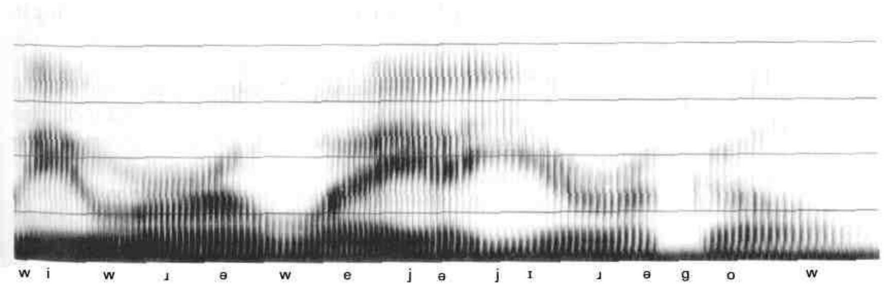
Towards Responsible Speech Processing

Isabel Trancoso

inesc id lisboa

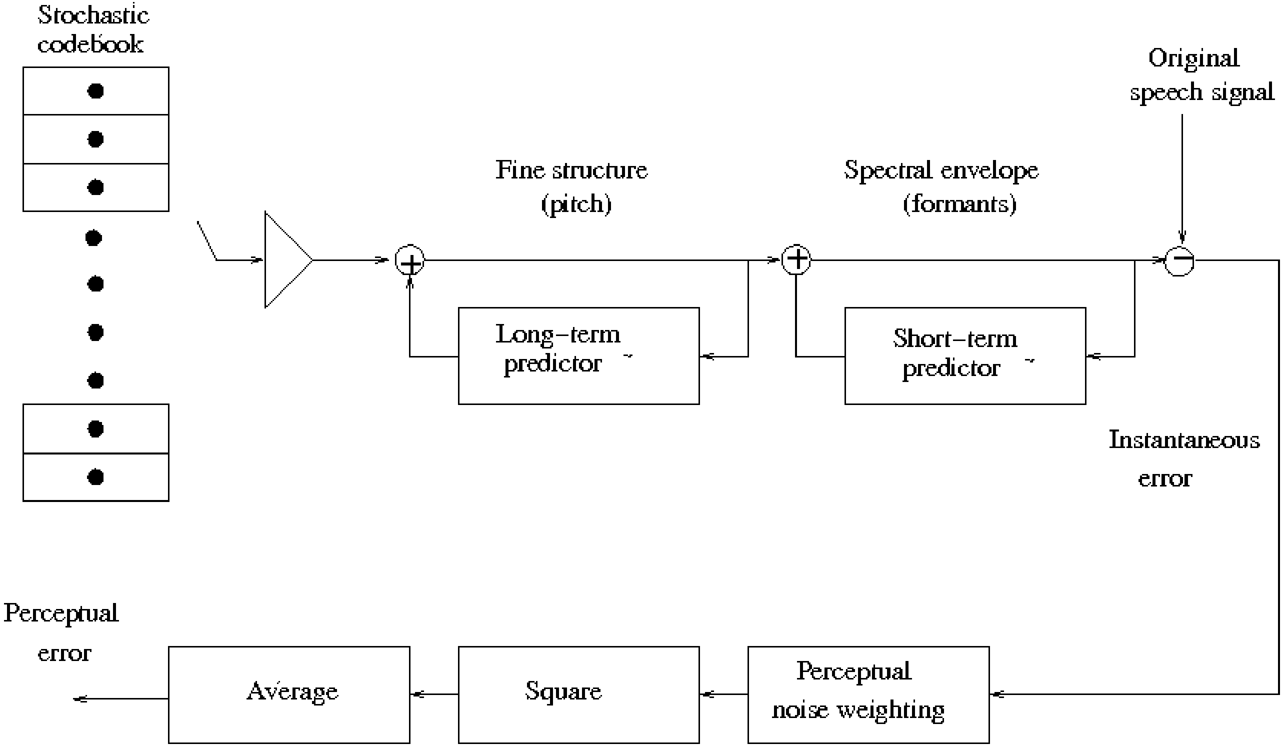
TÉCNICO LISBOA

From the old HP2116C@IST…
… to a CRAY supercomputer @AT&T Bell Labs



http://home.cc.umanitoba.ca/%7Ekrussll/13/sec4/specgram.htm

CELP Coder


Stochastic codebook
Original speech signal
Fine structure (pitch)
Spectral envelope (formants)
Long−term predictor
Short−term predictor
Instantaneous error
Perceptual error
Average
Square
Perceptual noise weighting

In the 1980s

determination
technology
aspects
understanding
synthesiser
experimental
computer
knowledge-based
experiments
Voice recognition
conversion
intonation
dialogue
German
discrimination
simulation
acoustic-phonetic
warping
Text-to-speech synthesis
spectral
Speaker-independent
clustering
communication
vocabulary
syllables
study
representation
output
segmentation
signal
without
English
issues
isolated word recognition
control
Chinese
conversion system
adaptive
recognition
telephone
quality
technique
feature
prosodic
residual
model
rule
LPC
hindi
New
digit
analysis
phoneme
human
Real-time
voice
DTW
base
linguistic
performance
service
phonetic
synthesis
auditory
TES
use
data
ASR
measure
vowel
speech
vocal
tract
speech recognition
text-to-speech system
French
information
formant
aid
acoustic
coder
Word
dynamic
method
Time
pitch
coding
network
automatic
effects
text
design
application
estimation
Italian
sounds
system
speech synthesis
Markov
word recognition
vector quantization
research
linear
simple
language
parameters
approach
Automatic speech recognition
patterns
algorithm
input
result
modelling
text-to-speech
comparison
implementation
written
recognition system
prediction
large vocabulary
dutch
Spoken
robust
second
independent
extraction
isolated
hidden
Japanese
coarticulation
vocoder
predictive
grapheme-to-phoneme

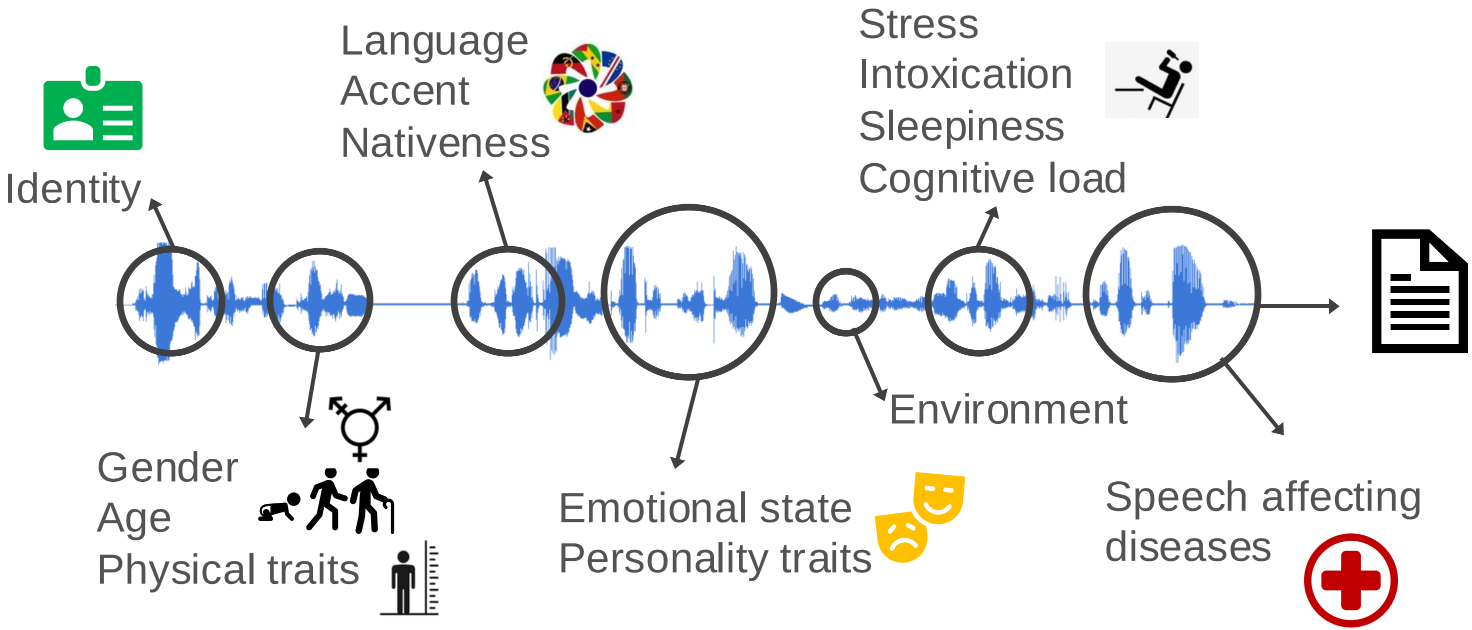
In the 2020s

speech
recognition
learning
dialogue
self-supervised
models
evaluation
automatic
language
large
transformers
representations
data
model
chatgpt
speaker
classification
detection
asr
robust
machine
benchmarking
benchmark
open-domain
generation
towards
based
diarization
conversational
memorizing
analysis
via
leveraging
pre-training
publicly
tasks
deep
utility
systems
representation
synthetic
children's
timing
datasets
adapters
text
subjective
generative
empathetic
prosody
features
era
annotating
exploring
chatbots
human
approach
truth
selection
available
using

Pillars of Responsible AI

- Robustness & Safety
- Fairness & Inclusion
- Explainability
- Privacy & Security
- Sustainability
- Accountability & Governance
- User Agency, Trust & Wellbeing

Info in speech



Identity

Language
Accent
Nativeness

Stress
Intoxication
Sleepiness
Cognitive load

Gender
Age
Physical traits

Emotional state
Personality traits

Environment

Speech affecting diseases

Pillars of Responsible Speech Processing

- Robustness & Safety
- **Fairness & Inclusion**
- Explainability
- Privacy & Security
- Sustainability
- Accountability & Governance
- User Agency, Trust & Wellbeing

Pillars of Responsible Speech Processing

- Robustness & Safety
- Fairness & Inclusion
- **Explainability**
- Privacy & Security
- Sustainability
- Accountability & Governance
- User Agency, Trust & Wellbeing

Pillars of Responsible Speech Processing

- Robustness & Safety
- Fairness & Inclusion
- Explainability
- **Privacy & Security**
- Sustainability
- Accountability & Governance
- User Agency, Trust & Wellbeing

Pillars of Responsible Speech Processing

- Robustness & Safety
- Fairness & Inclusion
- Explainability
- Privacy & Security
- **Sustainability**
- Accountability & Governance
- User Agency, Trust & Wellbeing

Fairness & Inclusion

- Models tend to reflect stereotypes present in their training data; Internet-trained models have internet-scale biases
- Bias along the dimensions of accent, race, gender, age, ...
  - M. Adda-Decker and L. Lamel. Do speech recognizers prefer female speakers? Interspeech 2005.
  - R. Tatman. Gender and Dialect Bias in YouTube’s Automatic Captions. EthNLP@EACL 2017.
  - D. Harwell. The accent gap. Washington Post, 2018.
  - L. Lima. Empirical analysis of bias in voice-based personal assistants. Companion of The WWW Conference, 2019.
  - A.Koenecke, Racial disparities in speech recognition, Proc. National Academy of Sciences, 2020
  - A. Kulkarni et al., Unveiling Biases while Embracing Sustainability, Interspeech 2024
  - S. Feng et al., Towards inclusive automatic speech recognition, Computer Speech and Language, 2024

Fairness & Inclusion

- Models tend to reflect stereotypes present in their training data; Internet-trained models have internet-scale biases
- Bias along the dimensions of accent, race, gender, age, ...
    - » M. Adda-Decker and L. Lamel. Do speech recognizers prefer female speakers? Interspeech 2005.
    - » R. Tatman. Gender and Dialect Bias in YouTube's Automatic Captions. EthNLP@EACL 2017.
    - » D. Harwell. The accent gap. Washington Post, 2018.
    - » L. Lima. Empirical analysis of bias in voice-based personal assistants. Companion of The WWW Conference, 2019.
    - » A.Koenecke, Racial disparities in speech recognition, Proc. National Academy of Sciences, 2020
    - » A. Kulkarni et al., Unveiling Biases while Embracing Sustainability, Interspeech 2024
    - » S. Feng et al., Towards inclusive automatic speech recognition, Computer Speech and Language, 2024
  - ❑ Child speech was recognized worst

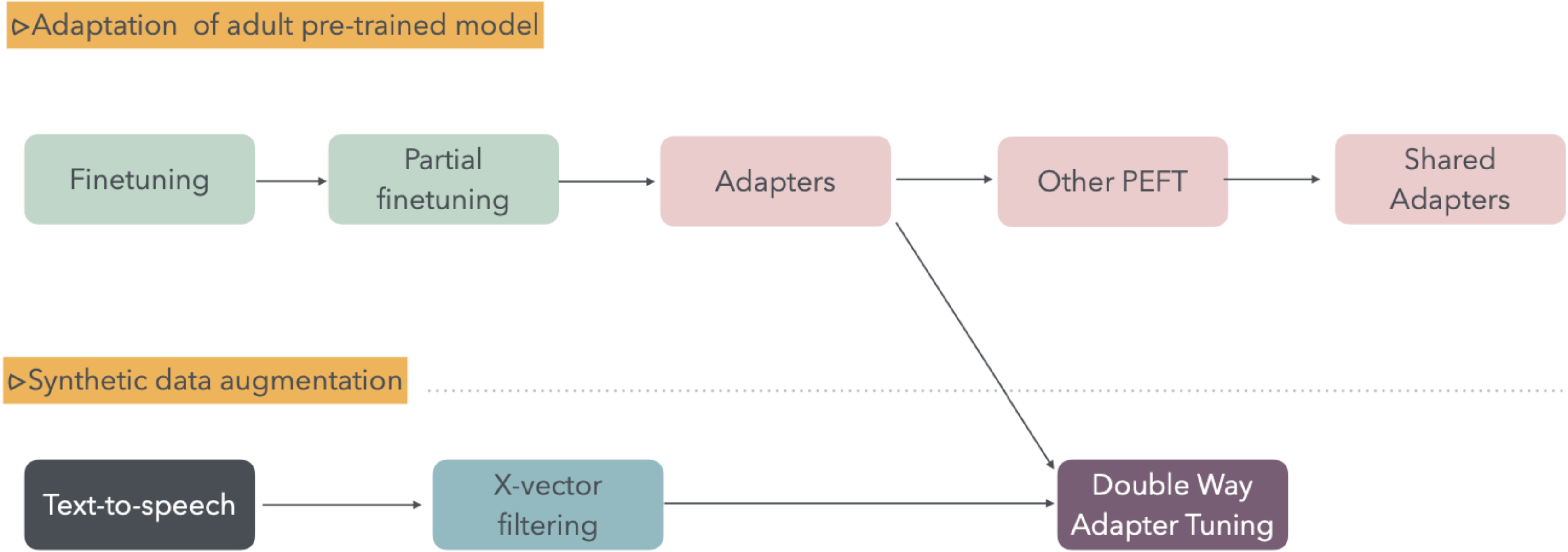
Towards improved ASR for children

PhD Thesis of Thomas Rolland, supervised by Alberto Abad

- Introduction to Partial fine-tuning: A comprehensive evaluation of end-to-end children’s automatic speech recognition adaptation (IS 2024, Thursday, SS-8)
- Exploring adapters with conformers for children’s automatic speech recognition (ICASSP 2024)
- Shared-Adapters: A novel Transformer-based parameter efficient transfer learning approach for children’s automatic speech recognition (IS 2024, Tuesday, A8-O4)
- Improved children’s automatic speech recognition combining adapters and synthetic data augmentation (ICASSP 2024)

TAPAS PRR Plano de Recuperação e Resiliência REPÚBLICA PORTUGUESA Financiado pela União Europeia NextGenerationEU

Roadmap towards improving ASR for children



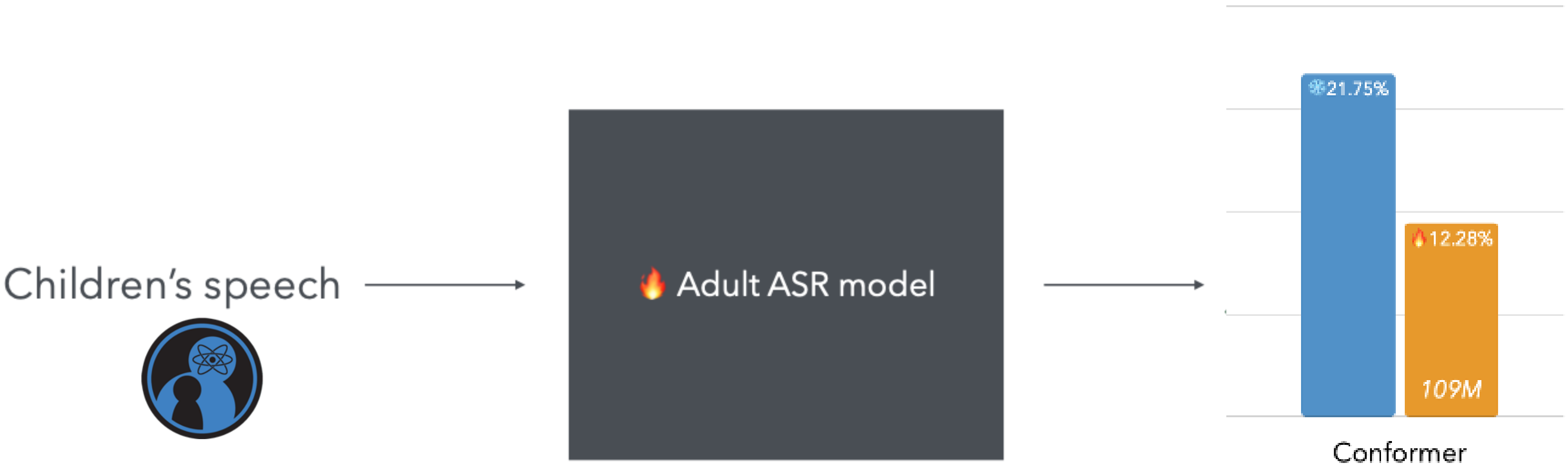
▷Adaptation of adult pre-trained model

Finetuning → Partial finetuning → Adapters → Other PEFT → Shared Adapters

▷Synthetic data augmentation

Text-to-speech → X-vector filtering → Double Way Adapter Tuning

Finetuning



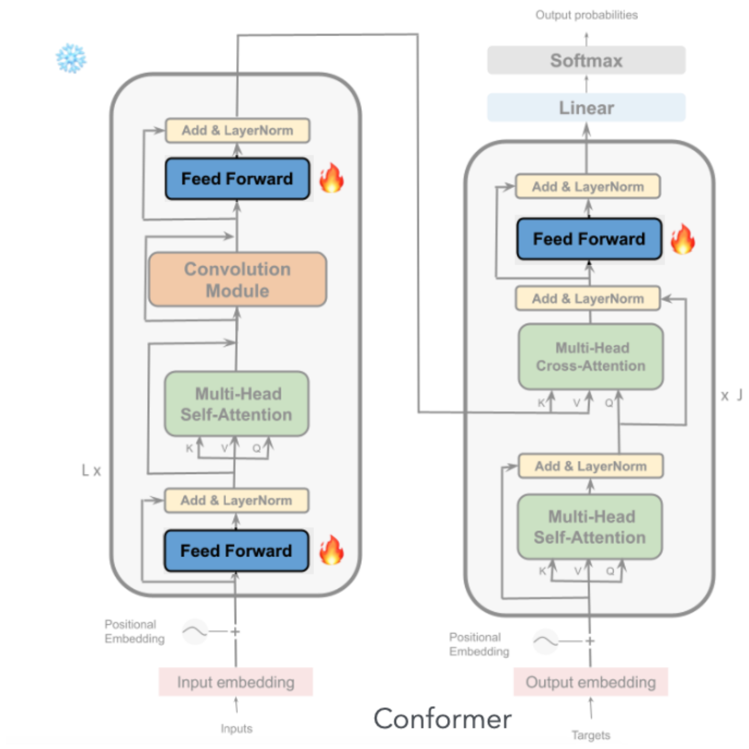
Children's speech

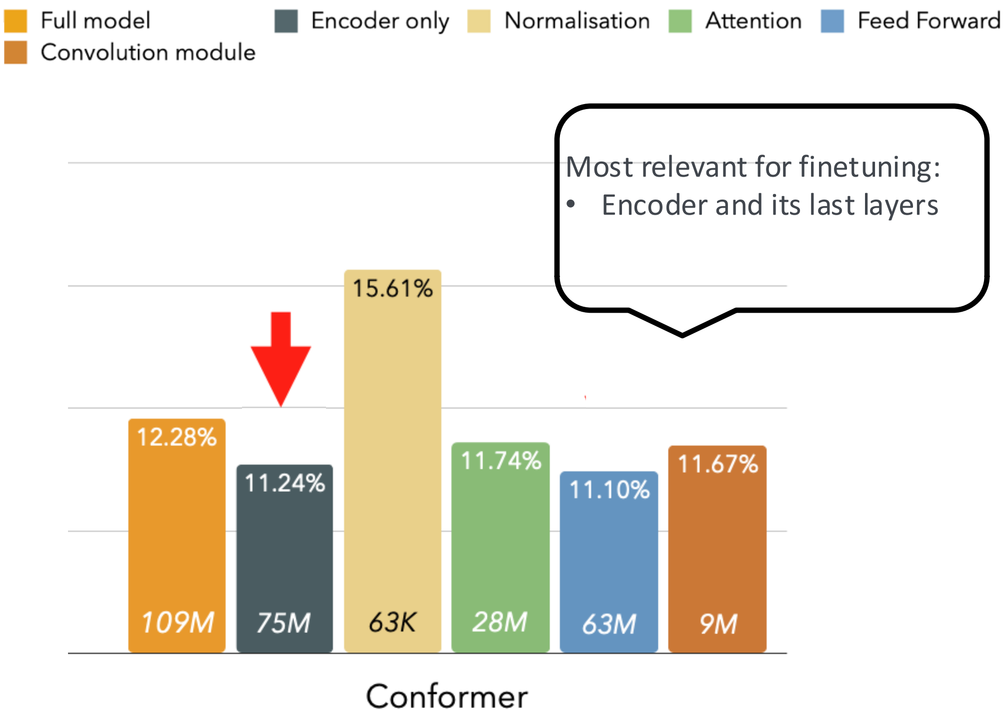
🔥 Adult ASR model

My Science Tutor
MyST
(Ward et al. 2013)

(Gulati et al. 2020)
https://huggingface.co/speechbrain/asr-conformer-transformerlm-librispeech

Partial finetuning

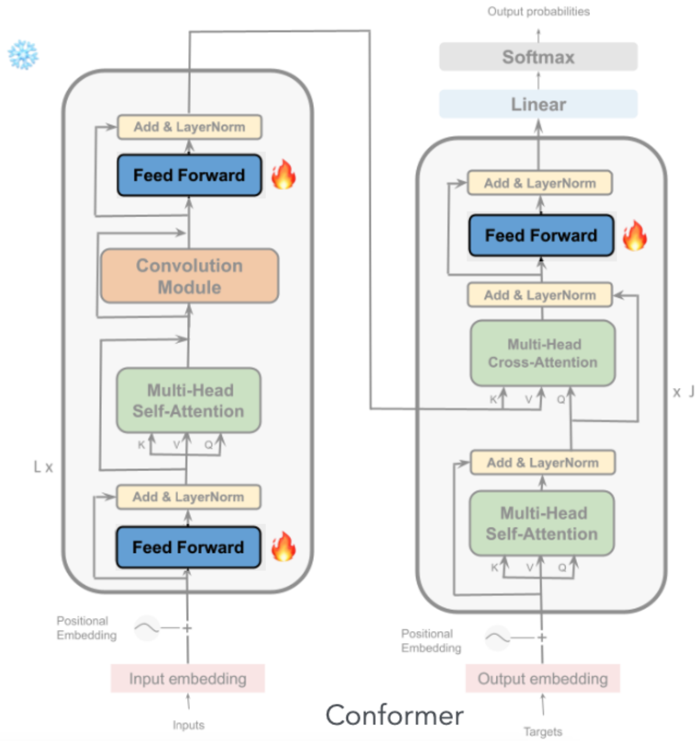


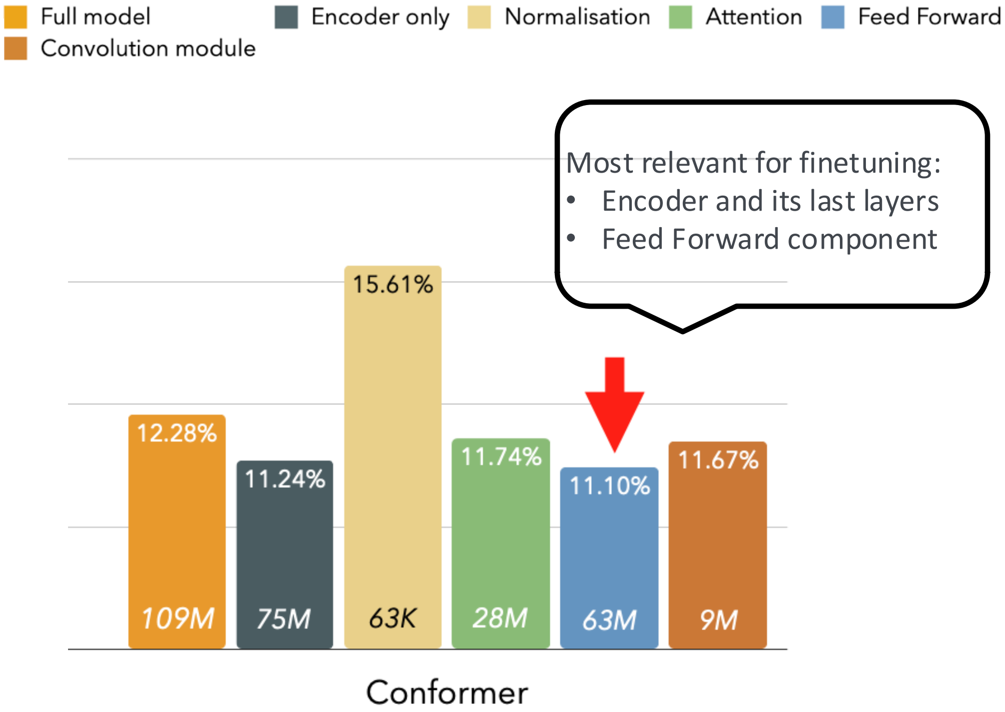


Most relevant for finetuning:

- Encoder and its last layers

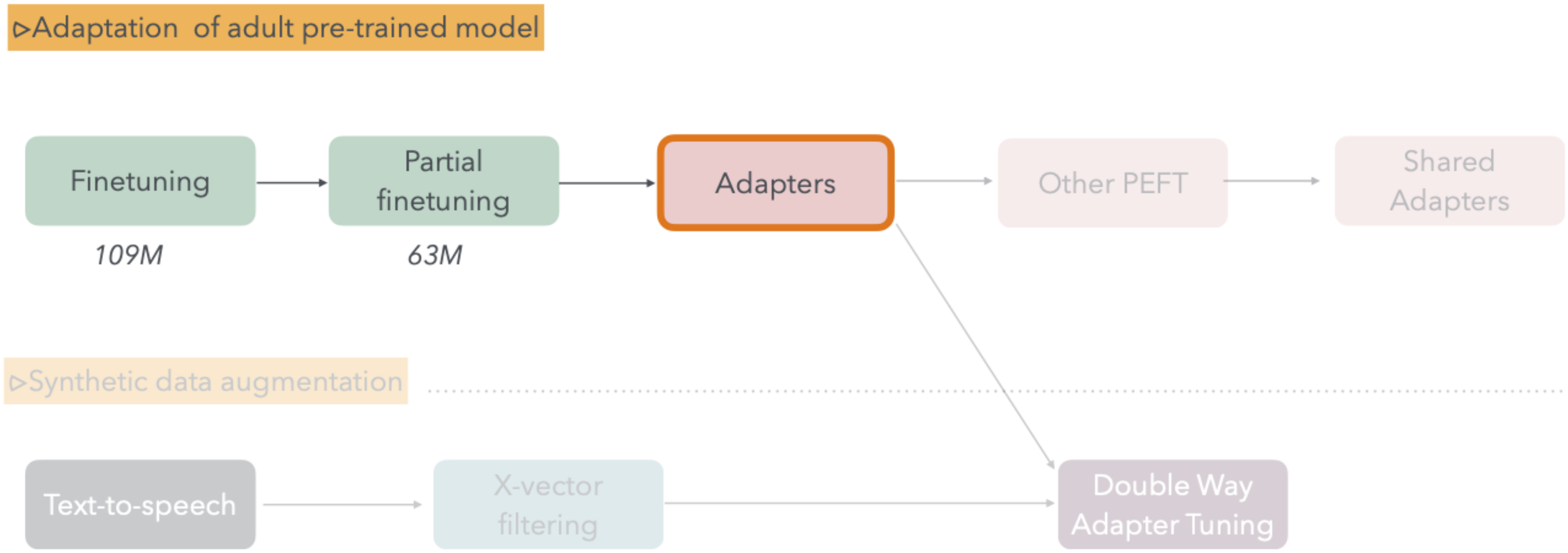
Partial finetuning





Most relevant for finetuning:
- Encoder and its last layers
- Feed Forward component

Roadmap towards improving ASR for children



▷Adaptation of adult pre-trained model

Finetuning

109M

Partial finetuning

63M

Adapters

Other PEFT

Shared Adapters

▷Synthetic data augmentation

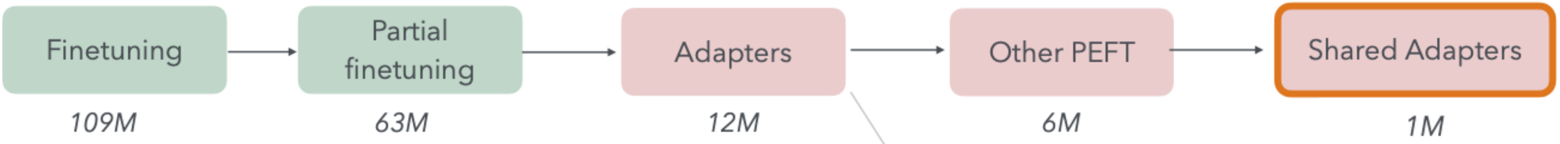
Text-to-speech

X-vector filtering

Double Way Adapter Tuning

Roadmap towards improving ASR for children

▹Adaptation of adult pre-trained model


Finetuning → Partial finetuning → Adapters → Other PEFT → Shared Adapters

109M

63M

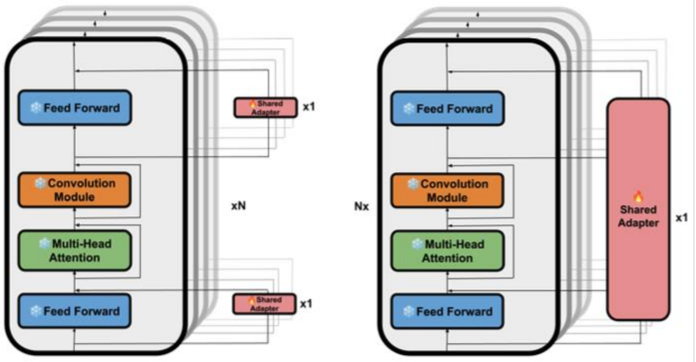
12M

6M

1M

▹Synthetic data augmentation

Text-to-speech → X-vector filtering

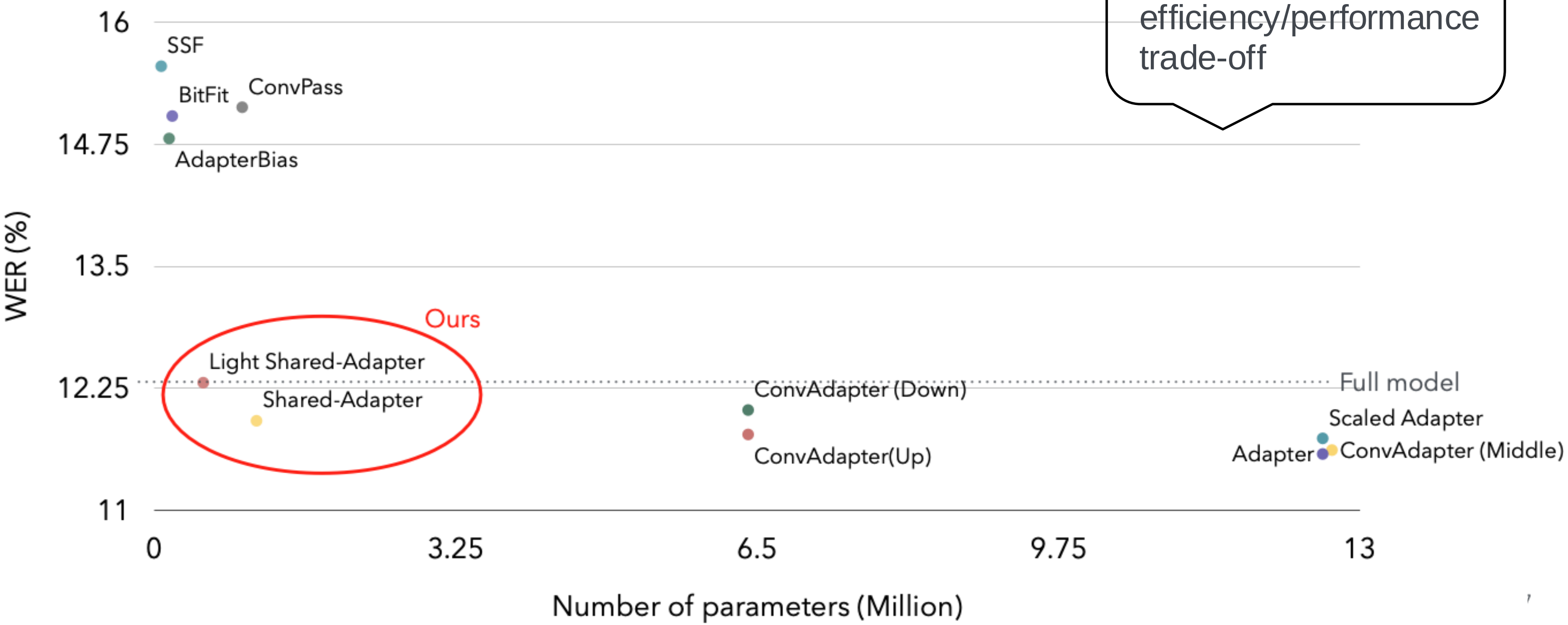


Shared-Adapter

Light Shared-Adapter

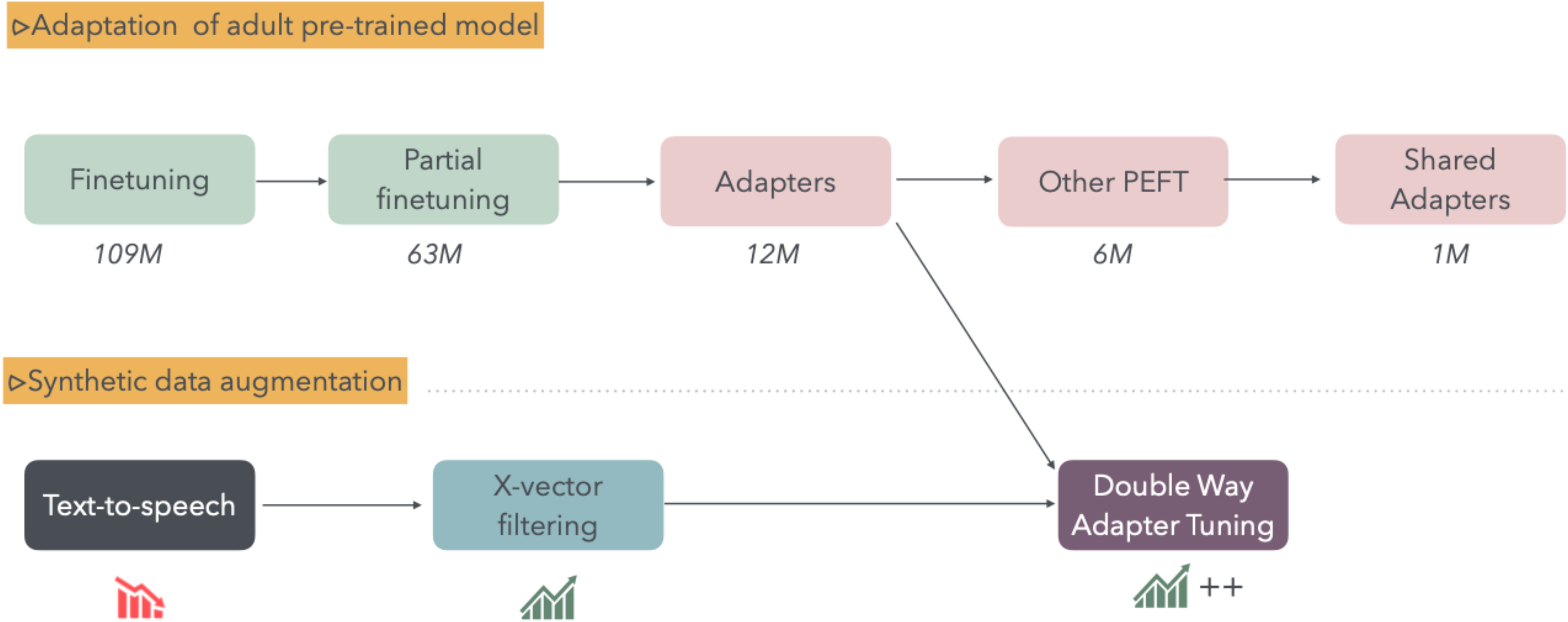
Shared adapters

Shared-Adapters:
best parameter
efficiency/performance
trade-off



SSF (Lian et al., 2022) ; BifFit (Zaken et al., 2022); ConvPass (Li et al., 2023); AdapterBias (Fu et al., 2022); ConvAdapter (Yang et al., 2023); Scaled Adapter (He et al., 2022).

Roadmap towards improving ASR for children



▷Adaptation of adult pre-trained model

Finetuning

109M

Partial finetuning

63M

Adapters

12M

Other PEFT

6M

Shared Adapters

1M

▷Synthetic data augmentation

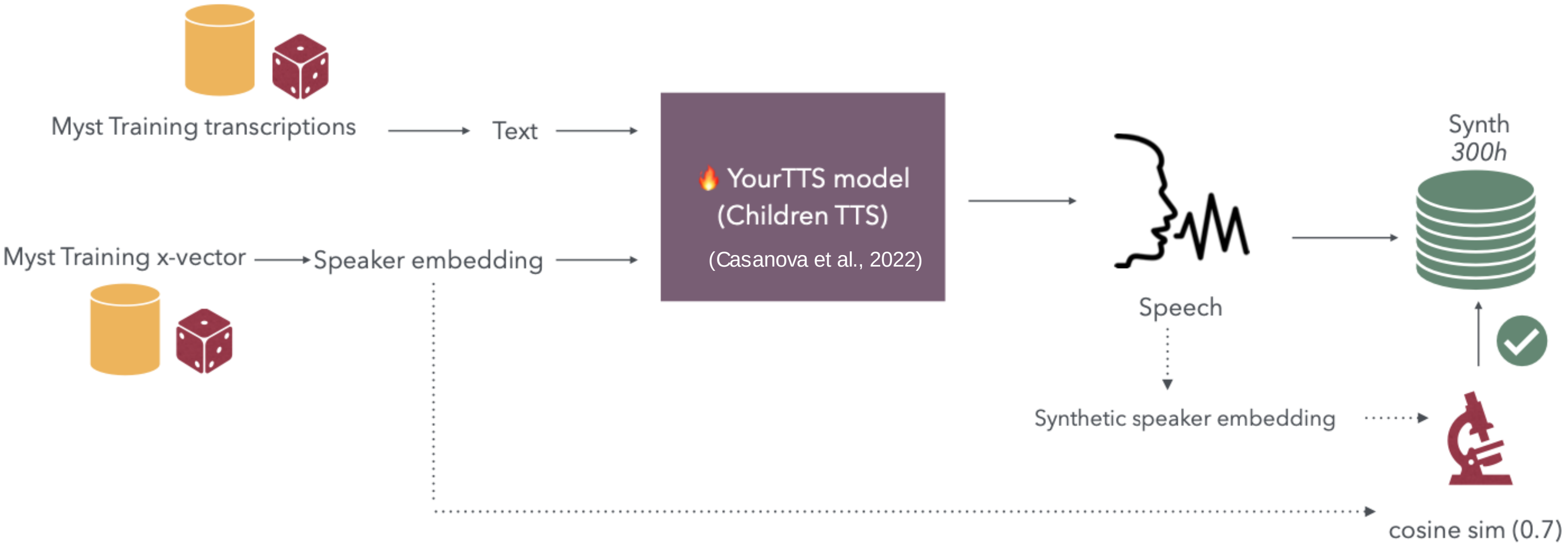
Text-to-speech

X-vector filtering

Double Way Adapter Tuning

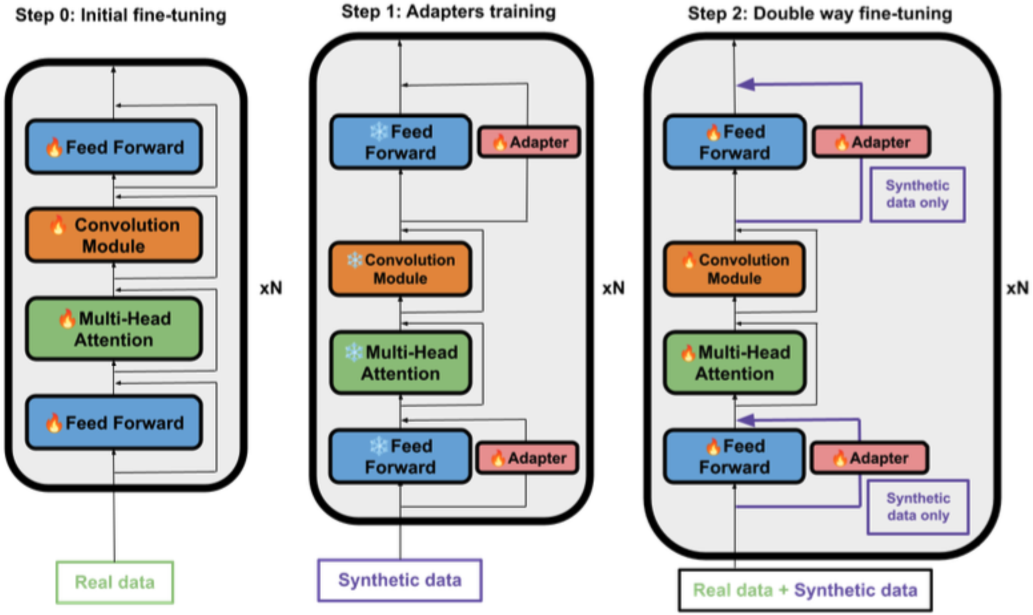
++

Synthetic data augmentation

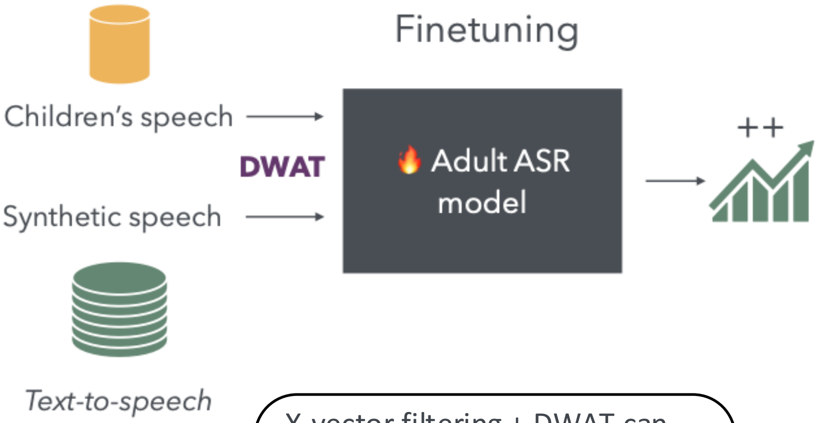


Synthetic data augmentation

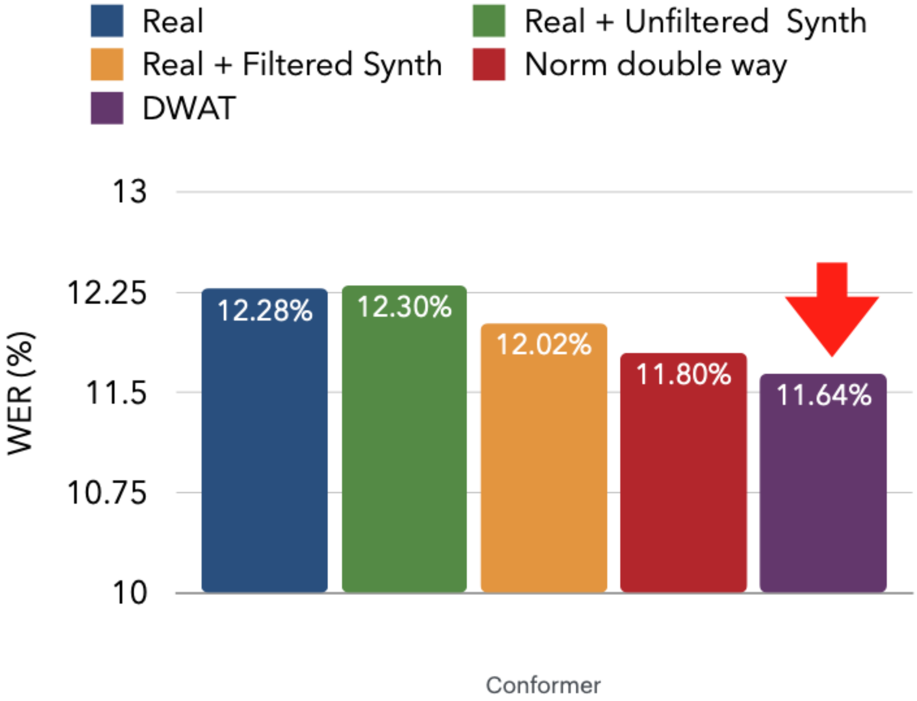
Double Way Adapter Transfer (DWAT)



Synthetic data augmentation



X-vector filtering + DWAT can reduce the mismatch between real and synthetic data, and control the quality and speakers' variability of the synthetic utterances.



Towards improved ASR for children

Fine-tuning:

- Essential for good children’s ASR performance

Selective fine-tuning:

- Encoder and its last layers
- Feed Forward component

Can such bias mitigation strategies be adopted to other biases?

Additive fine-tuning:

- Shared-Adapters, the best parameter efficiency / performance trade-off

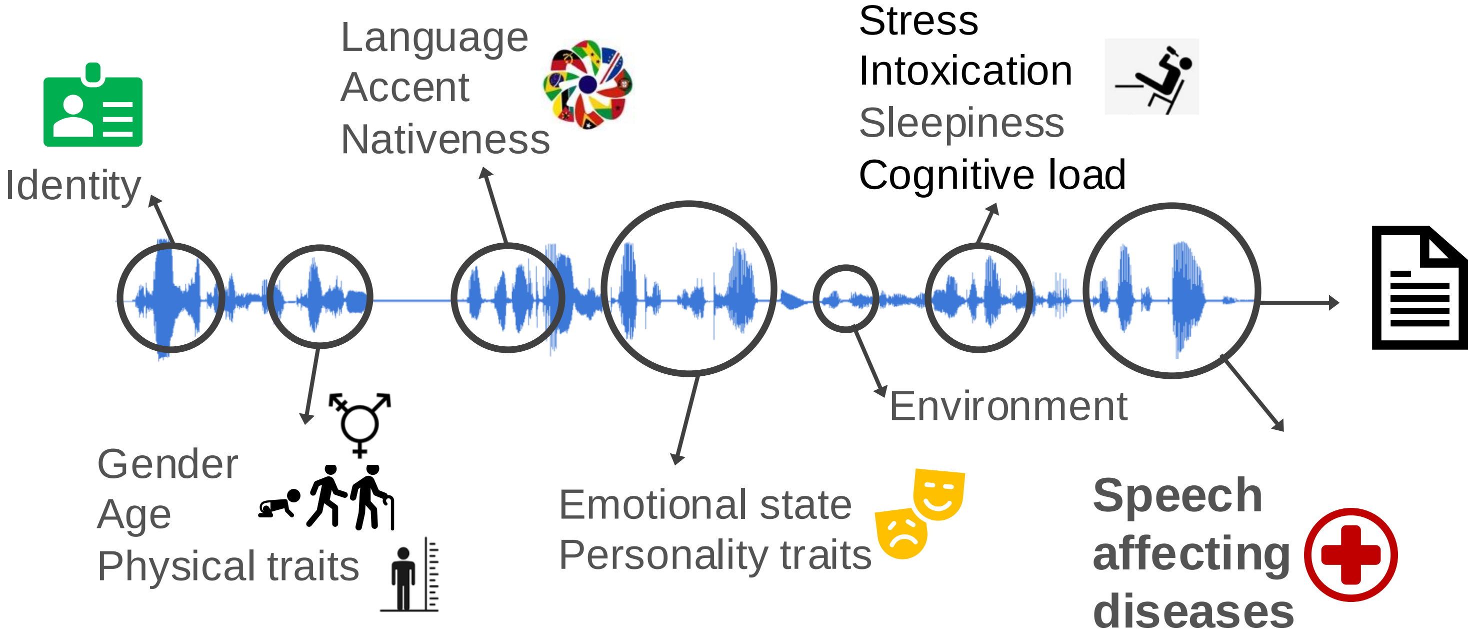
Synthetic data augmentation:

- Can enhance fine-tuning
- Must address domain mismatches between real and synthetic speech data

Explainability

- Choosing the most accurate and explainable model
  - The Great AI Debate@NIPS 2017
- Interpretable Machine Learning (Molnar et al., 2020)
- Local, global & mixed explanations
- Particularly relevant for domains such as criminal justice or healthcare

Info in speech



Identity

Language
Accent
Nativeness

Stress
Intoxication
Sleepiness
Cognitive load

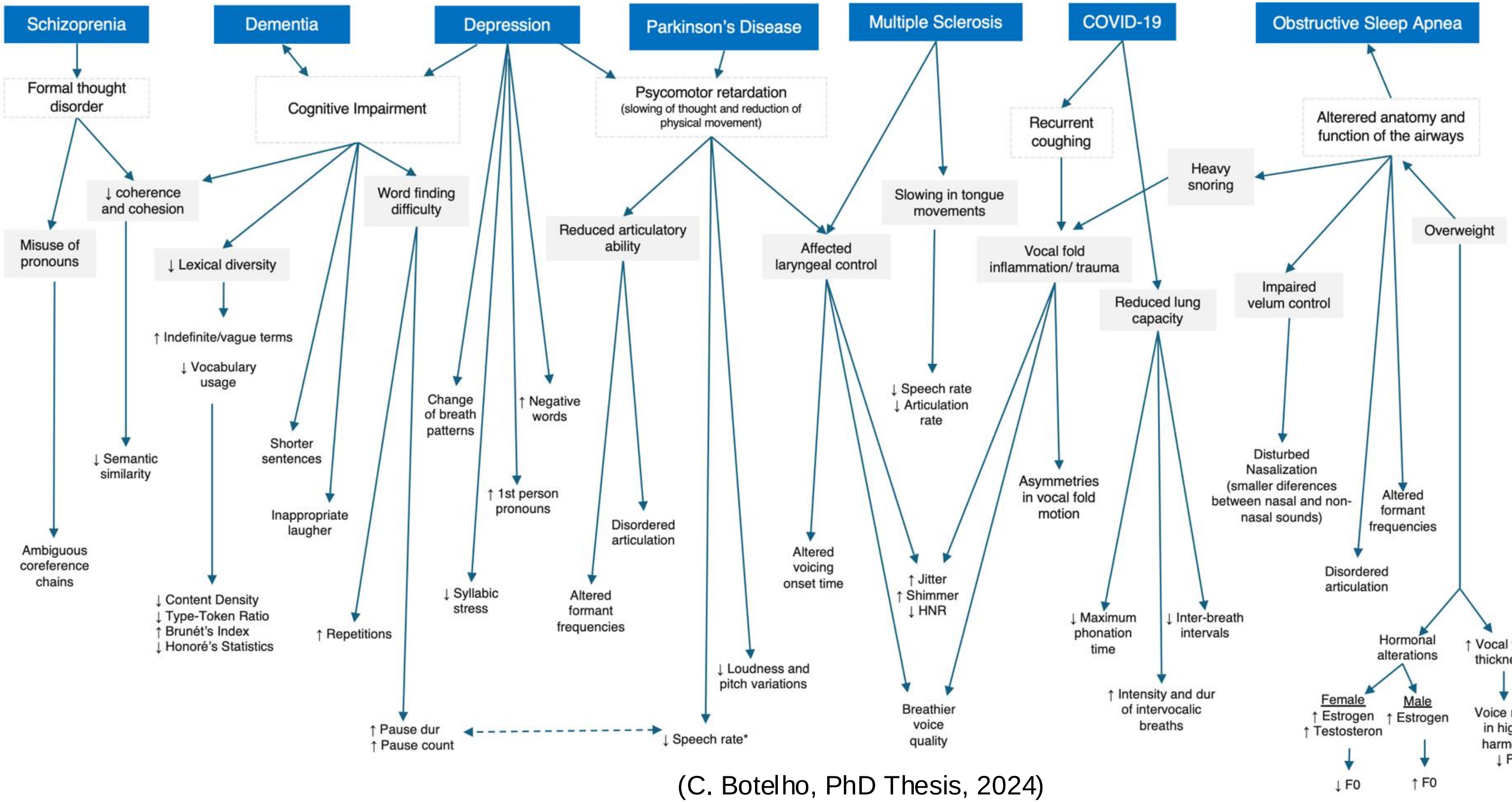
Gender
Age
Physical traits

Emotional state
Personality traits

Environment

Speech affecting diseases

Speech affecting diseases

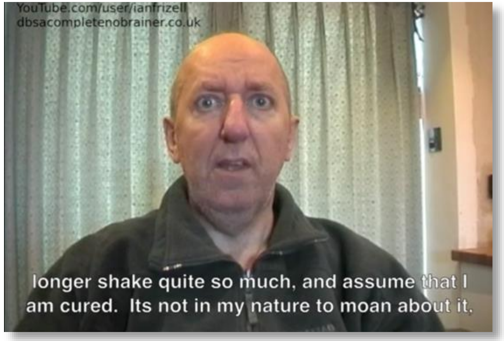


(C. Botelho, PhD Thesis, 2024)

Data scarcity

- Collection in clinical facilities, lack of longitudinal studies, different conditions
- Crowdsourced collection (e.g. COVID-19, CLAC)
- In-the-wild collection (e.g. WSM) → VLOGs
  - PhD of Joana Correia

survive so I'm kind of sad about that I
feel empty depression flex

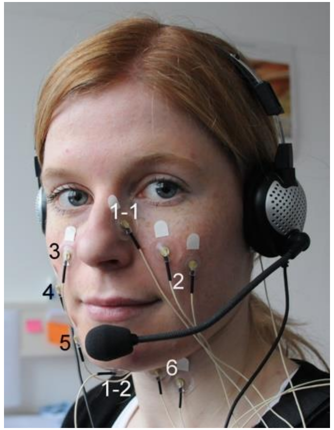


longer shake quite so much, and assume that I
am cured. Its not in my nature to moan about it.

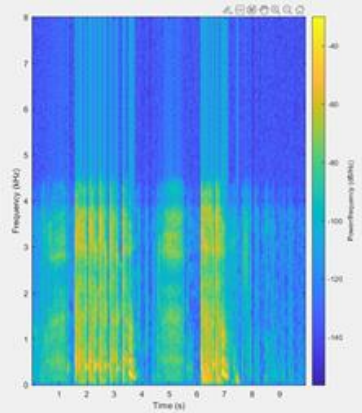
Beyond Speech

- Other non-invasive and invasive modalities
- Other body sounds (respiratory sounds, snoring, coughing)

(Botelho et al., 2021)



(Botelho et al, 2020; Diener et al. 2020)



(Solera et al., 2021)

Explainability

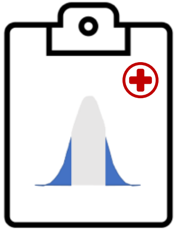
- PhD thesis of Catarina Botelho, supervised by I. Trancoso, A. Abad, T. Schultz
  - Macro-descriptors for Alzheimer’s disease detection using large language models (IS 2024, Tuesday, SS-5B)
  - Towards reference speech characterization for health applications (IS 2023)
  - Challenges on studies of pathological speech in longitudinal and cross-domain corpora (IS 2022)

Definition of reference speech

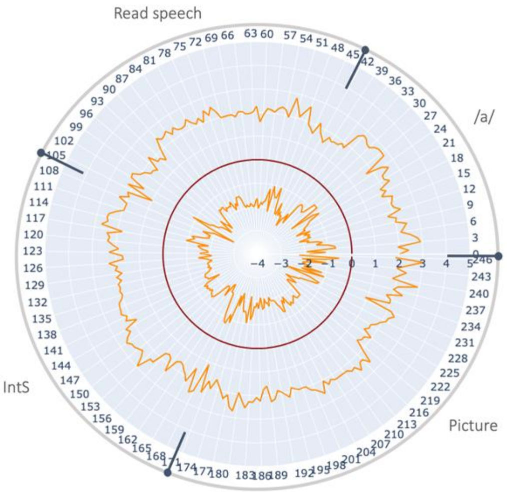
Extract features

Remove outliers

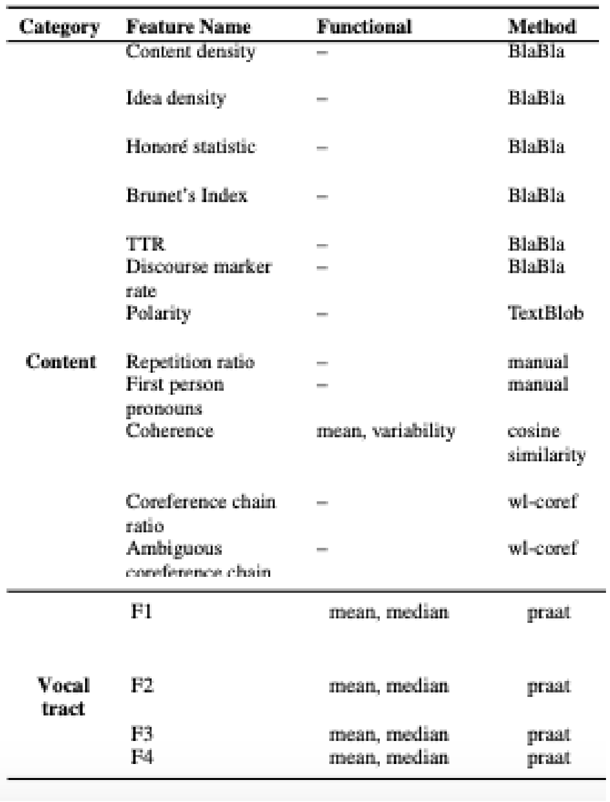
Partition reference data



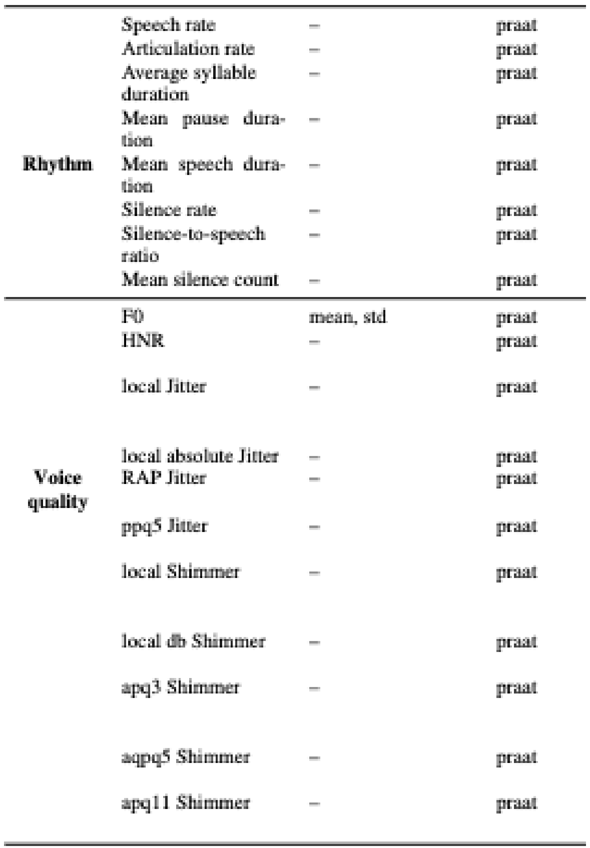
Reference Intervals



Features

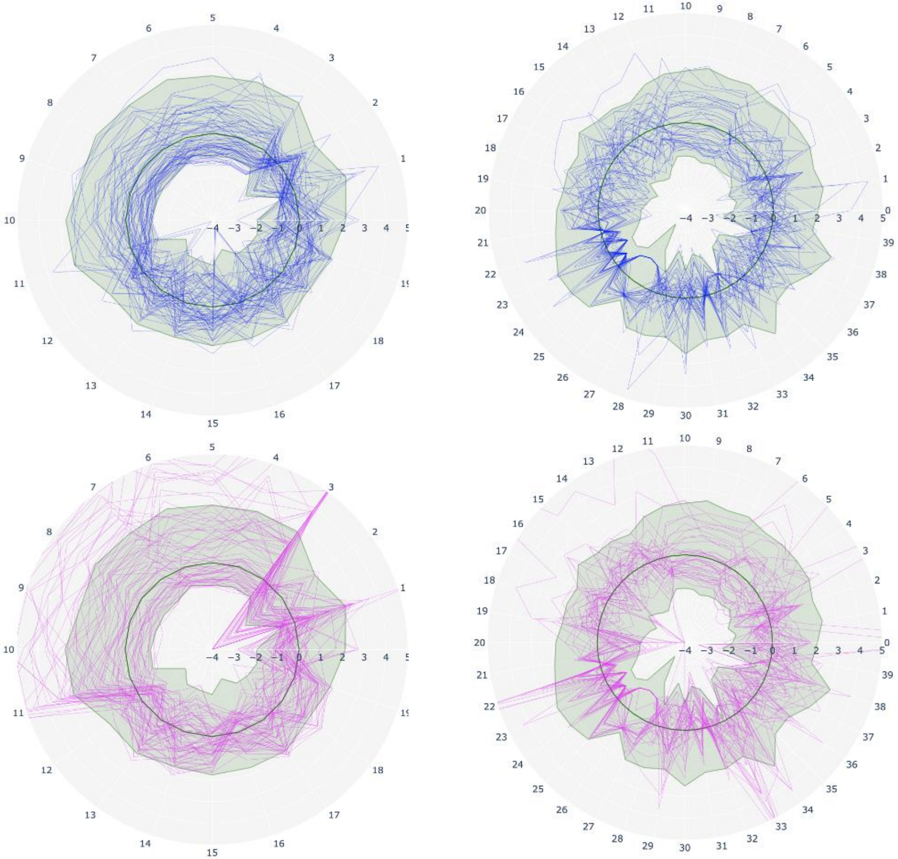


Category	Feature Name	Functional	Method
Content	Content density	–	BlaBla
	Idea density	–	BlaBla
	Honoré statistic	–	BlaBla
	Brunet's Index	–	BlaBla
	TTR	–	BlaBla
	Discourse marker rate	–	BlaBla
	Polarity	–	TextBlob
	Repetition ratio	–	manual
	First person pronouns	–	manual
	Coherence	mean, variability	cosine similarity
	Coreference chain ratio	–	wl-coref
	Ambiguous coreference chain	–	wl-coref
Vocal tract	F1	mean, median	praat
	F2	mean, median	praat
	F3	mean, median	praat
	F4	mean, median	praat



Rhythm	Speech rate	–	praat
	Articulation rate	–	praat
	Average syllable duration	–	praat
	Mean pause duration	–	praat
	Mean speech duration	–	praat
	Silence rate	–	praat
	Silence-to-speech ratio	–	praat
	Mean silence count	–	praat
Voice quality	F0	mean, std	praat
	HNR	–	praat
	local Jitter	–	praat
	local absolute Jitter	–	praat
	RAP Jitter	–	praat
	ppq5 Jitter	–	praat
	local Shimmer	–	praat
	local db Shimmer	–	praat
	apq3 Shimmer	–	praat
	aqpq5 Shimmer	–	praat
	apq11 Shimmer	–	praat

Radar plots



Sustained vowels (Female)

Picture description (Female)

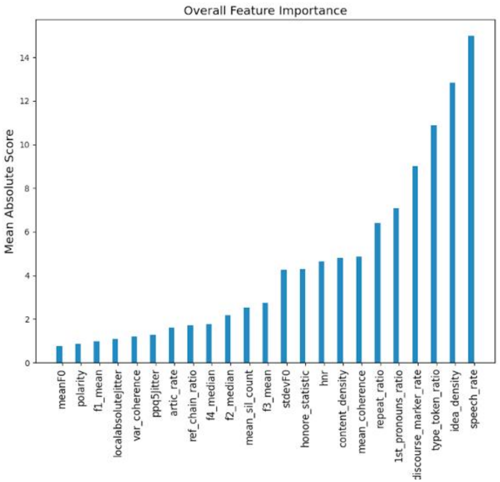
Datasets:

- CLAC (RIs)
  (Haulcy and Glass, 2021)
- PC-GITA (PD)
  (Orozco-Arroyave et al., 2014)
- ADReSS (AD)
  (Luz et al., 2020)

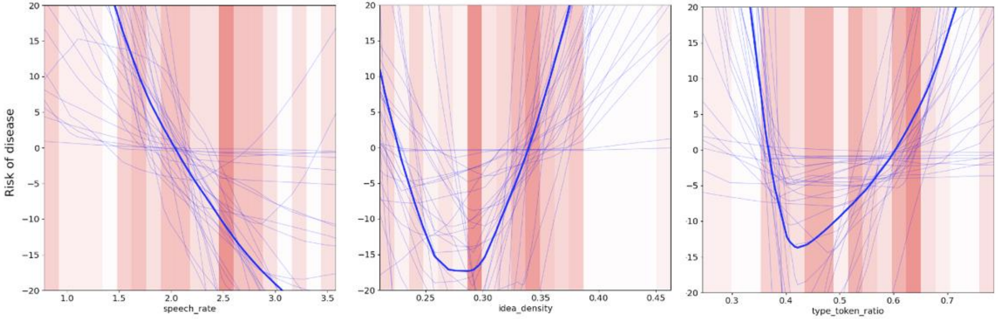
Neural Additive Model (NAMs)

- Linear combination of neural networks, each attending to a single feature, that are trained jointly using backpropagation (Agarwal et al., 2021)

Male subjects



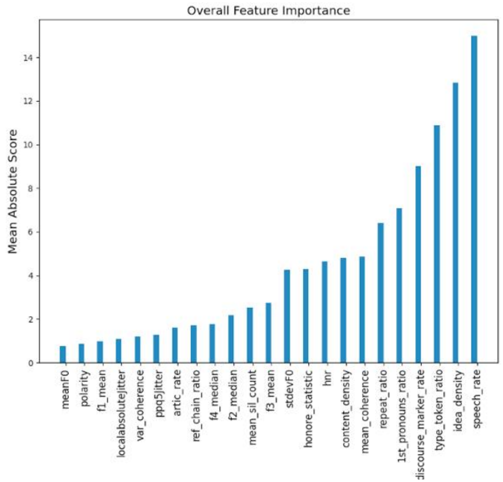
Accuracy on test male recordings: 77.3%



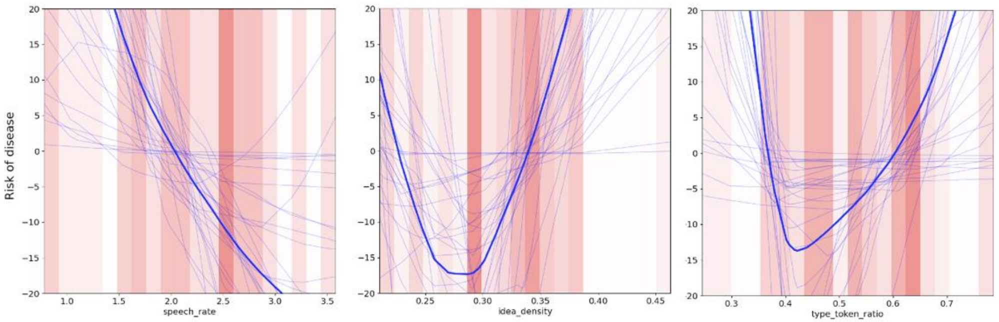
Neural Additive Model (NAMs)

- Linear combination of neural networks, each attending to a single feature, that are trained jointly using backpropagation (Agarwal et al., 2021)

Male subjects

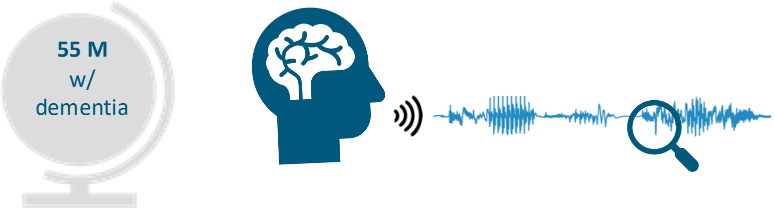


Accuracy on test male recordings: 77.3%



Not a posteriori explanations

Macro-descriptors for AD detection using LLMs



- ↓ Coherence
- ↓ Lexical diversity
- ↑ Word finding difficulties
- ↓ Sentence Length

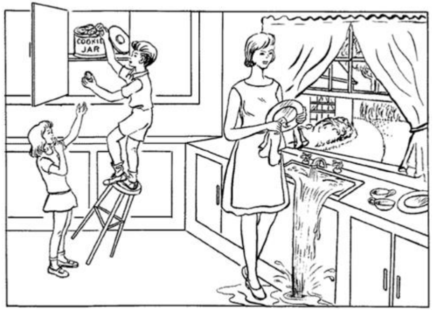
- Are LLMs already able to perform AD detection from speech transcriptions?
- Can we leverage the potential of LLMs to capture **macro-descriptors** that describe and help differentiate between the speech of healthy/AD subjects?

LLMs

- Mistral-7BInstruct-v0.2 (Jiang et al., 2023)
- Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024)
- GPT-3.5-Turbo (Ouyang et al.,2022)

Data

- ADReSS
  - 78 AD + 78 Control

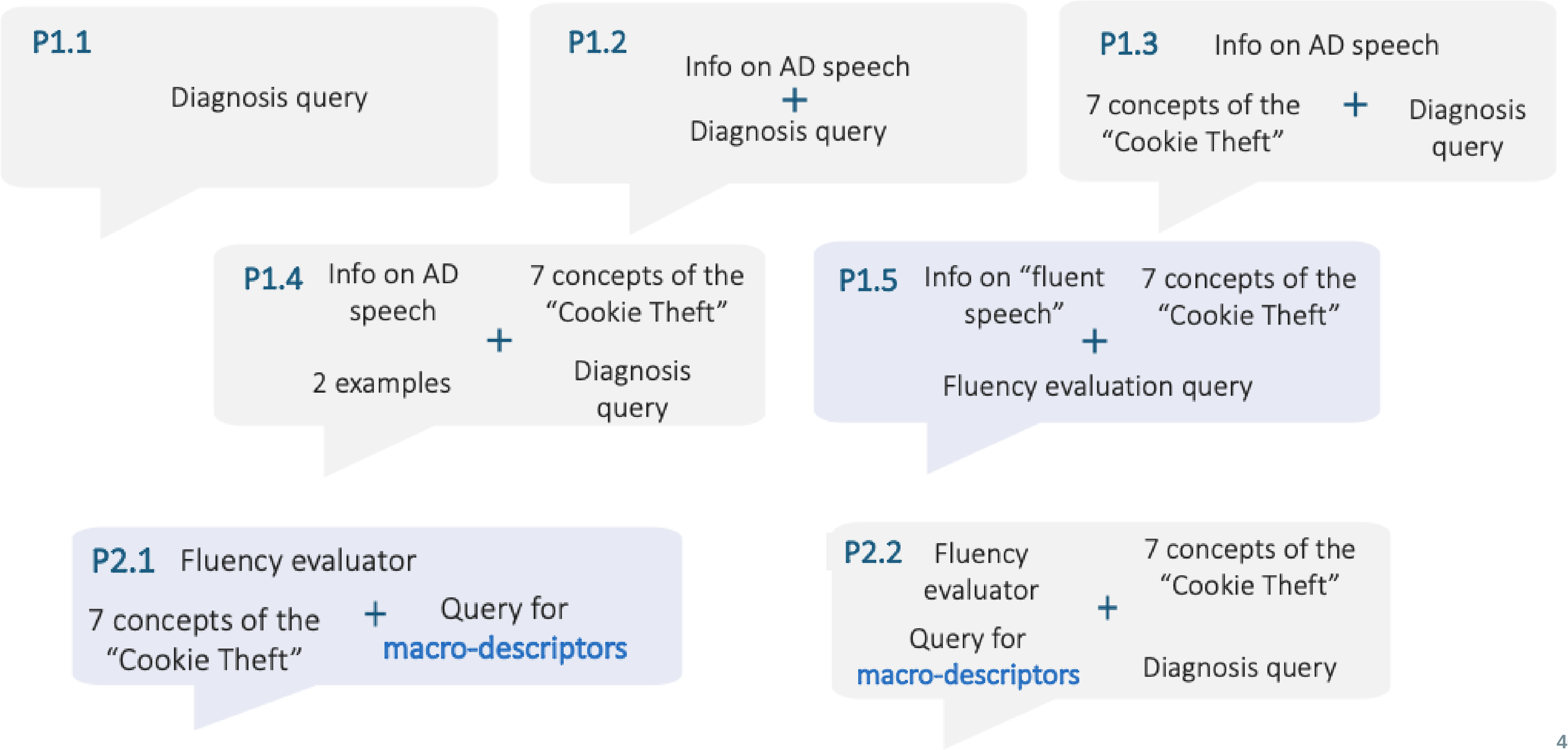


Transcriptions

- Manual
- Automatic (best of 5 ASR models):
  - whisper-large (Radford et a., 2023)
    - WER: 26.9 %
  - wav2vec2-large-robust-ft-swbd-300h (Hsu et al, 2021)
    - WER: 37.9%
    - wav2vec failed to output a transcription for 6 files

- Example:
  - manual: “uh well this here”
  - whisper: this here
  - wav2vec: uhe this yeur

Prompting strategies



P1.1

Diagnosis query

P1.2

Info on AD speech
+
Diagnosis query

P1.3 Info on AD speech

7 concepts of the “Cookie Theft” + Diagnosis query

P1.4 Info on AD speech

2 examples

+

7 concepts of the “Cookie Theft”

Diagnosis query

P1.5 Info on “fluent speech”

7 concepts of the “Cookie Theft”

+

Fluency evaluation query

P2.1 Fluency evaluator

7 concepts of the “Cookie Theft” + Query for macro-descriptors

P2.2 Fluency evaluator

Query for macro-descriptors

+

7 concepts of the “Cookie Theft”

Diagnosis query

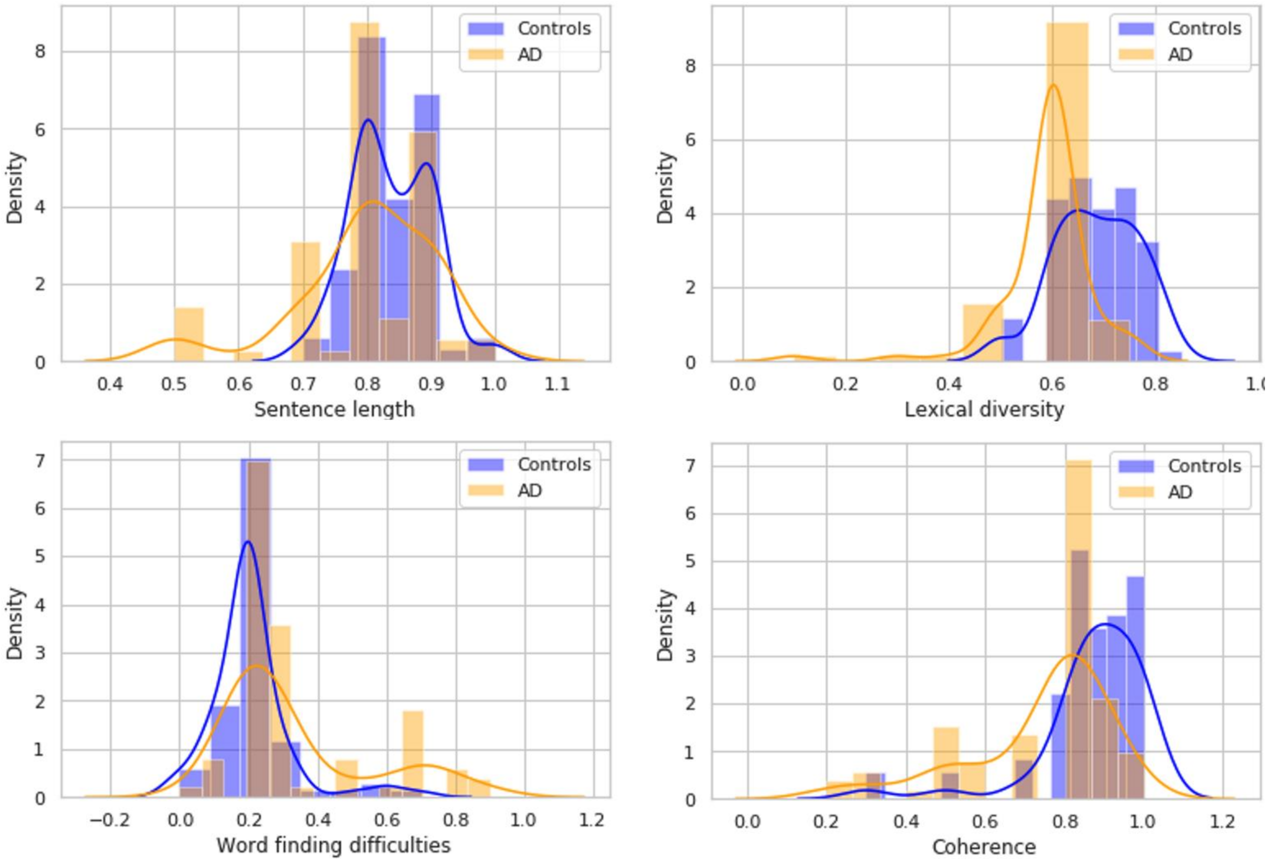
Distributions of the macro-descriptors

Annotations by Mistral
Transcriptions by Whisper
Prompt P2.2

Transcription

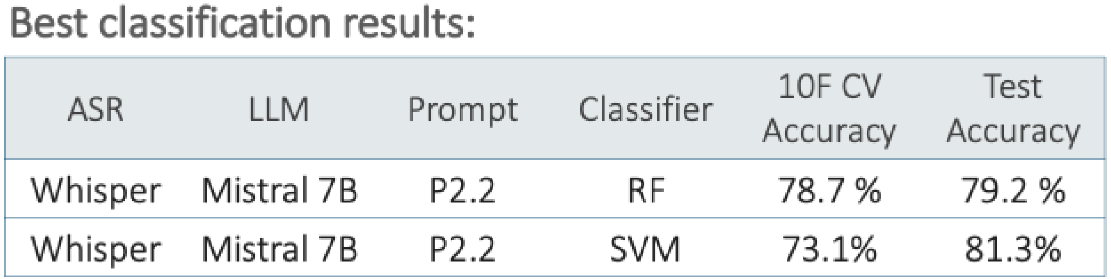
I don’t see nothing but some roots. It’s like somebody took some pencils or something and went up and down those things. Oh, I see a girl standing there or something. Some little knots or something on there. Oh, a lot of it around here. Some kind of little flower. And a sun. And a sun. And a girl is there. And there’s something else over there. There’s another girl. Look like... Look like some old girl is in there. I don’t see nothing but some marks and things. Look to me about the same, except them things up there…

Coherence 0.3
Word Finding Difficulties 0.8
Lexical Diversity 0.5
Sentence Length 0.6
AD Prediction: YES
Confidence: HIGH



Potential of LLMs for AD detection

- Support Vector Machine
- Linear Discriminant Analysis
- 1-Nearest Neighbour
- Decision Tree
- Random Forest



Best classification results:

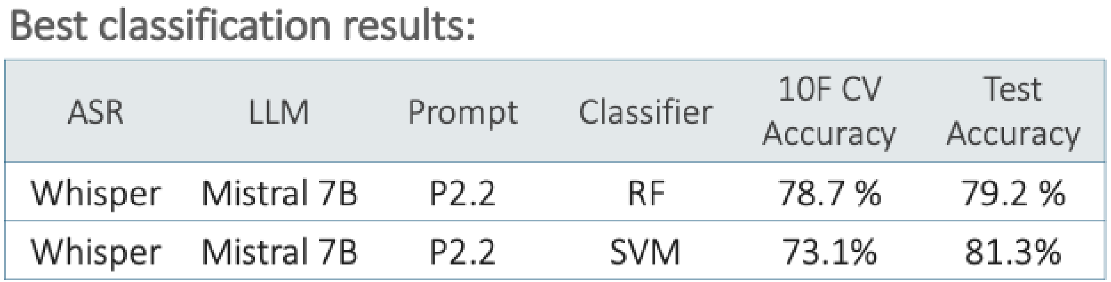
ASR	LLM	Prompt	Classifier	10F CV Accuracy	Test Accuracy
Whisper	Mistral 7B	P2.2	RF	78.7 %	79.2 %
Whisper	Mistral 7B	P2.2	SVM	73.1%	81.3%

Potential of LLMs for AD detection

- Support Vector Machine
- Linear Discriminant Analysis
- 1-Nearest Neighbour
- Decision Tree
- Random Forest

Employing LLMs as extractors of macro-descriptors for AD compares favourably with the direct prediction of AD by the LLM:

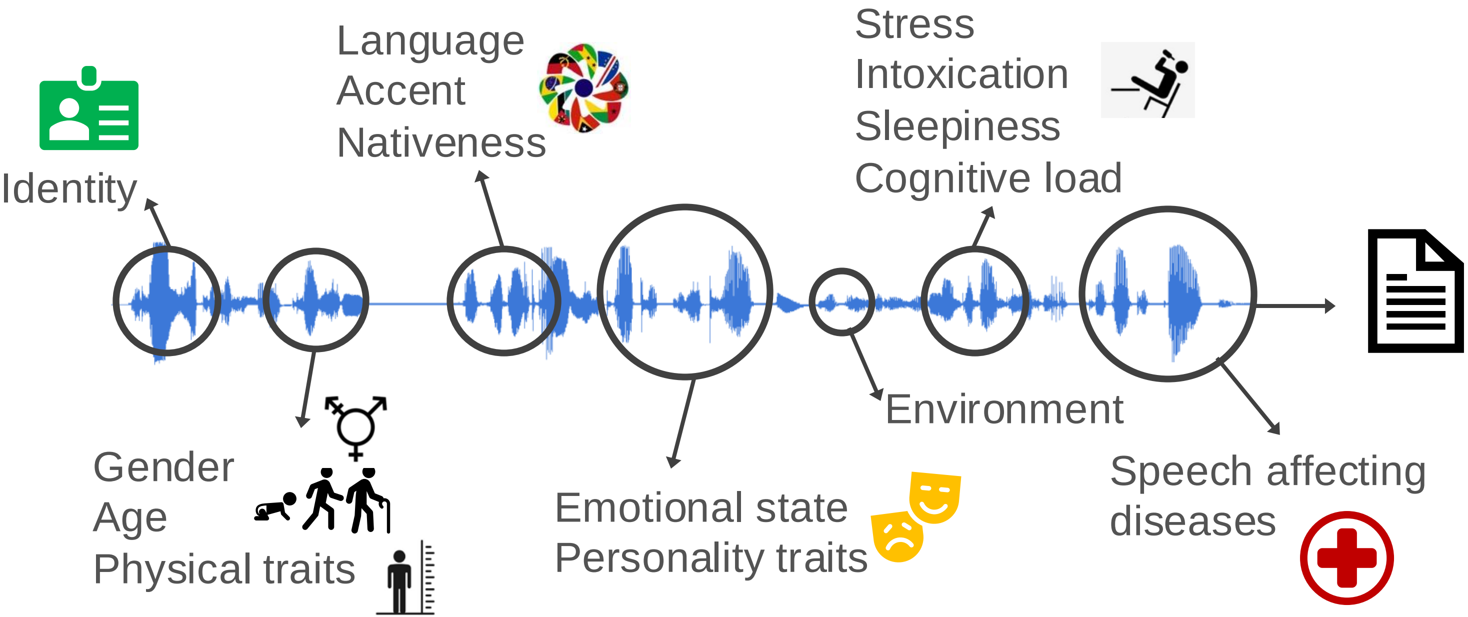
- ↑ performance
- ↓ failed predictions
- ↑ **interpretability**



Best classification results:

ASR	LLM	Prompt	Classifier	10F CV Accuracy	Test Accuracy
Whisper	Mistral 7B	P2.2	RF	78.7 %	79.2 %
Whisper	Mistral 7B	P2.2	SVM	73.1%	81.3%

Privacy and Security



Identity

Language
Accent
Nativeness

Stress
Intoxication
Sleepiness
Cognitive load

Gender
Age
Physical traits

Emotional state
Personality traits

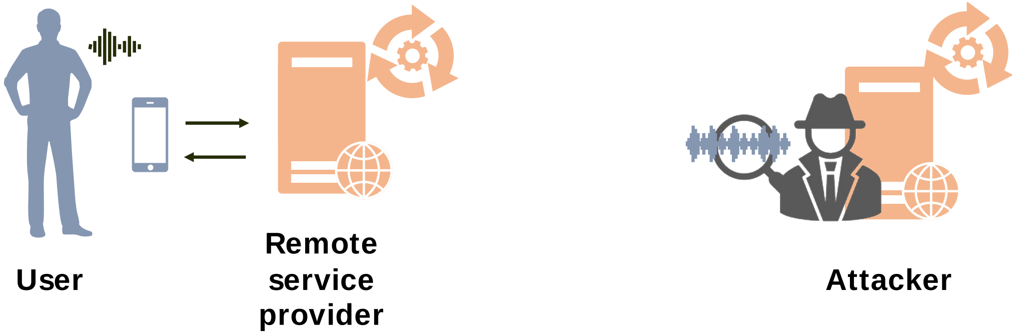
Environment

Speech affecting
diseases

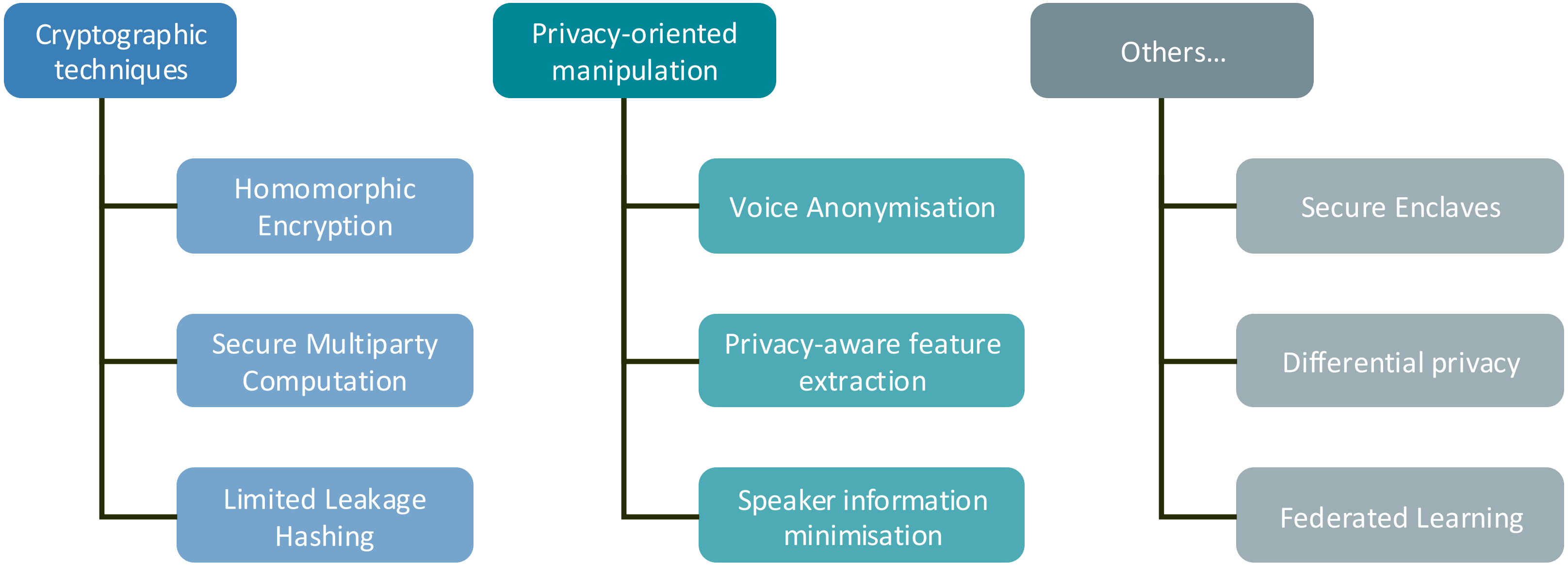
Vulnerabilities: Profiling & Impersonation
ISCA SIG Security and Privacy in Speech Communication

Privacy-preserving ML for remote speech processing

- PhD thesis of Francisco Teixeira, supervised by I. Trancoso, A. Abad & B. Raj
  - Privacy-oriented Manipulation of Speaker Representations (IEEE Access, 2024)
  - Privacy-preserving Automatic Speaker Diarization (ICASSP 2023)
  - Towards end-to-end private Automatic Speaker Recognition (IS 2022)

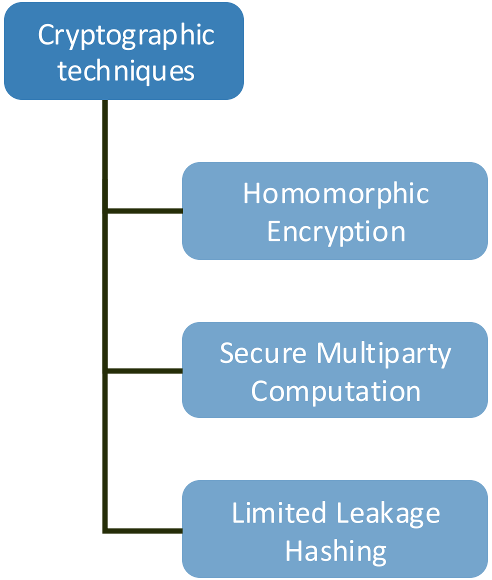


Privacy in Remote Speech Processing



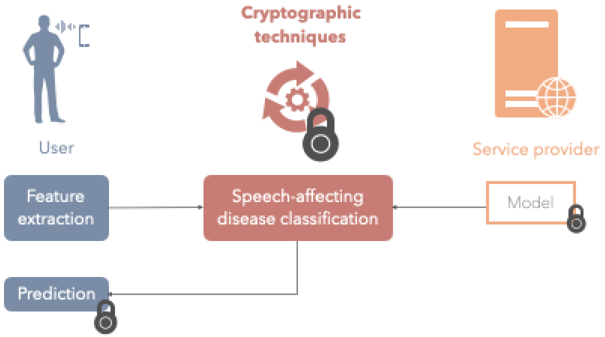
Cryptographic techniques

- Suited to tasks where it is difficult to disentangle speaker and task-related information
- Require the collaboration of the user and the service provider
- Provide confidentiality and formal privacy guarantees
- High computational and communication costs



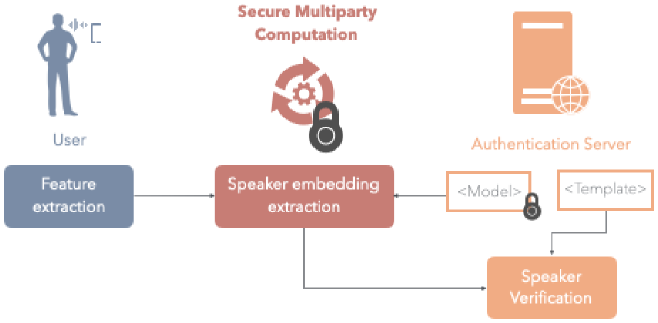
Cryptographic techniques

- Privacy-preserving Support Vector Machine w/ Radial Basis Function kernel:
  - Relied on Homomorphic Encryption, Secure Multiparty Computation and Secure Modular Hashing
  - Application to Disease detection (PD, OSA)
  - No performance degradation compared to baseline
  - 2000x slower than a non-encrypted classifier



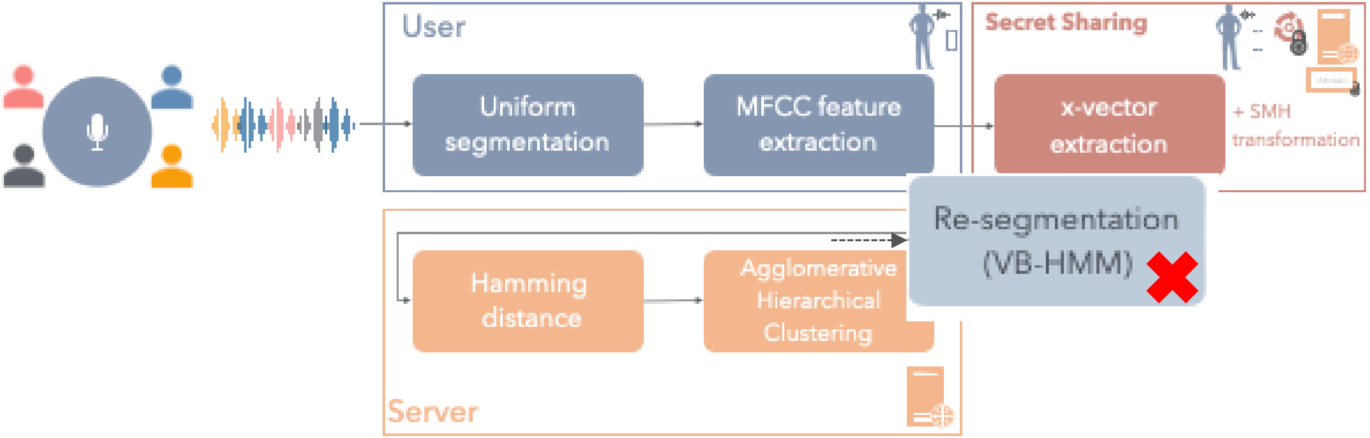
Cryptographic techniques

- Privacy-preserving speaker embedding extraction (x-vectors)
  - Relied only on Secret Sharing protocols, involving 2, 3 & 4 parties
  - Applied to speaker verification (using using cosine similarity scores)
  - No performance degradation compared to baseline
  - Only computationally feasible if involving at least a trusted 3rd party



Cryptographic techniques

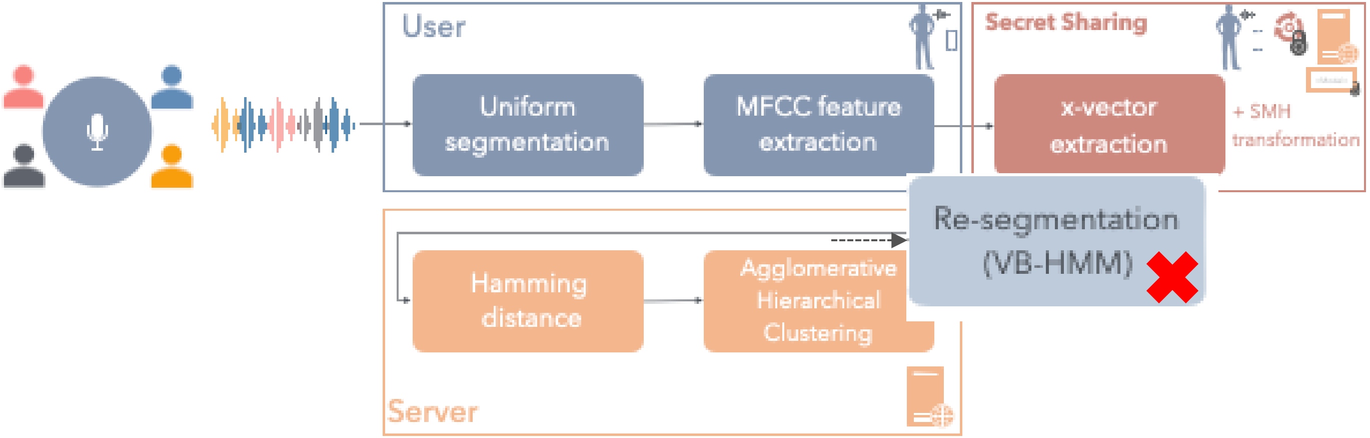
- Application to Automatic Speaker Diarization (ASD) (N. Rayant et al., 2021)
- Degradation of around 10% in DER from original baseline
- PP-Diarization of 4 minutes takes 5-7 minutes using 3-party protocol.



Cryptographic techniques

- Application to Automatic Speaker Diarization (ASD) (N. Rayant et al., 2021)
- Degradation of around 10% in DER from original baseline
- PP-Diarization of 4 minutes takes 5-7 minutes using 3-party protocol.

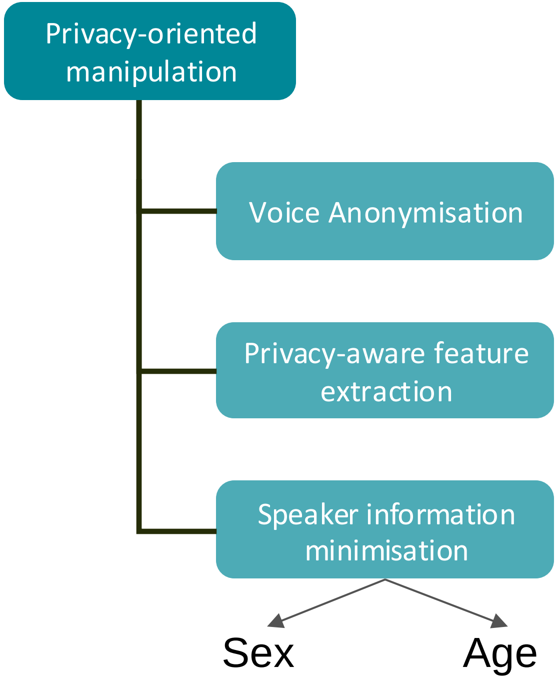
Usable for low-complexity tasks, but still impractical for high-complexity applications



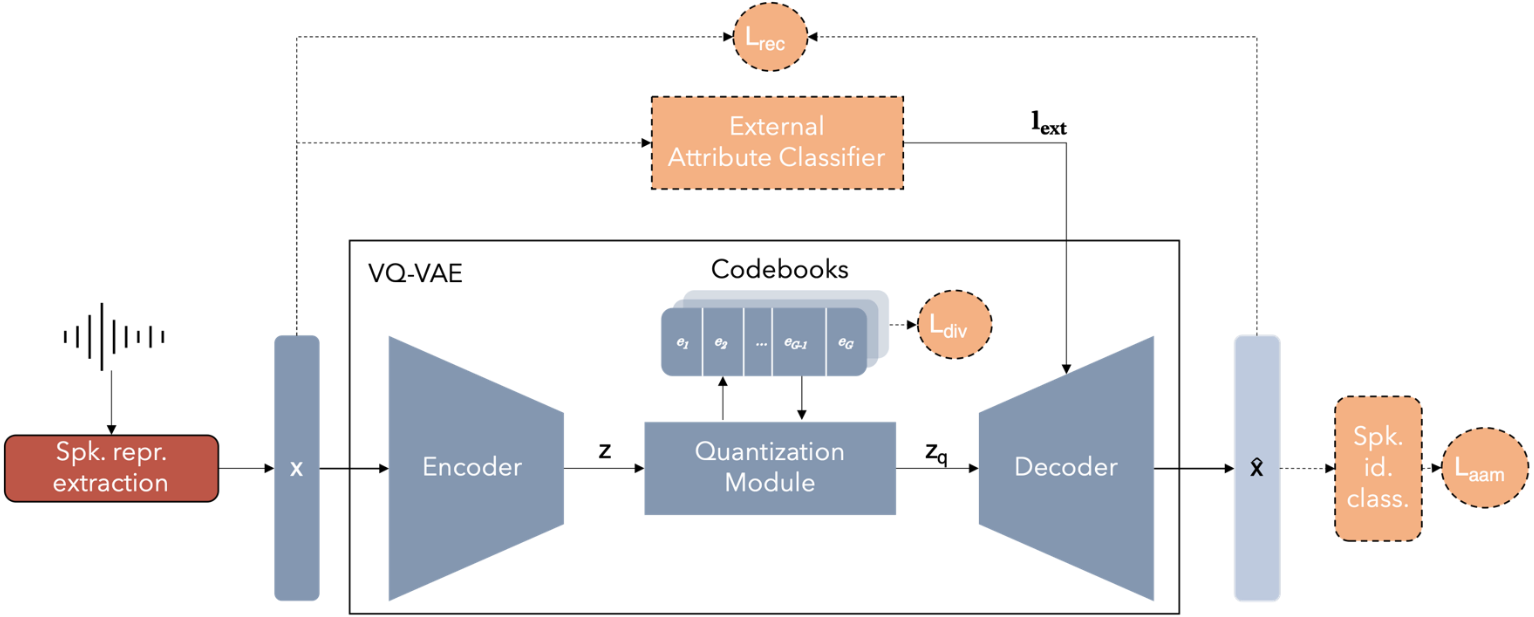
Privacy-oriented manipulation

Speaker information minimisation

- Remove or obfuscate task-unrelated information
- User-centred: can be performed directly on the user's device
- Empirical guarantees of privacy
- Low computational costs

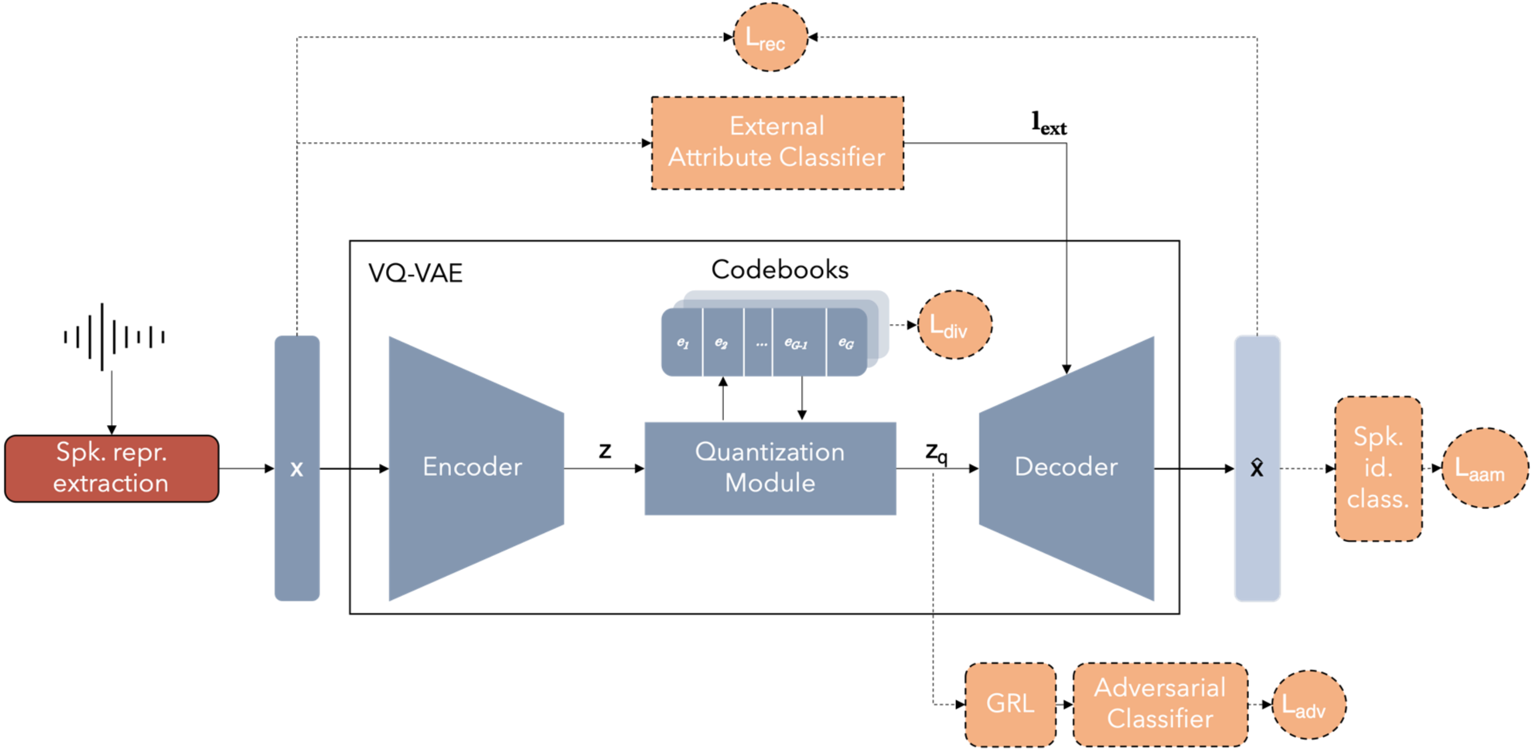

Privacy-oriented manipulation
Voice Anonymisation
Privacy-aware feature extraction
Speaker information minimisation
Sex
Age

Privacy-oriented manipulation of speaker representations

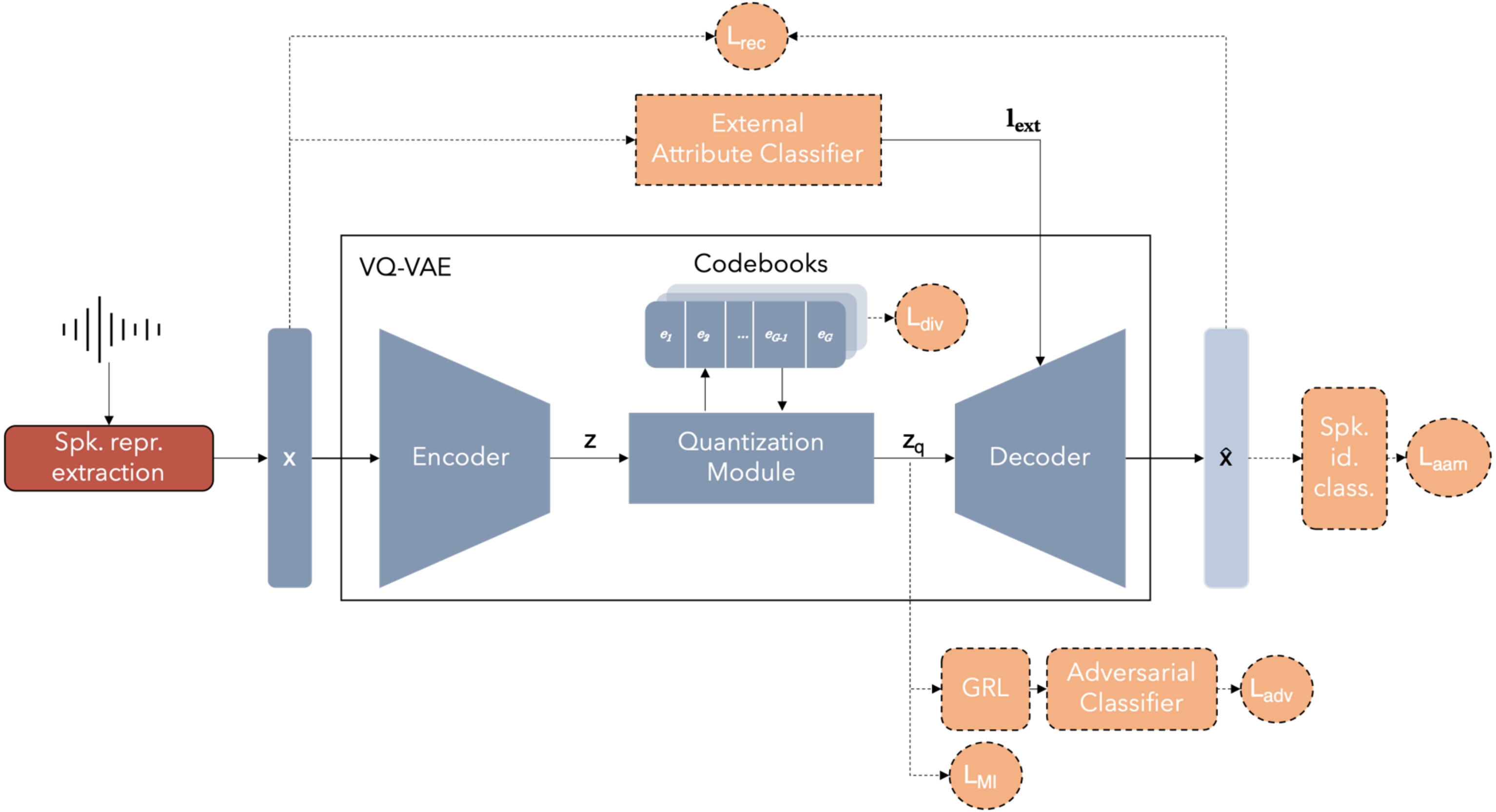


(Van Den Oord et al., 2017; P.-G. Noé et al., 2021)

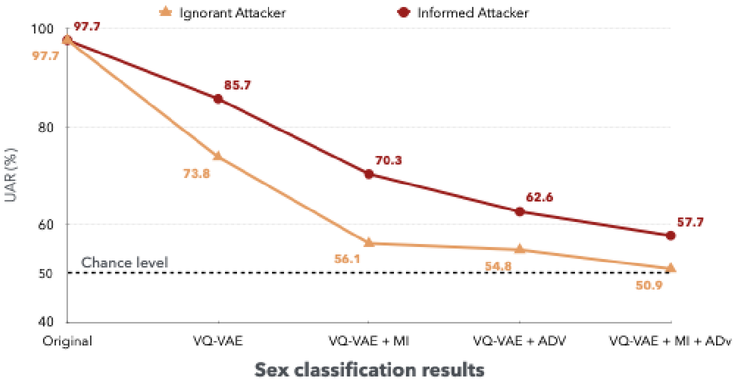
Privacy-oriented manipulation of speaker representations

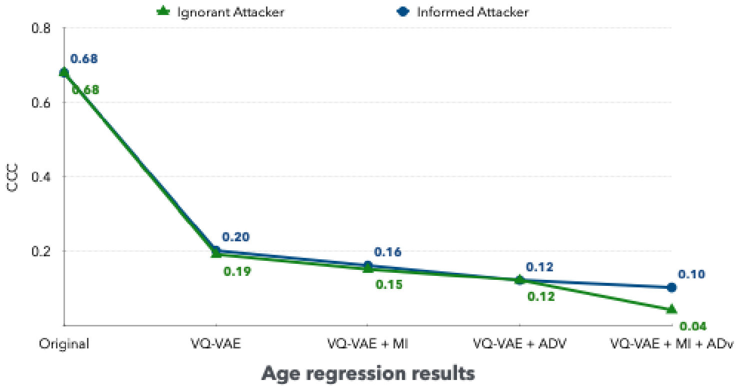


Privacy-oriented manipulation of speaker representations

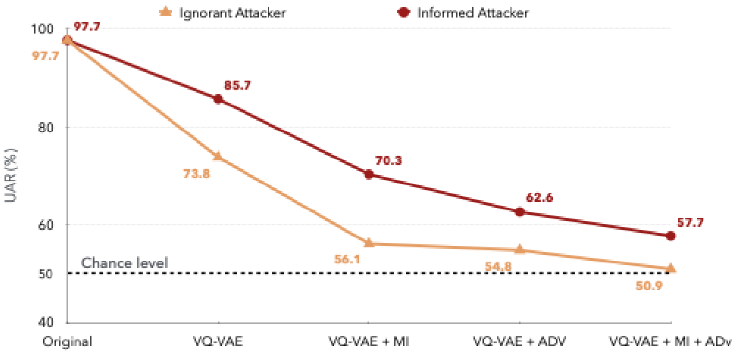


Privacy-oriented manipulation of speaker representations

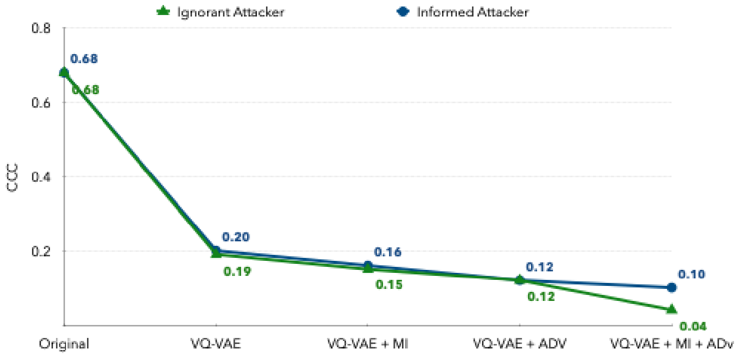




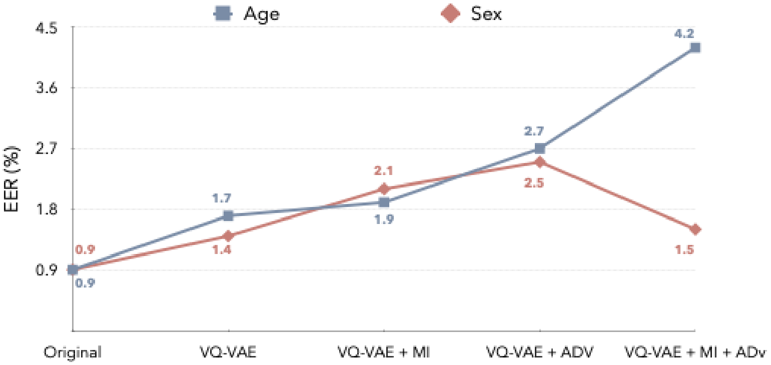
Privacy-oriented manipulation of speaker representations



Sex classification results



Age regression results



Speaker verification results

Trade-off between privacy and task performance (although approach only tested for ASV).

Can this type of manipulation be explored for voice anonymization?

VQ-VAE — Sex information manipulation

Original (Male)

🔊

Male2Female

🔊

“Genderless”

🔊

Original (Female)

🔊

Female2Male

🔊

“Genderless”

🔊

🔊 🔊

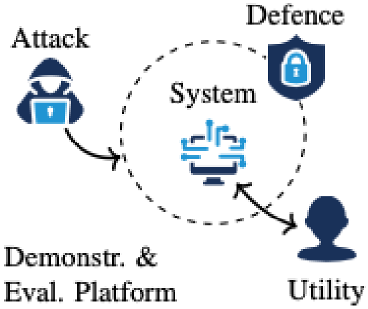
VQ-VAE — Age information manipulation

Original (adult M)	7	12	14	20	40	80
🔊	🔊	🔊	🔊	🔊	🔊	🔊

Original (adult F)	7	12	14	20	40	80
🔊	🔊	🔊	🔊	🔊	🔊	🔊
			🔊			

Challenges - Privacy for Smart Speech Technology (PSST)

- Marie Skodowska-Curie Action - Doctoral Networks (DN-JD)
- PSST is recruiting 12 PhD students. Contact us at: info@psst-doctoralnetwork.eu
  - Protection against deepfakes in speech
  - Speech anonymisation for privacy-preserving emotion recognition
  - Disentangled representations for selective attribute suppression
  - Transparent Exchange of Speaker Attributes
  - Revealing social relationships in conversations
  - Robust attack models and tools for the credible evaluation of anonymisation and attribute suppression
  - Privacy impact assessment for comprehensive attacks exploiting audio, speech, and metadata
  - Attacking information bottlenecks – Theoretical metrics and bounds of privacy
  - Robust privacy-preserving industrial voice interfaces
  - Detection of speech-affecting diseases in anonymized speech
  - Utility of Speech Samples as Privacy-Preserving, Transparent and Reusable Model-Updates for Distributed Learning
  - Methods for subjective and objective evaluation of privacy



Sustainability

- The size of SOTA NLP language models has doubled every 3-4 months
- Reporting is usually limited to compute resources used strictly for training
  - Thousands of petaFLOP/s-day range
- Forecasting the carbon footprint of inference is harder:
  - 3 billion tokens would have to be generated for inference costs to catch up with training costs (Lakim et al., 2022)
  - At some point during its beta, GPT-3 was reported to generate 4.5 billion words per day

    https://openai.com/index/gpt-3-apps/

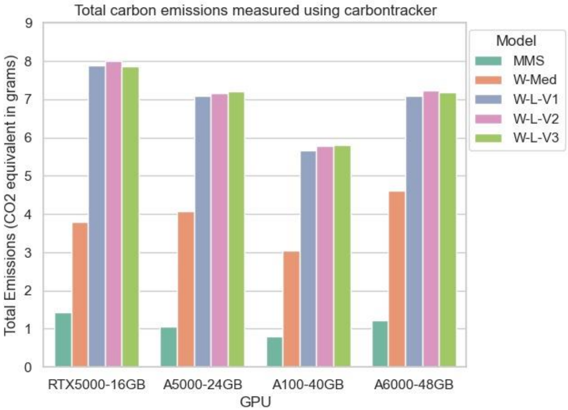
Sustainability

- Collaboration with Ajinkya Kulkarni and Miguel Couceiro
  - Unveiling Biases while Embracing Sustainability: Assessing the Dual Challenges of Automatic Speech Recognition Systems (IS 2024, Thursday, SS-7)

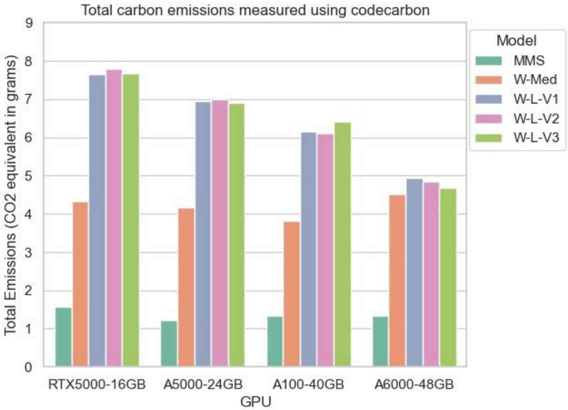
Sustainability study

- 5 ASR systems
  - Massive Multilingual Speech Model by Meta AI, 2023 (Pratap et al., 2024)
    - MMS (~1 B)
  - Whisper by Open AI, 2022 (Radford et al., 2022)
    - Medium (0.769 B), Large-v1 (1.550 B), Large-v2 (1.550 B) and Large-v3 (1.550 B)
- 3 different platforms to measure the carbon emission intensity and energy consumption
  - Codecarbon (https://codecarbon.io/), Carbontracker (https://carbontracker.org/), Eco2ai (S. Budennyy et al., 2022)
- Inference of ASR on 20 mins of speech utterances across 4 NVIDIA GPUs, x 3 times
  - RTX-5000-16GB, RTX-A5000-24GB, A100-40GB, A6000-48GB
- Cloud service provider
  - Choice of region, time of day, preference for data centers with lower PUE (Dodge et al., 2022)
  - Based in Tamil, Nadu, India, 32GB of RAM, 7 CPU cores

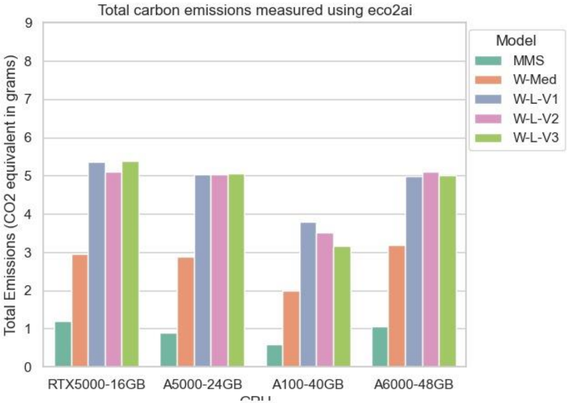
Sustainability study - Results



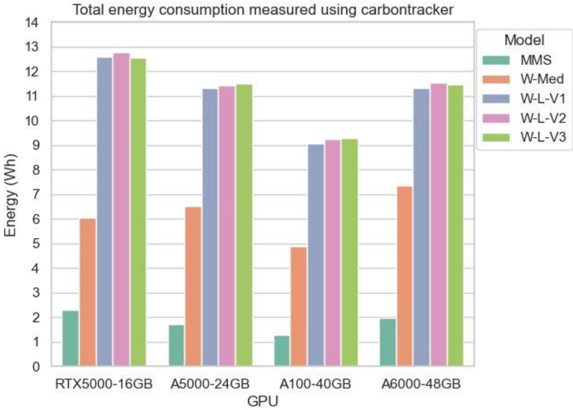
Total carbon emissions measured using carbontracker



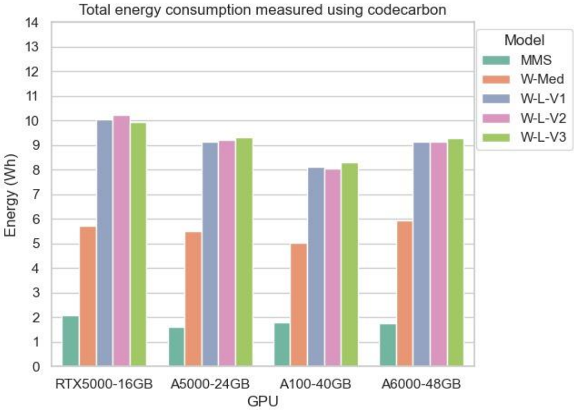
Total carbon emissions measured using codecarbon



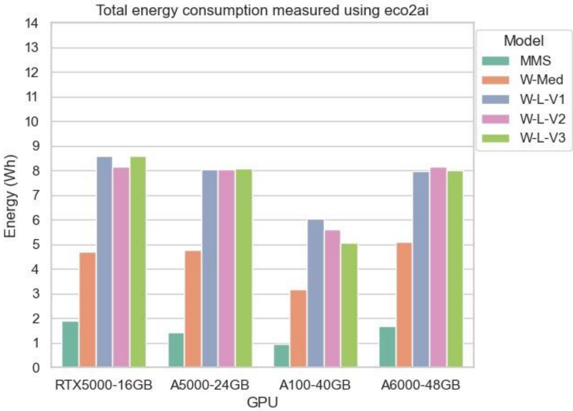
Total carbon emissions measured using eco2ai



Total energy consumption measured using carbontracker



Total energy consumption measured using codecarbon



Total energy consumption measured using eco2ai

Sustainability study - Discussion

- Clear advantage of MMS over Whisper variants
  - MMS features multiple Transformer blocks, each enhanced with a language-specific adapter, that can be dynamically loaded and swapped during inference.
- Whisper Medium > Whisper Large variants
  - Whisper large variants have 2 x number of parameters
  - Similar behaviour of the 3 Whisper Large variants

Language-specific adapters can help save carbon emissions.

Mixture of Experts are energy efficient architectures (Lakim et al., 2022)

Sustainability study - Discussion

- Clear advantage of MMS over Whisper variants
  - MMS features multiple Transformer blocks, each enhanced with a language-specific adapter, that can be dynamically loaded and swapped during inference.
- Whisper Medium > Whisper Large variants
  - Whisper large variants have 2 x number of parameters
  - Similar behaviour of the 3 Whisper Large variants
- Slight advantage of NVIDIA GPU A100-40GB over other NVIDIA GPUs

Language-specific adapters can help save carbon emissions. Mixture of Experts are energy efficient architectures (Lakim et al., 2022)

Wide GPU bandwidth seems to have a positive impact in both carbon emissions and energy consumption.

Sustainability study - Discussion

- Clear advantage of MMS over Whisper variants
  - MMS features multiple Transformer blocks, each enhanced with a language-specific adapter, that can be dynamically loaded and swapped during inference.
- Whisper Medium > Whisper Large variants
  - Whisper large variants have 2 x number of parameters
  - Similar behaviour of the 3 Whisper Large variants
- Slight advantage of NVIDIA GPU A100-40GB over other NVIDIA GPUs
- All platforms show similar trends for the 5 ASR
  - Slightly optimistic view provided by eco2ai

Language-specific adapters can help save carbon emissions. Mixture of Experts are energy efficient architectures (Lakim et al., 2022)

Wide GPU bandwidth seems to have a positive impact in both carbon emissions and energy consumption.

Need for a comprehensive sustainability analysis of ASR systems that considers diversity:

✓ performance metrics
✓ implementations
✓ evaluation methodologies

Towards 1-bit LLMs

- Post-Training Quantization (PTQ) vs. Quantization-Aware Training (QAT) (Hutson, 2024)
- BitNet 1.58b (Wang et al., 2023)
  - QAT: 1, 0, -1
  - Binarized 3B LLaMa model
- BiLLM (Huang et al., 2024)
  - QAT: 1-bit for most weights, 2-bit for salient weights
  - Binarized 13B LLaMa model
- OneBit (Xu et al., 2024)
  - QAT + PTQ
  - Binarized 13B LLaMa model

Towards 1-bit LLMs

- Post-Training Quantization (PTQ) vs. Quantization-Aware Training (QAT) (Hutson, 2024)
- BitNet 1.58b (Wang et al., 2023)
  - QAT: 1, 0, -1
  - Binarized 3B LLaMa model
  - 2.71x as fast, 72% less GPU memory, 94% less GPU energy
- BiLLM (Huang et al., 2024)
  - QAT: 1-bit for most weights, 2-bit for salient weights
  - Binarized 13B LLaMa model: PP=15 (PP=5, unquantized)
  - 10% memory
- OneBit (Xu et al., 2024)
  - QAT + PTQ
  - Binarized 13B LLaMa model: PP=9 (PP=5, unquantized)
  - 10% memory

Towards 1-bit LLMs

- Post-Training Quantization (PTQ) vs. Quantization-Aware Training (QAT) (Hutson, 2024)
- BitNet 1.58b (Wang et al., 2023)
  - QAT: 1, 0, -1
  - Binarized 3B LLaMa model
  - 2.71x as fast, 72% less GPU memory, 94% less GPU energy
- BiLLM (Huang et al., 2024)
  - QAT: 1-bit for most weights, 2-bit for salient weights
  - Binarized 13B LLaMa model: PP=15 (PP=5, unquantized)
  - 10% memory
- OneBit (Xu et al., 2024)
  - QAT + PTQ
  - Binarized 13B LLaMa model: PP=9 (PP=5, unquantized)
  - 10% memory

Potential advantages of custom hardware

Towards 1-bit LLMs

- Post-Training Quantization (PTQ) vs. Quantization-Aware Training (QAT) (Hutson, 2024)
- BitNet 1.58b (Wang et al., 2023)
  - QAT: 1, 0, -1
  - Binarized 3B LLaMa model
  - 2.71x as fast, 72% less GPU memory, 94% less GPU energy
- BiLLM (Huang et al., 2024)
  - QAT: 1-bit for most weights, 2-bit for salient weights
  - Binarized 13B LLaMa model: PP=15 (PP=5, unquantized)
  - 10% memory
- OneBit (Xu et al., 2024)
  - QAT + PTQ
  - Binarized 13B LLaMa model: PP=9 (PP=5, unquantized)
  - 10% memory

Potential advantages of custom hardware

Potential advantages in terms of privacy-preserving ML ?

Pillars of Responsible Speech Processing

- Robustness & Safety
- **Fairness & Inclusion**
- **Explainability**
- **Privacy & Security**
- **Sustainability**
- Accountability & Governance
- User Agency, Trust & Wellbeing

References

- R. Agarwal et al., Neural additive models: Interpretable machine learning with neural nets, Advances in Neural Information Processing Systems, 34:4699–4711, 2021.
- C. Botelho, et al., Toward silent paralinguistics: Speech-to-EMG – retrieving articulatory muscle activity from speech, Interspeech 2020.
- C. Botelho et al., Visual Speech for Obstructive Sleep Apnea Detection", Interspeech 2021.
- S. Budennyy et al., eco2ai: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI, Doklady Mathematics, 2022
- E. Casanova et al., YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone, PMLR 2022.
- B. Desplanques et al., ECAPA- TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, Interspeech 2020.
- L. Diener et al., Towards silent paralinguistics: Deriving speaking mode and speaker ID from electromyographic signals, Interspeech 2020.
- J. Dodge et al., Measuring the Carbon Intensity of AI in Cloud Instances, ACM Conference on Fairness, Accountability, and Transparency, 2022.
- C.-L. Fu et al., AdapterBias: Parameter-efficient token-dependent representation shift for adapters in NLP tasks, NAACL 2022.
- L. Gelin et al., End-to-end acoustic modelling for phone recognition of young readers, Speech Communication, 2021.
- A. Gulati et al., Conformer: Convolution-augmented Transformer for Speech Recognition, Interspeech 2020.

References

- R. Haulcy and James Glass, CLAC: A Speech Corpus of Healthy English Speakers, Interspeech 2021.
- J. He et al., Towards a unified view of parameter-efficient transfer learning, International Conference on Learning Representations, 2022.
- W. Hsu et al., Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training, Interspeech 2021.
- T. Hu et al., Synt++: Utilizing imperfect synthetic data to improve speech recognition, ICASSP 2022.
- W. Huang et al., BiLLM: Pushing the limit of post-training quantization for LLMs, arXiv:2402.04291, 2024.
- M. Hutson, 1-bit LLMs Could Solve AI’s Energy Demands, IEEE Spectrum, May 2024.
- A. Jiang et al., Mistral 7B. arXiv 2023.
- A. Jiang et al., Mixtral of experts, arXiv 2024.
- I. Lakim et al., A Holistic Assessment of the Carbon Footprint of Noor, a Very Large Arabic Language Model, BigScience 2022.
- Y. Li et al., Evaluating parameter-efficient transfer learning approaches on SURE benchmark for speech under- standing, ICASSP, 2023.
- D. Lian et al., Scaling & shifting your features: A new baseline for efficient model tuning, Advances in Neural Information Processing Systems, vol. 35, pp. 109–123, 2022.

References

- S. Luz et al., Alzheimer’s Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge, Interspeech 2020.
- J. Mendonça and I. Trancoso, VoxCeleb-PT–a dataset for a speech processing course, IberSPEECH 2022.
- C. Molnar et al., Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges, ECML PKDD 2020 Workshops, Springer 2020.
- A. Nagrani et al., Voxceleb: Large-scale speaker verification in the wild. Computer Speech & Language, 60:101027, 2020.
- A. Nautsch et al. The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding, Interspeech 2019
- P.-G. Noé et al., Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation, Interspeech 2021.
- L. Ouyang et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems, 2022.
- J. Orozco-Arroyave et al., New Spanish speech corpus database for the analysis of people suffering from Parkinson’s disease, LREC 2014.
- V. Pratap et al., Scaling Speech Technology to 1000+ Languages, Journal of Machine Learning Research, 2024.

References

- A. Radford et al., Robust speech recognition via large- scale weak supervision, Proc. of the 40th International Conference on Machine Learning, 2023.
- M. Ravanelli et al., SpeechBrain: A General-Purpose Speech Toolkit, arXiv:2106.04624, 2021.
- N. Ryant et al., The Third DIHARD Diarization Challenge, Interspeech 2021.
- P. Shivakumar and S. Narayanan, End-to-end neural systems for automatic children speech recognition: An empirical study, Computer Speech & Language, 2022.
- R. Solera-Ureña et al., Transfer Learning-Based Cough Representations for Automatic Detection of COVID-19, Interspeech 2021.
- N. Tawara et al. Age-VOX-Celeb: Multi-Modal Corpus for Facial and Speech Estimation, ICASSP 2021.
- A. Van Den Oord et al. Neural discrete representation learning, Advances in neural information processing systems, 2017.
- H. Wang et al., Bitnet: Scaling 1-bit transformers for large language models. arXiv:2310.11453, 2023.
- W. Wang et al., Towards data selection on TTS data for children's speech recognition, ICASSP 2021.
- W. Ward et al., My science tutor: A conversational multimedia virtual tutor, Journal of Educational Psychology, vol. 105, 2013.
- Y. Xu et al., OneBit: Towards Extremely Low-bit Large Language Models. arXiv:2402.11295, 2024.
- L.-J. Yang et al., Parameter-Efficient Learning for Text-to-Speech Accent Adaptation, Interspeech 2023.
- E. Zaken et al., BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, ACL 2022.

AI Act

The Regulation on Artificial Intelligence is dense (Nautch et al., 2019) and very complex

- 180 recitals
- 113 articles
- 13 annexes
- 459 pages

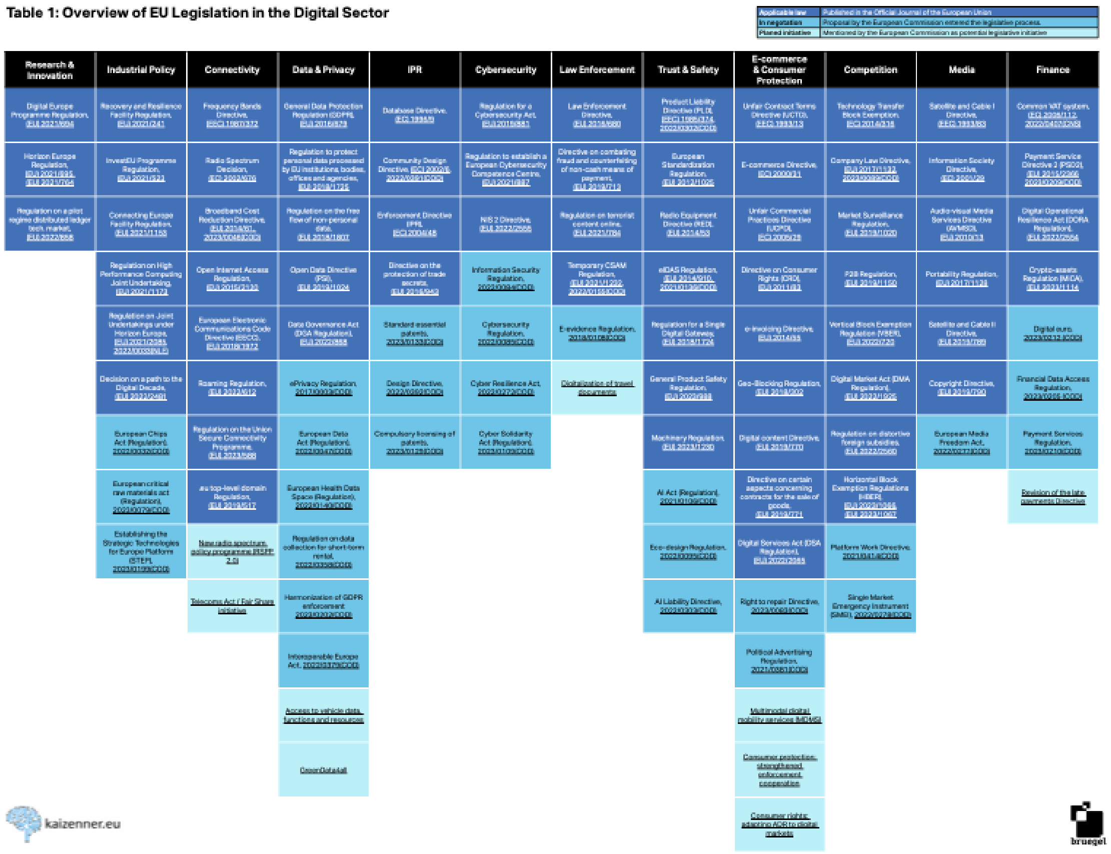


Table 1: Overview of EU Legislation in the Digital Sector

Applicable law	Published in the Official Journal of the European Union
In negotiation	Proposal by the European Commission entered the legislative process.
Planned initiative	Mentioned by the European Commission as potential legislative initiative

Research & Innovation	Industrial Policy	Connectivity	Data & Privacy	IPR	Cybersecurity	Law Enforcement	Trust & Safety	E-commerce & Consumer Protection	Competition	Media	Finance
Digital Europe Programme Regulation, (EU) 2021/694	Recovery and Resilience Facility Regulation, (EU) 2021/241	Frequency Bands Directive, (EEC) 1987/372	General Data Protection Regulation (GDPR), (EU) 2016/679	Database Directive, (EC) 1996/9	Regulation for a Cybersecurity Act, (EU) 2019/881	Law Enforcement Directive, (EU) 2016/680	Product Liability Directive (PLD), (EEC) 1985/374, 2022/0302(COD)	Unfair Contract Terms Directive (UCTD), (EEC) 1993/13	Technology Transfer Block Exemption, (EC) 2014/316	Satellite and Cable I Directive, (EEC) 1993/83	Common VAT system, (EC) 2006/112, 2022/0407(CNS)
Horizon Europe Regulation, (EU) 2021/695, (EU) 2021/764	InvestEU Programme Regulation, (EU) 2021/523	Radio Spectrum Decision, (EC) 2002/676	Regulation to protect personal data processed by EU institutions, bodies, offices and agencies, (EU) 2018/1725	Community Design Directive, (EC) 2002/6, 2022/0391(COD)	Regulation to establish a European Cybersecurity Competence Centre, (EU) 2021/887	Directive on combating fraud and counterfeiting of non-cash means of payment, (EU) 2019/713	European Standardization Regulation, (EU) 2012/1025	E-commerce Directive, (EC) 2000/31	Company Law Directive, (EU) 2017/1132, 2023/0089(COD)	Information Society Directive, (EC) 2001/29	Payment Service Directive 2 (PSD2), (EU) 2015/2366, 2023/0209(COD)
Regulation on a pilot regime distributed ledger tech. market, (EU) 2022/858	Connecting Europe Facility Regulation, (EU) 2021/1153	Broadband Cost Reduction Directive, (EU) 2014/61, 2023/0046(COD)	Regulation on the free flow of non-personal data, (EU) 2018/1807	Enforcement Directive (IPR), (EC) 2004/48	NIS 2 Directive, (EU) 2022/2555	Regulation on terrorist content online, (EU) 2021/784	Radio Equipment Directive (RED), (EU) 2014/53	Unfair Commercial Practices Directive (UCPD), (EC) 2005/29	Market Surveillance Regulation, (EU) 2019/1020	Audio-visual Media Services Directive (AVMSD), (EU) 2010/13	Digital Operational Resilience Act (DORA Regulation), (EU) 2022/2554
	Regulation on High Performance Computing Joint Undertaking, (EU) 2021/1173	Open Internet Access Regulation, (EU) 2015/2120	Open Data Directive (PSI), (EU) 2019/1024	Directive on the protection of trade secrets, (EU) 2016/943	Information Security Regulation, 2022/0084(COD)	Temporary CSAM Regulation, (EU) 2021/1232, 2022/0155(COD)	eIDAS Regulation, (EU) 2014/910, 2021/0136(COD)	Directive on Consumer Rights (CRD), (EU) 2011/83	P2B Regulation, (EU) 2019/1150	Portability Regulation, (EU) 2017/1128	Crypto-assets Regulation (MiCA), (EU) 2023/1114
	Regulation on Joint Undertakings under Horizon Europe, (EU) 2021/2085, 2022/0033(NLE)	European Electronic Communications Code Directive (EECC), (EU) 2018/1972	Data Governance Act (DGA Regulation), (EU) 2022/868	Standard essential patents, 2023/0133(COD)	Cybersecurity Regulation, 2022/0085(COD)	E-evidence Regulation, 2018/0108(COD)	Regulation for a Single Digital Gateway, (EU) 2018/1724	e-invoicing Directive, (EU) 2014/55	Vertical Block Exemption Regulation (VBER), (EU) 2022/720	Satellite and Cable II Directive, (EU) 2019/789	Digital euro, 2023/0212 (COD)
	Decision on a path to the Digital Decade, (EU) 2022/2481	Roaming Regulation, (EU) 2022/612	ePrivacy Regulation, 2017/0003(COD)	Design Directive, 2022/0392(COD)	Cyber Resilience Act, 2022/0272(COD)	Digitalization of travel documents	General Product Safety Regulation, (EU) 2023/988	Geo-Blocking Regulation, (EU) 2018/302	Digital Market Act (DMA Regulation), (EU) 2022/1925	Copyright Directive, (EU) 2019/790	Financial Data Access Regulation, 2023/0205 (COD)
	European Chips Act (Regulation), 2022/0032(COD)	Regulation on the Union Secure Connectivity Programme, (EU) 2023/588	European Data Act (Regulation), 2022/0047(COD)	Compulsory licensing of patents, 2023/0129(COD)	Cyber Solidarity Act (Regulation), 2023/0109(COD)		Machinery Regulation, (EU) 2023/1230	Digital content Directive, (EU) 2019/770	Regulation on distortive foreign subsidies, (EU) 2022/2560	European Media Freedom Act, 2022/0277(COD)	Payment Services Regulation, 2023/0210(COD)
	European critical raw materials act (Regulation), 2023/0079(COD)	.eu top-level domain Regulation, (EU) 2019/517	European Health Data Space (Regulation), 2022/0140(COD)				AI Act (Regulation), 2021/0106(COD)	Directive on certain aspects concerning contracts for the sale of goods, (EU) 2019/771	Horizontal Block Exemption Regulations (HBER), (EU) 2023/1066, (EU) 2023/1067		Revision of the late payments Directive
	Establishing the Strategic Technologies for Europe Platform (STEP), 2023/0199(COD)	New radio spectrum policy programme (RSPP 2.0)	Regulation on data collection for short-term rental, 2022/0358(COD)				Eco-design Regulation, 2022/0095(COD)	Digital Services Act (DSA Regulation), (EU) 2022/2065	Platform Work Directive, 2021/0414(COD)		
		Telecoms Act / Fair Share initiative	Harmonization of GDPR enforcement 2023/0202(COD)				AI Liability Directive, 2022/0303(COD)	Right to repair Directive, 2023/0085(COD)	Single Market Emergency Instrument (SMEI), 2022/0278(COD)		
			Interoperable Europe Act, 2022/0379(COD)					Political Advertising Regulation, 2021/0381(COD)			
			Access to vehicle data, functions and resources					Multimodal digital mobility services (MDMS)			
			OpenData4all					Consumer protection: strengthened enforcement cooperation			
								Consumer rights: adapting ADR to digital markets			

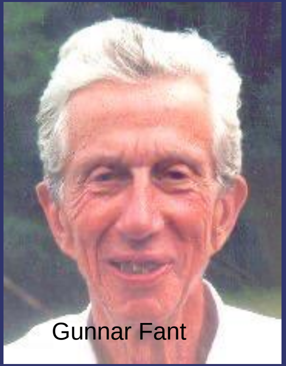
kaizenner.eu

bruegel

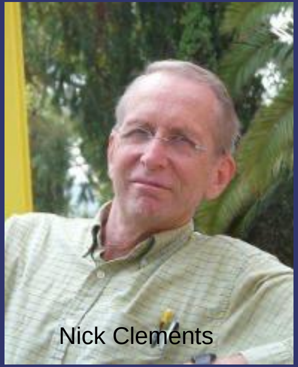
inesc id
lisboa

Thank you!
Obrigada!

international speech
communication association


Gunnar Fant

Ganesh Ramaswamy


Nick Clements

Maria Uther
Mark Huckvale
Steve Renals
Thomas Hain
Ji Ming
Simon King
Andrew Breen
Ben Milner
Martin Russell
Anna Barney
Denis Johnston
Steve Young
Shona D'Arcy
Simon Worgan
Michael McTear
Philip Jackson
Peter Jancovic

Thank you all!

inesc id lisboa

Thank you!
Obrigada!

to my family...

Thank you!

Human Language Technology@INESC-ID

ACCELERAT.AI

PRR Plano de Recuperação e Resiliência

REPÚBLICA PORTUGUESA

Financiado pela União Europeia NextGenerationEU

Alumni

Mentors (& Friends)

is awarded to
MATTHEW P. BLACK, MING LI, ANGELIKI
NGBOK LEE, SHRIKANTH S. NARAYANAN
of outstanding results and enthusiasm shown in

Thank you!
Obrigada!