**Chair of Health Informatics**

# Clinical Application Project: Acoustic and Conversational Markers of ADHD in Parent-Child Dialogue

Winter Semester 2025

**Moritz Lenhard, Konstantin Starke**

Supervisor: Andreas Triantafyllopoulos

September 1, 2025

# 1 Intoduction

Attention-Deficit/Hyperactivity Disorder (ADHD) is a neurodevelopmental condition characterized by symptoms such as inattention, hyperactivity, and impulsivity. Beyond its cognitive and behavioral manifestations, ADHD can also influence how children interact with others in social settings. For instance, children with ADHD may interrupt conversations more frequently, have difficulty waiting their turn, or struggle to modulate the volume and timing of their speech. These differences in social communication can impact interactions with parents, peers, and teachers, making everyday conversations more challenging.

The goal of this clinical application project is to analyze parent–child interactions to determine ADHD–control group differences by examining two domains: conversational aspects (e.g., how long each participant speaks) and acoustic aspects (e.g., the speaker's loudness).

This clinical application project comprised three stages: (1) annotating Jenga parent–child interactions to mark who spoke when and for how long; (2) reviewing key conversational and acoustic features and analyzing them with Python tools (e.g. Pandas, Seaborn, scikit-learn); and (3) summarizing the results this report.

# 2 Background and Related Work

Previous research has mainly focused on analyzing ADHD through speech using methods such as the FMSS (Five-Minute Speech Sample) and the PFMSS (Preschool FMSS), which has been employed to assess expressed emotion and related communicative patterns. [1, 2, 3]

Prior clinical voice work found ADHD-related perceptual differences (louder, hoarser, more strained or breathy voices) and a lower sustained-vowel F0, with few other acoustic effects in a small sample [4]; their recordings were sustained vowels produced in clinic (holding an /a/ sound), whereas ours are spontaneous parent–child dialogue.

A new approach was suggested by Spiesberger et al. to use paralinguistic and linguistic features to predict ADHD diagnosis. In their paper, they focus on the parents speech samples. [5]

# 3 Setup of Experiment

The dataset used in this analysis consists of 115 combined video and audio recordings of conversations between parents and their children. 60 of these recordings were placed into the experimental group (child has ADHD), the other 55 were placed into the control group (child has no ADHD). Contents of the recordings are the interactions of the parent and child while playing the game "Jenga".

The dataset was further processed by annotating the time a person started speaking and also for how long. In this context only speaker segments that contributed to the interaction between parent and child were annotated. This means interactions with instructors or phone calls were disregarded.

# 4 Conversational aspects

## 4.1 Objective

The first part of the analysis encompasses basic conversational aspects, meaning quantifying interaction dynamics between participants based on speech timing data (start time and duration of a segment). This involves identifying and analyzing patterns in turn-taking, pauses, overlaps, interruptions etc. to better understand communication behavior.

## 4.2 Methodology

The analysis is based on the speech timing data from which we derive additional conversational features. For each feature, the effect size and p-value between the ADHD group and the control group are calculated for each participant (parent, child) using Cohen's d and the Mann-Whitney U test respectively. The following paragraphs will give an overview of these features and explain what they mean and how they were computed. After that, the most relevant features will be presented.

**Segment count** represent the number of times a person started to verbally communicate like speaking or laughing. This feature was already given by the segmented dataset.

**Duration** is the duration of time of a given segment during the conversation. This feature was also already given by the segmented dataset.

**Pause duration** is the total duration of time a person did not communicate during the conversation. This was calculated by subtracting the start time from the n-th segment from the end time of the segment n-1 of the same person. If a segment did not have a previous segment, it was disregarded. This means that the time until a person first spoke was not counted as a pause.

**Speaker overlap** splits into two features: overlap count and overlap duration. Overlap count is the amount of times a person started to communicate while another person was already doing so. Overlap duration is the total duration of time a person overlapped over the other person. Note that overlap duration is only attributed to the person who starts the overlap as this distinction helps identify patterns of conversational assertiveness.

**Interruptions** build upon the overlap data. Interruptions differ from overlap by acknowledging how the person who was already communicating reacted to the overlap. We identified three main categories for interruptions: *No interruption attempt*, *failed interruption* and *successful interruption*. To classify these three categories, we use a heuristic.
To disregard backchanneling and instances where both participants start communicating at the same time, segments with a length less then one second and overlaps where the start time of both segments are within half a second to each other, are marked as a no interruption attempt.
If the interrupter stops speaking before the interrupted and the interrupted does not stop speaking within one second after the overlap, the segment is marked as a failed interruption.
Overlaps that are neither in one of the other two other interruption categories are marked as a successful interruption.
In combination with the overlap duration, the duration of interruption time for each interruption type are also calculated.

**Conversational flow** is the analysis how frequently the conversational spotlight switches between participants or how long it stays on one person. For this two new segment classifications are introduced: Switch and repeat. Segments that are shorter than one second (backchanneling) or marked as a failed interruption are ignored as they do not lead to change in the conversational focus. Segments that are marked as successful interruptions or no attempted interruptions where the speaker changed compared to the previous segment are marked as switches. If the speaker is the same as the previous segment, it is marked as a repeat.
From this data, the amount of consecutive repeats and switches can be calculated by adding up consecutive segments which are marked as a repeat or switch respectively.

## 4.3 Limitations

The heuristic nature of the interruption classification means that slight changes in the parameter values can change the outcome drastically. We picked the mentioned parameters as natural classification criteria for interruptions and backchanneling. A better approach would be to use the semantic meaning of what was spoken for a more robust classification.
It is also important to mention that the feature values are not normalized to the length of the conversations which could lead to misrepresentation of the data. For example a shorter conversation probably has a smaller segment count compared to longer conversations but is treated equally in the analysis.

## 4.4 Key Findings.

After calculating the Cohen's d and p-value for every feature, none of the features reached statistical significance (all p-values were above 0.05) and most features have just a minor effect size. However, to give an overview of the results, the features with the most promising differences in effect size are presented. A comprehensive overview of features and their corresponding effect sizes and p-values can be seen in Table 1 and Table 2.

| Feature name | Cohen's d | p-value |
|---|---|---|
| Segment count | 0.19 | 0.51 |
| Segment duration | 0.07 | 0.83 |
| Overlap count | 0.25 | 0.83 |
| Overlap duration | 0.18 | 0.96 |
| Pause duration | -0.04 | 0.39 |
| N interruption count | 0.18 | 0.56 |
| N interruption duration | 0.13 | 1.0 |
| S interruption count | 0.14 | 0.62 |
| S interruption duration | 0.27 | 0.66 |
| F interruption count | 0.31 | 0.68 |
| F interruption duration | 0.33 | 0.78 |
| Repeat count | 0.30 | 0.19 |
| Switch count | -0.22 | 0.24 |

Table 1: Effect sizes and p-values for conversational features for children.

| Feature name | Cohen's d | p-value |
|---|---|---|
| Segment count | -0.14 | 0.37 |
| Segment duration | -0.17 | 0.25 |
| Overlap count | 0.19 | 0.92 |
| Overlap duration | 0.33 | 0.45 |
| Pause duration | -0.06 | 0.53 |
| N interruption count | -0.17 | 0.27 |
| N interruption duration | 0.21 | 0.88 |
| S interruption count | 0.39 | 0.39 |
| S interruption duration | 0.51 | 0.17 |
| F interruption count | 0.30 | 0.15 |
| F interruption duration | 0.41 | 0.20 |
| Repeat count | 0.09 | 0.88 |
| Switch count | -0.22 | 0.26 |

Table 2: Effect sizes and p-values for conversational features in parents.

Overlap duration is visibly longer in parents in the ADHD group compared to the control group with a Cohen's d value of 0.33 (Figure 1). Even though it is only a small difference, the trend can be interpreted as parents tending to talk over the child more often or longer.

The duration of successful interruption underlines the previous statistic. With an effect size of 0.51 it has a more noticeable difference. In general, these previous two findings may indicate that parents of children with ADHD are more inclined to talk over and interrupt the

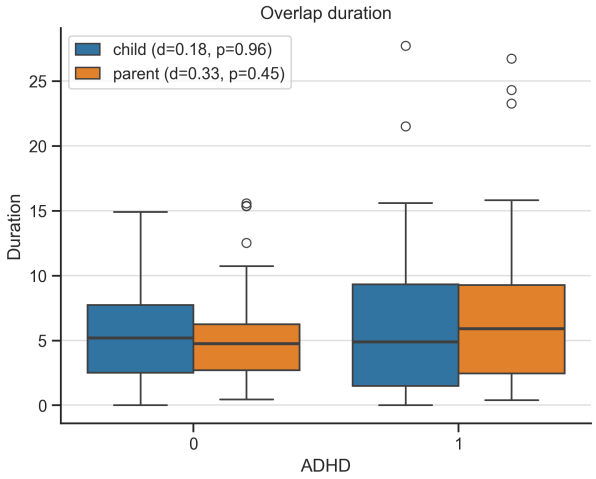


Figure 1: Distribution of overlap durations per participant.

child.

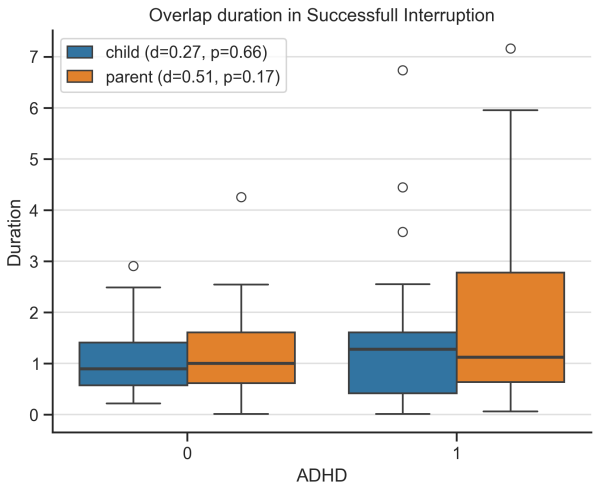


Figure 2: Distribution of successful interruption overlap durations per participant.

The repeat count is a features that tends to be greater for children with ADHD compared to the control group with a Cohen's d of 0.3.

# 5 Acoustic Aspects

We analyze segment-level acoustics using the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [6], a standardized set of 88 low-level descriptors covering pitch (F0), loudness, spectral shape, and voice quality. Our working table is at the *segment* level - each sample is one speaking segment with start/end time, duration, and role (child/parent) from the Jenga interactions. Diagnostic labels exist at the *participant* level (*id*) and are merged onto all rows; consequently, observations

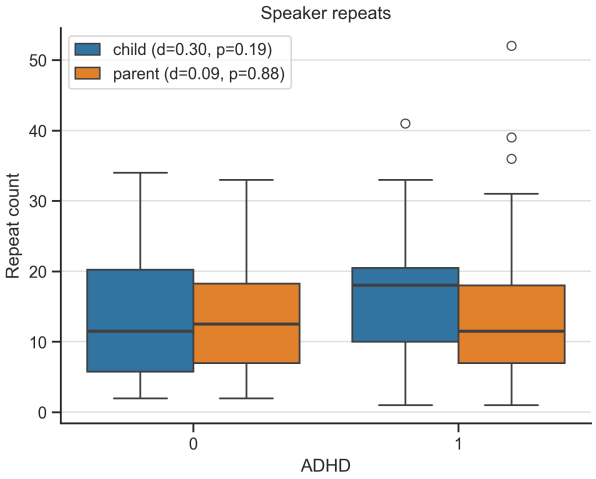


Figure 3: Distribution of repeats per participant.

from the same *id* are dependent. All statistical tests and models therefore respect this per-id structure (e.g., by aggregating to per-id summaries or using group-aware validation). The focus here is not feature extraction itself, but whether these established descriptors capture ADHD-control differences in our data.

## 5.1 Group Differences

**Objectives.** We test whether per-participant acoustic features differ between children with ADHD and controls. We want to identify acoustic features that significantly differ between the two groups. We quantify discriminative strength via effect sizes.

**Methodology.** To examine group differences between children with ADHD and controls, we used the acoustic features extracted with the eGeMAPS. Since the diagnostic label is defined at the participant level, we aggregate segment-level features to per-participant medians, computed separately for child and parent speech.

For each feature, we performed Mann–Whitney U (MWU) tests to compare ADHD and control groups, using an alpha level of 0.05. We use Cohen's d as a measure of effect size, focusing on the features with the strongest discriminative power. By convention we define

$$d = (\mu_{\text{ADHD}} - \mu_{\text{Control}})/s_{\text{pooled}} \quad (1)$$

so $d < 0$ indicates lower ADHD means. We assume the feature values are normally distributed.

**Key Findings.** Among children, the discriminative signal is concentrated: 9 of 88 eGeMAPS features exhibit medium-to-large group differences ($|d| > 0.5$), and 8 of these survive significance testing ($p < 0.05$).

These features cluster into three interpretable groups: (i) *intonation dynamics* - flatter F0 contours (lower mean rising/falling slopes), reduced short-term variability, and a narrower pitch range; (ii) *phonatory stability* - lower

| Feature name | Cohen's d | p-value |
|---|---|---|
| HNRdBACF_sma3nz_stddev Norm | -0.85 | <0.001 |
| jitterLocal_sma3nz_amean | -0.83 | <0.001 |
| VoicedSegmentsPerSec | -0.83 | <0.001 |
| F0semitoneFrom27.5Hz_ sma3nz_meanRisingSlope | -0.71 | 0.004 |
| F0semitoneFrom27.5Hz_ sma3nz_stddevNorm | -0.66 | 0.007 |
| F0semitoneFrom27.5Hz_ sma3nz_pctlrange0-2 | -0.60 | 0.029 |
| F0semitoneFrom27.5Hz_ sma3nz_meanFallingSlope | -0.60 | 0.012 |
| F0semitoneFrom27.5Hz_ sma3nz_stddevFallingSlope | -0.56 | 0.015 |

Note: $d = (\mu_{\text{ADHD}} - \mu_{\text{Control}})/s_{\text{pooled}}$; negative values indicate lower ADHD means.

Table 3: Effect sizes and p-values for acoustic features for children with medium to large effects ($|d| > 0.5$, $p < 0.05$)

jitter and lower variability of Harmonics-to-Noise Ratio (HNR), and (iii) *temporal organization* - fewer voiced segments per second, suggestive of longer continuous voiced stretches.

## 5.2 Temporal Dynamics

**Hypothesis.** As the child–parent interaction unfolds, acoustic differences between children with ADHD and controls become more pronounced. Concretely, the absolute group effect size $|d|$ for eGeMAPS features increases from early to late normalized-time bins, with the largest divergences expected in the final third of the conversation.

**Methodology.** For each participant, conversational time was min–max normalized to $[0, 1]$ using segment starts, $\tilde{t} = (t - t_{\min})/(t_{\max} - t_{\min})$. The normalized time was partitioned into $n = 3$ equal-width bins. Within each bin and for each acoustic feature, we aggregated all segments from one participant by the median to obtain one participant-level summary . For every feature and bin, we compared ADHD and control groups at the participant level using Cohen's $d$ (Eq. 1).

**Key Findings.** Across participant-level medians (per-id, within-bin), the magnitude of group differences tends to increase over time for *children*, while remaining smaller and flatter for *parents*. In Figure 4, the child distribution shifts upward from Bin 1 to Bin 3, with more high-value outliers in the final bin; the parent distributions are consistently lower and show only a mild increase, if any. This pattern is descriptive support for the hypothesis that ADHD–control differences amplify as the interaction unfolds, especially in child speech.

Figure 7 also reinforces this picture: it shows the per-bin top-10 features by $|d|$, as well as the mean $|d|$ top-10 features, which for children climbs $0.44 \rightarrow 0.63 \rightarrow 0.68$ across bins, while parents stay lower and uneven ($0.34 \rightarrow 0.18 \rightarrow 0.29$). In Bin3, the most prominent features cluster into loudness descriptors, alongside F0 dynamics/variability and voice-quality/temporal organization markers. This late-bin profile echoes the findings of the Section 5.1 - where jitter, HNR variability, VoicedSegmentsPerSec, and F0 dynamics showed medium-to-large effects, while additionally highlighting a *loudness* group that becomes more discriminative toward the end of the interaction despite not ranking among the strongest global effects earlier.

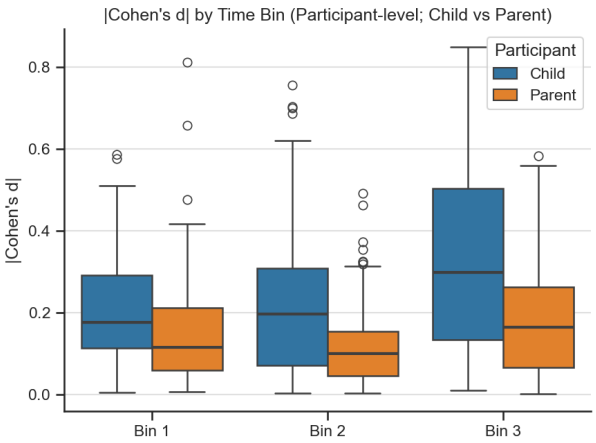


Figure 4: Distribution of eGeMAPS effect sizes per time bin

**Limitations.** The temporal pattern is descriptive; no formal test of a diagnosis×time effect has been performed, so the observed increase should be interpreted cautiously.

Min–max normalization of time to $[0, 1]$ per participant removes absolute time: recordings differ substantially in length (mean $\approx 387$ s; some $\approx 111$ s), so equal-width normalized bins correspond to very different absolute durations (with three bins: $\approx 129$ s vs. $\approx 37$ s per bin), confounding "late in the interaction" with "later in absolute minutes."

Binning by *segment start time* ignores segment duration; many short bursts are weighted the same as long monologues.

Finally, we aggregate to per-id *medians*. While this controls pseudoreplication, it discards within-person variability and makes conclusions sensitive to the chosen summary statistic.

## 5.3 Predictive Modeling

**Task and Inputs.** We formulate ADHD diagnosis as a binary classification problem at the *participant* level. The inputs are acoustic descriptors from the eGeMAPS

set. All numeric eGeMAPS features are used as predictors after removing meta-columns (id, participant label, timestamps), with no hand-crafted aggregation or scaling applied. To respect subject-level labels and prevent leakage, cross-validation is *group-aware* by participant ID.

**Evaluation & Tuning.** We use nested cross-validation to obtain unbiased performance estimates and tune the Random Forest. The *outer* loop is a 5-fold *StratifiedGroupKFold* split by participant ID, so all segments from a participant appear in exactly one fold, preventing leakage. Inside each outer training split, a *3-fold StratifiedGroupKFold* grid search selects hyperparameters using F1 as the scoring metric. The grid spans standard controls: *n_estimators*, *max_depth*, *min_samples_split*, *min_samples_leaf*, *max_features*, and *class_weight*. After inner selection, the model is refit on the outer-train fold and evaluated on the held-out fold. Finally, we refit on all data using fold-wise consensus hyperparameters.

**Model.** We employ scikit-learn's *RandomForestClassifier* (RF) as a strong nonparametric baseline. RFs capture nonlinear decision boundaries and higher-order feature interactions via ensembles of decision trees, while remaining robust to monotone transformations and feature scaling - well suited to correlated eGeMAPS descriptors without additional preprocessing. Trees are grown on bootstrap samples and split on feature subsamples, and classification is obtained by majority vote across trees. Although class sizes are only mildly imbalanced (60 ADHD vs. 55 control), we include class weighting in the tuning grid; the final configuration selected `class_weight='balanced'`. This provides a competitive, leakage-safe baseline that minimizes assumptions about feature distributions and keeps the pipeline simple.

**Metrics & Reporting.** Performance is reported as F1 on the *outer* folds, with the mean $\pm$ standard deviation as the primary summary. We also report the train–test *generalization gap* (mean train F1 minus mean test F1) to quantify overfitting. Across the 5 outer folds, mean train F1 was $0.796$ and mean test F1 was $0.644$ (gap $= 0.152$), with high between-fold variability on test (std $\approx 0.193$).

| Fold | Train F1 | Test F1 | Gap |
|---|---|---|---|
| 1 | 0.862 | 0.762 | 0.100 |
| 2 | 0.713 | 0.400 | 0.313 |
| 3 | 0.792 | 0.952 | -0.160 |
| 4 | 0.871 | 0.571 | 0.300 |
| 5 | 0.740 | 0.533 | 0.207 |
| **Mean** | **0.796** | **0.644** | **0.152** |
| **Std** | **0.063** | **0.193** | — |

Table 4: Outer-fold F1 scores and train–test gaps under nested, group-aware CV.

To complement the F1 metric, we show a precision–recall (PR) curve (Figure 5); it lies only modestly above the prevalence baseline (AP = 0.595) and peaks at $F_1 = 0.699$ near recall $\sim 0.75$ at threshold $\sim 0.49$, indicating some discriminative signal but limited headroom.

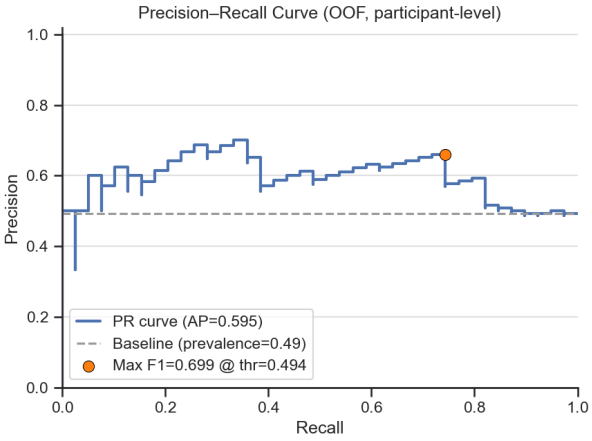


Figure 5: Precision–Recall curve for the final model

**Results summary.** Test F1 varied widely across outer folds ($0.400$–$0.952$), with the extremes (Fold 2 at $0.400$, Fold 3 at $0.952$) underscoring sensitivity to fold composition; the overall mean test F1 was $0.644$ (std $\approx 0.193$). The PR curve shows limited headroom (AP$= 0.595$; best F1$\approx 0.70$ in a narrow region), suggesting only modest discriminative signal and threshold sensitivity rather than an easy win from calibration. Inner-loop tuning converged on a conservatively regularized Random Forest: *class_weight='balanced'*, *max_depth=5*, *max_features='log2'*, *min_samples_leaf=7*, *min_samples_split=4*, *n_estimators=100*.

**Model interpretability (SHAP).** We interpret the fitted RF with SHAP [7], using *TreeExplainer* on the final refit model for the ADHD class. Figure 6 shows their signed effects. The top contributors were loudness descriptors, spectral change/tilt , MFCCs, and mean F0. Higher loudness percentiles and higher equivalent sound level generally increase the ADHD score, whereas lower loudness tends to decrease it; spectral–flux and spectral–slope features exhibit mixed signed impacts across samples. This profile aligns with the Section 5.2, where loudness became more discriminative late in the interaction, while micro-perturbation markers with group-level effects (jitter, HNR variability) are not among the model's strongest drivers.

**Limitations & robustness.** The sample is small, which inflates uncertainty and likely contributes to the large fold-to-fold spread in test F1 ($0.400$–$0.952$) and the nontrivial train–test gap. The PR curve's modest headroom (AP$= 0.595$) and jagged shape point to threshold sensitivity and small-$N$ effects. Correlations among

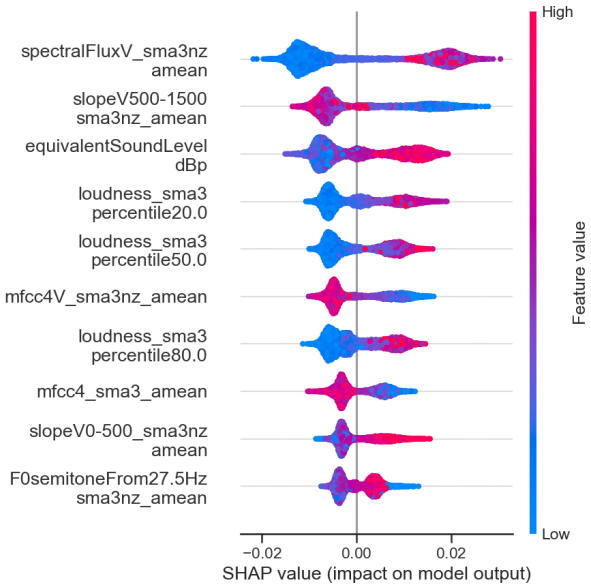


Figure 6: Beeswarm plot for top 10 features

eGeMAPS descriptors mean that importance and decision logic can spread across near-redundant features; SHAP rankings should therefore be read as feature *families* rather than single markers. Finally, results reflect one refit with consensus hyperparameters; both performance and attributions can vary with random seeds and comparable settings, so the reported numbers should be viewed with appropriate caution.

# 6 Results and Discussion

The conversational analysis did not reveal significant differences in the extracted features. This may be partly due to the limitations of heuristic analysis, which could be improved by creating a transcript, as well as the absence of normalization procedures that could have enhanced comparability across samples. Finally, the most pronounced effect size differences emerged primarily on the parents' side, suggesting that parental speech characteristics may play a more substantial role in distinguishing ADHD-related patterns than those of the children.

Effects of acoustics concentrate in *children*: several eGeMAPS features show medium–large group differences (5.1, Table 3), led by altered *F0 dynamics* and *phonatory stability* (flatter contours; lower jitter/HNR variability) and fewer voiced segments per second.

Across normalized time, child-level group differences increase from early to late bins while parent effects remain smaller (Fig. 4); in the final bin, *loudness* descriptors become more discriminative alongside F0 dynamics/voice-quality markers (Fig. 7), a descriptive pattern that warrants formal diagnosis×time testing (5.2).

A group-aware Random Forest shows modest but variable performance (mean test $F1 = 0.644 \pm 0.193$; Table 4); the PR curve sits only slightly above prevalence (AP$= 0.595$) (Fig. 5), and SHAP highlights loudness and spectral descriptors over HNR variability despite HNR's large group difference (Fig. 6).

A potential limitation of the dataset is the background noise from the Jenga bricks during the experimental task, which may interfere with the recordings and introduce artifacts. Such noise can mask or distort relevant speech features, thereby reducing the accuracy and reliability of the analysis. This was not looked into in this report but was noticed during the segmentation part of the clinical research project.

**Future research.** Formally test the temporal dynamics hypothesis of diagnosis×time interaction(5.2). Re-run the analysis on absolute-time bins and/or include total recording length as a covariate to separate relative from absolute time effects and compare alternative aggregators to assess stability. These insights could also inform subsequent predictive analyses.

# Disclaimer on the use of AI

LLMs were used to formulate or rewrite parts of this document. It was not used to generate information about the given topic. Generated text was examined by the authors for correctness.

# References

[1] E. Perez, M. Turner, A. Fisher, J. Lockwood, and D. Daley, "Linguistic analysis of the preschool five minute speech sample: What the parents of preschool children with early signs of adhd say and how they say it?" *PLoS ONE*, vol. 9, no. 9, 2014, cited by: 10; All Open Access, Gold Open Access, Green Open Access.

[2] D. Daley, E. Sonuga-Barke, and M. Thompson, "Assessing expressed emotion in mothers of preschool ad/hd children: Psychometric properties of a modified speech sample," *British Journal of Clinical Psychology*, vol. 42, no. 1, p. 53 – 67, 2003, cited by: 106.

[3] L. Psychogiou, D. M. Daley, M. J. Thompson, and E. J. S. Sonuga-Barke, "Mothers' expressed emotion toward their school-aged sons: Associations with child and maternal symptoms of psychopathology," *European Child and Adolescent Psychiatry*, vol. 16, no. 7, p. 458 – 464, 2007, cited by: 54.

[4] A.-L. Hamdan, R. Deeb, A. Sibai, C. Rameh, H. Rifai, and J. Fayyad, "Vocal characteristics in children with attention deficit hyperactivity disorder," *Journal of Voice*, vol. 23, no. 2, pp. 190–194, 2009.

[5] A. A. Spiesberger, A. Triantafyllopoulos, A. Kathan, A. Semertzidou, C. Gawrilow, T. Reinelt, W. A. Rauch, and B. Schuller, ""so . . . my child . . . " – how child adhd influences the way parents talk," in *Interspeech 2024*, 2024, pp. 2010–2014.

[6] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions." Curran Associates, Inc., 2017. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

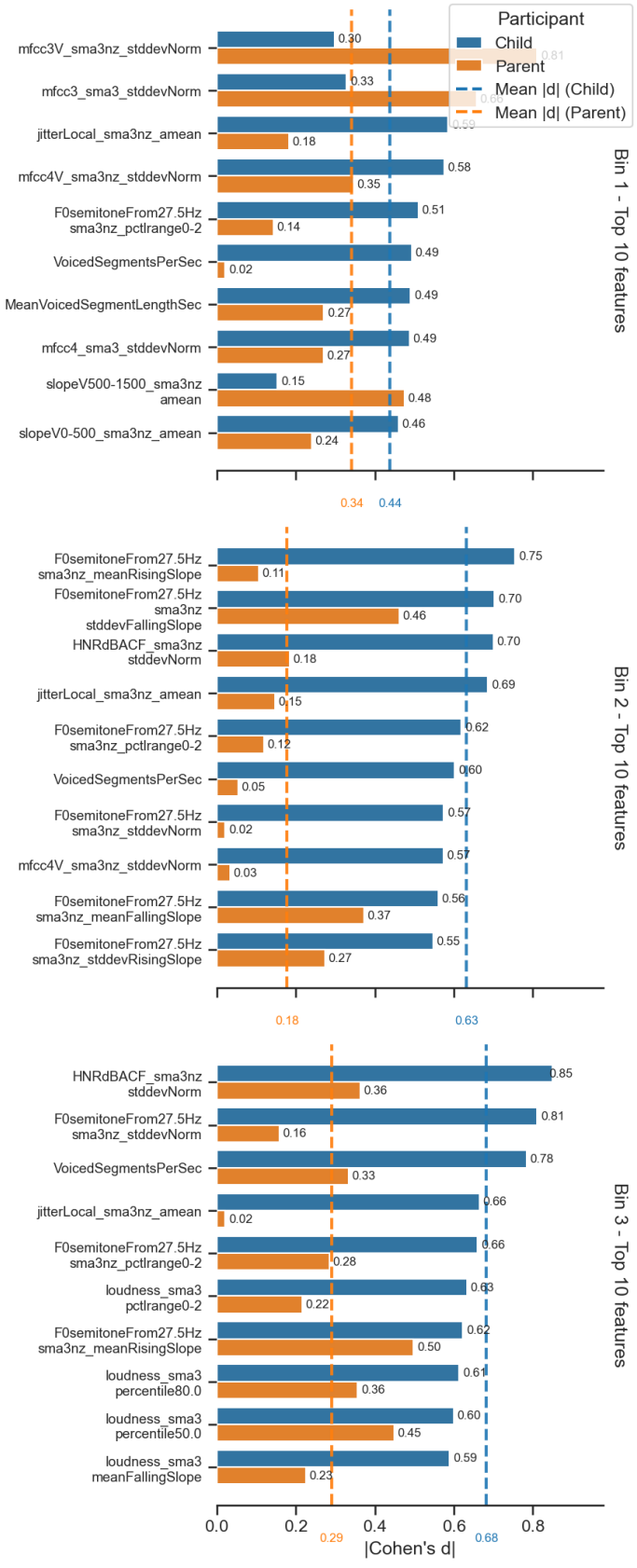


Figure 7: Top 10 acoustic features per time bin by |Cohen's d|