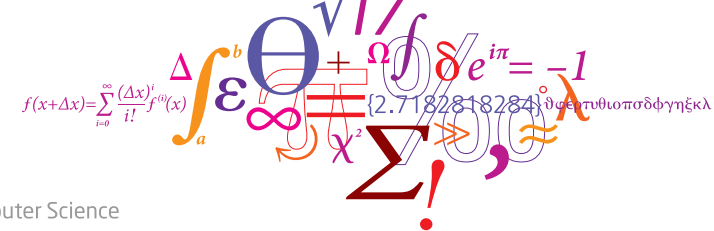
Few-shot Learning for Speech Processing and Automatic Transcriptions of Clinical-child Conversation in Danish

Speech Processing for clinical conversations

Sneha Das

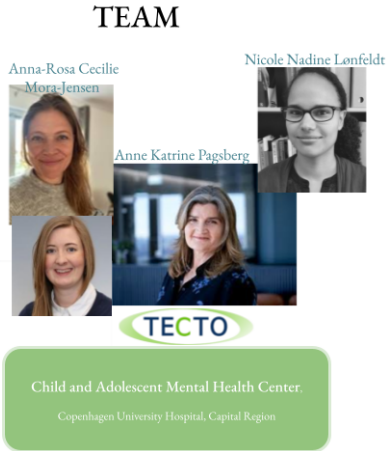
Department of Applied Mathematics and Computer Science, Technical University of Denmark



DTU Compute
Department of Applied Mathematics and Computer Science

Wrist Angel: Using wearables to predict OCD-events

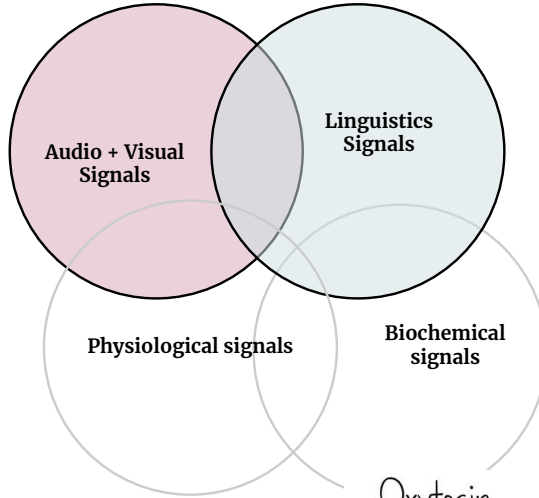
Multiple signal modalities for OCD management (intervention, feedback)



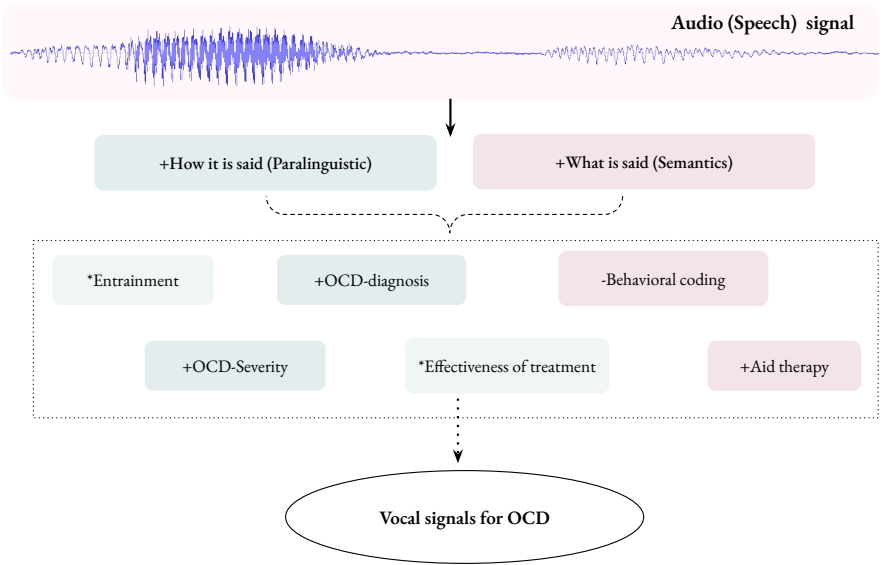
WristAngel: Research for Intervention and Management of OCD

* 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
Oxytocin

Audio (Speech) in OCD Management



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.

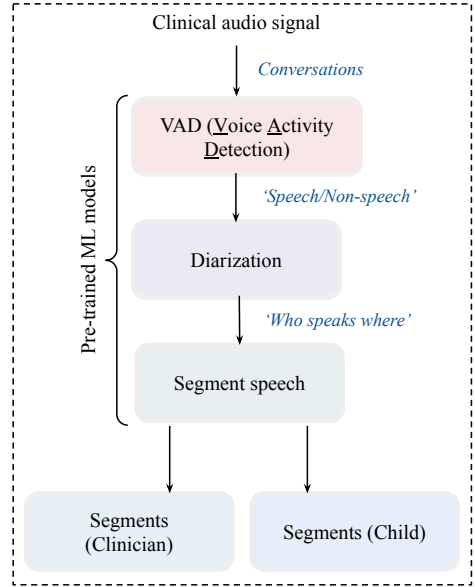
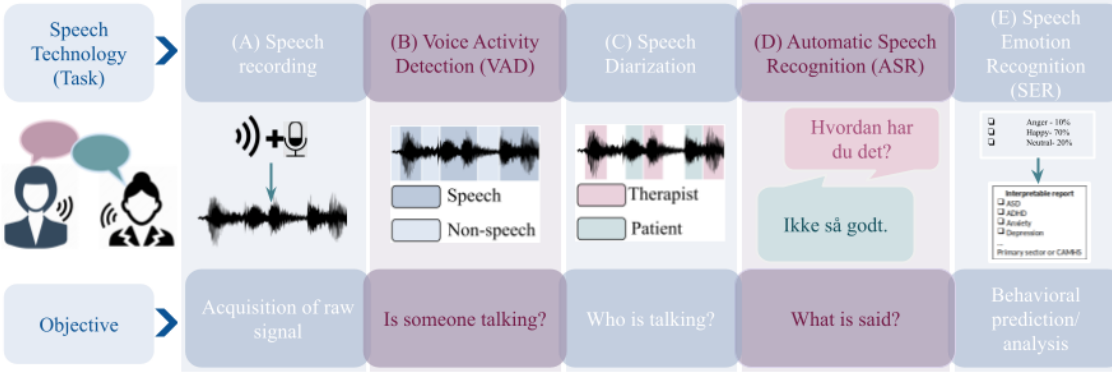


Figure: ML pre-processing pipeline.

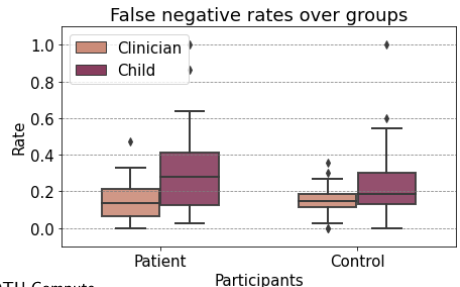
Figure 1: Speech tasks

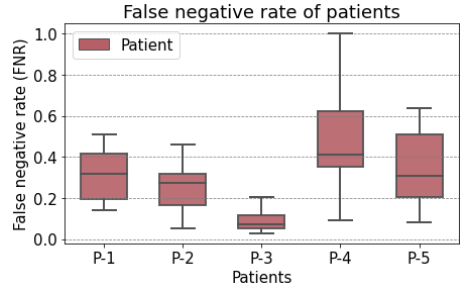


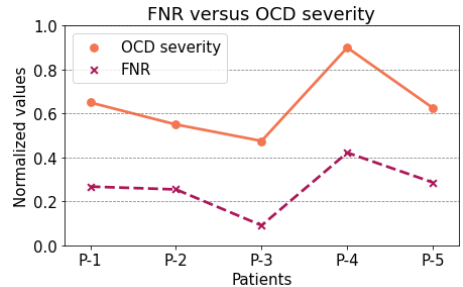
Speech pre-processing

Challenges:

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







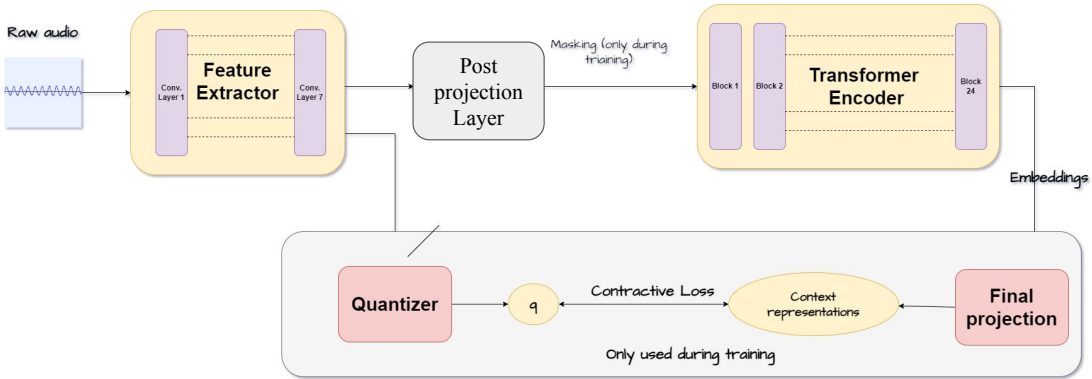
Automatic Speech Recognition and Transcriptions

- Clinical documentation
- Screening, diagnosis, management.

Automatic Speech Recognition and Transcriptions

1. State-of-the-art Models $\rightarrow$ English + Adults
2. State-of-the-model for Danish $\rightarrow$ 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

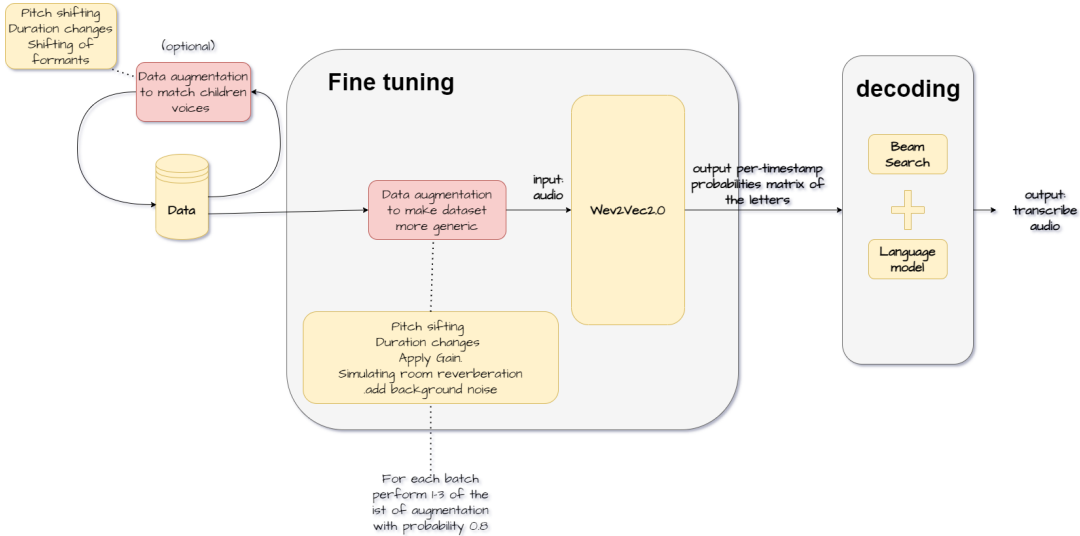
To aid in generalisation

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

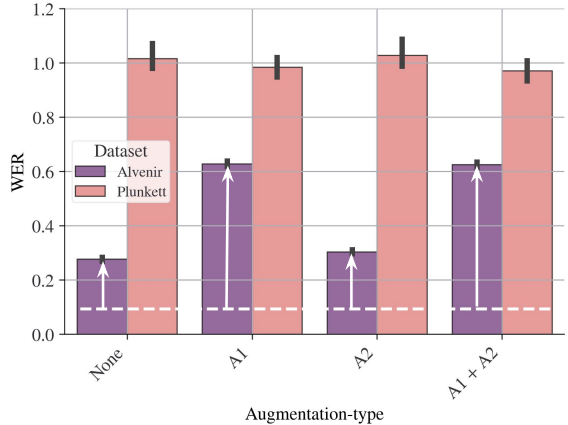
To aid in transfer to children

- Formant-shift
- Pitch modification
- Duration modification

Data augmentation (Synthetic data)



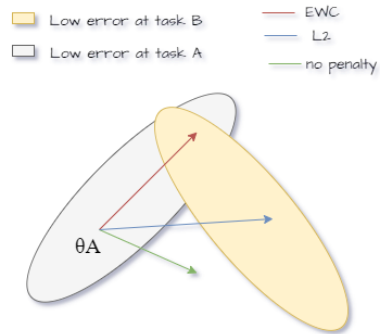
Data augmentation



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

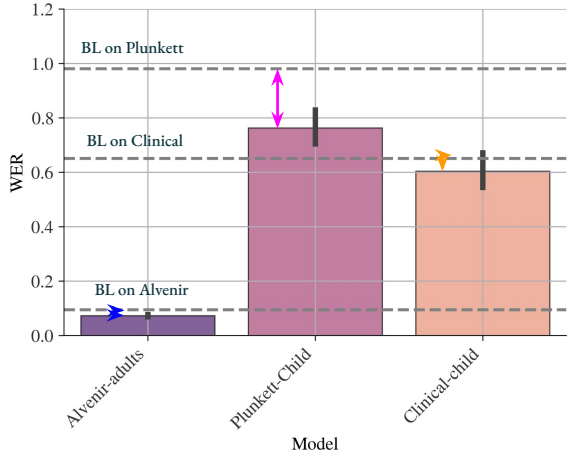
How to avoid Catastrophic forgetting?

- 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.

Summary

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

Thankyou!

Email: sned@dtu.dk