Research in Developmental Disabilities 157 (2025) 104906

Contents lists available at ScienceDirect

Research in Developmental Disabilities

journal homepage: www.elsevier.com/locate/redevdis

Automated segmentation of child-clinician speech in naturalistic clinical contexts

Giulio Bertamini [a,b,c,*], Cesare Furlanello [d], Mohamed Chetouani [c], David Cohen [a,c], Paola Venuti [b]

[a] Department of Child and Adolescent Psychiatry, Pitié-Salpêtrière University Hospital - Sorbonne University, 47-83 Bd de l'Hôpital, Paris, Île-de-France 75013, France
[b] Laboratory of Observation, Diagnosis, and Education, Department of Psychology and Cognitive Science - University of Trento, Via Matteo del Ben, 5B, Rovereto, TN 38068, Italy
[c] Institute of Intelligent Systems and Robotics, Sorbonne University, Pyramide - T55, 4 Pl. Jussieu 65, Paris, Île-de-France 75005, France
[d] HK3 Lab, Piazza Manifattura, 1, Rovereto, TN 38068, Italy

ARTICLE INFO

Keywords:
Metric learning
Automated speech segmentation
Naturalistic clinical contexts
Non-invasive
Child-clinician interaction

ABSTRACT

Background: Computational approaches hold significant promise for enhancing diagnosis and therapy in child and adolescent clinical practice. Clinical procedures heavily depend n vocal exchanges and interpersonal dynamics conveyed through speech. Research highlights the importance of investigating acoustic features and dyadic interactions during child development. However, observational methods are labor-intensive, time-consuming, and suffer from limited objectivity and quantification, hindering translation to everyday care.

Aims: We propose a novel AI-based system for fully automatic acoustic segmentation of clinical sessions with autistic preschool children.

Methods and procedures: We focused on naturalistic and unconstrained clinical contexts, which are characterized by background noise and data scarcity. Our approach addresses key challenges in the field while remaining non-invasive. We carefully evaluated model performance and flexibility in diverse, challenging conditions by means of domain alignment.

Outcomes and results: Results demonstrated promising outcomes in voice activity detection and speaker diarization. Notably, minimal annotation efforts —just 30 seconds of target data— significantly improved model performance across all tested conditions. Our models exhibit satisfying predictive performance and flexibility for deployment in everyday settings.

Conclusions and implications: Automating data annotation in real-world clinical scenarios can enable the widespread exploitation of advanced computational methods for downstream modeling, fostering precision approaches that bridge research and clinical practice.

* Corresponding author at: Department of Child and Adolescent Psychiatry, Pitié-Salpêtrière University Hospital - Sorbonne University, 47-83 Bd de l'Hôpital, Paris, Île-de-France 75013, France.

E-mail addresses: giulio.bertamini@unitn.it (G. Bertamini), cesare.furlanello@hk3lab.ai (C. Furlanello), mohamed.chetouani@sorbonne-universite.fr (M. Chetouani), david.cohen@aphp.fr (D. Cohen), paola.venuti@unitn.it (P. Venuti).

https://doi.org/10.1016/j.ridd.2024.104906
Received 22 May 2024; Received in revised form 4 December 2024; Accepted 28 December 2024
Available online 18 January 2025

1. Introduction

1.1. AI in mental health

Computational techniques such as Artificial Intelligence (AI) hold significant potential to enhance clinical procedures in mental health. They can improve quantification and objectivity in developmental research while addressing key challenges that limit its translational applications (Bickman, 2020; Shatte et al., 2019). AI has shown promising evidence in both diagnostics and recognition of mental disorders (Low et al., 2020), as well as in predicting clinical outcomes (Lutz et al., 2019; Zilcha-Mano, 2018). Beyond these areas, AI's potential also extends to clinical processes themselves. For instance, AI could automatically transcribe patient-clinician dialogues during therapy sessions, allowing therapists to concentrate entirely on their patients. Moreover, computational analyses of large datasets can uncover subtle markers that may be challenging for humans to detect (Miner et al., 2020). Importantly, research suggests that clinicians may not be aware of their own behavioral and emotional responses or adaptations. In fact, at least some of these responses appear to be unconscious and might require the acquisition of procedural skills (Zilcha-Mano, 2017; Archinard et al., 2000). These factors underscore the relevance of AI applications also in the clinical training for mental health professionals.

1.2. Benefits to child psychotherapy

Clinical research in child mental health relies on observational methods, which are non-invasive but subjective, time-consuming, and labor-intensive. Computational approaches could enhance diagnostic procedures and allow the systematic monitoring of therapeutic paths, a critical aspect in the context of Neuro-Developmental Conditions (NDCs) (Lord et al., 2022) and psychotherapy (Lambert, 2017). For example, digital phenotyping offers scalable and ecological tools for early autism screening (Perochon et al., 2023). Current intervention research also highlights the need for precision approaches to improve psychotherapy efficacy by uncovering the underlying mechanisms of change (Taubner et al., 2023). Clinical practice in child development often occurs in unconstrained settings from assessment to treatment, involving background noise and unstructured, free interactions. Further, patients are often unable to use wearable sensors or adhere to strict protocols (Godel et al., 2023). The therapeutic process itself also unfolds within the patientclinician dyadic interaction (Anonymous ref), which is increasingly considered as a main therapy mediator (Albaum et al., 2022; Mössler et al., 2019; Green & Garg, 2018). Therefore, the computational analysis of interpersonal dynamics in child psychotherapy may deepen our understanding of therapeutic mechanisms and support outcome optimization. Although psychotherapy is inherently language-based, few studies have explored its acoustic aspects, and most have been restricted to adults (Flemotomos et al., 2021; Miner et al., 2020; Weiste & Peräkylä, 2014).

1.3. Significance of speech in assessments

Speech is a central element of human communication, expressing thoughts, intentions, and emotions. Itsimportance for child development and mental health is widely supported by research (Bourvis et al., 2018; Fusaroli et al., 2017). Speech is a readily collectible signal, challenging to consciously alter, and a direct expression of emotions. However, predictive models for speech processing still require improvement, with specificity being a major challenge (Godel et al., 2023; Pokorny et al., 2016), underscoring the need for advanced studies in dyadic communication (Marschik et al., 2022). While child speech features were linked to symptom severity in autism (Ahn et al., 2020), existing models generally fail in achieving good diagnostic performance based solely on acoustic signatures (Rybner et al., 2022; Chi et al., 2022). Measuring social communication and interaction patterns, such as turn-taking, represents a key area of focus (Gupta et al., 2016; Leclère et al., 2014; Messinger et al., 2010). These dynamics may, in fact, reflect conditionspecific adaptations (Cho et al., 2019; Bone et al., 2014; Saint-Georges et al., 2011). Furthermore, the development of conversational turn-taking has been associated with early language development, diagnosis, and treatment outcomes (Bourvis et al., 2018; Romeo et al., 2021; Donnelly & Kidd, 2021).

To enable such fine-grained analyses, it is crucial to perform an accurate speech segmentation and process sufficient amounts of data (Ouss et al., 2020). However, accumulating large high-quality datasets with rigorous setups often remains unfeasible in clinical contexts (Pokorny et al., 2016), and challenges of naturalistic data collection impact signal quality. Another important need is the development of scalable yet accurate models to reduce the time required for data annotation. Data annotation, even when automated, poses several challenges. It often relies on engineering methods that do not generalize to other contexts (Eni et al., 2020; Li et al., 2019; Cohen et al., 2013). Moreover, in therapeutic settings, child-clinician communication usually extends beyond language. Non-linguistic vocalizations, partial linguistic verbalizations (e.g., word approximations), child-directed speech (motherese), onomatopoeic sounds, and other communicative utterances like laughter form the foundational building blocks of the exchange. Any DL system purposed to detect such complex interactive patterns needs to be trained on these unique and challenging types of data, addressing data scarcity as a main condition (Li et al., 2021; Ebrahimpour et al., 2020). Current recommendations for clinical research in speech processing still advocate for using separate microphones during data acquisition (Low et al., 2020), which is often impractical in real-world settings outside laboratory environments. Furthermore, recent research highlighted the specificity and challenges of infant speech for automated systems, emphasizing the need for tailored solutions (Xu et al., 2023; Lahiri et al., 2023). In light of recent promising results in the field of speech-based autism diagnosis in controlled settings (Briend et al., 2023), automating acoustic segmentation represents a priority. Such advancements could have a direct impact on everyday clinical care and intervention in ecological settings.

1.4. Aim

To pave the way for the broad and accurate use of computational methods to study patient-clinician acoustic interaction, we explored the feasibility of non-invasive automated speech processing of naturalistic signals. Our study faced challenges such as data scarcity, low-quality audio signals, and ambient noise in unconstrained clinical settings.We also evaluated the adaptability of the proposed architecture to previously unseen conditions using domain alignment techniques (Farahani et al., 2021). Specifically, we employed a DL model based on Siamese Neural Networks (SNNs, Shorfuzzaman and Hossain 2021) for metric learning to perform second-by-second speech annotation of video-recorded clinical sessions. These sessions included Naturalistic Developmental Behavioral Intervention (NDBI) (Vivanti & Zhong, 2020) and diagnostic assessments with preschool autistic children. Our objective was to perform Voice Activity Detection (VAD) and Speaker Diarization (SD) to automatically segment the signal, which serves as a foundation for downstream modeling. This enables the non-invasive extraction of quantitative metrics, crucial for analyzing patient-clinician communication dynamics during psychotherapy. We evaluated model robustness and adaptability over increasingly challenging conditions, including different dyads, environments, devices, age groups, and languages. Model adaptability was tested using a limited resources approach. For both VAD and SD, our models were trained on about two hours of data, and adapted with only 30 seconds of target domain data.

2. Material and methods

2.1. Clinical datasets and annotation procedures

Clinical data were extracted from video-recorded sessions of NDBI and clinical evaluations of autistic preschool children conducted at the Laboratory of Observation, Diagnosis, and Education (ODFLab) of the Department of Psychology and Cognitive Science of the University of Trento. The language spoken by both patients and clinicians was Italian.

Manual annotation were performed categorizing data into four distinct labels:

- 0 (clinician): segments perceived as containing only the clinician's voice
- 1 (child): segments perceived as containing only the child's voice
- (mixed): segments containing both child's and clinician's voices, either overlapping or subsequent
- (noise): segments perceived as not containing any human voice

Manual annotations were conducted based on the majority of the content perceived within each 1-second segment, i.e. 16000 frames. To maximize available data, ensure feasibility of human annotation, and precisely detect overlapping segments, data were labeled second-by-second. This approach increased the variability of vocalization types, with particular focus on non-linguistic vocalizations. Further, it reduced the possibility to learn content-specific features for the classification tasks. Our primary goal was to enhance the identification of segments containing a single voice amid ambient noise. Recognizing that clinical communication dynamics are often actively supported by clinicians, with frequent and rapid speaker shifts, we selected the shortest time window that met our criteria. In addition to this, autistic preschoolers often exhibit reduced communication and single-syllable vocalizations (Paul et al., 2010). Preliminary analyses of shorter (0.5 s) and longer (2 s) segments revealed challenges: selecting shorter segments caused model convergence issues, while longer segments increased the number of mixed-content segments, which were unsuitable for metric learning approaches.

Manual annotations were assisted by a specialized Python script, enabling accurate, second-by-second data labeling. The script allowed repeated playback of each segment, along with adjacent ones, to provide additional context in case of uncertainty. For example, segments containing acoustic information only at the beginning or end posed challenge for human recognition. This method facilitated label review in ambiguous cases. If uncertainties persisted, videos were examined to provide visual information. Segments that could not be confidently classified were excluded to prevent bias in training data.

The *Clinical dataset* comprised 34 clinical sessions for 30 male autistic children (chronological age range 24–60 months; developmental age-equivalent range 14–45 months) and 8 clinicians (one male and seven females), with an average duration of 5 minutes (sd=5) for each child, totaling 146 minutes of data, or 8770 1-second samples. Of these, 5872 contained human voice: 2689 were clinician-only, and 3183 were child-only. Recordings were made in two rooms using identical bird's-eye cameras and environmental microphones. Annotations prioritized clinical evaluations where both clinician and child vocalizations were likely to occur, maximizing usable data for speaker diarization. Random moments during therapy sessions were also selected. Audio recordings exhibited variability and imbalance at the class, session, and child levels.

The *Clinical preschool test set* was created by randomly sampling audio from four children in the *Clinical dataset*, representing 15 % of the total number of children. This sampling strategy, detailed in the Data Analysis Plan (DAP) in 2.3, ensured robust representation at the child level. The 15 % proportion balanced sufficient data for training with effective model testing. With 30 children, this approach supported model convergence while testing on diverse data. Repeated cross-validation in the DAP further ensured that different combinations of infants were assessed at each external loop, generating a distribution for model evaluation.

Two additional clinical test sets were annotated. The *Clinical school-age test set* included two school-age male children and two female clinicians recorded during autism therapy in a different clinical environment. It contained 430 1-second segments. Both children and clinicians were different from those in the training data. The *Clinical adolescence test set* comprised recordings of a 15-year-old male adolescent during a clinical interview with two clinicians (1 male and 1 female), for a total of 485 1-second audio segments.

The clinicians, clinical setting, and camera setup differed from those in the *Clinical dataset*, although the same microphone was used.

These additional clinical datasets strengthened model evaluation by incorporating diverse acquisition conditions and patient age groups.

2.2. External datasets

Three distinct test sets were derived from open-source publicly available corpora. External data sources are fundamental to test the generalization of DL models to previously unseen conditions. In fact, DL models are prone to overfitting training data, which can lead to poorer performance in real-world application scenarios despite rigorous validation procedures. The *Non-clinical school-age test set* included adult speakers from the Emotional Voices Database (EmoV-DB, Adigwe et al. 2018), which contains recordings with emotional valence from four people (two males and two females). We focused on *neutral* and *amused* valences, which were more consistent with our clinical scenarios. Additionally, five speeches from famous people (two females and three males) were also included from a public speaker recognition dataset. Therefore a total of nine adults were included. Child segments were obtained from a dataset of 11 school-aged children (six males and five females) engaged in free speech or sentence reading. This dataset is in English and children are both native and non-native speakers (Kennedy et al., 2017).

The *Infant cry test set* was derived from a corpus featuring cry recordings of 786 newborns, captured in the first week of life in hospital or home environments and recorded by smartphones (Budaghyan et al., 2023).

The *Non-clinical adolescence test set* consisted of speech samples from 838 adolescents aged between 13 and 17. The language was Icelandic (Mena et al., 2021).

The *Clinical preschool test set*, *Non-clinical school-age test set*, *Non-clinical infant cry test set*, and *Non-clinical adolescence test set* were randomly sampled from their original data source during each experiment of our DAP. Finally, two publicly available datasets containing different sources of ambient noise were used as noise samples for VAD. Noise samples were included in datasets derived from open data and in the *Clinical adolescence test set*, since it did not include many ambient noise examples. The *Noise set* includes samples from two corpora: 546 segments recorded in an hospital environment (Ali et al., 2023) and 1917 samples of different ambient noise sources from which categories including human vocalizations, e.g., laugh, were excluded (Bansal & Garg, 2022).

Incorporating external non-clinical datasets enhanced model evaluation, allowing us to assess classification performance across different contexts, age groups, languages, and vocalization types. A summary of datasets included in this study is provided in Table 1.

Table 1
Dataset description.

Name	Language	Speaker number	Demographics	Samples (1-second)	Data source
Clinical dataset Clinical preschool	Italian	30 children 8 clinicians	Preschool	8770 total 5872 voiced 2898 noise 2689 clinician 3183 child 794 mixed	Laboratory clinical data
Clinical school-age	Italian	2 children 2 clinicians	School-age	430 total 342 voiced 88 noise 194 clinician 104 child 44 mixed	Clinical data from different context
Clinical adolescence	Italian	1 adolescent 2 clinicians	Adolescent	485 total 441 voiced 44 noise 171 clinician 225 adolescent	Laboratory clinical data
Non-clinical school age	English	11 children 9 adults	School-age	1838 total 1030 adults 808 child	Speeches, pronouncing sentences, Adigwe et al. (2018); Kennedy et al. (2017)
Infant cry		786 newborns	Infant	DAP: 500 randomly sampled 1-second segments from infants DAP: 500 randomly sampled 1-second segments from adults in Non-clinical school age	First cry at birth, Budaghyan et al. (2023)
Non-clinical adolescent	Icelandic	838 adolescents	Adolescent	DAP: 500 randomly sampled 1-second segments from adolescents DAP: 500 randomly sampled 1-second segments from adults in Non-clinical school age	Pronouncing sentences, Mena et al. (2021)
Noise		2463 noise segments		DAP: 500 randomly sampled noise segments	Environmental noise, Ali et al. (2023); Bansal and Garg (2022)

2.3. Data analysis plan

The DAP designed to evaluate our models consisted in a nested *M x N x K* cross-validation procedure. In this setup, the M external loops were replicated to split data in training and left-out test portions. Afterwards, a *N x K* repeated cross-validation schema was applied to the training portion, with N replicates of a *K-fold* cross-validation. For each external iteration $m = 1…M$, we randomly extracted four children from the *Clinical dataset*, i.e., 15 % of the 30 total patients, to form the *Clinical preschool test set* (shorthand of *Clinical preschool test set.m*), and removing them from the material used for the train-validation in the *N x K* part of the DAP. Notably, since the audio segments correspond to distinct children, the DAP is robust at the child-level. Consequently, the internal cross-validation procedure was performed on the remaining 26 children.

The *Clinical school-age test set* and *Clinical adolescence test set* were entirely used for testing, while the *Non-clinical school-age test set*, *Non-clinical infant cry test set*, and *Non-clinical adolescence test set* were also constructed stochastically across the *M* iterations to increase variability. In details: The *Non-clinical school-age test set* included 500 randomly sampled segments from children and adults from the corresponding datasets.

The *Infant cry test set* contained 500 randomly extracted cry samples.

The *Non-clinical adolescence test set* was built by sampling 500 adolescent segments.

Random sampling was performed prioritizing segments coming from different speakers. Adult segments for the *Infant cry test set* and *Non-clinical adolescence test set* were drawn from the remaining 500 samples that were not used in the construction of the *Non-clinical school-age test set*.

For the *N x K* internal cross-validation loop, we used K= 10 to train different models optimized for validation data, which were evaluated across all test sets. The entire process was repeated *M* times, and the results were aggregated. After initial extensive experimentation, the results reported in this study were based on a 20 × 1 × 10 DAP, yielding a distribution of 200 evaluations. We chose to increase the number of external loops in order to maximize the variability of the resampled sets. Each random extraction was balanced with respect to data categories. A set of seeds was defined to ensure reproducibility and to perform paired comparisons to evaluate model adaptation. For VAD analysis, noise samples coming from the public corpora were randomly injected in the *Non-clinical school-age test set*, *Non-clinical infant cry test set*, *Non-clinical adolescence test set*, and *Clinical adolescence test set*. In general, if the source

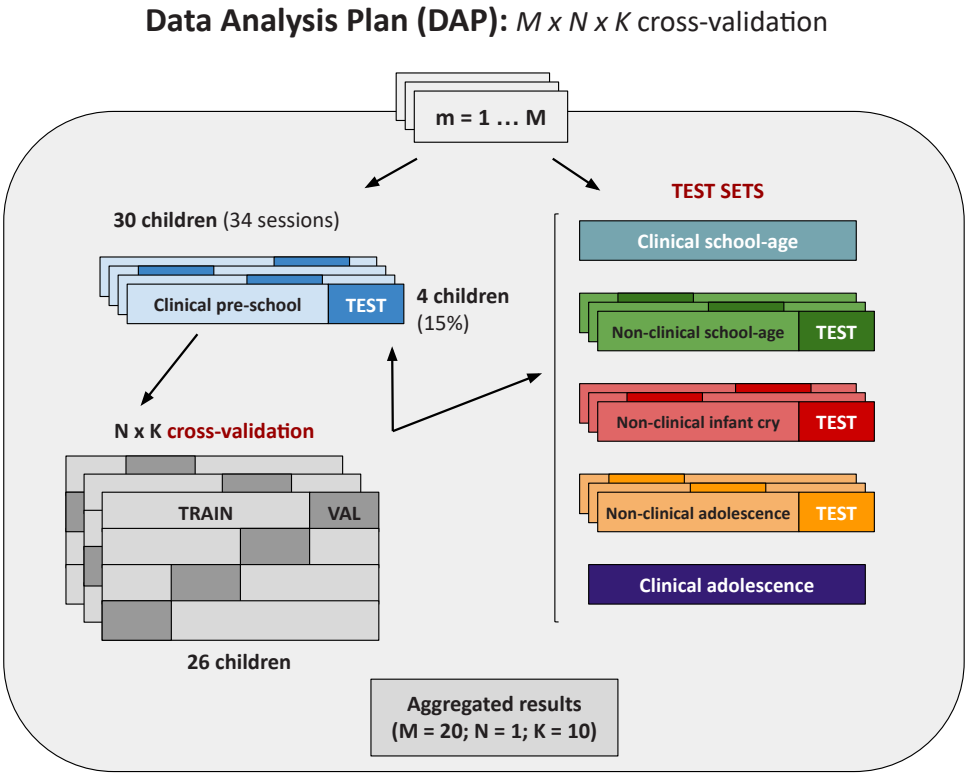

Fig. 1. Data Analysis Plan (DAP). **Data Analysis Plan (DAP):** M x N x K cross-validation. Pipeline for the *MxNxK* nested procedure for model training and validation over the test sets. At each of the M= 20 external iterations *Clinical preschool test set*, *Non-clinical school-age test set*, *Non-clinical infant cry test set*, and *Non-clinical adolescence test set* are constructed, then the model is trained and evaluated within a 10-fold repeated cross-validation, with 10 % of data used for validation during training. At each internal iteration, the trained model was evaluated across all test tests, and performance indexes were aggregated, obtaining a distribution of 200 observations for seven evaluation metrics: balanced accuracy, F1-score, Matthews Correlation Coefficient (MCC), sensitivity, specificity, precision, and Area Under the Curve (AUC).

audio signal was longer than one second, the portion of data to be extracted was randomly sampled at each iteration to increase variability. Conversely, in case of signals with a duration lower than one second, the signal was center-padded. The DAP is schematically represented in Fig. 1. Finally, we evaluated inter-rater reliability for the ground annotations, as well as the impact of gender-imbalance by a post-hoc additional analysis.

2.4. Model development and adaptation

Acoustic features consisted in Mel-Frequency Cepstral Coefficients (MFCCs) extracted from audio spectrograms. The Siamese model employed a parallel architecture with three branches and shared weights. It performed consecutive 2D dilated convolutions with increasingly wider receptive fields to implement hierarchical metric learning. Training data consisted of randomly generated triplets of elements (*anchor-positive-negative*) and the embedding space was optimized using triplet loss (Hermans et al., 2017). Classification was performed based on similarity between embedded and test elements in the latent space using *L2*-distance. Fig. 2 illustrates the model architecture and pipeline.

After model evaluation on the available test sets, we performed domain alignment to assess the model's ability to rapidly adapt to different scenarios (Farahani et al., 2021). Taking advantage of the flexibility of Siamese architectures (Atanbori & Rose, 2022; Wang & Zhai, 2020), we designed an adaptation procedure by modifying the process of triplet generation during training. In the same DAP setting, the adaptation procedure was based on training with a small number of elements randomly sampled from the new target test set, with speaker-balance. For each *m*-th external iteration, 30 samples were extracted and excluded from the corresponding test set. We chose to experiment with such numerousness considering that annotating 30 seconds of data would be highly convenient in clinical contexts. During triplet generation, a random number of these elements was included as triplet elements during each training step. To prevent overfitting the target domain, the number of injected triplets was randomly sampled from [2, 4, 6] at each training step with equal probabilities, and included within a batch of 20 triplets in each training step. Triplet injection was performed on training data

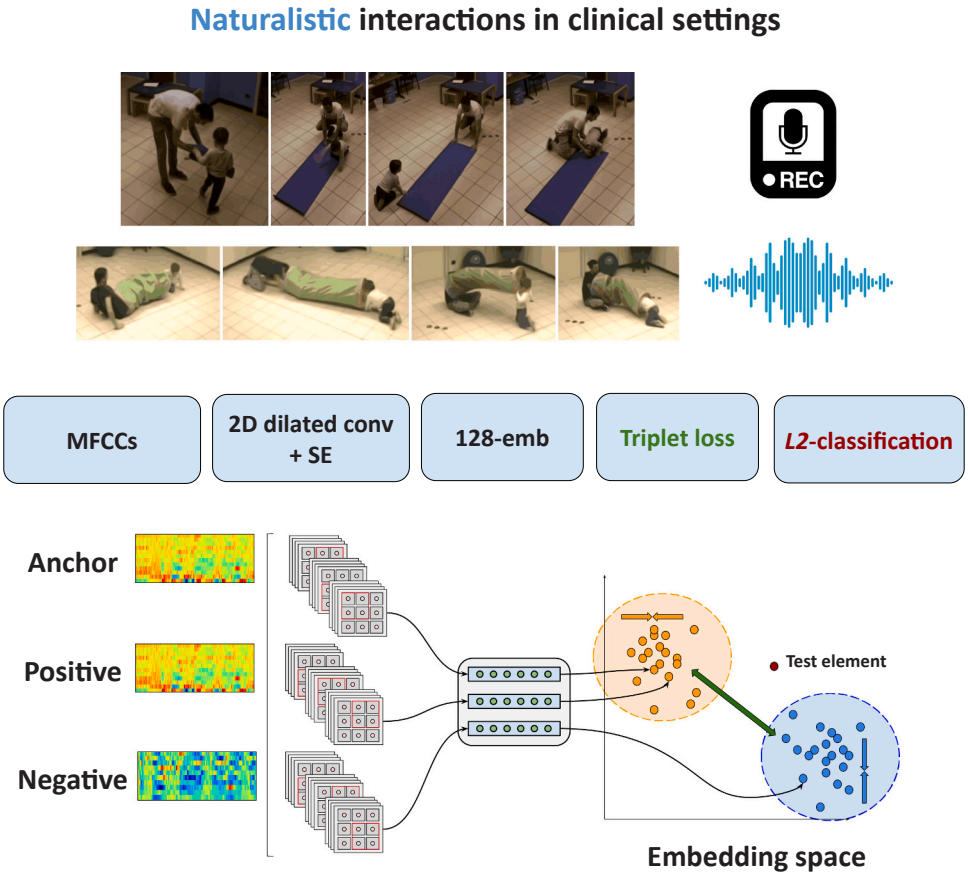

Fig. 2. Model architecture. **Naturalistic interactions in clinical settings.** Graphical representation of the system pipeline. The audio signal is non-invasively acquired by environmental microphones. After pre-processing, the set of MFCC features were extracted. The SNN was trained with randomly generated *anchor-positive-negative* triplets, using an architecture based on dilated convolutions incorporating a squeeze-and-excitation unit for dynamic channel-wise feature recalibration. The latent space is optimized through triplet loss. After training, test data are projected into the embedding space to perform *L2* distance-based classification.

only.

The adapted model was then evaluated against the baseline model on the reference test set across all seven evaluation metrics. Data normality was assessed with the Shapiro-Wilk test. The appropriate inferential test was applied to test for improvement significance, either a two-tailed paired *t*-test or Wilcoxon signed-rank test, was applied based on the normality of the performance metric distribution. Bonferroni correction was applied to control for multiple comparisons (Abdi, 2007).

Detailed technical information about model deployment, model architecture, feature extraction, acquisition conditions, and distance-based classification are provided in the Supplementary Materials. Additionally, the Supplementary Materials include a complementary non-clinical analysis based solely on publicly available external data. This analysis ensures full reproducibility of model training steps and can serve as a baseline to evaluate how architectures trained on non-specific data adapt to clinical scenarios.

2.5. Inter-rater reliability

The reliability of the initial manual annotations on the *Clinical dataset* was assessed by randomly sampling 100 1-second segments, balanced for category (child, clinician), which were re-annotated by two independent raters. These annotations were made based solely on the audio frames themselves, without any additional context such as previous or next seconds, or visual information, which made the task particularly challenging. Cohen's k between the two additional raters was 0.91, indicating a very high level of agreement. Fliess's k between the three sets of annotations was 0.87, which also reflects a strong agreement, even under the difficult annotation conditions.

3. Results

Our SNN models were applied to both VAD and SD tasks. The first model was trained to detect audio segments containing human voice versus segments containing ambient noise only. The second model was trained to recognize whether the vocalization came from the child patient or the adult clinician.

To assess model robustness across challenging datasets, we applied domain adaptation whenever model performance fell below a satisfactory threshold, i.e., average balanced accuracy lower than 80 %. Domain adaptation was based on only a minimal portion of target test data, i.e., 30 seconds of target data.

Model adaptation was therefore performed for SD on the *Clinical school-age test set*, *Non-clinical infant cry test set*, *Non-clinical adolescence test set*, and *Clinical adolescence test set*. Notably, only two alignments were necessary: one for children and one for adolescents. In fact, the model was able to simultaneously adapt to multiple conditions at the same time, based on the age range. The threshold for statistical significance was set to $p = 0.025$, as each adaptation metric compared the baseline model to its corresponding adapted model across the two domain alignments.

Fig. 3 represents the distribution of the fundamental frequency (*F0*) of both adult and child speech across all datasets.

3.1. Aim 1: voice activity detection

The results for VAD are reported in Table 2. Metrics are aggregated over the $M = 20$ splits and the $N = 1$ and $K = 10$ repeated cross-validation steps, yielding a total of 200 evaluations. Overall, the model performed consistently well, with performance ranging from good to optimal across all conditions, effectively differentiating human speech from segments containing only ambient noise. Furthermore, the model demonstrated strong generalization to various unseen sources of ambient noise, which were not present during training on the *Clinical dataset*.

3.2. Aim 2: speaker diarization

The results of SD and model adaptation are detailed in Table 3. Classification performance on the *Clinical preschool test set* was satisfactory, achieving high specificity and precision. To improve generalization, domain adaptation was

applied to the *Clinical school-age test set*, *Non-clinical school-age test set*, *Infant cry test set*, *Non-clinical adolescence test set*, and *Clinical adolescence test set*. Importantly, only two age-based domain alignment steps were sufficient for performance to generalize: one for datasets involving infants to school-age children and another for adolescents. Metrics were aggregated over the $M = 20$ splits, and the $N = 1$ and $K = 10$ repeated cross-validation steps, resulting in 200 evaluations.

With just 15 seconds of target data per class, adaptation to children improved balanced accuracy by an average of 23 %, F1-score by an average of 35 %, and MCC by an average of 88 % across the test sets. Despite slightly reduced accuracy across clinical test sets, the performance remained satisfactory.

Compared to the baseline, the adapted models demonstrated significantly better performance across all evaluation metrics while controlling for multiple comparisons. Performance consistently ranged from good to optimal across all tested conditions in the task of differentiating children's and adults' speech. Furthermore, high precision and specificity indicated the capability to reduce false positives.

Fig. 4 provides a visual comparison of adaptation results over the two clinical test sets (*Clinical school-age test set* and *Clinical adolescence test set*). The clinical datasets, captured in naturalistic and unconstrained clinical settings with high levels of ambient noise, underscore the robustness of the adapted models.

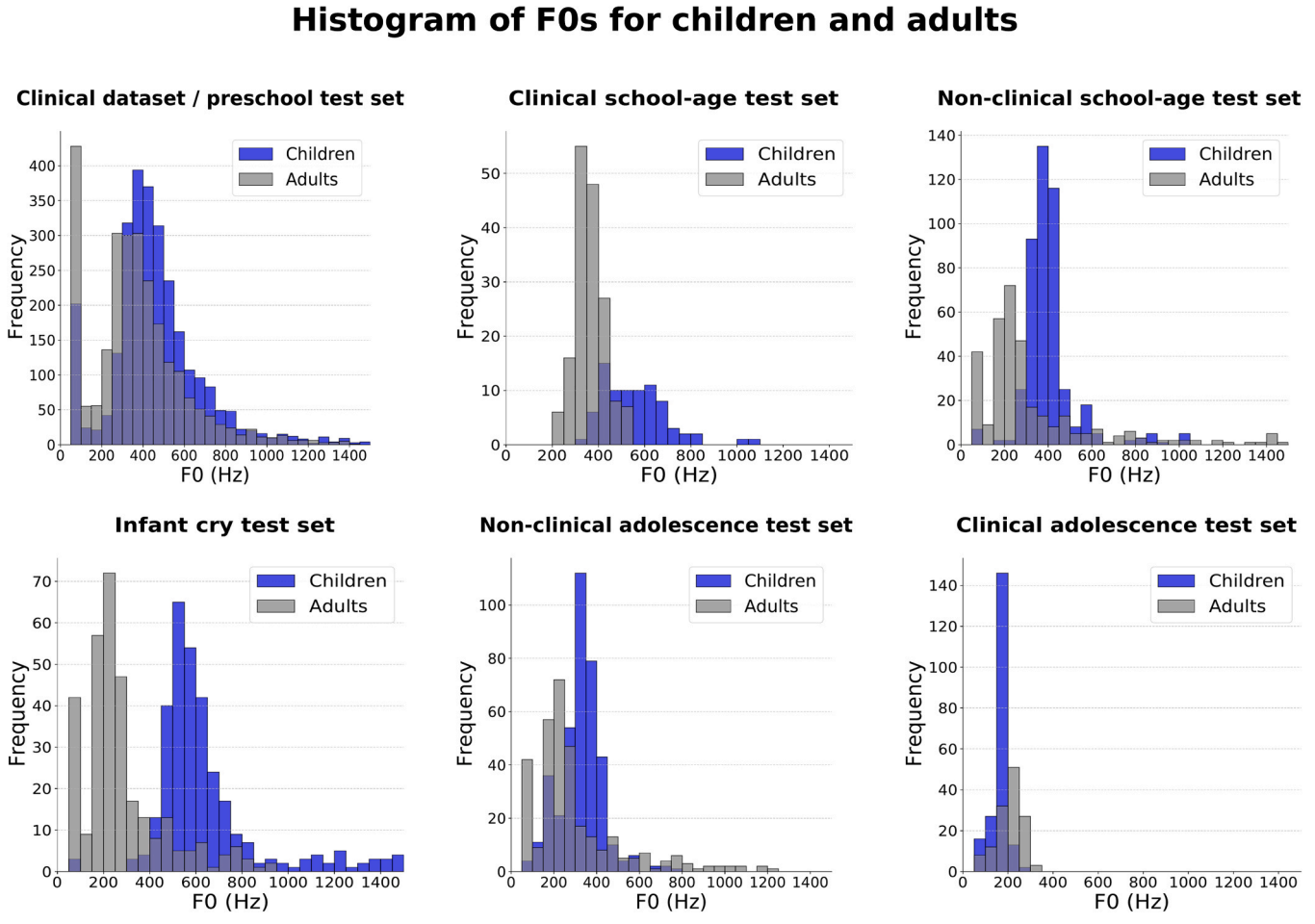

Fig. 3. *F0* frequency distribution. **Histogram of F0s for children and adults.** Distribution of *F0* for children and adults across all data sets. For the external datasets, 500 randomly sampled child/adolescent and adult segments were included.

Table 2
Voice activity detection.

Metric	Mean (Sd)
Clinical preschool test set	
Bal. Acc.	0.92 (0.04)
F1	0.95 (0.01)
MCC	0.81 (0.05)
Sensitivity	0.94 (0.02)
Specificity	0.90 (0.08)
Precision	0.97 (0.02)
AUC	0.98 (0.01)
Clinical school-age test set	
Bal. Acc.	0.93 (0.01)
F1	0.97 (0.01)
MCC	0.85 (0.02)
Sensitivity	0.96 (0.02)
Specificity	0.90 (0.04)
Precision	0.97 (0.01)
AUC	0.98 (0.01)
Non-clinical school-age test set	
Bal. Acc.	0.83 (0.02)
F1	0.83 (0.01)
MCC	0.66 (0.03)
Sensitivity	0.87 (0.02)
Specificity	0.78 (0.04)
Precision	0.80 (0.03)
AUC	0.90 (0.01)
Infant cry test set	
Bal. Acc.	0.83 (0.02)
F1	0.84 (0.02)
MCC	0.66 (0.03)
Sensitivity	0.89 (0.03)
Specificity	0.77 (0.04)
Precision	0.80 (0.03)
AUC	0.90 (0.02)
Non-clinical adolescence test set	
Bal. Acc.	0.83 (0.02)
F1	0.84 (0.01)
MCC	0.66 (0.03)
Sensitivity	0.88 (0.02)
Specificity	0.78 (0.04)
Precision	0.80 (0.03)
AUC	0.90 (0.02)
Clinical adolescence test set	
Bal. Acc.	0.81 (0.02)
F1	0.80 (0.03)
MCC	0.62 (0.04)
Sensitivity	0.84 (0.05)
Specificity	0.78 (0.04)
Precision	0.75 (0.03)
AUC	0.88 (0.02)

Model performance for voice activity detection. Evaluation metrics included: balanced accuracy, F1-score, MCC, sensitivity, specificity, precision, and AUC. Metrics are aggregated over the M= 20 external resamples, representing different data splits, and the N = 1 and K= 10 cross-validation iterations (totaling 200 evaluations). Adaptation involved injecting 30 1-second segments of target data during triplet generation. Mean and standard deviation are reported for each metric.

3.3. Post-hoc gender-based analysis

To address potential biases due to gender imbalance in the training data, a post-hoc evaluation was conducted using the *Non-clinical school-age test set*. Four subsets were created based on combinations of child-adult gender categories: male-male (mm), male-female (mf), female-male (fm), and female-female (ff). The adapted model from our DAP was evaluated on these subsets. Results demonstrated consistently high classification performance acrossall gender pairings Classification balanced accuracy remained consistently high (mm=0.92, mf=0.90, fm=0.96, ff=0.95). These findings indicate that the model's performance remains robust and unbiased with

Table 3
Speaker diarization and model adaptation.

	Baseline	Adaptation (30 seconds)	Test	
Metric	Mean (sd)	Mean (sd)	V	p
Clinical preschool test set				
Bal. Acc.	0.80 (0.04)	No adaptation		
F1	0.80 (0.06)			
MCC	0.61 (0.07)			
Sensitivity	0.76 (0.10)			
Specificity	0.84 (0.08)			
Precision	0.85 (0.08)			
AUC	0.90 (0.04)			
Clinical school-age test set				
Bal. Acc.	0.76 (0.05)	0.82 (0.03)	1045.50	< 0.001
F1	0.69 (0.07)	0.76 (0.04)	1402	< 0.001
MCC	0.52 (0.09)	0.64 (0.06)	1207.5	< 0.001
Sensitivity	0.78 (0.13)	0.83 (0.09)	5602	< 0.001
Specificity	0.75 (0.14)	0.82 (0.09)	4337	< 0.001
Precision	0.65 (0.10)	0.72 (0.08)	4235	< 0.001
AUC	0.85 (0.04)	0.91 (0.02)	562	< 0.001
Non-clinical school-age test set				
Bal. Acc.	0.68 (0.08) 0.89 (0.04)		1	< 0.001
F1	0.60 (0.15) 0.88 (0.04)		3.50	< 0.001
MCC	0.38 (0.16) 0.78 (0.07)		1	< 0.001
Sensitivity	0.52 (0.17) 0.85 (0.07)		44.50	< 0.001
Specificity	0.84 (0.07) 0.92 (0.05)		1018.50	< 0.001
Precision	0.76 (0.09) 0.92 (0.04)		21.50	< 0.001
AUC	0.78 (0.08) 0.96 (0.02)		0	< 0.001
Infant cry test set				
Bal. Acc.	0.76 (0.09) 0.93 (0.03)		103.50	< 0.001
F1	0.73 (0.13) 0.93 (0.04)		173.50	< 0.001
MCC	0.53 (0.17) 0.86 (0.06)		99.50	< 0.001
Sensitivity	0.68 (0.19) 0.94 (0.06)		438	< 0.001
Specificity	0.84 (0.07) 0.92 (0.05)		1151	< 0.001
Precision	0.83 (0.06) 0.93 (0.03)		279.50	< 0.001
AUC	0.86 (0.07) 0.98 (0.02)		79	< 0.001
Non-clinical adolescence test set				
Bal. Acc.	0.66 (0.08)	0.84 (0.04)	57	< 0.001
F1	0.58 (0.17)	0.87 (0.05)	124.50	< 0.001
MCC	0.32 (0.15)	0.68 (0.07)	0	< 0.001
Sensitivity	0.48 (0.19)	0.85 (0.09)	0	< 0.001
Specificity	0.84 (0.07)	0.83 (0.07)	9616	0.94
Precision	0.82 (0.07)	0.90 (0.03)	61	< 0.001
AUC	0.78 (0.07)	0.93 (0.03)	30.50	< 0.001
Clinical adolescence test set				
Bal. Acc.	0.63 (0.06)	0.81 (0.03)	1	< 0.001
F1	0.59 (0.13)	0.84 (0.04)	0	< 0.001
MCC	0.27 (0.12)	0.63 (0.06)	1	< 0.001
Sensitivity	0.51 (0.17)	0.85 (0.08)	110	< 0.001
Specificity	0.75 (0.09)	0.78 (0.08)	7412.50	< 0.001
Precision	0.73 (0.06)	0.84 (0.04)	105	< 0.001
AUC	0.72 (0.06)	0.90 (0.03)	0	< 0.001

Model performance for speaker diarization and model adaptation. Evaluation metrics included: balanced accuracy, F1-score, MCC, sensitivity, specificity, precision, and AUC. Metrics are aggregated over the M= 20 external resamples, representing different data splits, and the N = 1 and K= 10 cross-validation iterations (totaling 200 evaluations). Adaptation involved injecting 30 1-second segments of target data during triplet generation. Mean and standard deviation are reported for each metric, along with the results of the Wilcoxon signed-rank test for paired comparisons. Threshold considered for statistical significance was p = 0.025 with Bonferroni correction (Abdi, 2007), considering two adaptation steps performed. The evaluations for model adaptation were conducted on the four test sets.

respect to gender combinations. Detailed metrics and further analysis are provided in the Supplementary Materials.

4. Discussion

This work leveraged a Siamese neural network architecture to automate speech segmentation of clinical audio data, focusing on voice and speaker detection in sessions involving autistic preschoolers. This approach demonstrates significant potential to advance the field of child psychotherapy and developmental assessment by providing a noninvasive, scalable method for analyzing patient-clinician interactions in naturalistic, unconstrained clinical settings. By enabling quantitative modeling of patient-clinician interaction dynamics, this method addresses a critical need in clinical research and practice. The importance of studying interaction dynamics

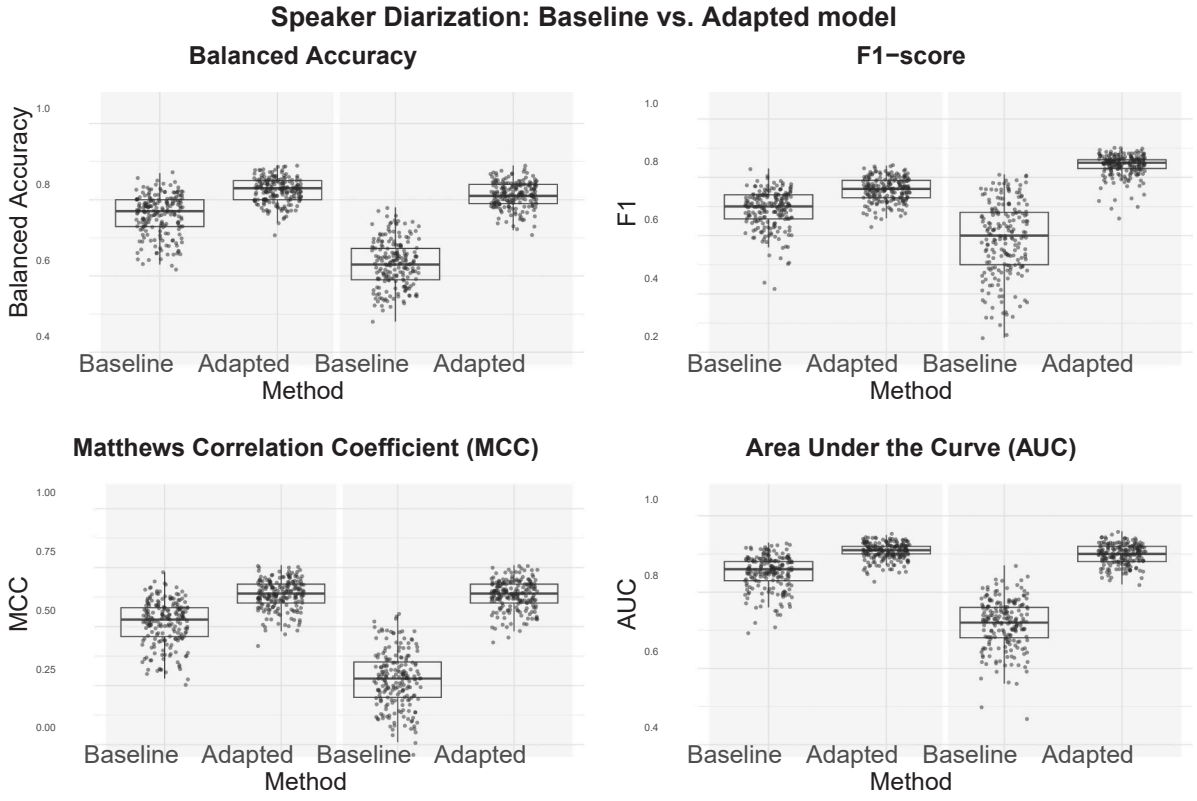

Fig. 4. Speaker diarization performance on clinical test sets. **Speaker Diarization: Baseline vs. Adapted model.** Comparison between baseline and adapted model for SD on *Clinical school-age test set* and *Clinical adolescence test set* over the nested cross-validation procedure of our DAP. Evaluation metrics included balanced accuracy, F1-score, MCC, and AUC. The figure illustrates the performance comparison between baseline and adapted models across various evaluation metrics and datasets. The inclusion of 30 seconds of target data for domain alignment significantly improved model performance across all evaluation metrics, as well as increasing model stability. Detailed information is reported in Table 3.

for diagnostic procedures and therapy responses has been emphasized in previous research, alongside the barriers to its widespread implementation (Washington & Wall, 2023; Flemotomos et al., 2021). In the context of child development, automated approaches can serve as foundational tools to implement quantitative, objective, and scalable methods, overcoming the limitations of observational research. Importantly, research also highlighted the crucial importance of longitudinal monitoring in the field of autism therapy (Lord et al., 2022), and psychotherapy (Norcross & Lambert, 2018; Lambert, 2017; Lambert et al., 2018) for precision approaches. Moreover, recent research has underscored the challenges associated with automated segmentation of clinical data involving autistic children, even in controlled setups with large datasets and sophisticated architectures (Lahiri et al., 2023).

The models presented in this work underwent rigorous evaluation through a nested cross-validation procedure across diverse test conditions. Our findings highlight the potential of automated systems to analyze acoustic real-world data coming from everyday, challenging clinical settings while remaining non-invasive.

The proposed models effectively detected voice activity across all test sets, with performance and stability confirmed by cross-validation. High precision and specificity were maintained across external datasets, confirming the model's ability to accurately identify positive examples while minimizing false positives. Reducing false positives is critical in this context, as they can distort speech analysis outcomes and lead to erroneous interpretations (Godel et al., 2023; Miner et al., 2020; Pokorny et al., 2016).

Our analysis indicates that the proposed architecture is well-suited for voice activity detection in unconstrained, ecological clinical conditions, even in the presence of significant ambient noise and limited data availability. Further, the model demonstrated robust generalization to previously unseen conditions, speakers, and technical setups without the need for adaptation.

With regard to speaker diarization, the model showed varying levels of performance. It achieved high performance on the test set recorded under conditions similar to those of training data, using the same devices, and involving preschoolaged children (*Clinical preschool test set*). Importantly, the model exhibited robustness at the individual child level, demonstrating its ability to generalize to unseen children.

The model achieved moderate performance on the *Non-clinical infant cry test set* and the *Clinical school-age test set*. The former consisted of infant cry samples, while the latter included annotations from school-age autistic children during therapy sessions. Both datasets were acquired under diverse recording conditions and in different environments. These results suggest the model's capacity to generalize, to some extent, across different age groups, environments, and technical setups. However, the nested cross-validation

procedure revealed higher variability in model stability, reflecting the increased complexity of the diarization task. Despite these challenges, precision and specificity remained satisfactory.

The model exhibited poorer performance on other test sets, indicating difficulties in generalizing across other age groups and languages. Specific challenges included accommodating school-aged English-speaking children, Englishspeaking adults, and adolescents speaking Italian and Icelandic (Fouquet et al., 2016; Markova et al., 2016; Nicollas et al., 2008).

Domain alignment involved the balanced integration of 30 seconds of target domain audio data in triplet generation during model training. During supervised metric learning for the Siamese neural network, a small subset of labeled examples from the new data was introduced to progressively incorporate knowledge from the target domain in the process, employing a few-shot learning approach (Atanbori & Rose, 2022; Wang & Zhai, 2020). Model adaptation significantly enhanced performance across all target datasets. Notably, only two adaptations based on target age were required to achieve very good performance. Precision and specificity generally remained high, except for the most challenging dataset of school-age autistic children, where performance, though slightly reduced, remained satisfactory. Model adaptation not only significantly improved all evaluation metrics but also reduced performance variance. The architecture demonstrated robust domain adaptation capability, achieving consistent good-to-optimal performance across diverse conditions. Remarkably, the training process, even with adaptation, required only a few epochs for convergence, completing in about five minutes. This efficiency enabled the architecture to generalize effectively to more challenging conditions with minimal data usage, covering scenarios ranging from newborns' crying to adolescent interactions. Research showed that infants' and children's voices exhibit distinct acoustic features due to immaturity of their vocal tracts, with these differences diminishing by adolescence. Nevertheless, when fine-tuned the successfully detected fine-grained features for adolescent-adult speaker diarization, even though it was primarily trained on children's voices (Fouquet et al., 2016; Markova et al., 2016; Nicollas et al., 2008). The model was also capable of detecting differences in spectrogram features between male children and female therapists, despite their potentially overlapping voice frequency ranges (Kent et al., 2021; Banik et al., 2015). In addition to this, clinicians often modulate their prosody during therapy with young children, employing child-directed speech to capture attention and engage them in the exchange (Saint-Georges et al., 2013; Mahdhaoui et al., 2009). Such modulation may further reduce the distinctions between adult and child voices at the spectrogram level, as suggested by previous research (Shute & Wheldall, 1999; Fernald & Kuhl, 1987; Warren-Leubecker & Bohannon, 1984).

A notable aspect of our work is the utilization of a small training dataset, comprising less than two hours of data for both tasks, which iseasilycollectible inmost clinicalscenarios. Domainadaptationrequired onlythirtysecondsofannotated target data, making the process efficient and straightforward. These factors suggest that our approach is well-suited for real-world clinical applications. Importantly, despite being trained on a limited preschool population under challenging conditions, the model demonstrated adaptability to newborns, school-aged children, and adolescents. Furthermore, our approach is agnostic to the linguistic content of audio segments. For instance, in autism intervention, language presence is often an outcome rather than a precondition. A key component of intervention is to assign communicative meaning to social partners' vocalizations (Camaioni, 2017) and establish reciprocal, circular communicative routines to support language development (Vivanti & Bottema-Beutel, & Turner-Brown, 2020). Notably, our system could also be applied to other relevant scenarios like infant-caregiver interaction in naturalistic settings.

Regarding the complementary reproducibility analysis conducted on publicly available non-clinical data, results supported the ability of the proposed architecture to learn suitable representations for detecting human speech and distinguishing between child and adult vocalizations. However, the models exhibited limitations in generalizing to realworld clinical scenarios characterized by high background noise, particularly in speaker diarization. This emphasizes the specificity of clinical data and emphasizes the need for collecting naturalistic samples to develop tailored solutions. In summary, our models may provide a robust and flexible foundation for the application of advanced computational techniques for downstream modeling to explore the dynamics of patient-clinician acoustic interactions, speech features, turn-taking patterns and affective content in child and adolescent clinical psychology. Advanced interaction metrics, such as interpersonal synchrony, could be easily and automatically extracted from large clinical samples (Ouss et al., 2020; Eyben et al., 2013; Lahiri et al., 2022; Li et al., 2022, 2021; Ochi et al., 2019). With minimal efforts, the model could also be quickly adapted to various conditions. Ultimately, these features may be exploited in predictive models of therapy response, to longitudinally monitor therapeutic paths, or applied in diagnostic procedures and therapeutic feedback.

Notably, our system exhibited high performance when compared to state-of-the-art solutions, despite being trained under challenging conditions, with significantly less data, and without relying on pre-trained models or multi-modal inputs (Xu et al., 2023; Lahiri et al., 2023).

The architecture addresses several challenges of automated approaches, including reducing false positives, distinguishing between adolescent and adult speech, recognizing child-directed speech, and identifying non-linguistic vocalizations. Its flexibility enables adaptation to diverse and challenging conditions using only a single environmental microphone, thus making it completely non-invasive (Godel et al., 2023; Lehet et al., 2021; Jones et al., 2019; Moffitt et al., 2022).

The current study is not without limitations. Foremost among these is the limited diversity and sample size for training, validation, and testing. All children in the analysis were males, reflecting the gender unbalance in autism research, particularly among preschoolers. This imbalance likely reduced generalization to female children and posed challenges to analysis. Male children typically exhibit lower fundamental frequencies in their speech compared to female children, resembling those of adults more closely. The complementary limitation holds for therapists. Considering the predominance of female psychologists, who typically have higher fundamental frequencies compared to males, their vocal signals would likely have been more similar to those of children. Consequently, while limiting generalization, the scenario of having predominantly male children and female adults likely created significant challenges for the diarization task (Fouquet et al., 2016; Markova et al., 2016; Nicollas et al., 2008). Our post-hoc analysis seems to support this hypothesis, identifying the male child-female adult condition as the most challenging. Conversely, the female child-male adult condition achieved the best performance despite, despite opposite gender-imbalance in training data. Additionally, variability in

evaluation procedures suggests potential underlying factors influencing embedding space construction that warrant further exploration. For instance, research showed that non-linguistic vocalizations may pose greater challenges compared to linguistic ones (Xu et al., 2023). Balancing these two classes during model training may be crucial, especially considering the high individual variability in the acoustic production of autistic children (Lahiri et al., 2023).

From a clinical perspective, while the architecture holds significant promise, its application is limited to contexts where vocal interaction, even if minimal, is present. This requirement may not always be met in neuro-developmental conditions.

Another limitation lies in the classification pipeline, that requires a two-step process: first identifying voiced segments and then performing speaker recognition. Further, in real-world applications some segments inevitably contain overlapping voices or mixed information, posing challenges to our second-level classification. Addressing this issue is a key area for future work. Potential solutions include implementing overlapping sliding window with majority-based classification, or devising a method to assess prediction confidence based on thresholding and data distribution in the embedding space. Clustering could be performed over the embedded data to both evaluate data training quality and identify challenging vocalizations. Afterwards, new data could be scored in terms of prediction confidence by evaluating their distance from the corresponding cluster centroids beyond their simple classification. Sliding windows could help in evaluating prediction stability over the same second, potentially helping in identifying segments containing mixed acoustic information. Alternatively, transformer architectures could also be employed to further incorporate temporal information, but requiring higher computational costs.

Future work may also consider expanding training data by incorporating additional sources and augmenting datasets to scale the existing data. Self-training approaches could also automatically increase available data. In our initial experiments, we applied various augmentation techniques to improve robustness, including signal filtering (noise reduction/injection), manipulation (time-pitch shifting), as well as techniques to augment the MFCC spectrogram (time-frequency masking). However, we did not observe significant improvements in model performance. We hypothesize that training data are already challenging in terms of ambient noise, reduced quantity, high variability in vocalization types, and temporal resolution.

5. Conclusions

This work represents an important step toward implementing adaptable, precise automated systems that significantly reduce human intervention in tasks like data labeling, annotation, and pre-processing in clinical research. Such advancements enable the application of computational techniques to extract quantitative features relevant to child and adolescent psychotherapy and clinical care, facilitating large-scale analyses and longitudinal monitoring. This addresses critical needs in clinical research, improving both the efficiency and impact of interventions (Lord et al., 2022; Lambert, 2017).

Ethics approval

This study was approved by the Research Ethics Board of the University of Trento (Protocol number: 2020–042) and complies with the principles laid down in the last version of the Declaration of Helsinki (2013).

Consent to participate

All participants were adequately informed about the objective of this research and signed an informed consent.

Consent for publication

All authors read and approved the final manuscript.

Funding

This research was supported by ERA Per Med Joint Transnational Call for Proposals (2021) for Multidisciplinary Research Projects on Personalized Medicine (grant. ID: 779282) - Development of Clinical Support Tools for Personalized Medicine Implementation for the project TECH-TOYS: Acquire digiTal biomarkErs in infanCy witH sensorized TOYS for early detection and monitoring of neurodevelopmental disorders (ERAPERMED2021-309).

CRediT authorship contribution statement

Giulio Bertamini: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Paola Venuti:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **David Cohen:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition. **Mohamed Chetouani:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition. **Cesare Furlanello:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of Competing Interest

This research does not involve any conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ridd.2024.104906.

Data availability

Raw source data can not be shared (personal biometric data of neurodivergent people). External data are publicly available. Source code and trained models will be released and open-source.

References

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics, 3*, 2007.

Adigwe, A., Tits, N., Haddad, K., Ostadabbas, S., Dutoit, T., 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems.

Ahn, Y., Moffitt, J., Tao, Y., Custode, S., Shyu, M., Perry, L., & Messinger, D. (2020). Objective measurement of social communication behaviors in children with suspected asd during the ados-2. *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 360–364. https://doi.org/10.1145/3395035.3425356

Albaum, C. S., Vashi, N., Bohr, Y., & Weiss, J. A. (2022). A systematic review of therapeutic process factors in mental health treatment for autistic youth. *Clinical Child and Family Psychology Review, 26*, 212–241. https://doi.org/10.1007/s10567-022-00409-0

Ali, S. N., Shuvo, S. B., Al-Manzo, M. I. S., Hasan, A., & Hasan, T. (2023). An end-to-end deep learning framework for real-time denoising of heart sounds for cardiac disease detection in unseen noise. *IEEE Access, 11*, 87887–87901. https://doi.org/10.1109/access.2023.3292551

Archinard, M., Haynal-Reymond, V., & Heller, M. (2000). Doctor's and patients' facial expressions and suicide reattempt risk assessment. *Journal of Psychiatric Research, 34*, 261–262. https://doi.org/10.1016/s0022-3956(00)00011-x

Atanbori, J., & Rose, S. (2022). MergedNET: A simple approach for one-shot learning in siamese networks based on similarity layers. *Neurocomputing, 509*, 1–10. https://doi.org/10.1016/j.neucom.2022.08.070

Banik, A., Arya, S., & Kant, A. (2015). Vocal parameters in children between 4 to 12 years of age: An attempt to establish a prototype database. *Intern J Scient Research Publications, 11*, 446–453.

Bansal, A., & Garg, N. K. (2022). Environmental sound classification: A descriptive review of the literature. *Intelligent Systems with Applications, 16*, Article 200115. https://doi.org/10.1016/j.iswa.2022.200115

Bickman, L. (2020). Improving mental health services: A 50-year journey from randomized experiments to artificial intelligence and precision mental health. *Administration and Policy in Mental Health and Mental Health Services Research, 47*, 795–843. https://doi.org/10.1007/s10488-020-01065-8

Bone, D., Lee, C. C., Black, M. P., Williams, M. E., Lee, S., Levitt, P., & Narayanan, S. (2014). The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *Journal of Speech, Language, and Hearing Research, 57*, 1162–1177. https://doi.org/10.1044/2014_jslhr-s-13-0062

Bourvis, N., Singer, M., Saint Georges, C., Bodeau, N., Chetouani, M., Cohen, D., & Feldman, R. (2018). Pre-linguistic infants employ complex communicative loops to engage mothers in social exchanges and repair interaction ruptures. *Royal Society Open Science, 5*, Article 170274. https://doi.org/10.1098/rsos.170274

Briend, F., David, C., Silleresi, S., Malvy, J., Ferré, S., & Latinus, M. (2023). Voice acoustics allow classifying autism spectrum disorder with high accuracy. *Translational Psychiatry, 13*. https://doi.org/10.1038/s41398-023-02554-8

Budaghyan, D., Onu, C.C., Gorin, A., Subakan, C., Precup, D., 2023.Cryceleb: A speaker verification dataset based on infant cry sounds. doi:10.48550/ARXIV.2305.00969.

Camaioni, L. (2017). The development of intentional communication: A re-analysis. In *New Perspectives in Early Communicative Development* (pp. 82–96). Routledge.

Chi, N. A., Washington, P., Kline, A., Husic, A., Hou, C., He, C., Dunlap, K., & Wall, D. P. (2022). Classifying autism from crowdsourced semistructured speech recordings: Machine learning model comparison study. *JMIR pediatrics and Parenting, 5*, Article e35406. https://doi.org/10.2196/35406

Cho, S., Liberman, M., Ryant, N., Cola, M., Schultz, R., & Parish-Morris, J. (2019). *Automatic Detection of Autism Spectrum Disorder in Children Using Acoustic and Text Features from Brief Natural Conversations*. In (pp. 2513–2517). Interspeech. https://doi.org/10.21437/Interspeech.2019-1452.

Cohen, D., Cassel, R., Saint-Georges, C., Mahdhaoui, A., Laznik, M., Apicella, F., & Chetouani, M. (2013). Do parentese prosody and fathers' involvement in interacting facilitate social interaction in infants who later develop autism? *Plos One, 8*, 61402. https://doi.org/10.1371/journal.pone.0061402

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development, 92*, 609–625. https://doi.org/10.1111/cdev.13511

Ebrahimpour, M., Schneider, S., Noelle, D., Kello, C., 2020. Infantnet: A deep neural network for analyzing infant vocalizations.

Eni, M., Dinstein, I., Ilan, M., Menashe, I., Meiri, G., & Zigel, Y. (2020). Estimating autism severity in young children from speech signals using a deep neural network. *IEEE Access, 8*, 139489–139500. https://doi.org/10.1109/access.2020.3012532

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM. https://doi.org/10.1145/2502081.2502224.

Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. (2021). A brief review of domain adaptation. In R. Stahlbock, G. M. Weiss, M. Abou-Nasr, C. Y. Yang, H. R. Arabnia, & L. Deligiannidis (Eds.), *Advances in Data Science and Information Engineering* (pp. 877–894). Cham: Springer International Publishing.

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. URL: https://www.sciencedirect.com/science/article/pii/0163638387900178 *Infant Behavior and Development, 10*, 279–293. https://doi.org/10.1016/0163-6383(87)90017-8.

Flemotomos, N., Martinez, V. R., Chen, Z., Singla, K., Ardulov, V., Peri, R., Caperton, D. D., Gibson, J., Tanana, M. J., Georgiou, P., Epps, J. V., Lord, S. P., Hirsch, T., Imel, Z. E., Atkins, D. C., & Narayanan, S. (2021). Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods, 54*, 690–711. https://doi.org/10.3758/s13428-021-01623-4

Fouquet, M., Pisanski, K., Mathevon, N., & Reby, D. (2016). Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood. *Royal Society Open Science, 3*, Article 160395. https://doi.org/10.1098/rsos.160395

Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D., & Gaigg, S. (2017). Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis. *Autism Research, 10*, 384–407.

Godel, M., Robain, F., Journal, F., Kojovic, N., Latrèche, K., Dehaene-Lambertz, G., & Schaer, M. (2023). Prosodic signatures of ASD severity and developmental delay in preschoolers. *npj Digital Medicine, 6*. https://doi.org/10.1038/s41746-023-00845-4

Green, J., & Garg, S. (2018). Annual Research Review: The state of autism intervention science: progress, target psychological and biological mechanisms and future prospects. *Journal of Child Psychology and Psychiatry, 59*, 424–443. https://doi.org/10.1111/jcpp.12892

Gupta, R., Bone, D., Lee, S., & Narayanan, S. (2016). Analysis of engagement behavior in children during dyadic interactions using prosodic cues. *Computer Speech and Language, 37*, 47–66. https://doi.org/10.1016/j.csl.2015.09.003

Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification.

Jones, R. M., Plesa Skwerer, D., Pawar, R., Hamo, A., Carberry, C., Ajodan, E. L., Caulley, D., Silverman, M. R., McAdoo, S., Meyer, S., Yoder, A., Clements, M., Lord, C., & Tager-Flusberg, H. (2019). How effective is lena in detecting speech vocalizations and language produced by children and adolescents with asd in different contexts? *Autism Research, 12*, 628–635. https://doi.org/10.1002/aur.2071

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., & Belpaeme, T. (2017). Child speech recognition in human-robot interaction: Evaluations and recommendations. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 82–90. https://doi.org/10.1145/2909824.3020229

Kent, R. D., Eichhorn, J. T., & Vorperian, H. K. (2021). Acoustic parameters of voice in typically developing children ages 4-19 years. *International Journal of Pediatric Otorhinolaryngology, 142*, Article 110614. https://doi.org/10.1016/j.ijporl.2021.110614

Lahiri, R., Feng, T., Hebbar, R., Lord, C., Kim, S.H., Narayanan, S., 2023. Robust self supervised speech embeddings for child-adult classification in interactions involving children with autism. doi:10.48550/ARXIV.2307.16398.

Lahiri, R., Nasir, M., Kumar, M., Kim, S., Bishop, S., Lord, C., & Narayanan, S. (2022). Interpersonal synchrony across vocal and lexical modalities in interactions involving children with autism spectrum disorder. *JASA Express Letters, 2*, Article 095202. https://doi.org/10.1121/10.0013421

Lambert, M. J. (2017). Maximizing psychotherapy outcome beyond evidence-based medicine. *Psychotherapy and Psychosomatics, 86*, 80–89. https://doi.org/10.1159/000455170

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy, 55*, 520–537. https://doi.org/10.1037/pst0000167

Leclère, C., Viaux, S., Avril, M., Achard, C., Chetouani, M., Missonnier, S., & Cohen, D. (2014). Why synchrony matters during mother-child interactions: A systematic review. *PLoS ONE, 9*, Article e113571. https://doi.org/10.1371/journal.pone.0113571

Lehet, M., Arjmandi, M. K., Houston, D., & Dilley, L. (2021). Circumspection in using automated measures: Talker gender and addressee affect error rates for adult speech detection in the language environment analysis (lena) system. *Behavior Research Methods, 53*, 113–138. https://doi.org/10.3758/s13428-020-01419-y

Li, J., Bhat, A., Barmaki, R., 2022. Dyadic Movement Synchrony Estimation Under Privacy-preserving Conditions.

Li, J., Hasegawa-Johnson, M., & McElwain, N. (2021). Analysis of acoustic and voice quality features for the classification of infant and mother vocalizations. *Speech Communication, 133*, 41–61. https://doi.org/10.1016/j.specom.2021.07.010

Li, M., Tang, D., Zeng, J., Zhou, T., Zhu, H., Chen, B., & Zou, X. (2019). An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder. *Computer Speech and Language, 56*, 80–94. https://doi.org/10.1016/j.csl.2018.11.002

Lord, C., Charman, T., Havdahl, A., Carbone, P., Anagnostou, E., Boyd, B., Carr, T., de Vries, P. J., Dissanayake, C., Divan, G., Freitag, C. M., Gotelli, M. M., Kasari, C., Knapp, M., Mundy, P., Plank, A., Scahill, L., Servili, C., Shattuck, P., Simonoff, E., Singer, A. T., Slonims, V., Wang, P. P., Ysrraelit, M. C., Jellett, R., Pickles, A., Cusack, J., Howlin, P., Szatmari, P., Holbrook, A., Toolan, C., & McCauley, J. B. (2022). The lancet commission on the future of care and clinical research in autism. *The Lancet, 399*, 271–334. https://doi.org/10.1016/s0140-6736(21)01541-5

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology, 5*, 96–116. https://doi.org/10.1002/lio2.354

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A. K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the trier treatment navigator (ttn). *Behaviour Research and Therapy, 120*, Article 103438. https://doi.org/10.1016/j.brat.2019.103438

Mahdhaoui, A., Chetouani, M., Zong, C., Cassel, R. S., Saint-Georges, C., Laznik, M. C., Maestro, S., Apicella, F., Muratori, F., & Cohen, D. (2009). *Automatic motherese detection for face-to-face interaction analysis* (pp. 248–255). Springer.

Markova, D., Richer, L., Pangelinan, M., Schwartz, D. H., Leonard, G., Perron, M., Pike, G., Veillette, S., Chakravarty, M. M., Pausova, Z., & Paus, T. (2016). Age- and sex-related variations in vocal-tract morphology and voice acoustics during adolescence. *Hormones and Behavior, 81*, 84–96. https://doi.org/10.1016/j.yhbeh.2016.03.001

Marschik, P. B., Widmann, C. A. A., Lang, S., Kulvicius, T., Boterberg, S., Nielsen-Saines, K., Bölte, S., Esposito, G., Nordahl-Hansen, A., Roeyers, H., Wörgötter, F., Einspieler, C., Poustka, L., & Zhang, D. (2022). Emerging verbal functions in early infancy: Lessons from observational and computational approaches on typical development and neurodevelopmental disorders. *Advances in Neurodevelopmental Disorders, 6*, 369–388. https://doi.org/10.1007/s41252-022-00300-7

Mena, C., Borsky, M., Mollberg, D.E., Guðmundsson, S.F., Hedström, S., Pálsson, R., Jónsson, Ó.H., Þorsteinsdóttir, S., Guðmundsdóttir, J.V., Magnúsdóttir, E.H., Þórhallsdóttir, R., Gudnason, J., 2021. Samromur children 21.09. URL: ⟨http://hdl.handle.net/20.500.12537/185⟩. CLARIN-IS.

Messinger, D. M., Ruvolo, P., Ekas, N. V., & Fogel, A. (2010). Applying machine learning to infant interaction: The development is in the details. URL: https://www.sciencedirect.com/science/article/pii/S0893608010001590 *Neural Networks, 23*, 1004–1016. https://doi.org/10.1016/j.neunet.2010.08.008.

Miner, A. S., Haque, A., Fries, J. A., Fleming, S. L., Wilfley, D. E., Wilson, G. T., Milstein, A., Jurafsky, D., Arnow, B. A., Agras, W. S., Fei-Fei, L., & Shah, N. H. (2020). Assessing the accuracy of automatic speech recognition for psychotherapy. *npj Digital Medicine, 3*. https://doi.org/10.1038/s41746-020-0285-8

Moffitt, J., Ahn, Y., Custode, S., Tao, Y., Mathew, E., Parlade, M., & Messinger, D. (2022). Objective measurement of vocalizations in the assessment of autism spectrum disorder symptoms in preschool age children. *Autism Research*. https://doi.org/10.1002/aur.2731

Mössler, K., Gold, C., Aßmus, J., Schumacher, K., Calvet, C., Reimer, S., & Schmid, W. (2019). The therapeutic relationship as predictor of change in music therapy with young children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 49*, 2795–2809. https://doi.org/10.1007/s10803-017-3306-y

Nicollas, R., Garrel, R., Ouaknine, M., Giovanni, A., Nazarian, B., & Triglia, J. M. (2008). Normal voice in children between 6 and 12 years of age: Database and nonlinear analysis. *Journal of Voice, 22*, 671–675. https://doi.org/10.1016/j.jvoice.2007.01.009

Norcross, J. C., & Lambert, M. J. (2018). Psychotherapy relationships that work III. *Psychotherapy, 55*, 303–315. https://doi.org/10.1037/pst0000193

Ochi, K., Ono, N., Owada, K., Kojima, M., Kuroda, M., Sagayama, S., & Yamasue, H. (2019). Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder. *PloS One, 14*, 0225377. https://doi.org/10.1371/journal.pone.0225377

Ouss, L., Palestra, G., Saint-Georges, C., Leitgel Gille, M., Afshar, M., Pellerin, H., & Cohen, D. (2020). Behavior and interaction imaging at 9 months of age predict autism/intellectual disability in high-risk infants with West syndrome. *Translational Psychiatry, 10*, 1–7. https://doi.org/10.1038/s41398-020-0743-8

Paul, R., Fuerst, Y., Ramsay, G., Chawarska, K., & Klin, A. (2010). Out of the mouths of babes: Vocal production in infant siblings of children with asd: Vocalizations in infant siblings. *Journal of Child Psychology and Psychiatry, 52*, 588–598. https://doi.org/10.1111/j.1469-7610.2010.02332.x

Perochon, S., Di Martino, J. M., Carpenter, K. L. H., Compton, S., Davis, N., Eichner, B., Espinosa, S., Franz, L., Krishnappa Babu, P. R., Sapiro, G., & Dawson, G. (2023). Early detection of autism using digital behavioral phenotyping. *Nature Medicine, 29*, 2489–2497. https://doi.org/10.1038/s41591-023-02574-3

Pokorny, F. B., Peharz, R., Roth, W., Zöhrer, M., Pernkopf, F., Marschik, P. B., & Schuller, B. (2016). Manual versus automated: The challenging routine of infant vocalisation segmentation in home videos to study neuro(mal)development. *in: Interspeech 2016*. ISCA. https://doi.org/10.21437/interspeech.2016-1341

Romeo, R. R., Leonard, J. A., Grotzinger, H. M., Robinson, S. T., Takada, M. E., Mackey, A. P., Scherer, E., Rowe, M. L., West, M. R., & Gabrieli, J. D. E. (2021). Neuroplasticity associated with changes in conversational turn-taking following a family-based intervention. URL: https://www.sciencedirect.com/science/article/pii/S187892932100058X *Developmental Cognitive Neuroscience, 49*, Article 100967. https://doi.org/10.1016/j.dcn.2021.100967.

Rybner, A., Jessen, E. T., Mortensen, M. D., Larsen, S. N., Grossman, R., Bilenberg, N., Cantio, C., Jepsen, J. R. M., Weed, E., Simonsen, A., & Fusaroli, R. (2022). Vocal markers of autism: Assessing the generalizability of machine learning models. *Autism Research, 15*, 1018–1030. https://doi.org/10.1002/aur.2721

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., Laznik, M. C., & Cohen, D. (2013). Motherese in interaction: at the cross-road of emotion and cognition?(a systematic review). *PloS One, 8*, Article e78103.

Saint-Georges, C., Mahdhaoui, A., Chetouani, M., Cassel, R., Laznik, M., Apicella, F., & Cohen, D. (2011). Do parents recognize autistic deviant behavior long before diagnosis? Taking into account interaction using computational methods. *PloS One, 6*, 22393. https://doi.org/10.1371/journal.pone.0022393

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine, 49*, 1426–1448. https://doi.org/10.1017/s0033291719000151

Shorfuzzaman, M., & Hossain, M. (2021). MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. *Pattern Recognition, 113*, Article 107700. https://doi.org/10.1016/j.patcog.2020.107700

Shute, B., & Wheldall, K. (1999). Fundamental frequency and temporal modifications in the speech of british fathers to their children. *Educational Psychology, 19*, 221–233. https://doi.org/10.1080/0144341990190208

Taubner, S., Ioannou, Y., Saliba, A., Sales, C. M. D., Volkert, J., Protić, S., Adler, A., Barkauskiene, R., Conejo-Cerón, S., Di Giacomo, D., Mestre, J. M., Moreno-Peral, P., Vieira, F. M., Mota, C. P., Henriques, M. I. R. S., Røssberg, J. I., Perdih, T. S., Schmidt, S. J., Zettl, M., Ulberg, R., & Heinonen, E. (2023). Mediators of outcome in adolescent psychotherapy and their implications for theories and mechanisms of change: a systematic review. *European Child and Adolescent Psychiatry*. https://doi.org/10.1007/s00787-023-02186-9

Vivanti, G., Bottema-Beutel, K., & Turner-Brown, L. (2020). *Clinical guide to early interventions for children with autism*. Springer. https://doi.org/10.1007/978-3-030-41160-2

Vivanti, G., & Zhong, H. (2020). Naturalistic Developmental Behavioral Interventions for Children with Autism. In G. Vivanti, K. Bottema-Beutel, & L. Turner-Brown (Eds.), *Clinical Guide to Early Interventions for Children with Autism* (pp. 93–130). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-41160-2_6.

Wang, J., & Zhai, Y. (2020). Prototypical siamese networks for few-shot learning. In *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication* (pp. 178–181). ICEIEC, IEEE. https://doi.org/10.1109/ICEIEC49280.2020.9152261.

Warren-Leubecker, A., & Bohannon, J. N. (1984). Intonation patterns in child-directed speech: Mother-father differences. *Child Development, 55*, 1379–1385. https://doi.org/10.2307/1130007. ⟨http://www.jstor.org/stable/1130007⟩

Washington, P., & Wall, D. P. (2023). A review of and roadmap for data science and machine learning for the neuropsychiatric phenotype of autism. *Annual Review of Biomedical Data Science, 6*, 211–228. https://doi.org/10.1146/annurev-biodatasci-020722-125454

Weiste, E., & Peräkylä, A. (2014). Prosody and empathic communication in psychotherapy interaction. *Psychotherapy Research, 24*, 687–701. https://doi.org/10.1080/10503307.2013.879619

Xu, A., Huang, K., Feng, T., Tager-Flusberg, H., Narayanan, S., 2023.Audio-visual child-adult speaker classification in dyadic interactions. doi:10.48550/ARXIV.2310.01867.

Zilcha-Mano, S. (2017). Is the alliance really therapeutic? revisiting this question in light of recent methodological advances. *American Psychologist, 72*, 311–325. https://doi.org/10.1037/a0040435

Zilcha-Mano, S. (2018). Major developments in methods addressing for whom psychotherapy may work and why. *Psychotherapy Research, 29*, 693–708. https://doi.org/10.1080/10503307.2018.1429691