

Background

- Common practice in Germany: Language Proficiency Assessment (LPA) for preschool children [1]
- Most applied LPA methods: test for children's vocabulary size, grammar skills [2, 3], and morphology [4]
- Speech fluency correlates with language proficiency, e.g. [5], [6], [7] → opportunity to enhance LPAs

Aim

- Final goal: (automatic) assessment of children's speech fluency
- Previous studies: (manual) annotation and extraction of fluency related features from audio files [8], human fluency ratings for the annotated data [9]
- This study: development of model to automatically predict human fluency ratings

Data & Methodology

- **Speech data** elicited with *WUSCHEL* [10], a game-based task in a custom-made app: children interact with virtual character, answer questions to progress through coherent scenes
- Age 4;6 - 5;6 years, 5 with L1 German, 5 with L2 German
- 28 scenes, 2 answers each => 560 segments (Ø 3.23 s of pure articulation time per segment)
- 7 features (Figure 1) extracted from annotations and speech signal [8]
- **Human ratings** from 32 German L1 raters, background in linguistics
- Online: rate perceived overall fluency of child in presented audio (no definition given, no special instructions)
- 1st round: 20 different stimuli in random order
2nd round: same stimuli in same order again
- 2 separate ratings of 20 different stimuli per rater => 1280 fluency ratings
- Build linear CLMM (Cumulative Link Mixed Model) to predict **human ratings** from annotated **speech data** (baseline)
- Non-linear relation between some features and human ratings (Figure 1)
- Model these features with penalised regression splines in further CLMMs [11]
- Apply all models to entire dataset, evaluate performance: Spearman's Rho (SR), Mean Absolute Error (MAE), Quadratic Weighted Kappa (QWK)

Results

- Non-linear effects can improve model fit (LR tests, $p \leq 0.011$)
- No improvement in prediction performance (Table 1)
- Predictions within ± 2 rating points (MAE = 1.35/1.36, Figure 2)
- QWK models (0.65) > QWK humans (0.51)

Non-linear factors	SR	MAE	QWK
None	0.72	1.36	0.65
No. Filler Particles (FPs)	0.72	1.35	0.65
No. Other Disfluencies (ODs)	0.72	1.36	0.65
No. FPs + No. ODs	0.72	1.36	0.65

Table 1: Prediction performance of the 4 models with given features as non-linear factors

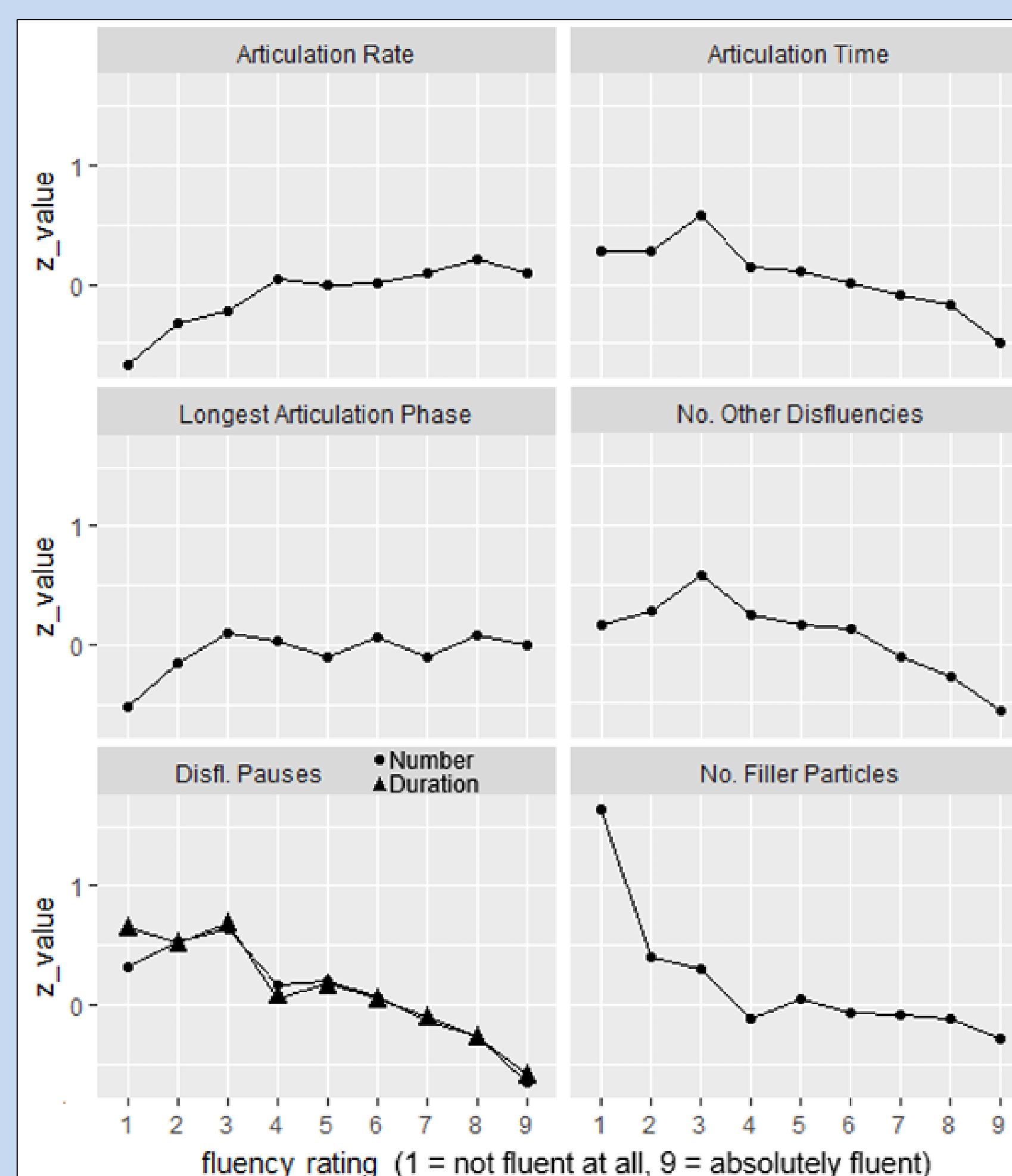


Figure 1: Features used in the CLMMs with their normalised mean z-value for stimuli awarded with certain rating

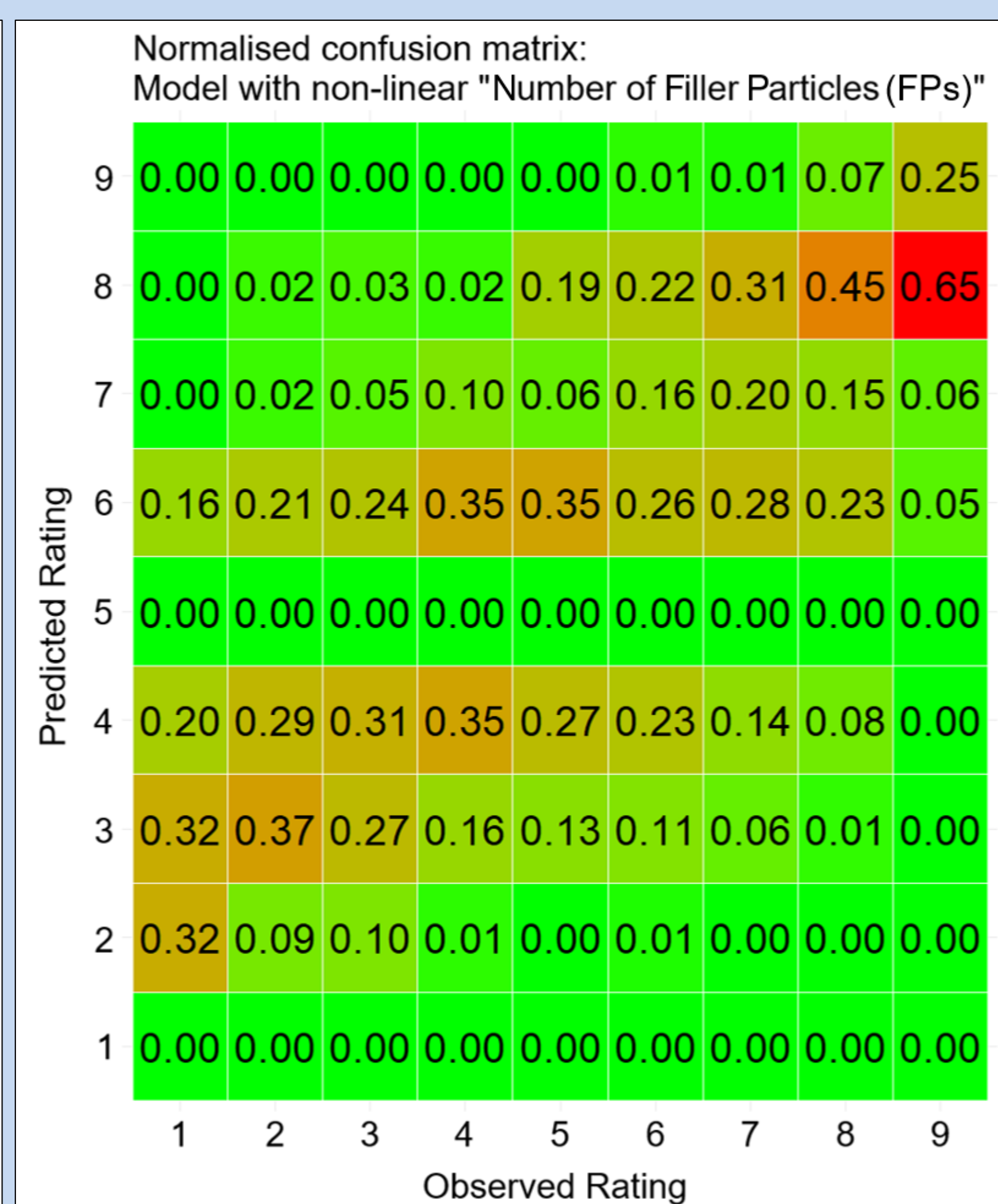


Figure 2: Confusion matrix of CLMM with No. FPs as non-linear factor, values are normalised by frequency of human ratings

Discussion

- Few low-end ratings
- Mostly short speech intervals → less room for differentiation
- Non-linear terms → better model fit, yet no better predictions
- Models comparable to human raters
- Increase robustness by using less fine-grained rating scale and aggregating ratings into a single rating per child

Next steps

- To date: 167 children recorded → stabilisation of dataset through further annotation and rating
- Automation of annotation process
- Ultimately: Fully automated, complete fluency assessment pipeline
- Integration into WUSCHEL LPA framework