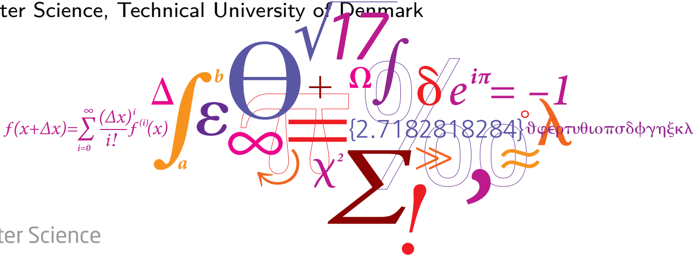
Feasibility of Few-shot Learning for the Automatic Transcriptions of Clinical-child Conversation in Danish

Speech Processing for clinical conversations

Sneha Das

Department of Applied Mathematics and Computer Science, Technical University of Denmark


$f(x+\Delta x)=\sum_{i=0}^{\infty}\frac{(\Delta x)^i}{i!}f^{(i)}(x)$
$\sqrt{17}$
$e^{i\pi}=-1$
{2.7182818284}
ϑφερτυθιοπσδφγηξκλ

DTU Compute
Department of Applied Mathematics and Computer Science

WristAngel: Research for Intervention and Management of OCD

* Progression and severity of disorder.

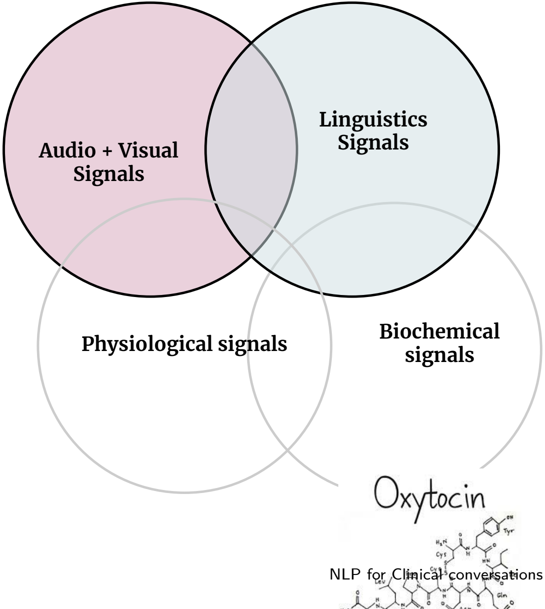
* Improve efficiency in CIB (Coding Interactive Behavior)

¤ Identify and predict impending OCD events.

¤ Aid in delivering cognitive behavioral therapy to patients.

¤ Provide useful interventions for management.

SPEECH



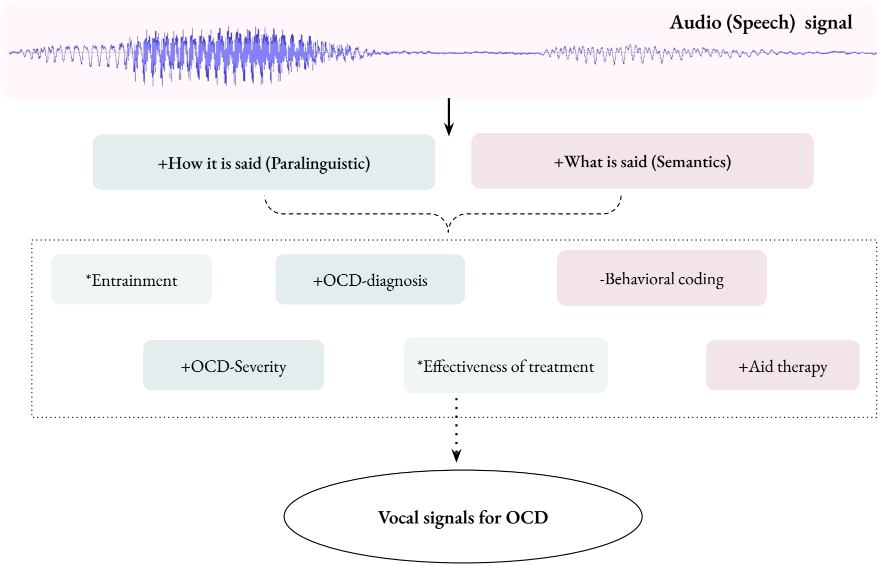
Audio + Visual Signals

Linguistics Signals

Physiological signals

Biochemical signals

Audio (Speech) in OCD Management



Audio (Speech) signal

+How it is said (Paralinguistic)

+What is said (Semantics)

*Entrainment

+OCD-diagnosis

-Behavioral coding

+OCD-Severity

*Effectiveness of treatment

+Aid therapy

Vocal signals for OCD

Speech preprocessing

1. Pre-processing: Conversations → speech segments.
2. Manual pre-processing: resource intensive
3. Approx. 13 minutes /per minute of annotation → 260 individual hours for annotating 10 minute long audio conversation for 120 audio samples.
4. Popular approach: ML pre-trained models pre-processing.

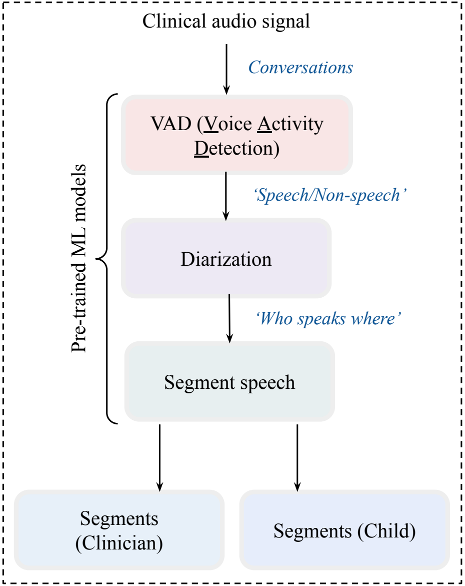
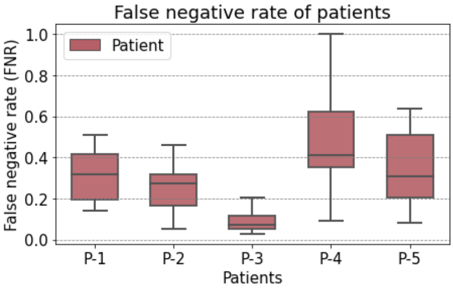


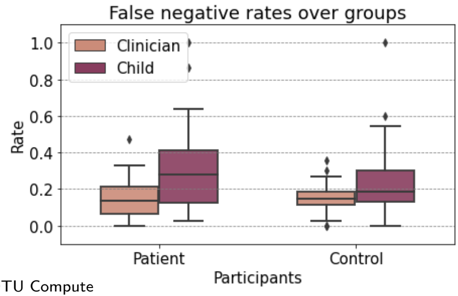
Figure: ML pre-processing pipeline.

Speech pre-processing

Challenges:

- Performance difference between clinicians and children.
- Errors (variance of error) higher for children in patient group.
- Correlation between error and OCD-severity score!!!





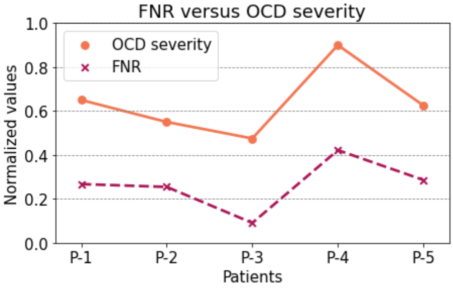
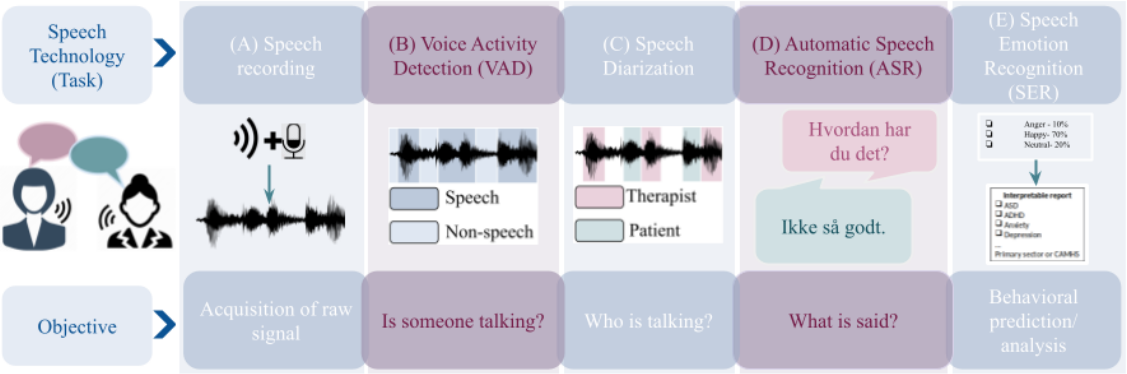


Figure 1: Speech tasks



Speech Technology (Task)	(A) Speech recording	(B) Voice Activity Detection (VAD)	(C) Speech Diarization	(D) Automatic Speech Recognition (ASR)	(E) Speech Emotion Recognition (SER)
Objective	Acquisition of raw signal	Is someone talking?	Who is talking?	What is said?	Behavioral prediction/ analysis

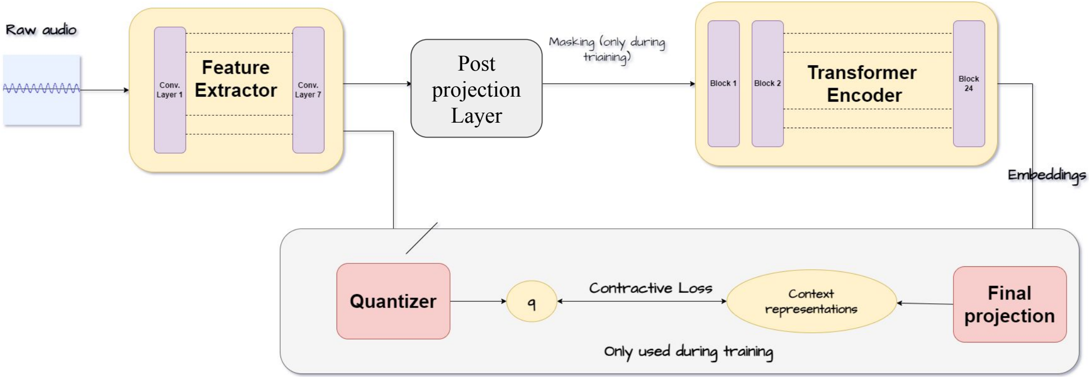
Automatic Speech Recognition and Transcriptions

- Clinical documentation
- Screening, diagnosis, management.

Automatic Speech Recognition and Transcriptions

1. State-of-the-art Models → English + Adults
2. State-of-the-model for Danish → Alvenir
3. Challenges:
   - Transcribe speech from children in Danish
   - Clinical conversations between clinician and child.
   - Do we have data?

Baseline and Wav2vec Model



What to do when no data?

Data-augmentation

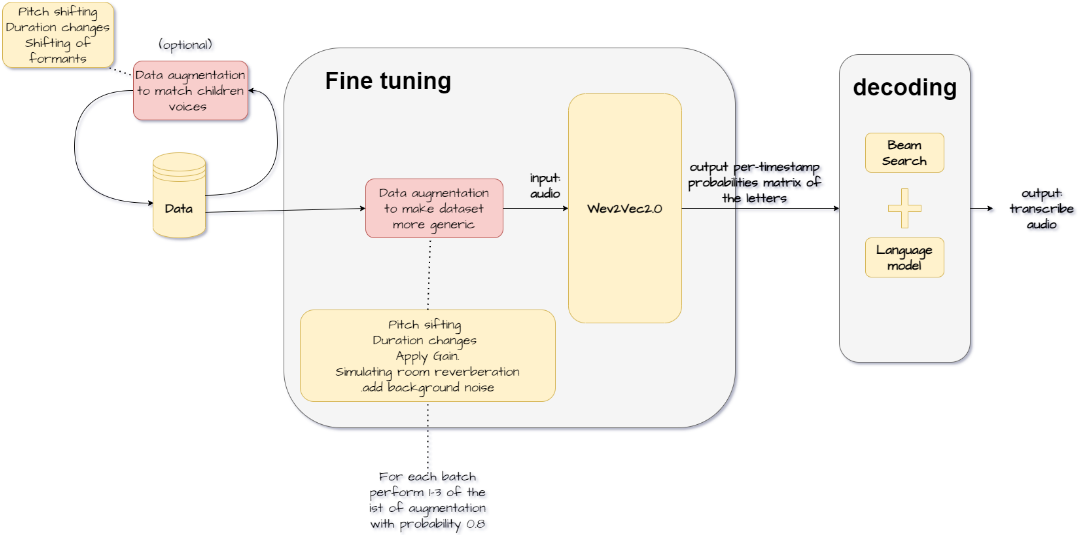
To aid in generalisation

- Gain change
- Reverberation
- Background noise
- pitch and duration modification

To aid in transfer to children

- Formant-shift
- Pitch modification
- Duration modification

Data augmentation



Pitch shifting
Duration changes
Shifting of
formants

(optional)

Data augmentation
to match children
voices

Data

Fine tuning

Data augmentation
to make dataset
more generic

input:
audio

Wev2Vec2.0

Pitch sifting
Duration changes
Apply Gain.
Simulating room reverberation
.add background noise

For each batch
perform 1-3 of the
ist of augmentation
with probability 0.8

output per-timestamp
probabilities matrix of
the letters

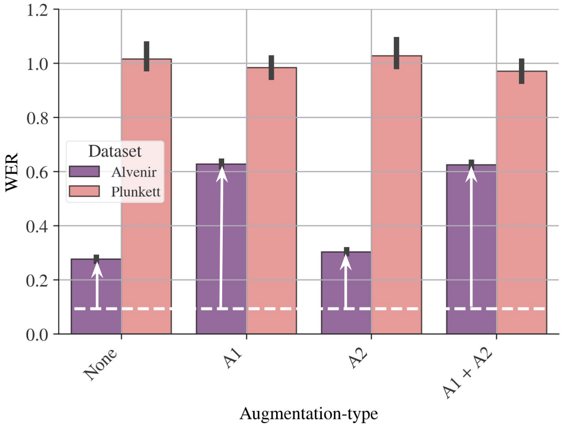
decoding

Beam
Search

Language
model

output:
transcribe
audio

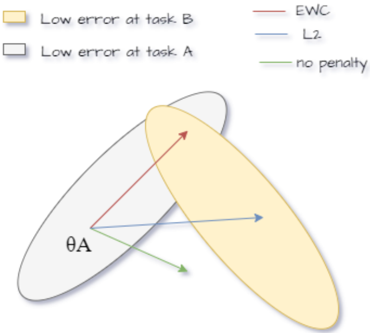
Data augmentation



- Testing on Alvenir + Plunkett
- Catastrophic forgetting → Not acceptable (!)

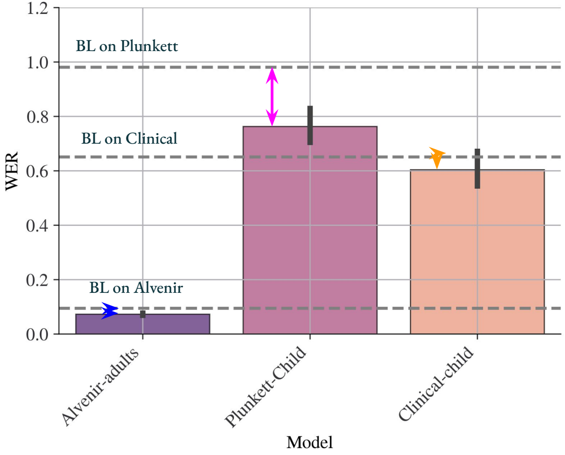
How to avoid Catastrophic forgetting?

- Weight freezing
  - Acoustic variability
  - Pronunciation variability
- Elastic weight consolidation: $L(\theta) = L_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$



Results

Performance of the best model[1]



[1] Garofalaki. M, Speech and natural language processing for clinical in-the-wild data 2023.

Affective-states from speech[1,2,5,6]

Applications:

- Entrainment
- Vocalization
- Behavioral coding

[1] S. Das et al, Towards Transferable Speech Emotion Representation: On loss functions for cross-lingual latent representations, ICASSP 2022.

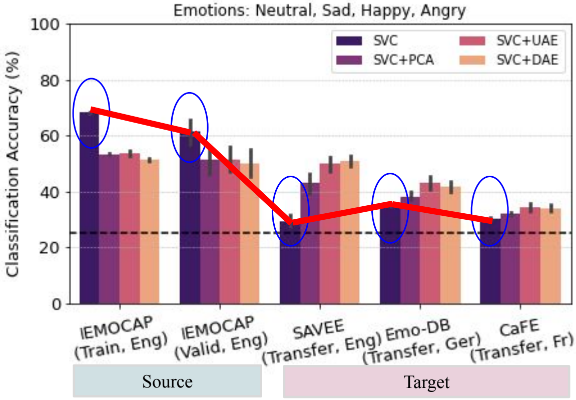
[2] S. Das et al, Continuous Metric Learning For Transferable Speech Emotion Recognition and Embedding Across Low-resource Languages, NLDL 2022.

[5] S. Das et al, Zero-shot Cross-lingual Speech Emotion Recognition: A Study of Loss Functions and Feature Importance, ISCA SPSC Symposium 2022.

[6] Clemmensen et al, Associations between OCD severity and vocal features in children and adolescents: A statistical and machine learning analysis plan, JMIR Protocols.

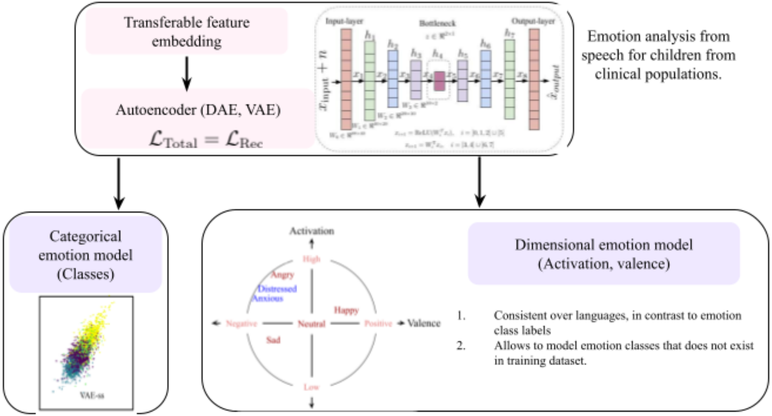
Audio-features → (Simple!) Emotion-recognition

- Input-features: descriptive features of speech features (f_0, tonality, intonation, etc) -features $R^{88\times 1}$ → Support vector machine (SVM) [Das, S, et al. 2021]



Transferability: What variable to condition on?

Emotion class (discrete) or dimensional model (continuous)?



Transferable feature embedding

Autoencoder (DAE, VAE)

$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Rec}}$

Emotion analysis from speech for children from clinical populations.

Categorical emotion model (Classes)

Dimensional emotion model (Activation, valence)

1. Consistent over languages, in contrast to emotion class labels
2. Allows to model emotion classes that does not exist in training dataset.

Universal emotion representation

Input Features

Computational model

Output == Activation

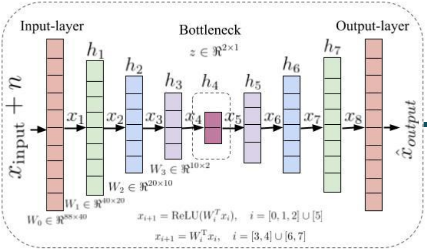
PITCH

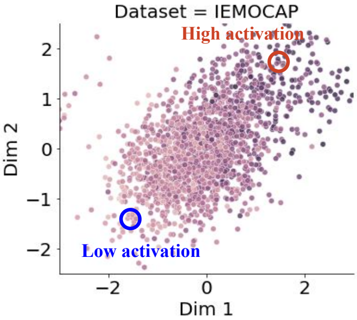
TONE

LOUDNESS

⋮

F0


Input-layer
Bottleneck
$z \in \Re^{2\times1}$
Output-layer
$x_{\text{input}} + \eta$
h_1
h_2
h_3
h_4
h_5
h_6
h_7
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
$\hat{x}_{output}$
$W_0 \in \Re^{88\times40}$
$W_1 \in \Re^{40\times20}$
$W_2 \in \Re^{20\times10}$
$W_3 \in \Re^{10\times2}$
$x_{i+1} = \text{ReLU}(W_i^T x_i), \quad i = [0,1,2] \cup [5]$
$x_{i+1} = W_i^{\text{T}} x_i, \quad i = [3,4] \cup [6,7]$



Summary

1. Speech-processing in psychiatry and psychology → accelerate and aid
2. Challenges:
   - Models are sensitive to language, age...
   - Lack of resources (data, labels)
3. How to adapt ASR modelled on adults to children with above challenges?
   - Augmentation
   - Continual learning → Elastic weight consolidation.
4. Performance on adults maintained
5. Performance on children improved by → 80%, 5%
6. (Large!) room to improve.

SCAN ME

Thankyou!

Email: sned@dtu.dk; Twitter: @dassneh