features:

Acoustic:

- MFCCs
- F0
- Formants
- HNR
- Jitter
- Shimmer
- Spectral Flux
- Voice onset time
- Articulation rate
- Average syllable duration
- Intelligibility
- ….

Linguistic:

- Content density
- Brunet Index
- Honeré's statistic
- Type-to-token ratio
- Usage of fillers
- Part of Speech (PoS) based features
- Coherence
- Perplexity
- ….

Automatic Disease Detection from speech

Pipeline

Feature Extraction

Deep Learning

Knowledge-based features:

Acoustic:

- MFCCs
- F0
- Formants
- HNR
- Jitter
- Shimmer
- Spectral Flux
- **Voice onset time** **
- Articulation rate
- Average syllable duration
- **Intelligibility** **
- ….

Linguistic:

- Content density
- Brunet Index
- Honeré's statistic
- Type-to-token ratio
- Usage of fillers
- Part of Speech (PoS) based features
- **Coherence** **
- **Perplexity** **
- ….

** Examples of works using NN-based approaches:

Vásquez-Correa, et al. (2017) *Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease,* Interspeech.

Quintas, et al (2022) *Automatic assessment of speech intelligibility using consonant similarity for head and neck cancer,* Interspeech.

Botelho, et al. (2023) *Towards reference speech characterization for health applications,* Interspeech.

Fritsch, et al. (2019), *Automatic diagnosis of Alzheimer's disease using neural network language models,* ICASSP.

Automatic Disease Detection from speech

Pipeline

Feature Extraction

Deep Learning

Knowledge-based features:

Acoustic:

- MFCCs
- F0
- Formants
- HNR
- Jitter
- Shimmer
- Spectral Flux
- Voice onset time
- Articulation rate
- Average syllable duration
- Intelligibility
- ….

Linguistic:

- Content density
- Brunet Index
- Honeré's statistic
- Type-to-token ratio
- Usage of fillers
- Part of Speech (PoS) based features
- Coherence
- Perplexity
- ….

Deep learning representations:

Acoustic:

- **Speaker embeddings**
- **PASE+ embeddings**
- ….

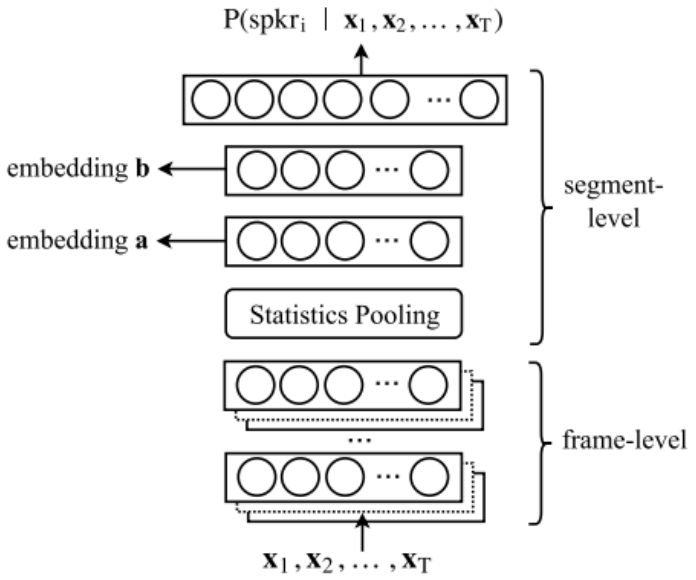
Linguistic:

- Sentence embeddings, eg. BERT, RoBERTa, etc.
- ….

Automatic Disease Detection from speech

Pipeline

Speaker-embeddings (x-vectors)

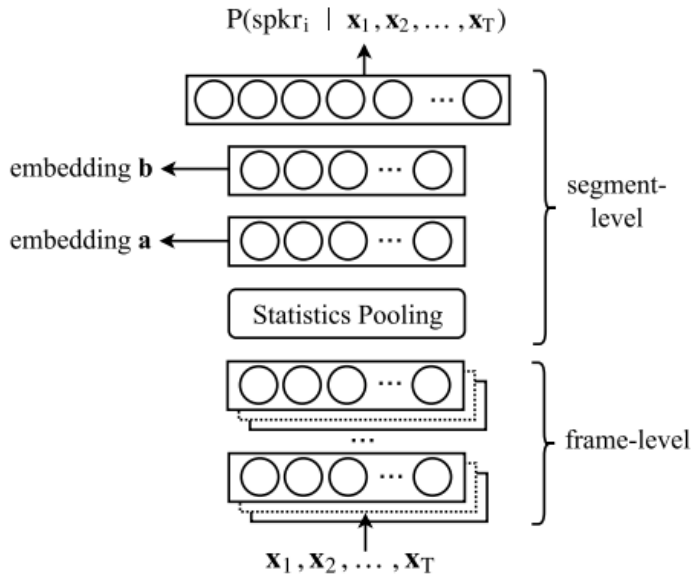


Snyder, et al. (2017) *Deep Neural Network Embeddings for Text-Independent Speaker Verification*, Interspeech.

Automatic Disease Detection from speech

Pipeline

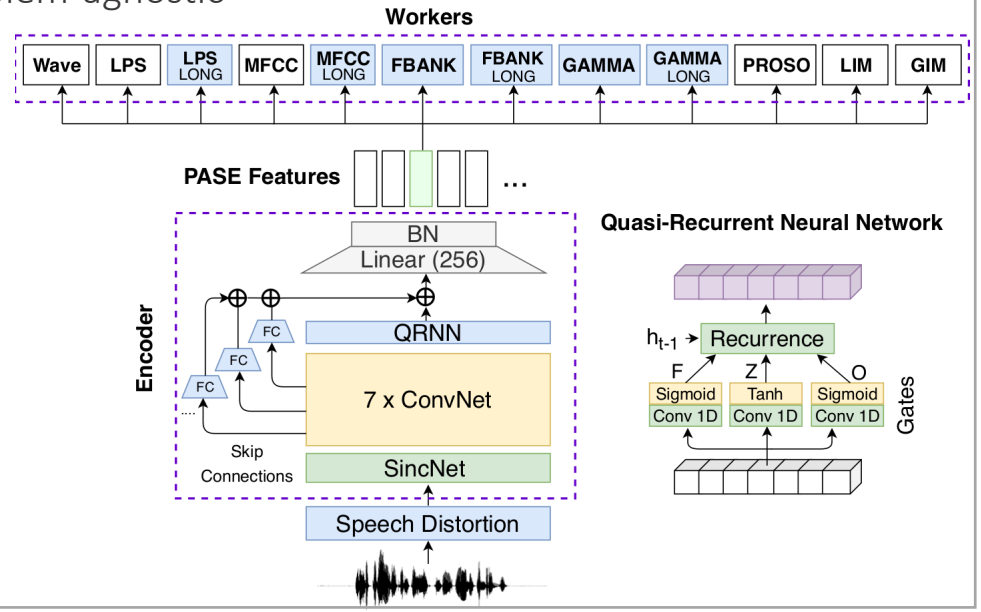
Speaker-embeddings (x-vectors)



Snyder, et al. (2017) *Deep Neural Network Embeddings for Text-Independent Speaker Verification*, Interspeech.

PASE+ Embeddings

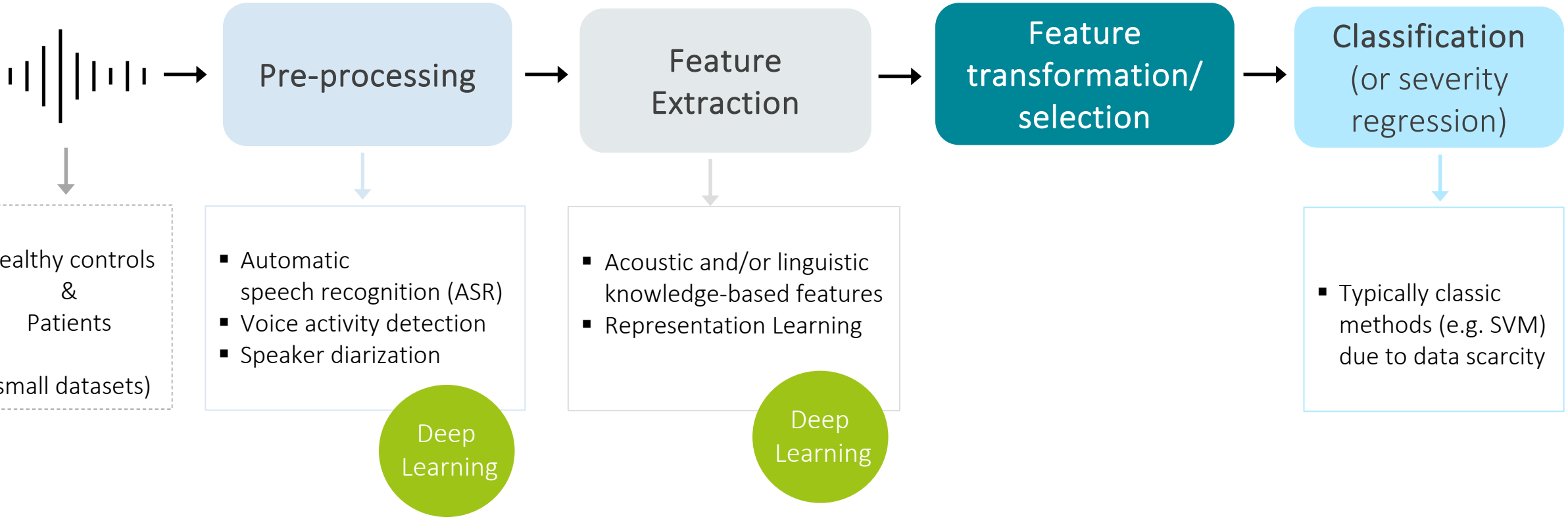
- Self-supervised
- Problem-agnostic



Pascual, et al. (2019) *Learning problem-agnostic speech representations from multiple self-supervised tasks*, Interspeech.
Ravanelli, et al (2020) *Multi-task self-supervised learning for robust speech recognition*, ICASSP.

Automatic Disease Detection from speech

Pipeline



Automatic Disease Detection from speech

- Results have been promising, but there are some challenges...

1. How to deal with data scarcity?
2. Can other modalities complement the speech signal for disease detection?
3. Can we monitor disease progression through time/make predictions into the future?
4. Can we provide interpretable reasoning that is useful to the medical community?
5. Are results generalizable? Can we transfer them across different datasets?

Automatic Disease Detection from speech

- Results have been promising, but there are some challenges...

1. How to deal with data scarcity?
2. Can other modalities complement the speech signal for disease detection?
3. Can we monitor disease progression through time/make predictions into the future?
4. Can we provide interpretable reasoning that is useful to the medical community?
5. Are results generalizable? Can we transfer them across different datasets?

How to deal with data scarcity?

- Why are data difficult to acquire?

How to deal with data scarcity?

- Why are data difficult to acquire?

- Some standard datasets

Disease	Corpus	Size [hours]	Language	Observations	Reference
Depression	DAIC-WOZ	~29 (patient side)	English	Clinical interviews	[J. Gratch et al., 2014]
Parkinson's disease	PC-GITA	~5	Spanish	Speaking exercises	[J.R. Orozco-Arroyave et al., 2014]
Common Cold	URTIC	~45	German	Speaking exercises	[N. Cummins et al., 2017]
Alzheimer's disease	Dementia bank	~4 (ADReSS version)	English	Speaking exercises	[S. Luz et al., 2020]
Cerebral palsy and ALS	TORGO	~23	English	Invasive	[F. Rudzicz et al., 2011]

How to deal with data scarcity?

How to deal with data scarcity?

Machine Learning-based strategies:

- Data augmentation
  - Introduce copies with perturbations
    - Eg. **Nam, et al. (2022).** *FilterAugment: An Acoustic Environmental Data Augmentation Method. ICASSP*
  - Synthesize data using GANs
    - Eg. **Deng, et al. (2017).** *Speech-based diagnosis of autism spectrum condition by generative adversarial network representations. International Digital Health Conference*

How to deal with data scarcity?

Machine Learning-based strategies:

- Data augmentation
- Intelligent labelling paradigms
  - Semi-supervised learning
  - Active learning
  - Cooperative learning

How to deal with data scarcity?

Machine Learning-based strategies:

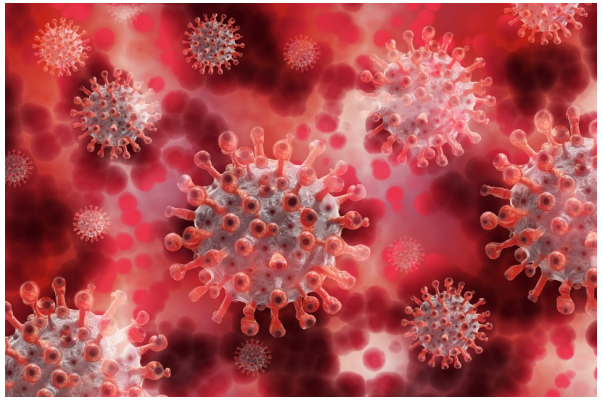
- Data augmentation
- Intelligent labelling paradigms
- Transfer learning

- Eg.:

R. Solera-Ureña, **C. Botelho**, F. Teixeira, T. Rolland, **A.Abad**, I. Trancoso, **(2021).** *Transfer Learning-Based Cough Representations for Automatic Detection of COVID-19", Interspeech.*

How to deal with data scarcity

Transfer Learning for COVID-19 detection from cough

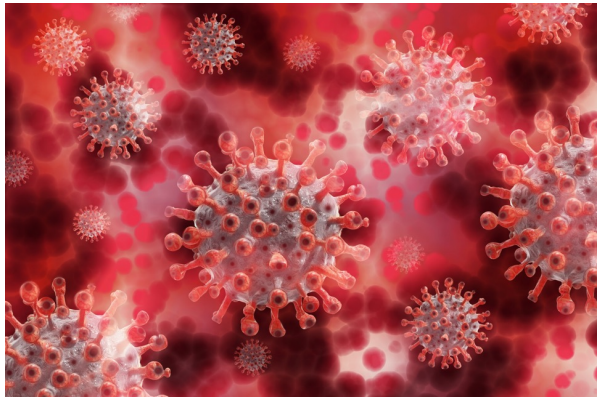


- ComParE 2021 COVID-19 **Cough C**hallenge

Schuller, et al. (2021). *The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates, Interspeech.*

How to deal with data scarcity

Transfer Learning for COVID-19 detection from cough



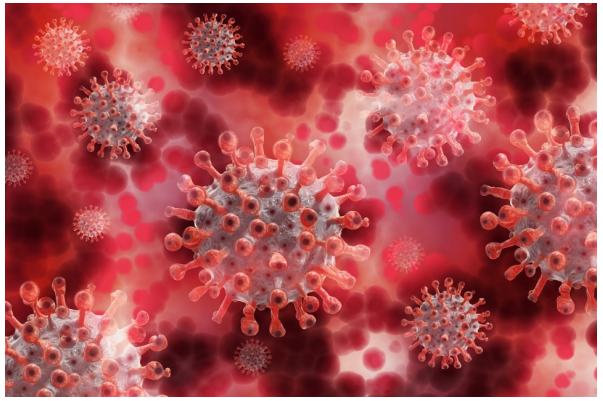
- ComParE 2021 COVID-19 **Cough C**hallenge
- Corpus: COVID-19 COUGH (University of Cambridge)

	Train		Dev		Test
	+	-	+	-	
$C19C_{original}$	71	215	48	183	208

Schuller, et al. (2021). *The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates, Interspeech.*

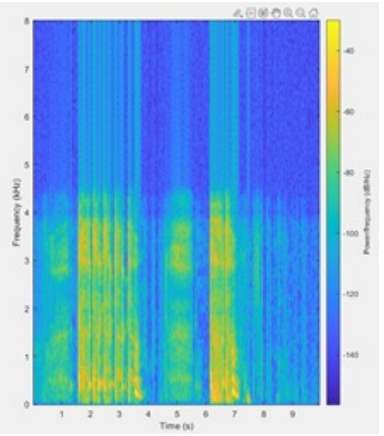
How to deal with data scarcity

Transfer Learning for COVID-19 detection from cough



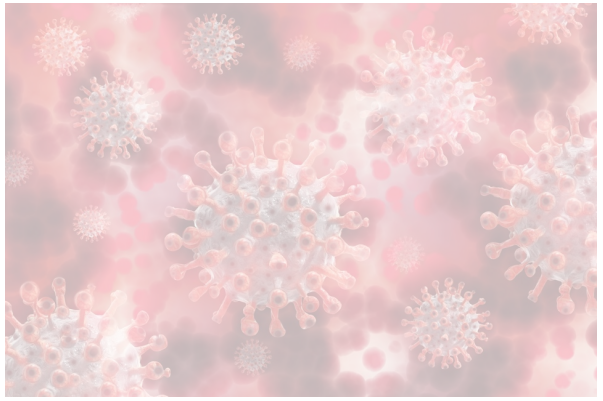
- ComParE 2021 COVID-19 **Cough C**hallenge
- Corpus: COVID-19 COUGH (University of Cambridge)

						Narrow band files				
	Train		Dev		Test	Train		Dev		Test
	+	-	+	-		+	-	+	-	+
$C19C_{original}$	71	215	48	183	208	13	0	8	0	8
$C19C_{fullband}$	58	215	40	183	-	removed all				-



How to deal with data scarcity

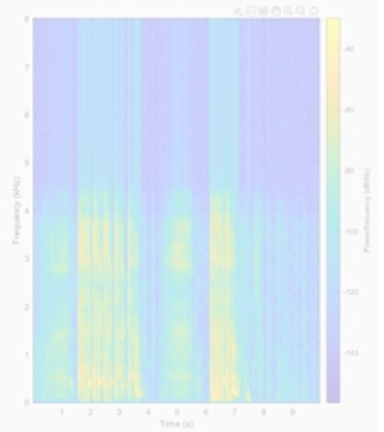
Transfer Learning for COVID-19 detection from cough



- ComParE 2021 COVID-19 **Cough C**hallenge
- Corpus: COVID-19 COUGH (University of Cambridge)

Method: **Transfer Learning**

	arrow band files								
	Train		Dev		Test	Train		Dev		Test
	+	-	+	-		+	-	+	-	+
$C19C_{original}$	71	215	48	183	208	13	0	8	0	8
$C19C_{fullband}$	58	215	40	183	-	removed all				-



How to deal with data scarcity

Transfer Learning for COVID-19 detection from cough

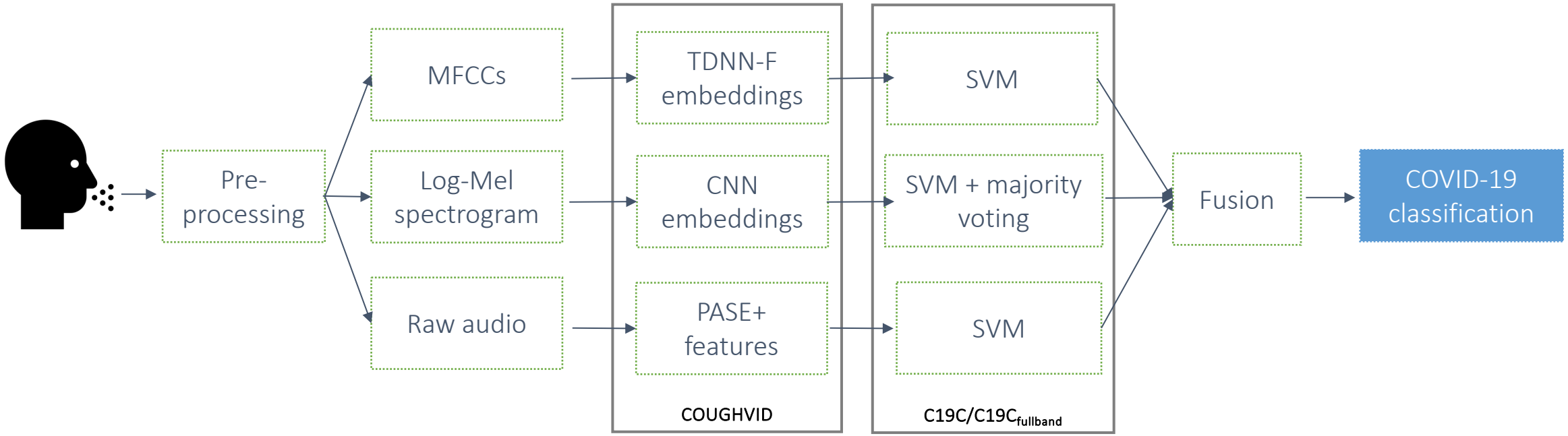
- Corpus for evaluating: COVID-19 COUGH (University of Cambridge)
- Auxiliary corpus for pre-training/fine-tuning: COUGHVID corpus (EPFL)
  - ~15k cough samples
  - ~10k with self-reported annotations COVID-19/symptomatic/healthy
  - Some samples annotated by pneumologists:
    - type of cough
    - cough severity
    - presence of audible dyspnea, wheezing, stridor, choking, and nasal congestion
    - diagnosis

Orlandic, et al. (2020). *The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms, preprint arXiv:2009.11644.*

How to deal with data scarcity

Transfer Learning for COVID-19 detection from cough

- Corpus for evaluating: COVID-19 COUGH (University of Cambridge)
- Auxiliary corpus for pre-training/fine-tuning: COUGHVID corpus (EPFL)



How to deal with data scarcity

Transfer Learning for COVID-19 detection from cough

- Corpus for evaluating: COVID-19 COUGH (University of Cambridge)
- Auxiliary corpus for pre-training/fine-tuning: COUGHVID corpus (EPFL)

TDNN-F embeddings

- Simplification of x-vector architecture
- Step 1: trained for age regression + gender classification
- Step 2: multitask classification
  - cough type (dry/wet);
  - presence of dyspnea (yes/no);
  - presence of wheezing (yes/no);
  - diagnosis (healthy cough/lower tract infection/upper tract infection/obstructive disease/COVID-19);
  - severity (healthy cough/mild/severe).

How to deal with data scarcity

Transfer Learning for COVID-19 detection from cough

- Corpus for evaluating: COVID-19 COUGH (University of Cambridge)
- Auxiliary corpus for pre-training/fine-tuning: COUGHVID corpus (EPFL)

TDNN-F embeddings

- Simplification of x-vector architecture
- Step 1: trained for age regression + gender classification
- Step 2: multitask classification
  - cough type (dry/wet);
  - presence of dyspnea (yes/no);
  - presence of wheezing (yes/no);
  - diagnosis (healthy cough/lower tract infection/upper tract infection/obstructive disease/COVID-19);
  - severity (healthy cough/mild/severe).

CNN embeddings

- VGGish model – CNN [1]
- Pre-trained for audio classification (5.4 Mh YouTube data)
- Fine-tuned on COUGHVID

[1]. Hershey, et al. (2017). *CNN architectures for large-scale audio classification representations, ICASSP.*

How to deal with data scarcity

Transfer Learning for COVID-19 detection from cough

- Corpus for evaluating: COVID-19 COUGH (University of Cambridge)
- Auxiliary corpus for pre-training/fine-tuning: COUGHVID corpus (EPFL)

TDNN-F embeddings

- Simplification of x-vector architecture
- Step 1: trained for age regression + gender classification
- Step 2: multitask classification
  - cough type (dry/wet);
  - presence of dyspnea (yes/no);
  - presence of wheezing (yes/no);
  - diagnosis (healthy cough/lower tract infection/upper tract infection/obstructive disease/COVID-19);
  - severity (healthycough/mild/severe).

CNN embeddings

- VGGish model – CNN
- Pre-trained for audio classification (5.4 Mh YouTube data)
- Fine-tuned on COUGHVID

PASE+ embeddings

- Trained on speech (LibriSpeech)

and

- Trained on coughs (COUGHVID)

How to deal with data scarcity

Transfer Learning for COVID-19 detection from cough

- Classification results in unweighted average recall (UAR): higher is better
- Best classification results on test set: 69.3% UAR

		dev	$dev_{fullband}$	test
TDNN-F embeddings	Trained on COUGHVID (step1)	68.8	63.6	-
	Fine-tuned on COUGHVID (step2)	68.1	62.3	-
CNN embeddings	Trained on YouTube data	66.9	62.4	-
	Fine-tuned on COUGHVID	71.2*	65.6	62.3
PASE+ embeddings	Trained on Librispeech	63.1	61.7	-
	Fine-tuned on COUGHVID	67.4	**66.8***	64.1
Calibrated Fusion	Fusion of predictions	**72.3***	66.1	**69.3**

How to deal with data scarcity?

Machine Learning-based strategies:

- Data augmentation
- Intelligent labelling paradigms
- Transfer learning

Alternative data collection strategies:

- Crowdsourcing
  - Coswara [N. Sharma, 2020]
  - ColiveVoice [https://www.colivevoice.org/en/]
  - CLAC [R. Haulcy, 2021]

How to deal with data scarcity?

Machine Learning-based strategies:

- Data augmentation
- Intelligent labelling paradigms
- Transfer learning

Alternative data collection strategies:

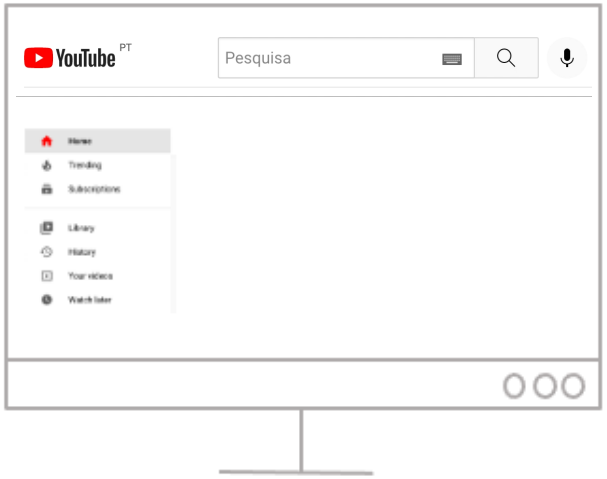
- Crowdsourcing
- **Mining in-the-wild data repositories**

  - Eg.:

J. Correia, F. Teixeira, **C. Botelho**, I. Trancoso, B. Raj, **(2021).** *The in-the-wild speech medical corpus, ICASSP 2021.*

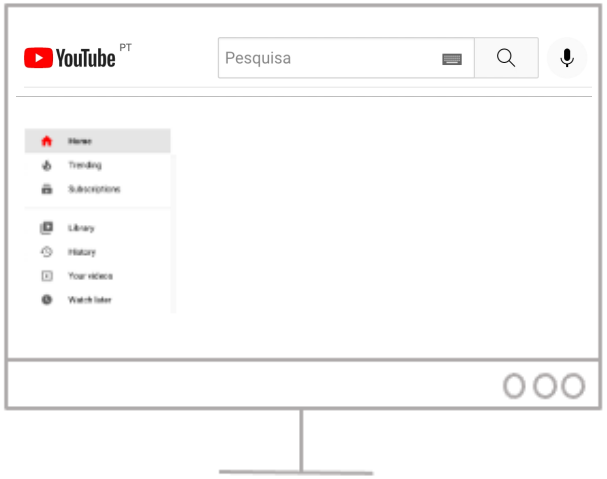
How to deal with data scarcity

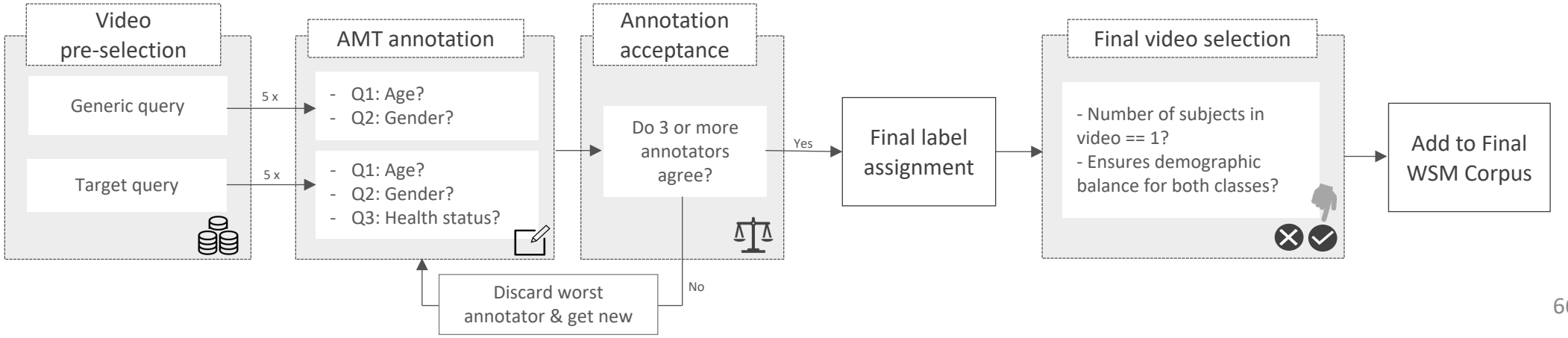
Mining in-the-wild repositories: WSM corpus



How to deal with data scarcity

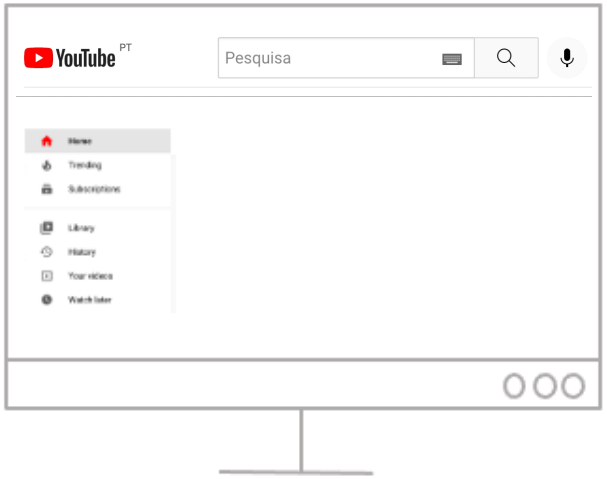
Mining in-the-wild repositories: WSM corpus



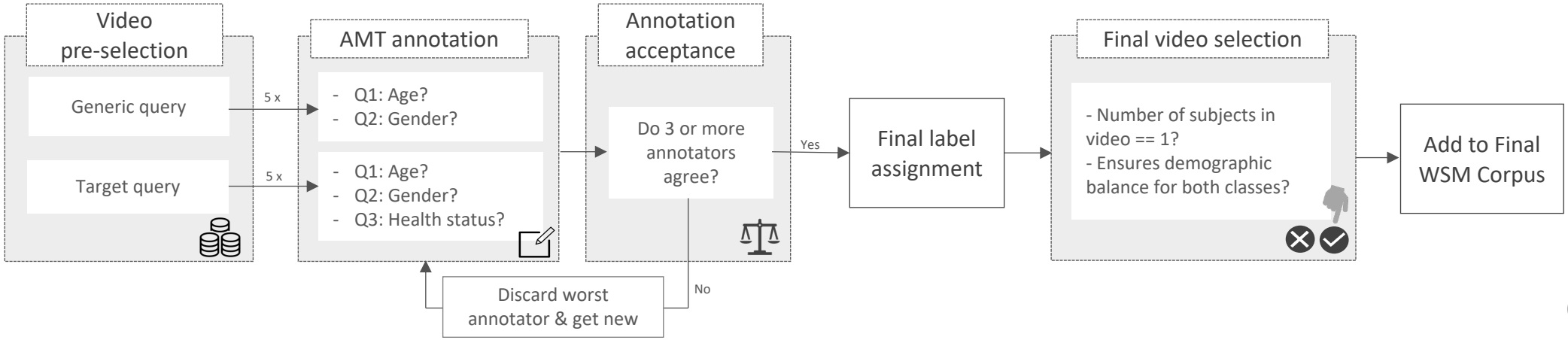


How to deal with data scarcity

Mining in-the-wild repositories: WSM corpus



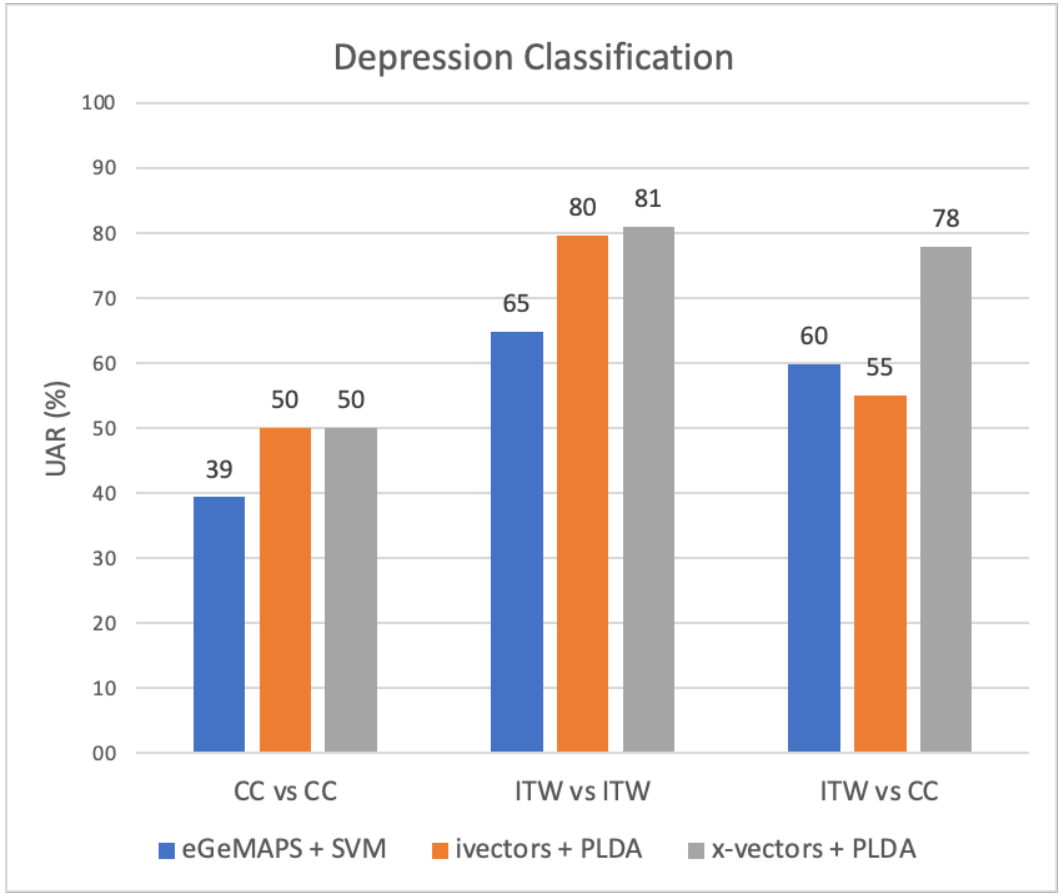
WSM Corpus dataset	Partition	Group	# Videos	# Hours	Age	Gender (m:f)
Depression	*train*	*D*	191	27.6	30 ±5	86 : 105
		HC	199	29.5	30 ±5	93 : 106
	devel	*D*	39	6.0	30 ±5	19 : 20
		HC	40	5.4	30 ±6	19 : 21
	test	*D*	37	6.7	29 ±5	18 : 19
		HC	37	7.8	29 ±5	18 : 19
Parkinson's disease	*train*	*PD*	157	18.5	45 ±10	79 : 78
		HC	155	20.7	43 ±13	76 : 79
	devel	*PD*	24	1.8	45 ±10	12 : 12
		HC	23	2.6	42 ±10	11 : 12
	test	*PD*	28	4.1	45 ±9	14 : 14
		HC	26	5.8	43 ±12	11 : 15



How to deal with data scarcity

Mining in-the-wild repositories: WSM corpus

- Is the self-reported health status a valid proxy for true health?

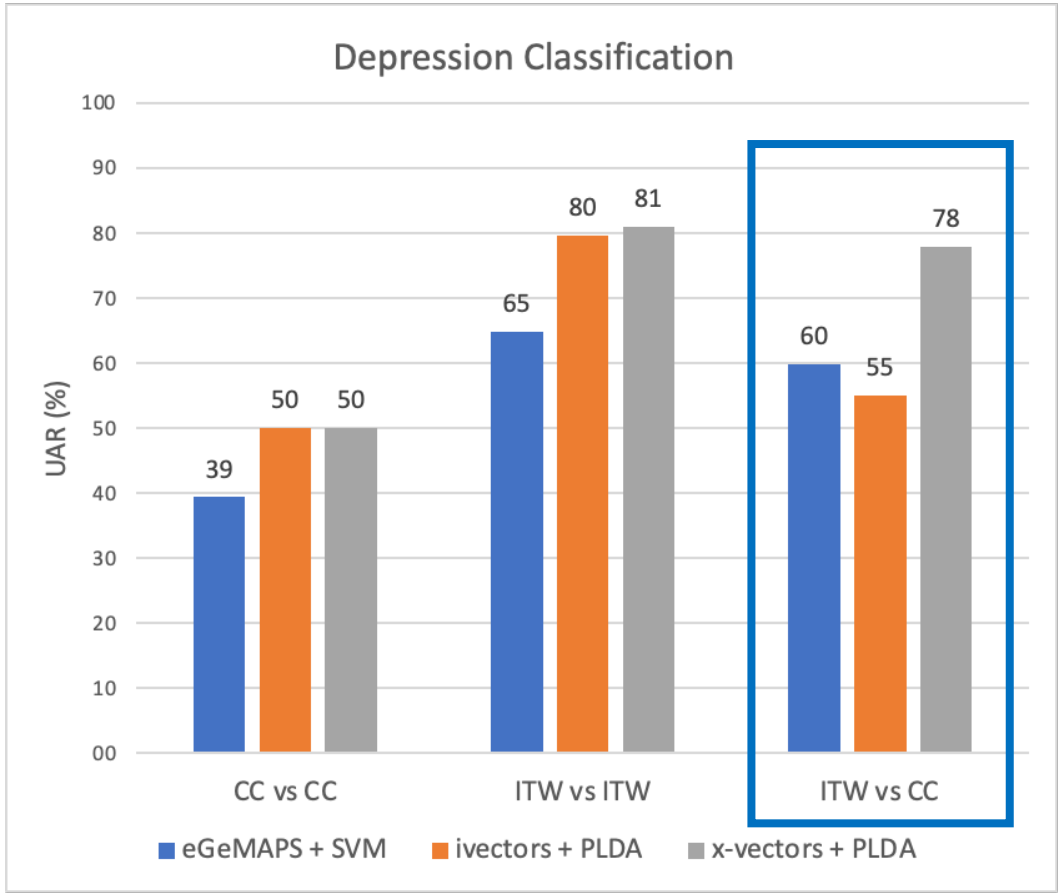


Correia (2021). *In-the-wild detection of speech affecting diseases.* PhD thesis, Carnegie Mellon University - University of Lisbon.

How to deal with data scarcity

Mining in-the-wild repositories: WSM corpus

- Is the self-reported health status a valid proxy for true health?



Correia (2021). *In-the-wild detection of speech affecting diseases.* PhD thesis, Carnegie Mellon University - University of Lisbon.

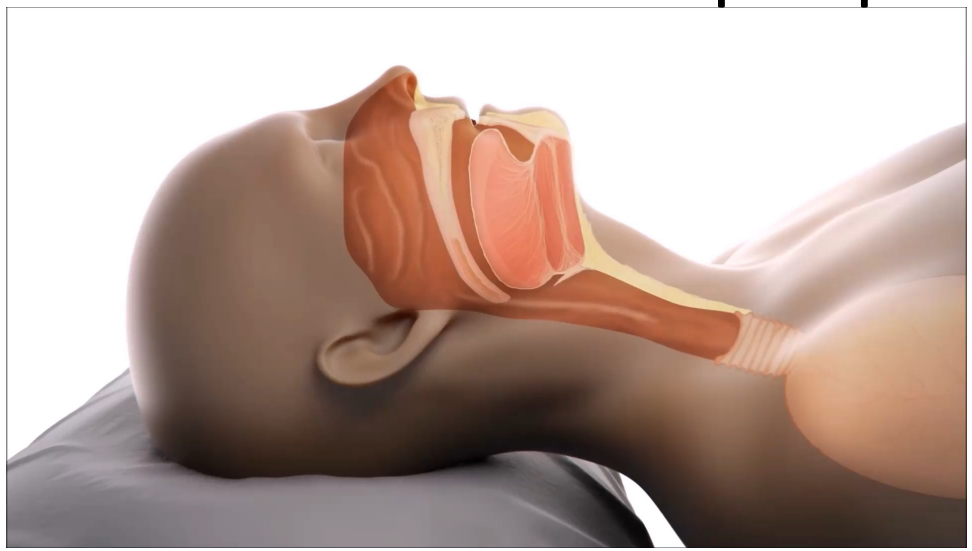
Automatic Disease Detection from speech

- Results have been promising, but there are some challenges...

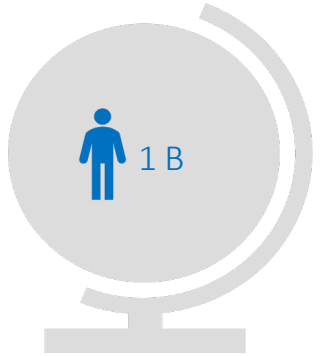
1. How to deal with data scarcity?
2. Can other modalities complement the speech signal for disease detection?
3. Can we monitor disease progression through time/make predictions into the future?
4. Can we provide interpretable reasoning that is useful to the medical community?
5. Are results generalizable? Can we transfer them across different datasets?

Modalities complementary to the speech signal

Obstructive sleep apnea detection



Source: Buteyko Clinic International, https://www.youtube.com/watch?v=cxEWHV67JIU



Associated with: Diabetes, cardiovascular diseases, cognitive impairment, fatigue, mood changes, ….

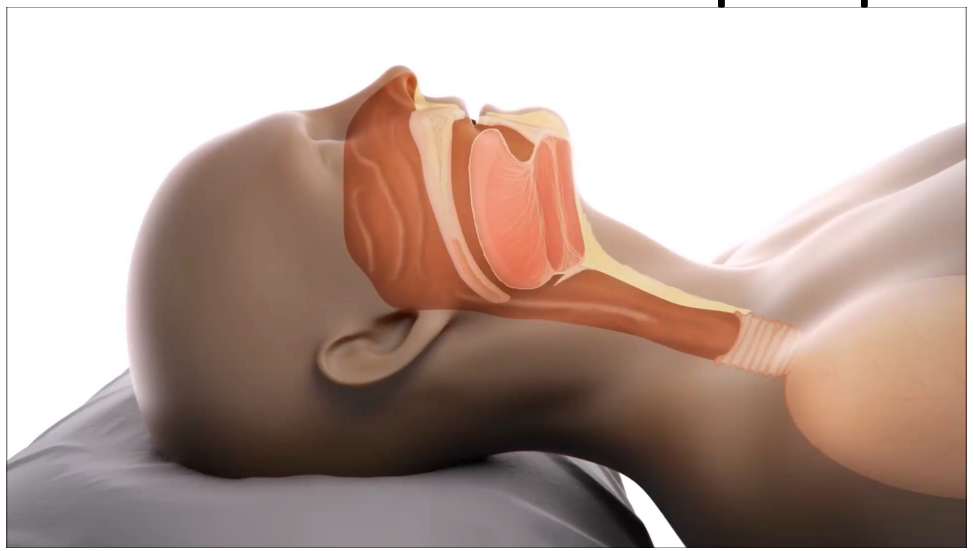
Diagnosis: Polysomnography

Risk factors: obesity, aging

Benjafield (2019). *Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis,* The Lancet Respiratory Medicine, 7(8):687-698.

Modalities complementary to the speech signal

Obstructive sleep apnea detection



Source: Buteyko Clinic International, https://www.youtube.com/watch?v=cxEWHV67JIU

- Anatomical alterations
- Altered muscle tone and pharyngeal wall properties

↓

Articulatory, phonation and resonance anomalies

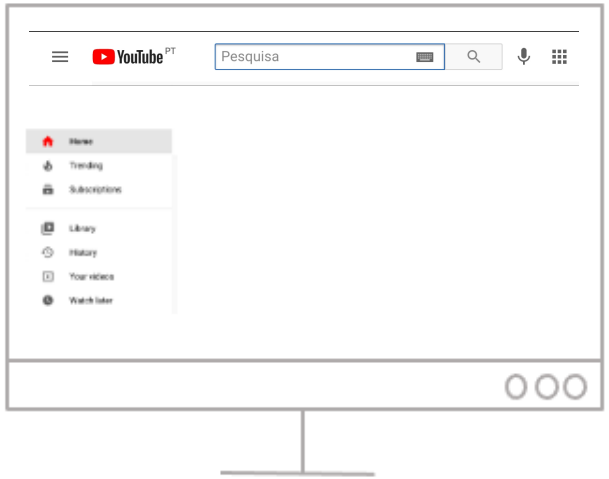
Associated with: Diabetes, cardiovascular diseases, cognitive impairment, fatigue, mood changes, ….

Diagnosis: Polysomnography

Risk factors: obesity, aging

Modalities complementary to the speech signal

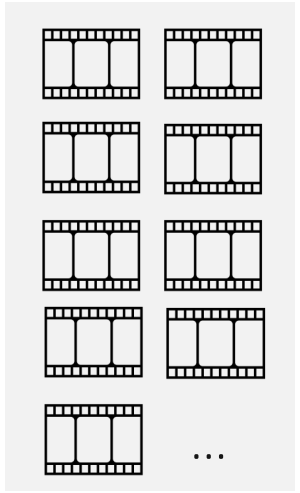
Obstructive sleep apnea detection



- Pilot corpus: WOSA
- 22 OSA subjects (m: 12, f: 19)
- 18 Controls (m: 9, f: 9)

Modalities complementary to the speech signal

Obstructive sleep apnea detection



Speech Activity Detection

Feature Extraction

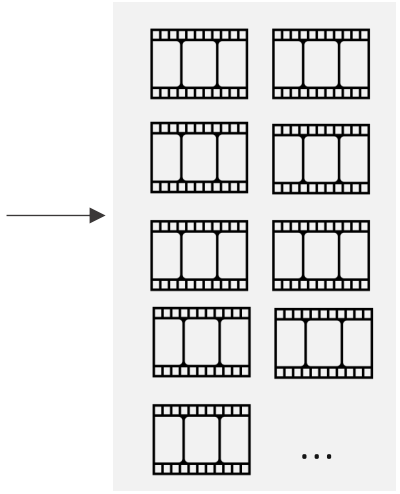
- X-vectors
- PASE+
- KB

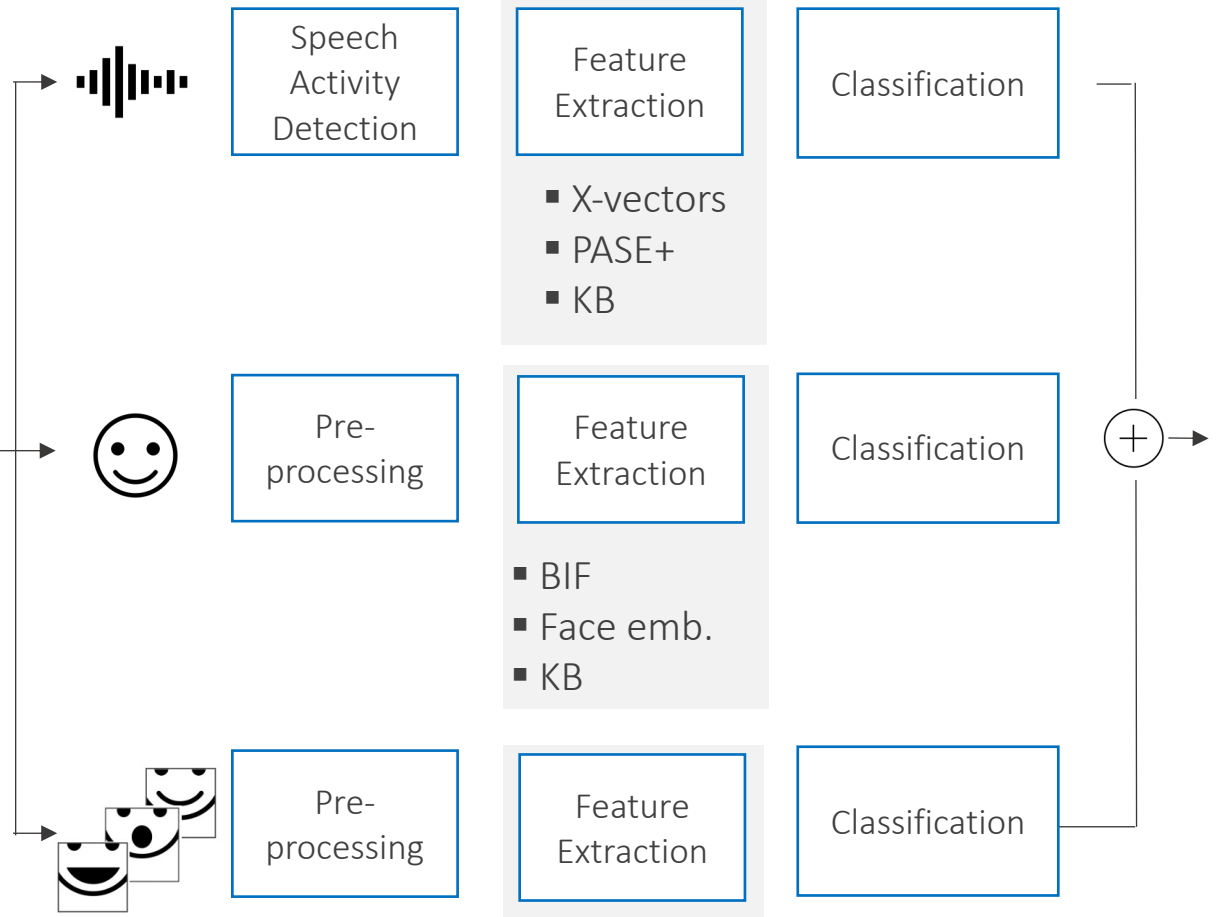
Classification

- Pilot corpus: WOSA
- 22 OSA subjects (m: 12, f: 19)
- 18 Controls (m: 9, f: 9)

Modalities complementary to the speech signal

Obstructive sleep apnea detection





- Pilot corpus: WOSA
- 22 OSA subjects (m: 12, f: 19)
- 18 Controls (m: 9, f: 9)

Modalities complementary to the speech signal

Obstructive sleep apnea detection

Modality	Experiment	Accuracy (sample) (%)	Majority Vote Accuracy (%)
Speech	X-vectors	65.1	67.5
Facial images	Raw images	76.3	77.5
Visual Speech	Lip reading emb.	69.8	80.0
Fusion	Late fusion		**82.5**

- Speech modalities achieves an accuracy of 67.5%
- Fusion of the modalities enables achieving an accuracy of 82.5%

Botelho et al. (2021). *Visual Speech for Obstructive Sleep Apnea Detection*, Interspeech.

Automatic Disease Detection from speech

- Results have been promising, but there are some challenges...

1. How to deal with data scarcity?
2. Can other modalities complement the speech signal for disease detection?
3. Can we monitor disease progression through time/make predictions into the future?
4. Can we provide interpretable reasoning that is useful to the medical community?
5. Are results generalizable? Can we transfer them across different datasets?

Automatic Disease Detection from speech

- Results have been promising, but there are some challenges...

1. How to deal with data scarcity?
2. Can other modalities complement the speech signal for disease detection?
3. Can we monitor disease progression through time/make prognosis?
4. Can we provide interpretable reasoning that is useful to the medical community?
5. Are results generalizable? Can we transfer them across different datasets?

Weiner et al. (2023). *Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews.* Interspeech.

Botelho et al. (2023). *Towards reference speech characterization for health application,* Interspeech.

Botelho et al. (2022). *Challenges of using longitudinal and cross-domain corpora on studies of pathological speech,* Interspeech.

Berisha et al. (2022). *Are reported accuracies in the clinical speech machine learning literature overoptimistic?,* Interspeech.

Automatic Disease Detection from speech

- Results have been promising, but there are some challenges...

1. How to deal with data scarcity?
2. Can other modalities complement the speech signal for disease detection?
3. Can we monitor disease progression through time/make prognosis?
4. Can we provide interpretable reasoning that is useful to the medical community?
5. Are results generalizable? Can we transfer them across different datasets?

Cross-lingual solutions?

Personalized references?

Comorbidities?

Weiner et al. (2023). *Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews.* Interspeech.

Botelho et al. (2023). *Towards reference speech characterization for health application,* Interspeech.

Botelho et al. (2022). *Challenges of using longitudinal and cross-domain corpora on studies of pathological speech,* Interspeech.

Berisha et al. (2022). *Are reported accuracies in the clinical speech machine learning literature overoptimistic?,* Interspeech.

Where to get started?

Where to get started?

Tools and libraries: Feature Extraction

- Librosa [https://librosa.org/doc/latest/index.html]
  - Python library for audio and music analysis and feature extraction

['ɹɒɹ.səl,maʊθ]

- Parselmouth [https://parselmouth.readthedocs.io/en/stable/, https://www.fon.hum.uva.nl/praat/]
  - Python library for the *Praat* software -- very popular in linguistics, data analysis and visualizations

openSMILE
audio feature extraction

- OpenSMILE [https://www.audeering.com/research/opensmile/, https://audeering.github.io/opensmile-python/]
  - Very popular features for paralinguistics (over 20k features!)
  - Most used configurations: ComPaRe2016 (~6k features) and eGeMAPS (88 features)
  - [Florian Eyben, Martin Wöllmer, Björn Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor". In: Proc. ACM Multimedia 2010, pp. 1459-1462.]

TorchAudio

- TORCHAUDIO [https://pytorch.org/audio/stable/index.html]
  - signal processing with PyTorch
  - I/O, signal processing, datasets, etc.

Where to get started?

Tools and libraries: Models

- Hugging Face [https://huggingface.co/]
  - Datasets: LibriSpeech, Common Voice, GigaSpeech, FLEURS, etc.
  - Pre-trained models: eg. Whisper (ASR), wav2vec (ASR), wavLM (ASR), ECAPA-TDNN (speaker-verification), etc.

- S3PRL [https://github.com/s3prl/s3prl]
  - Toolkit that provides SSL upstream pre-trained models for speech to fine-tine to downstream tasks.
  - Eg: TERA, HuBERT, wav2vec2.0, WAVLM, etc,

- SpeechBrain [https://speechbrain.github.io/]
  - Open-source and all-in-one speech toolkit based on PyTorch.

- ESPNET [https://github.com/espnet/espnet]
  - Toolkit dedicated to end-to-end speech processing (e.g. Speech Recognition, Text-to-Speech, Speech Enhancement)

- KALDI [https://kaldi-asr.org/]
  - Toolkit dedicated to hierarchical (HMM/DNN) speech recognition

What we have discussed today?

- Introduction to speech machine learning
- Automatic speech recognition overview
  - 1st Deep Learning success story in ASR
  - 2nd Deep Learning success story in ASR
- Automatic Disease Detection from Speech
  - Pipeline
  - Dealing with data scarcity
  - Complimentary modalities
  - Other challenges
- Where to get started: tools and libraries

Where to get started?

HLT @ INESC-ID

- HLT – The Human Language Technologies Lab

(Some) research topics

Major research projects

ACCELERAT.AI

Center for Responsible AI

Where to get started?

HLT @ INESC-ID

- HLT – The Human Language Technologies Lab

inesc id lisboa

TÉCNICO LISBOA

Thank you!

Obrigada!

/ˌɔβriˈɣaðɐ/