Typical vs. atypical: Combining auditory Gestalt perception and acoustic analysis of early vocalisations in Rett syndrome

Florian B. Pokorny[a,b,c], Katrin D. Bartl-Pokorny[a], Christa Einspieler[a], Dajie Zhang[a,d], Ralf Vollmann[c,e], Sven Bölte[f,g], Markus Gugatschka[a], Björn W. Schuller[h,i], Peter B. Marschik[a,c,d,f,*]

[a] iDN – interdisciplinary Developmental Neuroscience, Department of Phoniatrics, Medical University of Graz, Graz, Austria
[b] Machine Intelligence & Signal Processing group, MMK, Technical University of Munich, Munich, Germany
[c] Brain, Ears & Eyes – Pattern Recognition Initiative, BioTechMed-Graz, Graz, Austria
[d] Department of Child and Adolescent Psychiatry and Psychotherapy, iDN – interdisciplinary Developmental Neuroscience, University Medical Center Göttingen, Göttingen, Germany
[e] Department of Linguistics, University of Graz, Graz, Austria
[f] Center of Neurodevelopmental Disorders (KIND), Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden
[g] Child and Adolescent Psychiatry, Center for Psychiatry Research, Stockholm County Council, Sweden
[h] Chair of Embedded Intelligence for Health Care and Wellbeing, Augsburg University, Augsburg, Germany
[i] Group on Language, Audio & Music (GLAM), Department of Computing, Imperial College London, London, UK

* Corresponding author at: Research Unit iDN – interdisciplinary Developmental Neuroscience, Department of Phoniatrics, Medical University of Graz, Auenbruggerplatz 26, Graz, 8036, Austria.
E-mail addresses: peter.marschik@medunigraz.at, peter.marschik@ki.se (P.B. Marschik).

What this paper adds?

Our study adds to the state of knowledge that the early speech-language development of individuals with Rett syndrome (RTT) already bears atypicalities that are co-occurring with apparently typical verbal behaviour. Here, we provide a comprehensive delineation of this intermittent character of typical versus atypical early verbal behaviour in RTT on the basis of a comprehensive set of early RTT-associated vocalisations. To the best of our knowledge, this is the first study to describe early vocalisation atypicalities in RTT by combining auditory Gestalt perception in the framework of a listening experiment, with a precise acoustic vocalisation analysis. This allowed us to objectify listeners' auditory concepts of early vocalisation (a)typicality on the signal level. Our approach may pave the way for the generation of an objective model of early vocalisation atypicality in RTT or other developmental disorders usually diagnosed in or beyond toddlerhood. Such a model could be used for increasing the sensitivity of caregivers and healthcare professionals to identify early vocalisation atypicalities, and further to implement a probabilistic tool for an automated vocalisation-based earlier detection of RTT or other 'late diagnosed' developmental disorders.

1. Introduction

Rett syndrome (RTT, MIM 312750) is a severe progressive developmental disorder that almost exclusively affects females at a prevalence of approximately 1 in 10,000 live female births (Laurvick et al., 2006). More than 30 years after the first description of the clinical presentation of RTT by the Austrian neuropediatrician Andreas Rett (1966, 2016), mutations in the X-linked gene *MECP2* were identified as the main cause of the disease (Amir et al., 1999). There are, however, individuals with *MECP2* mutations showing no clinical signs and individuals with a RTT phenotype without *MECP2* mutations (Neul et al., 2010; Suter, Treadwell-Deering, Zoghbi, Glaze, & Neul, 2014). The latter can be, at least for a proportion, explained as variants or atypical forms of RTT related to mutations in other genes (e.g., *FOXG1*, *CDKL5*; Neul et al., 2010; Sajan et al., 2017). Thus, RTT remains a clinical diagnosis based upon four main clinical consensus criteria (Neul et al., 2010): (1) partial or complete loss of purposeful hand skills; (2) partial or complete loss of acquired spoken language; (3) gait abnormalities (dyspraxic or absent gait); and (4) stereotypic hand movements (e.g., repetitive washing-like movements, clapping movements, hand-to-mouth stereotypies). In addition to the main criteria, Neul et al. (2010) defined eleven supportive diagnostic criteria: breathing disturbances and bruxism when awake, sleep problems, abnormal muscle tone, vasomotor changes, scoliosis/kyphosis, growth retardation, small cold hands and feet, inappropriate laughing/screaming spells, reduced response to pain, and intense eye communication. An individual is diagnosed with typical RTT, if all four main criteria are met, or with atypical RTT, if at least two main criteria and five supportive criteria are met. According to these consensus criteria, we differentiate between three atypical forms of RTT, with the relatively milder preserved speech variant (PSV; Zappella variant) being one of them (Neul et al., 2010). PSV is typically associated with comparably better recovery of speech-language capacities and functional hand use after regression (Marschik, Einspieler, Oberle, Laccone, & Prechtl, 2009; Renieri et al., 2009). For parents, but also for healthcare professionals seeing the child during clinical routine examinations, early development in RTT might appear inconspicuous. Thus, the mean age of diagnosis is still at 2.7 years (Tarquinio et al., 2015). Often, the regression of already acquired functions raises parental concerns, and is the first warning signal that motivates parents to consult a specialist. There is, however, growing evidence that atypicalities in different developmental areas are already present prior to the onset of regression. Studies employing retrospective analysis of family videos of individuals who are later diagnosed with typical RTT or PSV have revealed deviances from typical development in motor functions and speech-language/socio-communicative capabilities during the first year of life (Bartl-Pokorny et al., 2013; Burford, Kerr, & Macleod, 2003; Einspieler, Kerr, & Prechtl, 2005a; Einspieler, Kerr, & Prechtl, 2005b; Einspieler, Freilinger, & Marschik, 2016; Einspieler, Marschik et al., 2014; Einspieler, Sigafoos et al., 2014; Leonard & Bower, 1998; Marschik et al., 2009, 2013; Marschik, Bartl-Pokorny et al., 2014; Marschik, Einspieler, & Sigafoos, 2012; Marschik, Kaufmann et al., 2012, Marschik, Pini et al., 2012; Marschik, Sigafoos et al., 2012; Marschik, Vollmann et al., 2014; Pokorny, Marschik, Einspieler, & Schuller, 2016; Tams-Little & Holdgrafer, 1996; Townend et al., 2015). In our previous work on aspects of early speech-language development, we have found that a proportion of individuals with typical RTT and PSV do not achieve certain early speech-language milestones (e.g., cooing, babbling, proto-words; Bartl-Pokorny et al., 2013; Marschik, Bartl-Pokorny et al., 2014; Marschik et al., 2013). Observed vocalisations occurred with an intermittent character of typical and atypical patterns (i.e., pressed, inspiratory, or high-pitched crying-like; Marschik et al., 2013; Marschik, Pini et al., 2012). In a previous listening experiment, Marschik, Einspieler et al. (2012) demonstrated that listeners were able to distinguish between the vocalisation sequences of typically developing infants and pre-selected atypical vocalisation sequences of infants later diagnosed with RTT. Interestingly, a vocalisation which was produced by an individual later diagnosed with RTT, and considered typical by speech-language pathologists, was not classified consistently in this experiment. These findings together with clinical and parental observations, indicate a need for more detailed characterisation of early verbal behaviour in RTT. To the best of our knowledge, the study presented here is the first that aims to comprehensively delineate specific features of the intermittent character of typical versus atypical early vocalisations in RTT. Our approach is unique in that it combines a listening experiment on an extensive sample of early RTT-associated vocalisations, with precise signal analytical descriptions. In this way, auditory analysis may be extended to include an acoustic perspective, shedding further light on aspects of early vocalisation development in RTT. With this study, we intended to pave the way for an approach in which the communicative environment of an infant with a not yet diagnosed genetic disorder can be actively involved in the promotion of an early identification of atypical behavioural patterns to facilitate earlier intervention. Such an approach seems to be promising as infants actively influence the communicative environment through their early verbal behaviours (cf. the concept of the self-generated environment as, for example, described by Esposito, Hiroi, and Scattoni (2017) for infants with autism spectrum disorder). In this regard, a higher amount and similar pattern of atypical vocalisations uttered by an infant might more likely lead to

parental concern about a developmental delay or deviation (e.g., reported for fragile X syndrome by Zhang et al., 2017). Specifically, in our study we aimed to: (i) explore whether early RTT-associated vocalisations can be consistently classified as typical or atypical by a group of experts in the field of early human development; (ii) assess the proportion of atypical verbal behaviour in pre-regressional RTT; (iii) detect auditory attributes that are closely related to vocalisation atypicality; and (iv) describe how atypicality in early RTT-associated vocalisations manifests in the acoustic signal domain.

2. Material and methods

Analyses in this study were based on a homogeneous set of early RTT-associated vocalisations (see Sections 2.1 and 2.2) that were auditorily evaluated by a number of participants in the framework of a listening experiment (see Section 2.3) and acoustically decomposed into signal level features (see Section 2.4). Ethics approval for analyses related to this study was obtained by the Institutional Review Board of the Medical University of Graz, Austria (EC no. 27-388 ex 14/15).

2.1. Material

We reviewed 88 clips of home video recordings of a female infant over the second half year of life, later diagnosed with PSV (*MECP2* mutation: large intragenic deletion c.378-43_964delinsGA; Marschik et al., 2009). The infant came from a monolingual German-speaking family. The material was recorded by the infant's parents in typical family settings such as playing situations, feeding situations, or bathing situations, using a Sony digital video (DV) camcorder. During the period of recording, the parents were not aware of their daughter's later RTT diagnosis. From the infant's 7th month of life we had 10 min video material available, from her 8th month 4 min, from her 9th month 9 min, from her 10th month 16 min, from her 11th month 9 min, and from her 12th month 13 min (total recording time = 61 min).

2.2. Segmentation and annotation

First, the material was manually segmented for infant vocalisations by the first author (FBP). The segmentation process was based on the criterion to assign a vocalisation to a distinct vocal breathing group (Lynch, Oller, Steffens, & Buder, 1995). Vegetative sounds, such as breathing sounds, smacking sounds, hiccups, etc., were not segmented and excluded from further analyses. Realisations of the (proto-)word /da/ (= German for /there/) that occurred in combination with index finger pointing in the 11th and 12th month of age, were the only linguistic vocalisations identified in our material. A total number of N = 363 (pre-)linguistic vocalisations were included in our study database. Seventy-three vocalisations stem from the recorded infant's 7th month of life, 26 from her 8th, 57 from her 9th, 73 from her 10th, 68 from her 11th, and 66 from her 12th month of life. The median vocalisation duration was 1.44 s (25th percentile = 1.2 s; 75th percentile = 2.07 s), the shortest vocalisation was 0.45 s in duration, the longest vocalisation had a duration of 12 s. The latter was a repetitive pattern of fully-resonant nuclei (i.e., vowel-like sounds) and vowel glides (i.e., vowel-like sounds or vowels with changes in vowel quality over time without an audible gap; Nathani, Ertmer, & Stark, 2006). Ninety-nine percent of vocalisations had a duration below 8.84 s. Four vocalisations exceeded this duration.

In the second step, FBP annotated all included vocalisations for the presence of background noise according to the following four mutually exclusive quality classes: (Q1) no background noise present; (Q2) stationary background noise present, such as the engine sound of a car or the sound of a vacuum cleaner; (Q3) transient background noise present, such as parental voice or television audio; and (Q4) stationary and transient background noise present simultaneously. From an acoustic point of view, Q1–Q4 were defined to follow a decreasing order of quality from optimal to suboptimal background conditions superimposing a segmented vocalisation. Two hundred and ten vocalisations (57.85%) were assigned to class Q1, 10 vocalisations (2.75%) to Q2, 125 vocalisations (34.44%) to Q3, and 18 vocalisations (4.96%) to Q4.

Data segmentation and segment annotation represented the data pre-processing steps essential for the further experimental and signal-analytical examinations in our study.

2.3. Listening experiment

In order for vocalisations to be evaluated according to auditory concepts of (a)typicality in early speech-language development, we set up a listening experiment.

2.3.1. Set-up, participants, and experimental design

The listening experiment was implemented in the form of a stand-alone application with a graphical user interface programmed in Matlab (www.mathworks.com). For playback we used professional studio headphones (Audio Technica ATH-M50). The volume was individually adjusted to a comfortable level once at the beginning of the experiment, and remained at this level for the entire experiment.

All 363 vocalisations from our study database were presented separately to five participants. These participants (henceforth also referred to as listeners) were experts in the fields of speech-language acquisition, developmental psychology, and/or developmental physiology.

The experiment consisted of six sessions, with each session containing only vocalisations from one month of the recorded infant's second half year of life. The sequence of sessions was arranged according to the recorded infant's age in ascending order. Thus, the

first session contained vocalisations from the infant's 7th month of life, the second session from the infant's 8th month of life, and so forth. The listeners were informed about the infant's age in months prior to each session, therefore they could apply their knowledge and experience on age-related vocalisation specificities. Within each session, the vocalisations were presented in randomised order including 10% duplicates for intra-rater reliability evaluation. Randomisation was implemented to avoid the duplicates being arranged in a consecutive order. For each vocalisation, the listeners were asked to rate whether they perceived the vocalisation as (I) 'typical' or (II) 'atypical'. Classification was made on the basis of the listeners' expert knowledge and their experience with pre-linguistic vocalisations of both typically developing infants and infants with various developmental disorders. In circumstances where the listener was not sure, he/she could rate (III) 'do not know'. If rated 'atypical', the listener had to further specify whether (i) rhythm, (ii) timbre, (iii) pitch, and/or (iv) any other (not explicitly specified) auditory attributes were predominantly perceived deviant from a typical age-adequate vocalisation. In this case, multiple selections were allowed. The participants were instructed to listen to a vocalisation as often as they needed to come to a final decision. The number of replays per vocalisation was automatically logged. The listeners were neither informed about replay logging nor about the existence of duplicates prior to their participation in the listening experiment.

2.3.2. Evaluation measures

The listeners' ratings for each vocalisation were organised into three evaluation measures: (1) the information factor; (2) the atypicality factor; and (3) the replay factor (see Formulary below). The information factor (Eq. (1)) indicates the proportion of 'typical' or 'atypical' ratings for each of the 363 vocalisations. In other words, this measure shows how many listeners were able to make a decision on a vocalisation unequal to 'do not know'. The atypicality factor (Eq. (2)) specifies the proportion of the rating for 'atypical' for each vocalisation. Both the information factor and the atypicality factor can take values from 0 to 1. An information factor of 0 would mean that all five listeners rated 'do not know' for a vocalisation. By contrast, all five listeners rating for 'typical' or 'atypical' on a vocalisation would lead to an information factor of 1. An atypicality factor of 0 would mean that none of the listeners rated 'atypical'. A vocalisation consistently rated as 'atypical' by all five listeners would obtain an atypicality factor of 1. Finally, the replay factor (Eq. (3)) describes the mean number of replays per vocalisation. The minimum value is 1 if all 5 participants listened to a vocalisation only once.

Formulary:
$K \ldots number \ of \ evaluators$
$k \ldots evaluator$
$p \ldots number \ of \ replays$
$r \ldots rating \in \{r_1, \ r_2, \ r_3\} \triangleq \{typical, \ atypical, \ do \ not \ know\}$
$s \ldots sample/vocalisation$

$$\text{information factor}_s = 1 - \left[\frac{1}{K} \sum_{k=1}^{K} r_{3,s,k}\right] \tag{1}$$

$$\text{atypicality factor}_s = \frac{1}{K} \sum_{k=1}^{K} r_{2,s,k} \tag{2}$$

$$\text{replay factor}_s = \frac{1}{K} \sum_{k=1}^{K} p_{s,k} \tag{3}$$

To investigate the correlations between the introduced evaluation measures and the outlier-afflicted vocalisation duration, as well as between the evaluation measures and the ordinal scaled audio background quality, we calculated Spearman's rank correlation coefficients. For intra-rater and inter-rater reliability evaluations we calculated unweighted Cohen's and Fleiss' kappas at a significance level of $\alpha = 0.05$.

2.4. Acoustic analysis

For the acoustic analysis we first extracted acoustic features from the uncompressed audio track of each vocalisation (44.1 kHz, 16 bit, 1 channel, PCM; see Section 2.4.1). Then we selected features related to the listeners' ratings, i.e., the vocalisations' categorisation into a typical and an atypical class (see Section 2.4.2).

2.4.1. Feature extraction

To extract the acoustic features we used the open-source tool kit openSMILE (Eyben, Wöllmer, & Schuller, 2010; Eyben, Weninger, Groß, & Schuller, 2013; www.audeering.com). To ensure reproducibility at high performance, we extracted the official baseline features of the 2013–2017 Interspeech Computational Paralinguistics Challenges (ComParE Challenges; Schuller et al., 2013, 2014, 2015, 2016, 2017). These ComParE features currently constitute one of the most extensive sets of standardised acoustic features available. It consists of 6373 static higher-order features. These features describe statistical functionals (e.g., arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles) calculated for the trajectories of a wide range of acoustic short-term low-level descriptors (e.g., fundamental frequency, zero-crossings rate, harmonics-to-noise ratio, Mel-
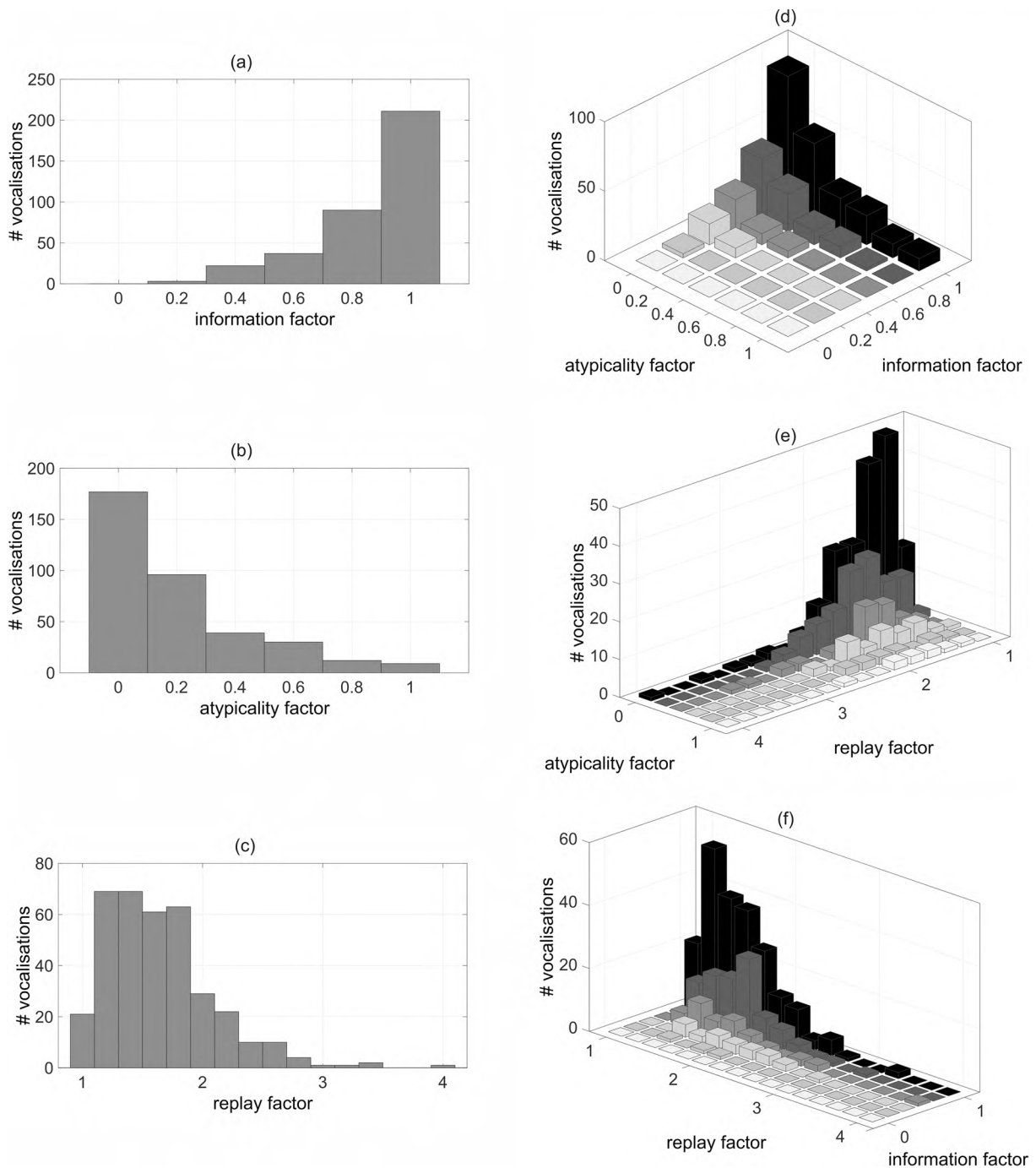
Fig. 1. Evaluation measures. Two-dimensional histograms showing the absolute number (#) of vocalisations for (a) the information factor, (b) the atypicality factor, and (c) the replay factor; (d)–(f) three-dimensional histograms for combinations of evaluation measures.

frequency cepstral coefficients, jitter, shimmer) and their derivatives within a vocalisation (Schuller et al., 2013).

2.4.2. Feature selection

To identify features related to the listeners' ratings, we applied a feature selection approach to the complete set of 6373 acoustic features. We included all vocalisations with an information factor of 1 (n = 211; see Fig. 1a). As a function of the atypicality factor, we defined six confidence scenarios. In the lowest confidence scenario 1, a vocalisation was considered atypical if its atypicality factor was 0.2 or higher (i.e., at least 1 of the 5 listeners rated the vocalisation as 'atypical'). This revealed 118 atypical versus 93 typical vocalisations in scenario 1. In the next higher confidence scenario 2, an atypicality factor of 0.4 or higher was required to

categorise a vocalisation as 'atypical'. This yielded a distribution of 64 atypical versus 147 typical vocalisations in scenario 2. Accordingly, the atypicality factor threshold was 0.6 for the confidence scenario 3 (40 atypical versus 171 typical vocalisations), 0.8 for the confidence scenario 4 (19 atypical versus 192 typical vocalisations), and 1 for the confidence scenario 5 (9 atypical versus 202 typical vocalisations). For the confidence scenario 6, we retained the 9 atypical vocalisations from scenario 5, and added 9 vocalisations that were rated as 'typical' consistently by all 5 listeners. These vocalisations were best matched to the nine atypical vocalisations for vocalisation duration, audio background quality, and infant age.

For all six scenarios, the worth of each acoustic feature was evaluated by measuring the Pearson's correlation between the feature values and the class values (corresponding to either typical or atypical) across all included vocalisations. Finally, we ranked the features and selected the ten features with the highest absolute correlation coefficient for each scenario.

3. Results

For better readability, the results are structured according to the methodological cascade.

3.1. Auditory domain

Two hundred and eleven of the complete set of 363 vocalisations had an information factor of 1 (see Fig. 1a). Ninety-three of these vocalisations were consistently rated as 'typical' (see Fig. 1d; information factor = 1, atypicality factor = 0), nine 'atypical' (see Fig. 1b and d; information factor = 1, atypicality factor = 1). In the latter, atypicality was related to the auditory attribute 'rhythm' with a percentage of 37.8%, to 'timbre' with 77.8%, to 'pitch' with 28.9%, and to 'other' with 73.3% (multiple selections were allowed for one vocalisation; see Section 2.3.1).

None of the vocalisations had an information factor of 0. However, three vocalisations were classified (as 'typical' or 'atypical') by only one of the five listeners (information factor = 0.2; see Fig. 1a).

For more than half of the vocalisations (186/363), at least one of the five listeners rated 'atypical' (atypicality factor > 0; see Fig. 1b).

A high proportion of vocalisations (312/363) were on average played back between one and two times by a listener before making a final decision (see Fig. 1c). The median replay factor was 1.6. The maximum replay factor was 4 and was obtained for 1 vocalisation. This vocalisation had a duration of 2.2 s and an audio background quality class Q3 due to parental voice that prominently superimposed upon the target vocalisation. It was perceived typical by three listeners, with the other two listeners providing a rating of 'do not know'. Detailed relations between the replay factor and the atypicality factor, and between the replay factor and the information factor are shown in Fig. 1e and f, respectively.

We did not observe differences in the evaluation measures as a function of the recorded infant's age. The standard deviations over the mean information factors, the mean atypicality factors, and the mean replay factors for vocalisations from the 7th to the 12th months of the recorded infant's life, were 0.05 or below.

Fair positive correlations (Portney & Watkins, 2000) were found between the information factor and the vocalisation duration (r = 0.38), and between the atypicality factor and the vocalisation duration (r = 0.36). Thus, longer vocalisations were more likely classified as compared to shorter ones. All 4 vocalisations with a duration above the 99th percentile had an information factor of 1. Most of the short vocalisations had a low atypicality factor. No correlation was found between the replay factor and the vocalisation duration (r = −0.04).

The audio background quality was not related to any of the evaluation measures information factor (r = −0.16), atypicality factor (r = −0.04), and replay factor (r = 0.06).

3.1.1. Intra-rater reliability
The intra-rater reliability evaluation on our listening experiment revealed an average Cohen's kappa of 0.66 when considering all three rating options 'typical', 'atypical', and 'do not know'. Vocalisation duplicates were consistently rated in 83.2% of cases (see Table 1a). The average Cohen's kappa increased to 0.88 when only considering the rating options 'typical' and 'atypical'. In 3.2% of cases, the listeners changed their opinion from 'typical' to 'atypical', or vice versa (see Table 1a).

Table 1
(a) **Intra-rater reliability.** Confusion matrix showing proportions of constellation occurrences between first ($_1$) and second ($_2$) ratings on vocalisation duplicates for the three rating options 'typical', 'do not know', and 'atypical'. (b) **Inter-rater reliability.** Reliabilities between all five raters in the form of Fleiss' kappas for three different category groupings, i.e., for considering 'typical', 'do not know', and 'atypical' as separate groups, and for considering the category 'do not know' combined with first the category 'atypical' and then the category 'typical'.

(a)				(b)	
	typical$_1$	do not know$_1$	atypical$_1$	grouping	K
typical$_2$	60.5%	6.0%	1.6%	typical | do not know | atypical	0.204
do not know$_2$	4.3%	6.5%	2.7%	typical | do not know or atypical	0.198
atypical$_2$	1.6%	0.5%	16.2%	typical or do not know | atypical	0.263

Table 2
Feature ranking. Low-level descriptors (see Section 2.4.1) underlying the selected ten best acoustic higher-order features, and the Pearson's correlation coefficients for scenarios 1–6 (see Section 2.4.2). Correlation coefficients are given as absolute values.

rank	scenario 1	scenario 2	scenario 3	scenario 4	scenario 5	scenario 6
1	$f0$ 0.514	$spec_{Rfilt}$ 0.53	$spec_{Rfilt}$ 0.447	$jitter_{local}$ 0.43	$mag_{specKurt}$ 0.495	$spec_{Rfilt}$ 0.84
2	$spec_{Rfilt}$ 0.504	$spec_{Rfilt}$ 0.527	$spec_{l1norm}$ 0.441	$jitter_{DDP}$ 0.429	$mag_{specKurt}$ 0.482	$spec_{Rfilt}$ 0.822
3	$spec_{Rfilt}$ 0.494	$spec_{Rfilt}$ 0.518	$spec_{Rfilt}$ 0.436	$jitter_{local}$ 0.417	$mag_{specKurt}$ 0.474	$jitter_{local}$ 0.821
4	$spec_{Rfilt}$ 0.49	$spec_{Rfilt}$ 0.497	$spec_{Rfilt}$ 0.435	$jitter_{DDP}$ 0.4	$mag_{specKurt}$ 0.473	$mag_{specKurt}$ 0.819
5	$spec_{Rfilt}$ 0.485	$spec_{Rfilt}$ 0.495	$jitter_{local}$ 0.415	$spec_{Rfilt}$ 0.4	$mag_{specSkew}$ 0.45	$spec_{Rfilt}$ 0.809
6	$spec_{l1norm}$ 0.48	$spec_{l1norm}$ 0.487	$spec_{Rfilt}$ 0.408	$spec_{Rfilt}$ 0.386	$spec_{Rfilt}$ 0.434	$jitter_{DDP}$ 0.808
7	$spec_{Rfilt}$ 0.478	$spec_{Rfilt}$ 0.484	$jitter_{DDP}$ 0.404	$spec_{Rfilt}$ 0.381	mag_{fBand} 0.431	$spec_{Rfilt}$ 0.807
8	$spec_{Rfilt}$ 0.467	$f0$ 0.479	$spec_{Rfilt}$ 0.401	$jitter_{DDP}$ 0.38	$mag_{specSkew}$ 0.41	$jitter_{DDP}$ 0.807
9	$spec_{l1norm}$ 0.467	$spec_{Rfilt}$ 0.477	$jitter_{local}$ 0.4	$jitter_{local}$ 0.377	$spec_{Rfilt}$ 0.41	$spec_{Rfilt}$ 0.805
10	$spec_{Rfilt}$ 0.466	$spec_{Rfilt}$ 0.476	$jitter_{local}$ 0.4	$jitter_{local}$ 0.377	$mag_{specKurt}$ 0.396	$spec_{Rfilt}$ 0.803

$f0$ = fundamental frequency, $jitter_{DDP}$ = first order derivative of local jitter, $jitter_{local}$ = local jitter, mag_{fBand} = spectral band energy, $mag_{specKurt}$ = spectral kurtosis, $mag_{specSkew}$ = spectral skewness, $spec_{l1norm}$ = l1-norm (rectilinear distance) of auditory spectrum components (loudness), $spec_{Rfilt}$ = relative spectral transform (RASTA)-filtered auditory spectrum.

3.1.2. Inter-rater reliability

The evaluation of the inter-rater reliability in our listening experiment yielded an overall Fleiss' kappa of 0.2 when considering all three rating options 'typical', 'atypical', and 'do not know' (see Table 1b). The minimum Cohen's kappa value between 2 listeners was 0.11, the highest 0.35. When only including the rating options 'typical' and 'atypical', the minimum Cohen's kappa value between 2 listeners was 0.15, the highest 0.46.

3.2. Acoustic domain

The ten best ranked acoustic features (out of 6373; see Section 2.4.1) for each confidence scenario are shown in Table 2. The vocalisation distributions on the basis of the top three acoustic features to distinguish between 'typical' and 'atypical' are illustrated in Fig. 2. The best placed features were based on the filtered auditory spectrum, the fundamental frequency, jitter (i.e., fluctuations in the length of the fundamental period), and on spectral kurtosis. These features acoustically describe spectral, prosodic, and voice quality characteristics of a speech signal (e.g., Schuller & Batliner, 2014) and are related to auditory attributes such as timbre and pitch. Fig. 3 presents spectrograms of vocalisations with an atypicality factor of 1. Each of these vocalisations had the respective highest relative relation to one of the auditory attributes 'rhythm' (Fig. 3a), 'timbre' (Fig. 3b), 'pitch' (Fig. 3c), and 'other' (Fig. 3d). The spectrogram in Fig. 3a reveals a sequence of changes between inspiratory and expiratory vocalisation components irregular in time structure ('rhythm'). The vocalisation shown in Fig. 3b exhibits a high proportion of inharmonic overtones over the second half of the vocalisation due to a pressed phonation ('timbre'). The spectrogram in Fig. 3c is characterised by a high-pitched crying-like phonation pattern over the first half of the vocalisation ('pitch'). The spectrogram in Fig. 3d does not reveal any harmonic spectral structures due to a hoarse phonation ('other').

4. Discussion

In previous studies on individuals with RTT, we have reported on the presence of atypical vocalisations co-occurring with apparently normal vocalisations during the first year of life (Marschik, Einspieler et al., 2012; Marschik, Pini et al., 2012; Marschik et al., 2013). Inspiratory, pressed, and high-pitched crying-like patterns were among the most prominent atypical characteristics (Marschik et al., 2013; Marschik, Pini et al., 2012). In the present study, we have extended our understanding of this intermittent character of early vocalisations by combining a listening experiment on an extensive vocalisation database with acoustic vocalisation analysis. The study presented here, has confirmed the intermittent character of typical and atypical early vocalisations. The high proportion of vocalisations with inconsistent ratings (71.9%) suggests that although the listeners were experts in the field of early human development and had consistent individual concepts of early vocalisation (a)typicality (high intra-rater reliability), their individual concepts did not entirely overlap (low inter-rater reliability). Moreover, 42% of vocalisations had an information factor below 1, indicating that the dichotomic classification of early vocalisations is not a trivial task. On the other hand, the high proportion of inconsistently rated vocalisations could indicate the potential presence of hidden atypicalities in these vocalisations. This can be related to the findings reported in our previous listening experiment (Marschik, Einspieler et al., 2012) that revealed uncertainty in 21% of 400 listeners when rating a vocalisation of a girl with RTT that was considered typical by speech-language pathologists.

Even though only 2.5% of the vocalisations included in our study were rated as 'atypical' by all five experts, for more than half of the total vocalisations at least one of the listeners thought to experience atypical verbal realisations. Interestingly, the mean number of replays was independent from a listener's rating behaviour (reflected by the information factor and the atypicality factor). In contrast, the duration of a vocalisation was found to be positively correlated with both the information factor and the atypicality factor. This raises the questions (i) whether a minimum vocalisation duration is required for accurate classification paradigms of pre-linguistic vocalisations based on the auditory Gestalt perception, and (ii) if early RTT-associated verbal atypicality can be more
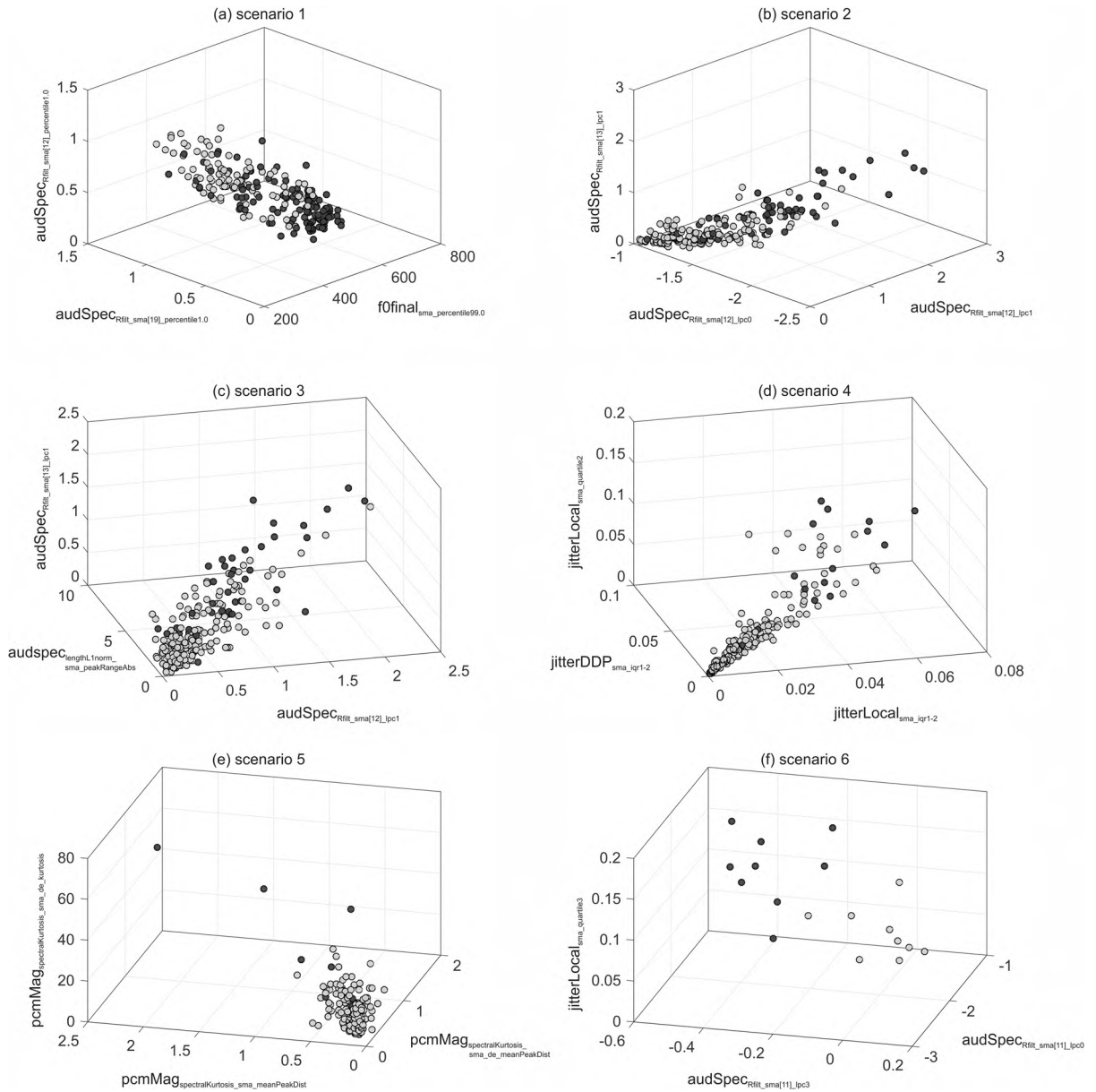
Fig. 2. Vocalisation distributions. Vocalisations considered typical (grey dots) and atypical (black dots) for (a)–(f) confidence scenarios 1–6 (see Section 2.4.2) in the three-dimensional space of the respective top three ranked acoustic features (see Table 2). abs = absolute value, audSpec = auditory spectrum, DDP = difference of differences of periods, de = derivative, dist = distance, f0 = fundamental frequency, iqr = inter-quartile range, lpc = linear predictive coding, pcmMag = pulse code modulation magnitude, Rfilt = relative spectral transform (RASTA)-filtered, sma = simple moving average.

accurately detected in more complex pre-linguistic vocalisation types, such as canonical babbling. Vocalisation duration also plays a role in the specification of perceived atypicality. For example, rhythmic atypicalities cannot be identified in very short vocalisations, which might explain the infrequent selection of this auditory attribute in the presented study. The high number of selections for 'other' auditory attributes could mean that the default attributes were insufficient to cover all facets of RTT-associated verbal atypicality, or listeners were unsure about how to precisely define a somehow experienced peculiarity. In our study, perceived atypical verbal behaviour was closely related to the auditory attribute 'timbre'. This finding was also reflected in the acoustic signal domain, in which the spectral structure-related filtered auditory spectrum (spec$_{Rfilt}$ in Table 2) was the most prominent acoustic descriptor of atypicality. In line with our previous studies (Marschik et al., 2013; Marschik, Pini, et al., 2012), vocalisations with inspiratory, pressed, and high-pitched crying-like character were among the most atypical vocalisations (see Section 3.2 and Fig. 3). In comparison, studies on vocalisation characteristics in infants later diagnosed with autism spectrum disorder (ASD) reported on deviances in the fundamental frequency compared to neurotypical controls (e.g., Esposito, del Carmen Rostagno, Venuti, Haltigan, & Messinger, 2014; Esposito & Venuti, 2010; Sheinkopf, Iverson, Rinaldi, & Lester, 2012). In our study, fundamental frequency and its auditory
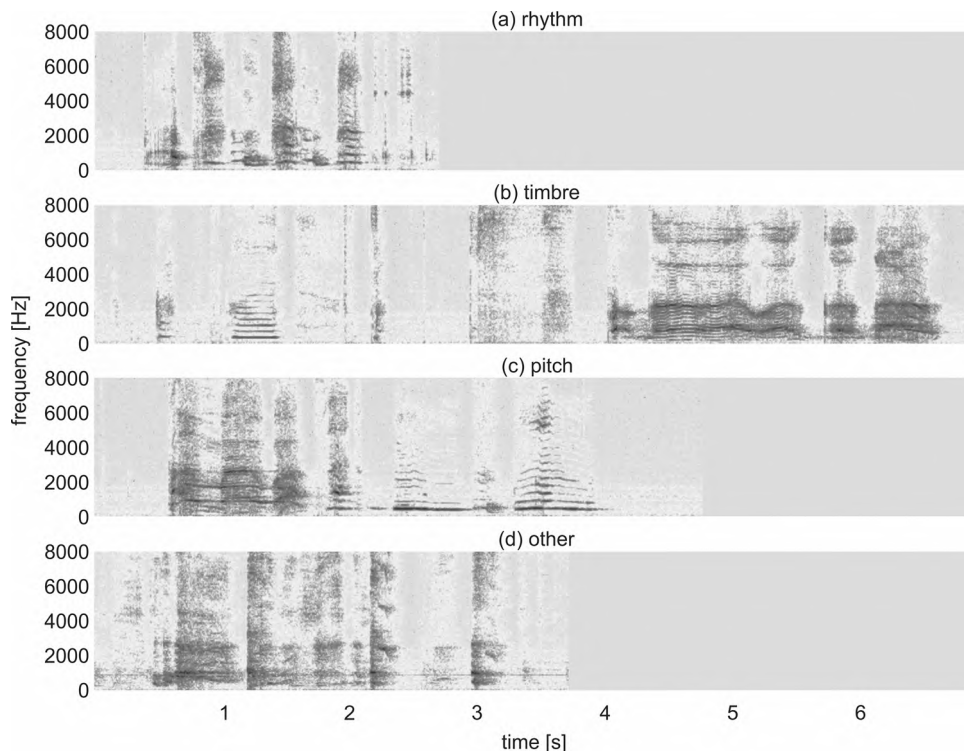
Fig. 3. Spectrograms. Spectrograms of vocalisations with an atypicality factor of 1 and the highest relative relation to the auditory attribute (a) 'rhythm', (b) 'timbre', (c) 'pitch', and (d) 'other'.

counterpart pitch, played only a secondary role in the differentiation between typical and atypical verbal behaviour. Interestingly, the fundamental frequency often remains the only measured descriptor in acoustic studies on early vocalisations (e.g., Esposito & Venuti, 2010; Sheinkopf et al., 2012; Wermke et al., 2016). The potential of combining a number of thoroughly selected acoustic features for specific classification tasks is indicated in Fig. 2. For example, in confidence scenario 6 (Fig. 2f), a precise separation between typical and atypical vocalisation clusters could be performed in the three-dimensional feature space, whereas the inclusion of less than three features would lead to a decreased separation accuracy.

While providing a number of novel insights into verbal phenomena associated with the early phenotype of RTT, the findings of our study must be interpreted with caution. We needed to deal with methodological limitations which are related to the existing database: First, the corpus for this study consisted of material from one individual with PSV. This does not allow for generalisations and needs to be seen in the light of phenotypic heterogeneity in typical RTT and RTT variants. For this study, we explicitly controlled for a number of variables such as homogeneity with respect to constant audio recording quality (the same camera was used for all recordings) and constant intrinsic voice characteristics across all of the material, beneficial for both auditory vocalisation assessment and objective acoustic vocalisation analysis. We furthermore selected data from a participant with the PSV of RTT over a participant with typical RTT, to have a higher number and a broader range of different vocalisation types for analysis. Second, the retrospective analysis of home video recordings involves general risks, e.g., the absence of particular behaviours in an available set of data (e.g., Marschik & Einspieler, 2011). Thus, it is likely that within our database the 'real' proportion of atypical verbal behaviour might not be captured as we anticipate that parents are less likely to record their children during distressed states. For this reason, vocalisations related to negative mood could be underrepresented in our dataset. Another limitation related to home videos is the presence of background noise. However, audio background quality was not associated with a listener's rating behaviour (reflected by the evaluation measures) in our listening experiment. Despite these limitations, retrospective home video analysis provides a unique opportunity for studying the early development of individuals with 'late diagnosed' developmental disorders under natural environmental conditions (Adrien et al., 1993; Crais, Watson, Baranek, & Reznick, 2006; Palomo, Belinchón, & Ozonoff, 2006; Saint-Georges et al., 2010).

5. Conclusions

The present study adds to the body of knowledge that the pre-regression period of individuals with RTT is not to be considered asymptomatic (e.g., Bartl-Pokorny et al., 2013; Burford et al., 2003; Einspieler, Sigafoos et al., 2014; Marschik et al., 2013; Tams-Little & Holdgrafer, 1996). Focussing on aspects of the speech-language domain, our findings confirm the intermittent occurrence of apparently typical and atypical early vocalisations in RTT (Marschik et al., 2013; Marschik, Pini et al., 2012). The novelty of our

approach is the combination of the auditory Gestalt perception of early RTT-associated vocalisations in a comprehensive listening experiment, with the extraction and analysis of signal-level descriptors to consider the vocalisations' 'acoustic ground truth', i.e., the vocalisations' objective representation in the acoustic domain. More than half of the extensive set of RTT-associated vocalisations were rated as 'atypical' by at least one listener. Perceived atypicality was mostly related to the auditory attribute 'timbre' and to the filtered auditory spectrum in the acoustic domain.

The knowledge gained from this study shall contribute to the generation of an objective model of early vocalisation atypicality in RTT. In the auditory domain, such a model may build a solid basis for increasing the sensitivity of healthcare professionals and, especially, parents − representing a developing infant's communicative environment actively influenced through early verbal behaviours − to detect early atypical vocalisation characteristics. This could be managed, for example, via training courses, online seminars, or even audio books, in which age-related samples of typical and atypical vocalisation characteristics are presented and auditory criteria for the identification of differences between these classes of vocalisations are taught. In such an approach, the awareness of specific atypicalities in early vocalisations in the communicative environment of infants, may lead to an earlier identification of RTT. In the acoustic domain, an objective model of early vocalisation atypicality may be used for an automatic audio-based machine learning approach to enable the earlier identification of individuals with RTT (Pokorny et al., 2016). Such a model could be further embedded into a more complex probabilistic model combining knowledge from various developmental domains (e.g., speech-language domain, motor domain, socio-communicative domain) for the earlier detection of developmental disorders with a current diagnosis in or beyond toddlerhood (Marschik et al., 2017).

Declaration of interest

The authors alone are responsible for content and writing of this article. They declare no conflicts of interest.

References

Adrien, J. L., Lenoir, P., Martineau, J., Perrot, A., Hameury, L., Larmande, C., et al. (1993). Blind ratings of early symptoms of autism based upon family home movies. *Journal of the American Academy of Child & Adolescent Psychiatry, 32*, 617–626.

Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., & Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2. *Nature Genetics, 23*, 185–188.

Bartl-Pokorny, K. D., Marschik, P. B., Sigafoos, J., Tager-Flusberg, H., Kaufmann, W. E., Grossmann, T., et al. (2013). Early socio-communicative forms and functions in typical Rett syndrome. *Research in Developmental Disabilities, 34*, 3133–3138.

Burford, B., Kerr, A. M., & Macleod, H. A. (2003). Nurse recognition of early deviation in development in home videos of infants with Rett disorder. *Journal of Intellectual Disability Research, 47*, 588–596.

Crais, E. R., Watson, L. R., Baranek, G. T., & Reznick, J. S. (2006). Early identification of autism: How early can we go? *Seminars in Speech and Language, 27*(3), 143–160.

Einspieler, C., Kerr, A. M., & Prechtl, H. F. (2005a). Abnormal general movements in girls with Rett disorder: The first four months of life. *Brain and Development, 27*, 8–13.

Einspieler, C., Kerr, A. M., & Prechtl, H. F. (2005b). Is the early development of girls with Rett disorder really normal? *Pediatric Research, 57*, 696–700.

Einspieler, C., Marschik, P. B., Domingues, W., Talisa, V. B., Bartl-Pokorny, K. D., Wolin, T., et al. (2014). Monozygotic twins with Rett syndrome: Phenotyping the first two years of life. *Journal of Developmental and Physical Disabilities, 26*, 171–182.

Einspieler, C., Sigafoos, J., Bartl-Pokorny, K. D., Landa, R., Marschik, P. B., & Bölte, S. (2014). Highlighting the first 5 months of life: General movements in infants later diagnosed with autism spectrum disorder or Rett syndrome. *Research in Autism Spectrum Disorders, 8*, 286–291.

Einspieler, C., Freilinger, M., & Marschik, P. B. (2016). Behavioural biomarkers of typical Rett syndrome: Moving towards early identification. *Wiener Medizinische Wochenschrift, 166*, 333–337.

Esposito, G., & Venuti, P. (2010). Developmental changes in the fundamental frequency (f0) of infants' cries: A study of children with autism spectrum disorder. *Early Child Development and Care, 180*, 1093–1102.

Esposito, G., del Carmen Rostagno, M., Venuti, P., Haltigan, J. D., & Messinger, D. S. (2014). Brief report: Atypical expression of distress during the separation phase of the strange situation procedure in infant siblings at high risk for ASD. *Journal of Autism and Developmental Disorders, 44*, 975–980.

Esposito, G., Hiroi, N., & Scattoni, M. L. (2017). Cry, baby, cry: Expression of distress as a biomarker and modulator in autism spectrum disorder. *International Journal of Neuropsychopharmacology, 20*, 498–503.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE: The Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia, MM 2010 (ACM, Florence, Italy, 2010), 1459–1462*.

Eyben, F., Weninger, F., Groß, F., & Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. *Proceedings of the 21st ACM International Conference on Multimedia, MM 2013 (ACM, Barcelona, Spain, 2013), 835–838*.

Laurvick, C. L., de Klerk, N., Bower, C., Christodoulou, J., Ravine, D., Ellaway, C., et al. (2006). Rett syndrome in Australia: A review of the epidemiology. *Journal of Pediatrics, 148*, 347–352.

Leonard, H., & Bower, C. (1998). Is the girl with Rett syndrome normal at birth? *Developmental Medicine and Child Neurology, 40*, 115–121.

Lynch, M. P., Oller, D. K., Steffens, M. L., & Buder, E. H. (1995). Phrasing in prelinguistic vocalizations. *Developmental Psychobiology, 28*, 3–25.

Marschik, P. B., & Einspieler, C. (2011). Methodological note: Video analysis of the early development of Rett syndrome − one method for many disciplines. *Developmental Neurorehabilitation, 14*, 355–357.

Marschik, P. B., Einspieler, C., Oberle, A., Laccone, F., & Prechtl, H. F. (2009). Case report: Retracing atypical development: A preserved speech variant of Rett syndrome. *Journal of Autism and Developmental Disorders, 39*, 958–961.

Marschik, P. B., Einspieler, C., & Sigafoos, J. (2012). Contributing to the early detection of Rett syndrome: The potential role of auditory Gestalt perception. *Research in Developmental Disabilities, 33*, 461–466.

Marschik, P. B., Kaufmann, W. E., Einspieler, C., Bartl-Pokorny, K. D., Wolin, T., Pini, G., et al. (2012). Profiling early socio-communicative development in five young girls with the preserved speech variant of Rett syndrome. *Research in Developmental Disabilities, 33*, 1749–1756.

Marschik, P. B., Pini, G., Bartl-Pokorny, K. D., Duckworth, M., Gugatschka, M., Vollmann, R., et al. (2012). Early speech-language development in females with Rett syndrome: Focusing on the preserved speech variant. *Developmental Medicine and Child Neurology, 54*, 451–456.

Marschik, P. B., Sigafoos, J., Kaufmann, W. E., Wolin, T., Talisa, V. B., Bartl-Pokorny, K. D., et al. (2012). Peculiarities in the gestural repertoire: An early marker for Rett syndrome? *Research in Developmental Disabilities, 33*, 1715–1721.

Marschik, P. B., Kaufmann, W. E., Sigafoos, J., Wolin, T., Zhang, D., Bartl-Pokorny, K. D., et al. (2013). Changing the perspective on early development of Rett syndrome. *Research in Developmental Disabilities, 34*, 1236–1239.

Marschik, P. B., Bartl-Pokorny, K. D., Tager-Flusberg, H., Kaufmann, W. E., Pokorny, F., Grossmann, T., et al. (2014). Three different profiles: Early socio-communicative capacities in typical Rett syndrome, the preserved speech variant and normal development. *Developmental Neurorehabilitation, 17*, 34–38.

Marschik, P. B., Vollmann, R., Bartl-Pokorny, K. D., Green, V. A., van der Meer, L., Wolin, T., et al. (2014). Developmental profile of speech-language and communicative functions in an individual with the Preserved Speech Variant of Rett syndrome. *Developmental Neurorehabilitation, 17*, 284–290.

Marschik, P. B., Pokorny, F. B., Peharz, R., Zhang, D., O'Muircheartaigh, J., Roeyers, H., et al. (2017). A novel way to measure and predict development: A heuristic approach to facilitate the early detection of neurodevelopmental disorders. *Current Neurology and Neuroscience Reports, 17*, 43.

Nathani, S., Ertmer, D. J., & Stark, R. E. (2006). Assessing vocal development in infants and toddlers. *Clinical Linguistics & Phonetics, 20*, 351–369.

Neul, J. L., Kaufmann, W. E., Glaze, D. G., Christodolou, J., Clarke, A. J., Bahi-Buisson, N., et al. (2010). Rett syndrome: Revised diagnostic criteria and nomenclature. *Annals of Neurology, 68*, 944–950.

Palomo, R., Belinchón, M., & Ozonoff, S. (2006). Autism and family home movies: A comprehensive review. *Journal of Developmental & Behavioral Pediatrics, 27*(2), 59–68.

Pokorny, F. B., Marschik, P. B., Einspieler, C., & Schuller, B. W. (2016). Does she speak RTT? Towards an earlier identification of Rett Syndrome through intelligent pre-linguistic vocalisation analysis. *Proceedings of the 17th Annual Conference of the International Speech Communication Association, Interspeech 2016 (ISCA, San Francisco, USA, 2016),* 1953–1957.

Portney, L. G., & Watkins, M. P. (2000). *Foundations of clinical research: Applications to practice.* Upper Saddle River, NJ: Prentice Hall.

Renieri, A., Mari, F., Mencarelli, M. A., Scala, E., Ariani, F., Longo, I., et al. (2009). Diagnostic criteria for the Zappella variant of Rett syndrome (the preserved speech variant). *Brain and Development, 31*, 208–216.

Rett, A. (1966). On a unusual brain atrophy syndrome in hyperammonemia in childhood. *Wiener Medizinische Wochenschrift, 116*, 723–726.

Rett, A. (2016). On a remarkable syndrome of cerebral atrophy associated with hyperammonaemia in childhood. *Wiener Medizinische Wochenschrift, 166*, 322–324.

Saint-Georges, C., Cassel, R. S., Cohen, D., Chetouani, M., Laznik, M.-C., Maestro, S., et al. (2010). What studies of family home movies can teach us about autistic infants: A literature review. *Research in Autism Spectrum Disorders, 4*(3), 355–366.

Sajan, S. A., Jhangiani, S. N., Muzny, D. M., Gibbs, R. A., Lupski, J. R., Glaze, D. G., et al. (2017). Enrichment of mutations in chromatin regulators in people with Rett syndrome lacking mutations in MECP2. *Genetics in Medicine, 19*, 13–19.

Schuller, B., & Batliner, A. (2014). *Computational paralinguistics: Emotion, affect and pesonality in speech and language processing.* West Sussex: Wiley & Sons, Ltd.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). THE INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. *Proceedings of the 14th Annual Conference of the International Speech Communication Association, Interspeech 2013 (ISCA, Lyon, France, 2013),* 148–152.

Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., et al. (2014). THE INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load. *Proceedings of the 15th Annual Conference of the International Speech Communication Association, Interspeech 2014 (ISCA, Singapore, 2014),* 427–431.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Honig, F., Orozco-Arroyave, J. R., et al. (2015). THE INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition. *Proceedings of the 16th Annual Conference of the International Speech Communication Association, Interspeech 2015 (ISCA, Dresden, Germany, 2015),* 478–482.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., et al. (2016). THE INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language. *Proceedings of the 17th Annual Conference of the International Speech Communication Association, Interspeech 2016 (ISCA, San Francisco, USA, 2016),* 2001–2005.

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., et al. (2017). THE INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring. *Proceedings of the 18th Annual Conference of the International Speech Communication Association, Interspeech 2017 (ISCA, Stockholm, Sweden, 2017),* 3442–3446.

Sheinkopf, S. J., Iverson, J. M., Rinaldi, M. L., & Lester, B. M. (2012). Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder. *Autism Research, 5*, 331–339.

Suter, B., Treadwell-Deering, D., Zoghbi, H. Y., Glaze, D. G., & Neul, J. L. (2014). Brief report: MECP2 mutations in people without Rett syndrome. *Journal of Autism and Developmental Disorders, 44*, 703–711.

Tams-Little, S., & Holdgrafer, G. (1996). Early communication development in children with Rett syndrome. *Brain and Development, 18*, 376–378.

Tarquinio, D. C., Hou, W., Neul, J. L., Lane, J. B., Barnes, K. V., O'Leary, H. M., et al. (2015). Age of diagnosis in Rett syndrome: Patterns of recognition among diagnosticians and risk factors for late diagnosis. *Pediatric Neurology, 52*, 585–591.

Townend, G. S., Bartl-Pokorny, K. D., Sigafoos, J., Curfs, L. M., Bölte, S., Poustka, L., et al. (2015). Comparing social reciprocity in preserved speech variant and typical Rett syndrome during the early years of life. *Research in Developmental Disabilities, 43*, 80–86.

Wermke, K., Teiser, J., Yovsi, E., Kohlenberg, P. J., Wermke, P., Robb, M., et al. (2016). Fundamental frequency variation within neonatal crying: Does ambient language matter? *Speech, Language and Hearing, 19*, 211–217.

Zhang, D., Kaufmann, W. E., Sigafoos, J., Bartl-Pokorny, K. D., Krieber, M., Marschik, P. B., et al. (2017). Parents' initial concerns about the development of their children later diagnosed with fragile X syndrome. *Journal of Intellectual & Developmental Disability, 42*, 114–122.