Vocal markers of neuropsychiatric conditions: assessing the generalizability of machine learning models and their clinical applicability

Alberto Parola [a, c], Astrid Rybner [b], Emil Trenckner Jessen [b], Marie Damsgaard Mortensen [b], Stine Nyhus Larse [b], Arndis Simonsen [d], Jessica Mary Lin [b], Yuan Zhou [e], Huiling Wang [f], Shiho Ubukata [g], Katja Koelkebeck [h,i], Vibeke Bliksted [d], Riccardo Fusaroli [b, l]

a Center for Language Technology, Copenhagen University, Denmark b Department of Linguistics, Cognitive Science and Semiotics, Aarhus University **c** Department of Psychology, University of Turin, Italy, **d** Psychosis Research Unit - Department of Clinical Medicine, Aarhus University, Aarhus, Denmark, **e** Institute of Psychology, Chinese Academy of Sciences, Beijing, China, **f** Department of Psychiatry, Kyoto University, Japan, **g** LVR-Hospital Essen, Department of Psychiatry and Psychotherapy, Hospital and Institute of the University of Duisburg-Essen, Germany,.

Introduction

BACKGROUND

- Voice atypicalities, e.g.longer pauses or flattened intonation, are a **distinctive feature of schizophrenia often associated with specific symptoms.**
- **Computerized voice analysis** is a promising tool for identifying vocal markers of neuropsychiatric disorders:
- **Machine learning (ML) models** have shown **high classification performance in predicting patients' diagnosis and symptoms** (De Boer et al., 2021; Cohen et al., 2021)

OPEN ISSUES

- However, recent works (Parola et al. ,2020, 2022 a,b) show that **generalizability of voice-markers is an issue.**
- **It is unclear whether voice-based ML models generalize to different samples and languages:** can we use a model trained on English to predict Danish data?
- The assessment of **generalization performance is crucial for clinical applicability.**

AIMS

We assessed the generalizability of voice-based ML models:

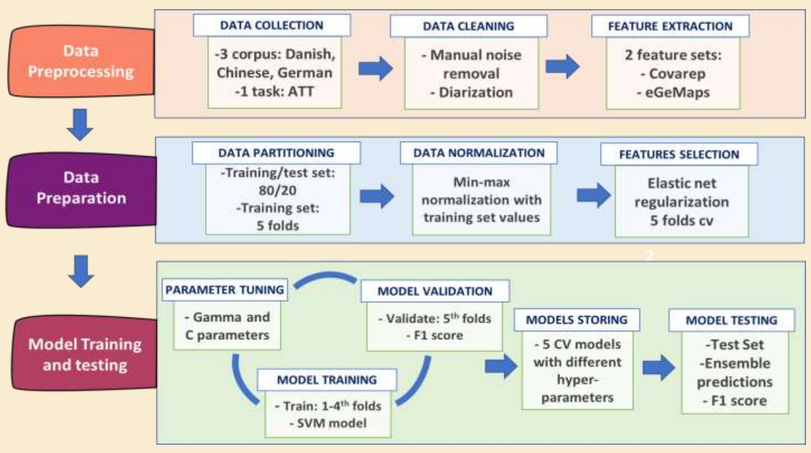
- **Q1:** How well do ML models generalize to different participants **speaking the same language?**
- **Q2:** How well do ML models generalize to participants **speaking a different native language?**
- **Q3:** Does combining models trained on different languages help improve generalization performance when predicting participants **speaking a different language?**
- **Q4: Does training models on a multilingual training set, i.e. combining participants speaking different languages,** help improve generalization performance?

Methods and design

PARTICIPANTS & PROCEDURE

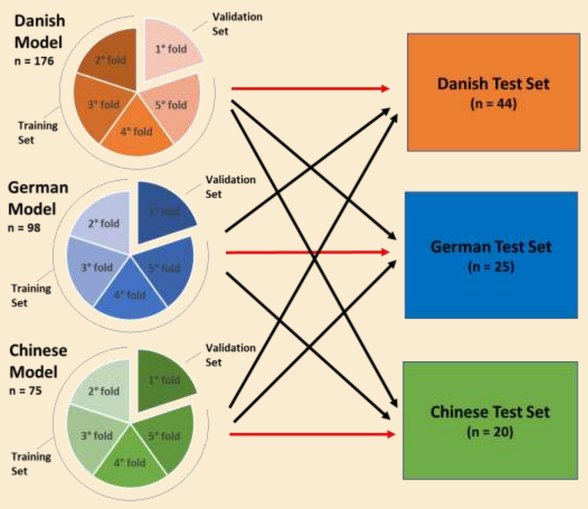
- **Participants: large cross-linguistic dataset (4 languages: Danish, German, Mandarin, Japanese)** involving 162 participants with SCZ (104 DK, 51 CH, 7 JP) and 172 matched controls (116 DK, 43 CH, 13 JP).
- **Speech task**: the **Animated Triangle Task, open-ended description of animated videos.**
- **Clinical data: SANS, SAPS, PANSS**

ML PIPELINE



- **A rigorous pipeline to minimize overfitting:**
- **Acoustic features:** Covarep, eGeMaps
- **Elastic net feature selection**
- **Train, validation and hold-out sets to avoid data leakage**
- **Five-fold CV on the training set only**
- **Stratified training sets (sex and diagnosis)**
- **Mixture of Experts (MoE) models**
- **Peformance: F1 metric**

MODEL TRAINING AND TESTING



We tested model generalizability on:

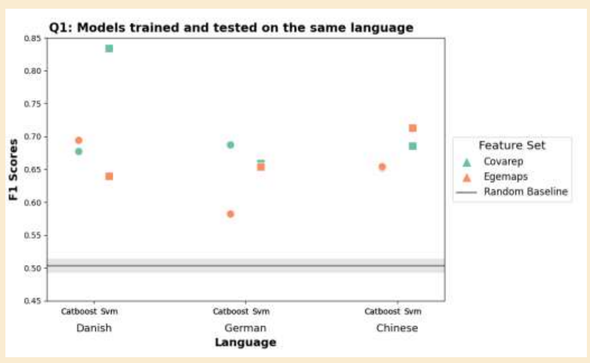
- **Q1: different participants, speaking the same language (hold-out test set);**
- **Q2: different participants, speaking a different language.**

We compared predictive performance:

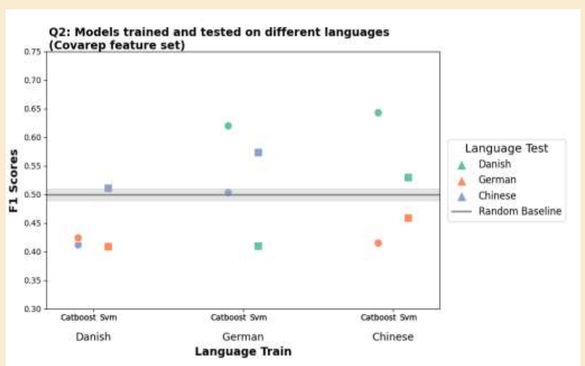
- **(i) models tested on a single language**
- **(ii) MoE models, i.e., ensemble of predictions of models trained on different languages**
- **(iii) multi-language models trained on multi-languages datasets.**

Results

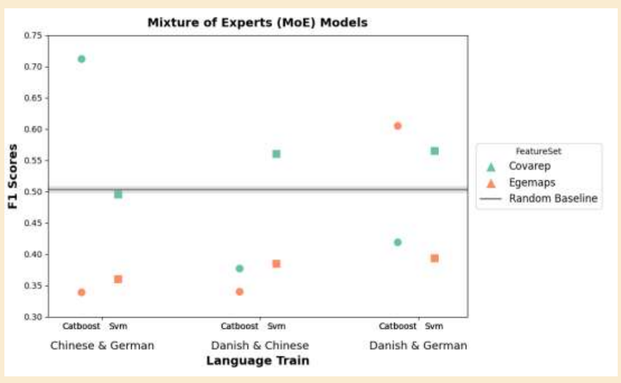
Q1: Train/Test same language



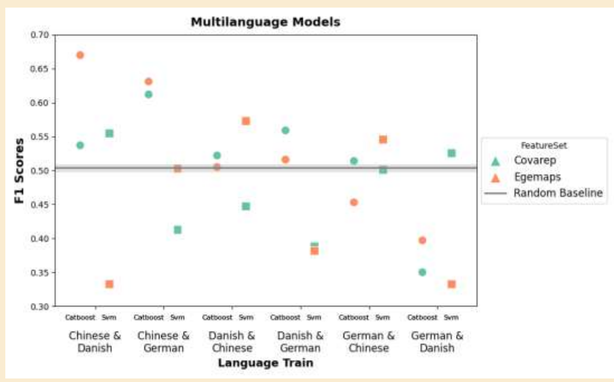
Q2: Train/Test different language



Q3 : Mixture of Experts Models



Q4: Multilanguage Models



Discussion

MAIN FINDINGS

- **Q1) Model performance comparable to state-of-the art findings** (F1 ~ 70%-80%) **when trained and tested on** participants speaking **the same language** (out-of-sample performance).
- **Q2)** Crucially the **ML models did not generalize well** - performance close to chance - **when trained in a language and tested on new languages**
- **Q3 and Q4) MoE and multi-language models show a slight increase of performance** (F1 up to 55%-60%), but still far from those requested for clinical applicability.
- **Cross-linguistic generalizability of voice-based ML models of schizophrenia is limited. If our first goal clinical appliicability, we need to account for this variability.**

RECOMMENDATIONS FOR FUTURE STUDIES

- 1) Larger open datasets to test: a) the generalizability of voice-based ML models across different speech tasks, heterogeneous clinical profiles, languages b) presence of bias
- 2) Rigorous pipeline to increase robustness and generalizability of ML models
- 3) Transfer learning: Tasks which allow a better transfer: e.g., emotional content, or relevant psychopathological dimensions.

References

- Parola, A., Simonsen, A., Lin, J. M., Zhou, Y., Huiling, W., Ubukata, S., ... & Fusaroli, R. (2022). Voice patterns as markers of schizophrenia: building a cumulative generalizable approach via cross-linguistic and meta-analysis based investigation. medRxiv.
- Parola, A., Lin, J. M., Simonsen, A., Bliksted, V., Zhou, Y., Wang, H., ... & Fusaroli, R. (2022). Speech disturbances in schizophrenia: assessing cross-linguistic generalizability of NLP automated measures of coherence. Schizophrenia Research.
- Parola, A., Simonsen, A., Bliksted, V., & Fusaroli, R. (2020). Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis. Schizophrenia research, 216, 24-40.
- De Boer, J. N., Voppel, A. E., Brederoo, S. G., Schnack, H. G., Truong, K. P., Wijnen, F. N. K., & Sommer, I. E. C. (2021). Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool. Psychological medicine, 1-11.

CONTACTS

Alberto Parola
alberto.parol@gmail.com