Does She Speak RTT? Towards an Earlier Identification of Rett Syndrome Through Intelligent Pre-linguistic Vocalisation Analysis

Florian B. Pokorny\textsuperscript{1,2,3}, Peter B. Marschik\textsuperscript{1,3,4}, Christa Einspieler\textsuperscript{1}, Björn W. Schuller\textsuperscript{5,6}

\textsuperscript{1}Research Unit iDN – interdisciplinary Developmental Neuroscience, Institute of Physiology, Center for Physiological Medicine, Medical University of Graz, Austria
\textsuperscript{2}Machine Intelligence & Signal Processing group, Technische Universität München, Germany
\textsuperscript{3}Brain, Ears & Eyes – Pattern Recognition Initiative (BEE-PRI), BioTechMed-Graz, Austria
\textsuperscript{4}Centre of Neurodevelopmental Disorders (KIND), Karolinska Institutet, Stockholm, Sweden
\textsuperscript{5}Chair of Complex & Intelligent Systems, University of Passau, Germany
\textsuperscript{6}Machine Learning Group, Department of Computing, Imperial College London, UK

florian.pokorny@medunigraz.at

\url{http://dx.doi.org/10.21437/Interspeech.2016-520}

Abstract

For many years, an apparently normal early development has been regarded as a main characteristic of Rett syndrome (RTT), a severe progressive neurodevelopmental disorder almost exclusively affecting girls/females. The speech-language domain represents a key domain for the clinical diagnosis of RTT, which usually happens around three years of age. Recent studies have built upon the assumption that this domain is already affected in the prodromal period. Aiming to find RTT-specific speech-language atypicalities on signal level as early acoustic markers, we analysed more than 16 hours of home video recordings of 4 girls later diagnosed with RTT and 4 typically developing girls aged 6 to 12 months. We segmented a total of 4 678 pre-linguistic vocalisations. A comprehensive set of acoustic features was extracted from the vocalisations as basis for the classification paradigm RTT versus typical development. A promising mean unweighted recognition accuracy of 76.5\% was achieved using linear kernel support vector machines and 4-fold leave-one-speaker-pair-out cross-validation. To the best of our knowledge, this is the first approach to automatically identify infants later diagnosed with RTT based on acoustic characteristics of pre-linguistic vocalisations. Our findings may build the basis for facilitating earlier identification and thus an avenue for an earlier entry into intervention.

Index Terms: Rett syndrome, early detection, infant vocalisation analysis, speech-language pathology

1. Introduction

Rett syndrome (RTT) is a severe neurodevelopmental disorder almost exclusively occurring in females [1, 2] with a prevalence of 1:5000 to 1:10000 live female births (rare disease) [3]. It was first described in 1966 by the Austrian neuropaediatrician Andreas Rett [4]. More than 30 years later, mutations in the X-linked gene encoding Methyl-CpG-binding protein 2 (MeCP2) were identified as the cause for RTT in most (but not necessarily in all) cases [5]. RTT has long been believed to be characterised by an apparently normal early development followed by a period of profound neurological regression. This regression affects – among other neurodevelopmental functions such as purposeful hand use or cognitive skills – the use of expressive language [6, 7, 8, 2, 9]. Thus, the detection of the speech language dysfunctions is one of the key domains for the diagnosis of RTT, which still is a clinical diagnosis at first, confirmed by genetic testing [2]. Recent observational studies have found increasing evidence that this domain is already affected in the pre-regression period, before diagnosis, challenging the paradigm of normal early development (e.g., [10, 11, 12, 13, 14]).

Besides delays in the achievement of certain speech-language milestones, or even their non-achievement in girls with RTT (e.g., [13]), verbal characteristics of infant vocalisations have been found to bear qualitative atypicalities (e.g., [11]). For six females with RTT, Marschik et al. [11] reported three different atypical vocalisation characteristics, of (i) pressed, (ii) inspiratory, and (iii) high-pitched crying-like quality. Interestingly, these vocalisations appeared intermittently with typical vocalisations as early as from the 7\textsuperscript{th} month of life onwards. Marschik et al. further stated: “The intermittent character of normal versus abnormal behaviors might contribute to an early identification of children with possible genetic mutations, and provides evidence that speech-language functions are abnormal from the very beginning.” [11, p.1]

In this study, we took on the challenge to start a very first attempt to itemise potential RTT-specific speech-language atypicalities on signal level. We aimed to make a step towards enabling an earlier identification of RTT on the basis of objective acoustic signal parameters in pre-linguistic vocalisations using machine learning methodology in order to facilitate an earlier entry into intervention for individuals with RTT.

2. Can we hear RTT?

In a case study (pilot listening experiment [Pokorny et al. 2016, in preparation]), we aimed to quantify the intermittent character of typical and atypical early vocalisations in RTT by means of listeners’ vocalisation assessments. Therefore, we presented more than 300 pre-linguistic vocalisations of a girl later diagnosed with RTT separately to five professionals in the fields of speech-language pathology, developmental psychology, and/or developmental physiology. We asked the experts to rate whether they perceived the vocalisations typical or atypical. Atypical vocalisations should be qualitatively further classified into: rhythm, timbre, pitch, or other acoustic parameters as predominantly deviant feature (multiple answers were allowed).

About half of the vocalisations were rated atypical by at least one listener. Only nine vocalisations were consenta-
neously rated as atypical exhibiting characteristics as described in [11]. Figure 1 shows spectrograms of (i) a pressed vocalisation with inharmonic overtone structure rated as atypical in timbre by all five listeners, (ii) a vocalisation with distinctive inspiratory phases rated as atypical in timbre by four listeners and in rhythm by three listeners, (iii) a high-pitched crying-like vocalisation rated as atypical in pitch by four listeners, and (iv) a vocalisation with harmonic overtone structure rated as typical by all five listeners.

However, the listening experiment’s low overall inter-rater reliability ($\kappa = 0.2$) indicated that atypicalities in early vocalisations may be hardly reliably identified by human listeners. This calls for machine-driven approaches to objectively define acoustic phenomena in early vocalisations in RTT.

## 3. Methods

In this study, we retrospectively focussed on pre-linguistic vocalisations of the second half year of life, as this striking and diagnostically relevant [15] period covers the transition from the use of first syllabic sounds and canonical babbling to the production of first meaningful words in typical development [16, 17, 18, 19, 20, 21].

### 3.1. Material

We reviewed more than 1 000 minutes of home video recordings of the second half year of life of 4 female infants later diagnosed with RTT and 4 typically developing (TD) female infants. The recordings were made by the infants’ parents in typical family settings (e.g., playing situations, feeding, bathing) and during special family events (e.g., birthday parties). At the time of recording, the parents of the individuals later diagnosed with RTT were not aware of their daughters’ medical condition. All eight participants stem from German-speaking families, who provided the audio-video material for the purpose of scientific analysis. The Institutional Review Board of the Medical University of Graz approved the method of retrospective audio-video analysis.

### 3.2. Segmentation

Manual segmentation of vocalisations was carried out using the video coding tool Noldus Observer XT. A vocalisation was defined as an utterance underlying a vocal breathing group [22]. We did not include vocalisations that could not be ascribed to a video’s participating infant with absolute certainty (e.g., in settings with more than one infant of about the same age present). We further excluded vegetative sounds such as breathing sounds, sneezes, hiccups, smacking sounds, etc. Relevant vocalisation detection in the video as well as raw segmentation were done by two female and two male research assistants following an intensive instruction by the first author. Prior to inclusion in this study, the first author verified each pre-selected vocalisation for validity and carried out the fine segmentation process. As itemised in Table 1, a total of 2 199 pre-linguistic vocalisations could be segmented within the footage of infants later diagnosed with RTT. The material of TD infants contained 2 479 pre-linguistic vocalisations. All segmented vocalisations were exported as audio tracks (44.1 kHz, 16 bit, 1 channel, PCM) for further analysis.

### 3.3. Analysis

To build the basis for vocalisation analysis/classification on signal level, feature extraction was carried out using the open-source tool kit openSMILE [23] in its current release [24]. Describing one of the most comprehensive standardised sets available, we extracted the official baseline features of the Interspeech Computational Paralinguistics Challenges (ComParE) used since 2013 (e.g., [25, 26]) from each vocalisation. The set comprises 6 373 higher-level features representing statistical functionals of a wide range of acoustic time-, spectral-, and/or energy-based short-term low-level descriptors’ trajectories and their derivatives.

In order to investigate the binary classification paradigm RTT versus typical development, we split our data into training, development, and test partitions. Due to the small number of infants per class (4 RTT versus 4 TD), we decided for evaluating classification performance via a 4-fold leave-one-speaker-pair-
out cross-validation scheme. For each of the four validation runs the training subsets contained the vocalisations of two infants per class, the development and test subsets each contained the vocalisations of one infant per class. The vocalisations of each infant were included in the test partition exactly one time. To ensure maximal class balancing within the test subsets for each of the four validation runs, we pairwise matched the numbers of segmented vocalisations per infant. Accordingly, the infant later diagnosed with RTT with the highest number of segmented vocalisations was grouped with the TD infant with the highest number of segmented vocalisations, etc. The detailed partitioning is given in Table 2. Training subsets with class imbalances exceeding the ratio 3:2 (training subset for third and fourth run, and training+development subset for third run) were upsampled applying (Weka’s implementation of) the synthetic minority oversampling technique (SMOTE) [27].

To study the influence of infant-specific feature value distributions on classification performance, on the one hand, we (Method A) did not normalise/standardise feature values speaker/infant-dependently prior to classification. On the other hand, we (Method B) performed speaker/infant normalisation, i.e., all features were speaker/infant-dependently normalised to the interval [0,1], and finally, we (Method C) applied speaker/infant standardisation, i.e., all features were speaker/infant-dependently standardised to have zero mean and unit variance, before passing feature data to the classifier.

Known not to be sensitive to feature overfitting, we applied linear kernel support vector machines (SVMs) as classifier by means of the widely used data mining tool kit Weka [28]. The kernel complexity parameter C was optimised for each of the four validation runs within \(\{1, 10^2, 10^3, 10^4, 10^5\}\) on the basis of the respective development subset. Subsequently, the training subset and the development subset were combined to a final training subset for validation on the basis of the respective test subset for each of the four runs. For SVM training, we selected the sequential minimal optimisation algorithm.

4. Results

For Method A, i.e., classification without preceding speaker/infant normalisation/standardisation, we achieved a mean class unweighted recognition accuracy (UA) ± standard deviation of 59.4% ± 11.4% over the four validation runs. 75.0% of the vocalisations of infants later diagnosed with RTT were correctly identified as class RTT, but the vocalisations of TD infants were classified about one half each as class TD and RTT. Method B, i.e., classification with preceding speaker/infant normalisation, performed significantly better (at a significance level of \(\alpha = 0.001\) using a one-sided z-test) than the other two methods. Here, 63.3% of the vocalisations of infants later diagnosed with RTT and 75.2% of the vocalisations of TD infants were correctly classified. In two of the four validation runs weighted and unweighted accuracies higher than 90% were achieved. Method C, i.e., classification with preceding speaker/infant standardisation, performed with significantly lower accuracy (at a significance level of \(\alpha = 0.001\) using a one-sided z-test) than the other methods. Here, 63.3% of the vocalisations of infants later diagnosed with RTT were correctly identified, but also 64.9% of the vocalisations of TD infants were (incorrectly) assigned to the RTT class. None of the four validation runs using Method C exceeded the level of random guessing leading to a mean UA of 49.8% ± 0.9%. Detailed results for the effective methods A and B including both mean class weighted and unweighted accuracies and standard deviations, as well as overall confusion matrices are given in Table 3.

5. Discussion

On our dataset, the basic feasibility of an automatic recognition of pre-linguistic vocalisations produced by infants later diagnosed with RTT was supported. However, data preprocessing in terms of infant-specific data normalisation/standardisation prior to classification was a crucial step. On our data, classification with preceding infant normalisation significantly outperformed classification with preceding infant standardisation and classification neither with normalisation nor with standardisation.

Cause for the significant differences among the three methods may be the intermittently occurring atypical vocalisations in infants later diagnosed with RTT that are, to date, hardly objectively documented on signal level. Another influential role may play the inhomogeneity of our material with respect to audio quality and background noise in the home videos.

From a neurodevelopmental and linguistic point of view, we implemented a straight-forward top-down approach by not differentiating between (i) vocalisations produced in different developmental stages (we integrated all pre-linguistic vocalisations produced during the second half year of life to one dataset, and the case with speaker normalisation (Method B). WA and UA are given in [%]. Values are rounded to one decimal place.
even though the exact age in months was known for each video clip), and (ii) different vocalisation types (e.g., quasi-resonant nuclei versus canonical babbling [16]) to a certain extent correlates with some outcome should be available to ensure that the data also contain a sufficient number of atypical vocalisations in case of RTT. In practical use, a prediction tool could be implemented working fully automatically on the basis of home video material, or e.g., on the basis of audio material recorded 24 hours with a microphone attached to an infant’s clothing. For such a tool, at least a reliable infant voice activity detection component would be required as the tool’s input stage. Anyway, an approach like ours should never be applied directly for diagnostic purposes, it should rather raise a probable cause to initiate a diagnostic cascade (i.e., neurological, neuropaediatric, and genetic testing).

6. Conclusions and Outlook

Families with infants with RTT usually undergo periods of uncertainty with respect to their children’s development until the diagnosis of RTT is made. In this study, we investigated the existence of potential acoustic signal level parameters in early vocalisations exploitable for facilitating an earlier identification of individuals with RTT in the context of the classification paradigm RTT versus typical development. We achieved promising recognition accuracies on the basis of 4,678 prelinguistic vocalisations of four infants later diagnosed with RTT and four TD infants when performing infant-dependent feature normalisation prior to classification. As far as we know, our study testified for the very first time that an objective approach to automatically identify infants with RTT in the first year of life based on vocalisation acoustics on signal level may be feasible and impact future earlier identification procedures.

Even though we built our study upon a considerable number of vocalisations per class, the overall number of infants per class was critically low. Therefore, we aim to expand our study by adding vocalisations of a high number of both infants later diagnosed with RTT and TD infants matched with respect to family language. We further aim to extend our age range of interest to the whole first year of life including the period that captures the first occurrences of melodic-modulated sounds/cooing in typical development [16, 17]. Based on a more extensive dataset, we shall treat different vocalisation types as well as vocalisations produced in different developmental stages in separate.

From a technological point of view, in future work a special focus should be put on the selection of acoustic features relevant or specific for RTT versus typical development or other neurodevelopmental disorders. A more detailed evaluation of different feature processing/feature normalisation/standardisation procedures should be carried out. Finally, the strengths and weaknesses of different classification approaches in context of this special area of application should be identified.

7. Acknowledgements

The authors acknowledge funding from the National Bank of Austria (OeNB; P16830), BioTechMed-Graz, the General Movements Trust, and the EU’s H2020 Programme via RIA #688835 (DE-ENIGMA). Special thanks go to Jorge Luis Moye, Sergio Roccabado, Adriana Villarroel, and Claudia Zitta for their assistance in the vocalisation segmentation process. Moreover, we want to thank Katrin D. Bartl-Pokorny for consultancy in linguistic matters. The authors are grateful to all parents who provided us with home video material for scientific analysis.
8. References


