Journal of Nonverbal Behavior
https://doi.org/10.1007/s10919-020-00332-4

ORIGINAL PAPER

Efficient Collection and Representation of Preverbal Data in Typical and Atypical Development

Florian B. Pokorny[1,2] · **Katrin D. Bartl-Pokorny**[1] · **Dajie Zhang**[1,3,4] · **Peter B. Marschik**[1,3,4,5] · **Dagmar Schuller**[6] · **Björn W. Schuller**[6,7,8]

Abstract
Human preverbal development refers to the period of steadily increasing vocal capacities until the emergence of a child's first meaningful words. Over the last decades, research has intensively focused on preverbal behavior in typical development. Preverbal vocal patterns have been phonetically classified and acoustically characterized. More recently, specific preverbal phenomena were discussed to play a role as early indicators of atypical development. Recent advancements in audio signal processing and machine learning have allowed for novel approaches in preverbal behavior analysis including automatic vocalization-based differentiation of typically and atypically developing individuals. In this paper, we give a methodological overview of current strategies for collecting and acoustically representing preverbal data for intelligent audio analysis paradigms. Efficiency in the context of data collection and data representation is discussed. Following current research trends, we set a special focus on challenges that arise when dealing with preverbal data of individuals with late detected developmental disorders, such as autism spectrum disorder or Rett syndrome.

Keywords Preverbal development · Data collection · Data representation · Infancy · Developmental disorders · Intelligent audio analysis

Introduction

Of all significant changes during infancy, the acquisition of verbal abilities is one of the most striking phenomena for parents, clinicians, and researchers. Several vocal transformations take place between a newborn's first cry and the production of first meaningful words, usually around the end of the first year of life (e.g., Nathani et al. 2006; Oller 1980, 2000; Papoušek 1994; Stark 1980, 1981; Stark et al. 1993). Vocal patterns with salient characteristics emerging in this preverbal period are—amongst others—cooing around the third month post-term age (Nathani et al. 2006; Oller 1980; Stark 1980) and canonical babbling

✉ Florian B. Pokorny
florian.pokorny@medunigraz.at

✉ Björn W. Schuller
schuller@IEEE.org

Extended author information available on the last page of the article

Published online: 02 March 2020

around the eighth month post-term age (Nathani et al. 2006; Oller 1980; Papoušek 1994; Stark 1980). Preverbal development is related to fundamental processes of infant brain development (e.g., Dehaene-Lambertz 2017) in combination with anatomical and voice-physiological changes (e.g., Holzki et al. 2018).

For almost 40 years, typical preverbal development has been intensively studied (e.g., Locke 1995; Oller 1980, 2000; Stark 1981; Stark et al. 1993). Besides seeking for appropriate schemes to phonetically categorize preverbal behavior (e.g., Nathani et al. 2006; Oller 1980; Papoušek 1994; Stark 1980, 1981), research has focussed on defining milestones of preverbal development, such as the above-mentioned onset of canonical babbling (Harold and Barlow 2013). Delays in reaching specific milestones or their non-achievement have been discussed as potential early indicators of atypical development (e.g., Lang et al. 2019; Lohmander et al. 2017; Oller et al. 1998). In fact, a number of developmental disorders are associated with deficits in the speech-language domain. Some of these disorders can be recognized at birth or even earlier. For example, infants with Down syndrome have a characteristic physical appearance and specific morphological features leading to clinical and genetic diagnosis (World Health Organization 2019). Other developmental disorders are lacking apparent early signs. These disorders are identified when certain physical features become apparent, developmental milestones are not achieved or their achievement is delayed, or behavioral/neurofunctional deviances reach a certain threshold to allow clinical diagnosis. The late clinical manifestation of these disorders currently leads to an accurate diagnosis of affected children usually not before toddlerhood (Baio et al. 2018; Christensen et al. 2016; Marschik et al. 2016; Sicherman et al. 2018; Tarquinio et al. 2015).

One of these best known late detected developmental disorders is autism spectrum disorder (ASD; American Psychiatric Association 2013; World Health Organization 2019). The current prevalence of ASD is 1 in 59 children in the USA (Baio et al. 2018) with a higher occurrence in males than in females (e.g., Baio et al. 2018), and a recurrence risk of up to 18% for younger siblings of children already diagnosed with ASD (e.g., Bhat et al. 2014; Bölte 2014). The exact etiology of ASD is still unknown (e.g., American Psychiatric Association 2013; Bhat et al. 2014; Bölte et al. 2019). Another late detected developmental disorder that is associated with deficits in the speech-language domain is Rett syndrome (RTT; World Health Organization 2019), occurring in about 1 in 10,000 live female births (Laurvick et al. 2006). De novo mutations in the X chromosome-linked gene *MECP2* were identified as its main cause (Amir et al. 1999). In most cases, affected male individuals do not survive the prenatal period (Tokaji et al. 2018). In both ASD and RTT, as well as in a number of other late detected developmental disorders, diagnosis criteria include deficits in the socio-communicative and speech-language domains (American Psychiatric Association 2013; Neul et al. 2010). Research is increasingly focusing on preverbal phenomena in these disorders in order to find early markers for an earlier identification of affected individuals. However, many of the existing studies were based upon limited datasets. For example, some individuals with RTT were found not to acquire certain preverbal capacities, such as cooing, babbling, or the production of proto-words (Bartl-Pokorny et al. 2013; Marschik et al. 2013, 2014a). The latter are word-like vocalizations that do not yet conform to target language concerning articulation and/or lexical meaning (Kauschke 2000; Papoušek 1994). The preverbal behavior of individuals with RTT was reported to have an intermittent character of apparently typical and atypical vocalization patterns, such as sequences of high-pitched crying-like phonation, phonation with an ingressive pulmonic airstream, or phonation with a pulmonic airstream of excessive pressure (e.g., Marschik et al. 2009, 2012a, 2013, 2014a; Pokorny et al. 2018). Repeatedly documented preverbal atypicalities in individuals with ASD are a late onset of canonical babbling, a low canonical babbling ratio, a

comparably low rate of vocalization, and monotonous intonation patterns (e.g., Chericoni et al. 2016; Patten et al. 2014; Paul et al. 2011; Roche et al. 2018).

While the first intensive efforts of preverbal sound categorization were made in the 1980s (e.g., Oller 1980; Stark 1980, 1981), the first acoustic data representations were used to describe recorded preverbal behavior at a signal level. For example, investigations by Kent and Murray (1982) were based on small sets of extracted acoustic features, such as vocalization duration, variations of the vocal tract source excitation, or formant frequencies. Bloom et al. (1999) calculated frequency–amplitude slopes in preverbal syllabic sounds as a measure of nasality. In addition, the fundamental frequency (F0), i.e., the lowest frequency of vocal fold oscillation, has always been a frequently extracted feature—in voice analytics in general, but also for acoustic preverbal sound characterization (e.g., Keating and Buhr 1978; Kent and Murray 1982; Petroni et al. 1994; Robb et al. 1989). A number of recent studies on crying vocalizations also reported on F0-related preverbal atypicalities in individuals with ASD (e.g., Esposito and Venuti 2010; Esposito et al. 2013; Sheinkopf et al. 2012).

With the rising age of high-performance computing, the request for efficient preverbal behavior analysis has also grown. The discipline of intelligent audio analysis (Schuller 2013), which deals with the combination of advanced audio signal processing and machine learning technology, provided the methodological tools for the automatic retrieval of preverbal sound-related (meta-)information. Specific tasks are, e.g., the automatic differentiation of preverbal vocalization types (e.g., Schuller et al. 2019), the automatic detection of infant distress (e.g., Chang and Li 2016; Chang et al. 2015; Lavner et al. 2016; Rodriguez and Caluya 2017; Schuller et al. 2018), or the automatic recognition of the medical condition of the infant who produces the vocalization (e.g., Orlandi et al. 2012; Pokorny et al. 2016a, 2017). However, different learning tasks require different learning strategies. These involve different 'optimal' representations of the collected preverbal data.

Building on our experience in collecting and analyzing preverbal data of typical and atypical development (e.g., Bartl-Pokorny et al. 2013; Marschik et al. 2012a, b, 2013, 2014a, b, 2017; Pokorny et al. 2016a, 2017, 2018), we provide a methodological overview of current strategies for the collection and representation of preverbal data for intelligent audio analysis purposes. Exemplified on the basis of empirical data, we will especially focus on application-oriented challenges and constraints that have to be considered when dealing with preverbal data of individuals with late detected developmental disorders. Finally, data collection and representation of preverbal behavior will be discussed in context of efficiency.

Data Collection

The process of data collection for acoustic studies on typical and atypical preverbal development involves the recording of preverbal behavior itself by means of a recording device and the acquisition of study-relevant meta-information, such as the participant's chronological age at the time of recording and his or her medical history. In particular, the participant's developmental outcome in terms of typical versus atypical has to be carefully documented. This, however, raises a crucial methodological issue for the collection of preverbal data of infants with a late detected developmental disorder. Dependent on a priori study criteria and potential future applications building on a study's findings, three possible strategies for the collection of preverbal data can be distinguished: data collection (1) 'in the

past', (2) in the lab, or (3) 'in the wild' (Fig. 1). Referring to a study's timeline, these data collection strategies can be categorized as either retrospective or prospective approaches, respectively.

Prospective data collection can be carried out in the lab or in the participant's natural environment, here referred to as 'in the wild'. Infants fulfilling specific inclusion criteria, can be systematically recruited and prospectively studied in a longitudinal study design with scheduled outcome assessments. Such procedures are well-suited for investigations in neurotypical cohorts, but also allow for the collection of preverbal data in late detected developmental disorders with known a priori risk factors, a high prevalence, and/or a high recurrence risk. This applies, for example, to ASD. Prospective high-risk ASD studies usually build upon the high familial recurrence of ASD (e.g., Bhat et al. 2014; Bölte 2014) and follow younger siblings of children with an existing ASD diagnosis. This is done in various research institutions and multi-center approaches mainly across Europe and the USA (e.g., Bölte et al. 2013; Ozonoff et al. 2011a).

A prospective collection of preverbal data of individuals with rare late detected developmental disorders with low familial recurrence is hardly possible. This applies, for example, to RTT being mainly caused by a rare, spontaneous genetic mutation (Amir et al. 1999). Participants must be sought out and recruited at a time at which their developmental outcome is already known, i.e., a diagnosis has already been made. Data are collected retrospectively, i.e., 'in the past', for example, via parents' associations, social networks, email blasts, or by cooperating with clinical expertise centers.

In the Past

In most cases, retrospectively collected data including audio of preverbal behavior are home videos provided by the participants' families (e.g., Boterberg et al. 2019; Einspieler and Marschik 2019; Marschik and Einspieler 2011; Roche et al. 2018). Audio–video

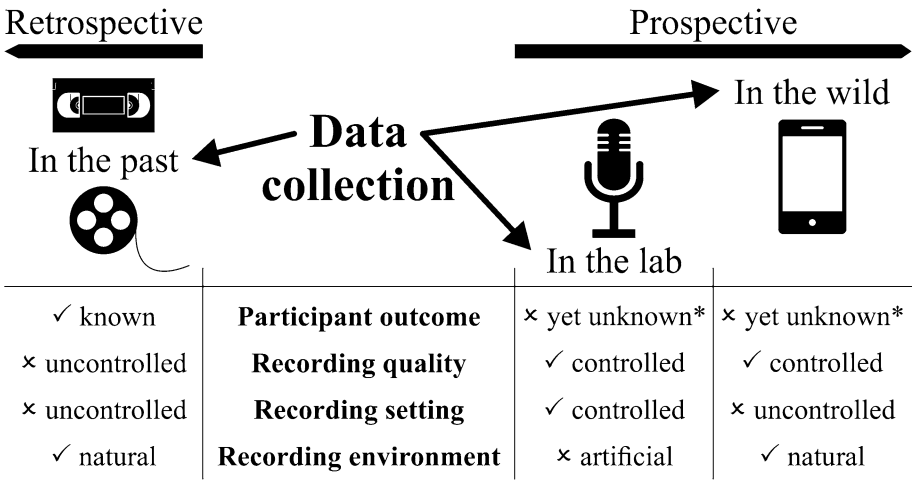

Fig. 1 Overview of strategies for collecting preverbal data and strategy comparison on the basis of study design criteria chronologically referring to the period of data collection. *For typically developing participants or participants with a late detected developmental disorder

recordings were typically taken by the participants' parents, who were not aware of their children's potential later diagnosis. Usually, they used standard video cameras or, in recent years, to a greater extent smartphones. On the basis of home videos, participants can be retrospectively studied in their natural environment during everyday situations, such as interactive playing, during typical family routines, such as bathing, diaper-changing, or feeding situations, or during special events, such as familial celebrations. However, neither recording quality nor recording setting are a priori controlled study design criteria. Inhomogeneous recording formats and microphone/camera positions, in combination with uncontrollable background noise conditions, cause a large acoustic variation within a home video dataset. For example, our Graz University Audiovisual Research Database for the Interdisciplinary Analysis of Neurodevelopment (GUARDIAN), contains home video data with a total runtime of several months. The earliest clips were taken in the 1960s. Their original video formats range from analogue standards, such as Super 8 or VHS, to state-of-the-art digital standards (Pokorny et al. 2016b). Studies based on preverbal data collected 'in the past' typically lack of standardization and reproducibility, and thus have limited comparability to other studies. Furthermore, the interactive setting that potentially influences an infant's preverbal behavior can only be partly evaluated as other persons or interactive (media) devices might be out of the recording device's range, e.g., out of the video camera's view. Another limitation is the potential absence of a specific behavior of interest within a collected dataset, such as the production of specific preverbal vocalization types. Moreover, the exact age of the recorded participant in a clip is often unknown and can only be roughly estimated. Regarding documentation of participant outcome, different diagnostic assessments are used (e.g., ADOS, ABC, or CARS for individuals with ASD) and often varying details on symptom severity and comorbidities are available. Nevertheless, data collection 'in the past' offers a unique opportunity to 'listen back' to preverbal phenomena. For the moment, it still represents one of the best strategies we have for studying the preverbal development of individuals with developmental disorders with a late clinical manifestation (e.g., Adrien et al. 1993; Crais et al. 2006; Marschik and Einspieler 2011; Palomo et al. 2006; Saint-Georges et al. 2010).

In the Lab

The prospective collection of preverbal data in a laboratory allows for controllable acoustic room conditions and a careful a priori definition of the optimal recording device/quality and the recording setting. The recording procedure is, thus, reproducible and enables the best possible data comparability. Moreover, participants can be invited according to a predefined study paradigm and appropriate time plan to collect preverbal data from exactly defined age windows. Exemplarily, the CRIED (Cry Recognition In Early Development) database of 5587 mood-annotated preverbal vocalizations from 20 individuals (Schuller et al. 2018) was collected at the Medical University of Graz, Austria, in the framework of a longitudinal study on typical neuro-functional and neuro-behavioral changes in early infancy (Marschik et al. 2017). Participants were recorded 7 times on a bi-weekly basis with the first session at 4 weeks $\pm$ 2 days and the last session at 16 weeks $\pm$ 2 days post-term age. During the sessions, the infants were lying awake in supine position in a cot with various recording devices including microphones positioned around (Marschik et al. 2017). Data collection in the lab generally implies that participants are recorded in an artificial environment. A popular laboratory setting in which some aspects of the participants' natural environment are simulated, is a parent–child-interaction setting. In such a setting,

parents are, e.g., instructed to play with their children with a predefined set of toys in a standardized recording room as they would do at home (e.g., Pokorny et al. 2017).

Apart from lab recordings targeting the collection of infant vocalizations, data originally recorded for other reasons can also represent a valuable source for compiling preverbal behavior analysis datasets. Such data could be, for example, audio–video recordings of developmental assessments routinely made for documentation or evaluation purposes.

In the Wild

The prospective collection of preverbal data under real-world conditions, i.e., in the participants' natural environment, permits predefining a desired recording quality. The only limiting requirement is that the recording device has to be placed 'in the wild'. This plays a role for the selection of an optimal device/microphone type, e.g., with respect to its dimensions or power supply. Typically, camcorders are used for audio–video data collection 'in the wild', while mobile wave recorders are used for audio data collection only. Alternatively, nowadays data can be easily collected via parental smartphones. In recent years, naturalistic data for studies on preverbal typical and atypical development have been frequently collected with the LENA® (Language Environment Analysis; https://www.lena.org [as of 5 April 2019]; Xu et al. 2008a, b) system (e.g., Oller et al. 2010; Swanson et al. 2018). The LENA® system enables the long-term recording of infants' vocalizations by means of a child-safe audio recording device attached to a vest (Xu et al. 2008a, b). Except for the position of the recording device, recording setting parameters, such as position, posture, and activity of the participant, are uncontrolled. Acoustic background noise conditions can only be marginally controlled, e.g., by giving parents respective instructions. All things considered, comparability of data collected 'in the wild' is very limited. Studies are only partially reproducible.

Of course, combinations of different data collection strategies are conceivable, such as data collection in the lab around a specific first age window of interest and data collection 'in the wild' around a specific second age window of interest. Besides the actual recording of preverbal data, both prospective approaches, i.e., data collection in the lab and 'in the wild', offer an easy way to acquire relevant additional information on the participants from the caregivers without running the risk of memory bias (e.g., Marschik 2014; Ozonoff et al. 2011b; Palomo et al. 2006; Zhang et al. 2017). However, in prospective approaches the participants' developmental outcomes are not yet known at the time of data collection and have to be evaluated in follow-up assessments.

Data Representation

For intelligent audio analysis purposes, an approriate acoustic representation has to be derived from the collected audio data that contain preverbal behavior. This process can be regarded as acoustic behavior modeling with the goal to gain distinct information characterizing a preverbal phenomenon. The appropriateness of a specific data representation highly depends on the intended type of subsequent analysis or learning task. Basically, the representation of acoustic data can be either based on predefined features ("Feature-Based representation" section) or automatically learned in context of a specific task ("Representation learning" section). However, some preprocessing steps need to be carried out before.

Preprocessing

First of all, preverbal behavior has to be identified and segmented within the recorded audio material. Segmentation describes the process of setting boundaries around identified preverbal behavior in order to create meaningful analyzable units. Depending on the intended analysis, classification task, or target application, meaningful units can be, for example, single phones, but also extensive phrases. In many cases, preverbal behavior is segmented into utterances that are denominated as preverbal vocalizations and generally lie in between the duration of a phone and a phrase (Lynch et al. 1995; Nathani and Oller 2001). There are two common procedures for utterance segmentation. The first procedure is based on setting segment boundaries at vocal pauses, i.e., phases without vocal activity, exceeding a predefined pause duration (e.g., Nathani and Oller 2001). Considering the physiological process of voice production as being linked to respiratory activity, the second segmentation procedure relies on the condition that each vocalization has to be assigned to a distinct vocal breathing group. Segment boundaries, thus, coincide with phases of ingressive breathing (e.g., Nathani and Oller 2001; Oller and Lynch 1992).

The LENA® system, for example, provides fully automatic vocalization segmentation and diarization alongside recorded audio. Diarization in this context means that the system automatically indicates if a vocalization was most probably produced by the infant of interest or, for example, by a caregiver (Xu et al. 2008a). However, automatic segmentation and diarization is prone to errors, especially if the audio material includes acoustic disturbances as presumed for data collected 'in the wild' or 'in the past' (Pokorny et al. 2016b). Therefore, preverbal data segmentation and diarization is still often done manually or at least semi-automated. In a semi-automated procedure, automatically created segments are manually checked for correctness. However, false negatives are missed out.

Another essential preprocessing step is the annotation of preverbal behavior. Each included segment needs to be labeled with study-relevant meta-information. On the one hand, there are behavior-independent variables implicating that the labels stay the same for all preverbal behaviors of, e.g., one and the same participant in one and the same recorded clip. Examples are participant-dependent variables, such as gender, family language, medical condition/later diagnosis, or age at the time of recording. In addition, data collected in non-standardized settings, i.e., 'in the wild' or 'in the past', usually need to be annotated for all factors that might have acoustically or physically influenced the recorded preverbal behavior, such as scene type (playtime, bathtime, mealtime, etc.), location (indoor vs. outdoor), presence and type of background noise, interactive setting, participant posture, participant physical restriction, or use of a pacifier. On the other hand, there are behavior-dependent variables, such as the number of syllabic elements within an infant vocalization, or the preverbal vocalization type based on a vocalization classification scheme, e.g., the Stark Assessment of Early Vocal Development-Revised (SAEVD-R; Nathani et al. 2006).

An optional signal-related data preprocessing step for subsequent intelligent audio analysis purposes is audio normalization, i.e., setting the audio signal's maximum amplitude to a defined level (e.g., Eyben et al. 2013b, c; Schuller et al. 2016). This is done to guarantee a comparable amplitude range across all included raw audio data of a dataset and can especially be beneficial when analyzing sets of data that were recorded with different recording devices at different recording levels. Dependent on the nature of the collected raw material, audio normalization is either done segment-wisely, i.e., in separate for each vocalization, or globally prior to the segmentation process.

Feature-Based Representation

The traditional audio representation for subsequent learning tasks builds upon the extraction of acoustic features from the collected and preprocessed audio data. A segment of preverbal behavior is transformed into a single multidimensional feature vector or a chronological series of feature vectors acoustically modeling the behavior's trajectory. Acoustic features are mathematical descriptors defined on the basis of a priori expert knowledge with the intention to characterize the content of an audio signal in a compact, informative, but preferably non-redundant way (Schuller 2013). Optimal acoustic features for preverbal behavior modeling or speech applications in general (e.g., Schuller and Batliner 2014) might differ from optimal acoustic features for other applications. Basically, the transformation of audio into a meaningful feature representation implies a reduction of information (Schuller 2013). The number of features typically extracted for subsequent speech-related learning tasks varies from less than 100 (e.g., Deng et al. 2017; Eyben et al. 2016) to several thousand (e.g., Schuller et al. 2013). Dependent on the applied machine learning algorithm, large feature vectors are either directly used as input representation for learning/classification, or reduced to an optimized feature subset by means of feature selection algorithms (Cai et al. 2018). 'Optimized' in this context means that features that do not or only marginally contain information relevant for the intended learning/classification task are identified and discarded.

Natural audio signals are usually time variant, i.e., they change over time (Deller et al. 1993). This also holds true for recorded preverbal behavior. Consequently, the extraction of acoustic features is usually carried out on the basis of short, window-function-weighted, overlapping time frames. Within each time frame, audio information is considered to be quasi-stationary (Deller et al. 1993). Features derived from this first level of signal subsampling are denominated as low-level descriptors (LLDs; Schuller 2013; Schuller and Batliner 2014).

There are a number of well-established acoustic LLDs, such as the F0. However, different mathematical and methodological ways of extracting one and the same feature exist. Moreover, the exact sequence of calculation steps including all set adjustments for deriving a specific feature from an audio signal is usually not specified in publications. This hampers the comparability of absolute feature values reported across different studies. Therefore, open-source feature extraction tools steadily grow in popularity in the research community. Such tools allow for reproducible feature calculation throughout different labs around the world. Furthermore, standard feature sets can be provided. A popular open-source feature extraction tool kit is openSMILE by audEERING™ GmbH (https://audeering.com/technology/opensmile/ [as of 8 April 2019]). openSMILE is based on C++, enables feature extraction both offline and in real-time, and comes along with standard feature sets well-proven for various application areas (Eyben et al. 2010, 2013a).

The so far most comprehensive standard feature set for openSMILE is the ComParE set. It is widely known, as it represented the official baseline feature set of the 2013–2019 Computational Paralinguistics ChallengEs (e.g., Schuller et al. 2013, 2018, 2019) carried out in connection with the Annual Conferences of the International Speech Communication Association (INTERSPEECH conferences). The ComParE set comprises 6373 acoustic supra-segmental features, so-called higher-level descriptors (HLDs). These are statistical functionals computed for the trajectories of a wide range of acoustic time-, energy-, and/or spectral/cepstral-based LLDs as well as their derivatives (Schuller et al. 2013). In contrast, the most current standard feature set for openSMILE

is the Geneva Minimalistic Acoustic Parameter Set (GeMAPS). It was launched in 2016 by Eyben et al. (2016) and represents a comparatively small set of only 62 frequency-, energy-, and spectral-related features. Its extended version, the eGeMAPS, contains 26 additional frequency- and spectral-related descriptors summing up to a total of 88 features. The features of the eGeMAPS were carefully selected based on (a) their theoretical and practical relevance for automatic voice analysis applications, including clinical applications, and (b) their proven value in previous studies in the related fields (Eyben et al. 2016). By default, LLD trajectories for both the ComParE set and the eGeMAPS are extracted on the basis of overlapping time frames of 60 ms at a step size of 10 ms. LLD contours are smoothed over an interval of three frames (unvoiced-voiced transitions in selected LLD contours excepted). HLDs are then calculated for the smoothed LLD contours of the entire input segment resulting in exactly one vector of 6373 or 88 feature values per segmented preverbal behavior, respectively. A data representation like this is suited for static learners. For dynamic learners, i.e., algorithms operating on the acoustic content's variations over time, HLD trajectories have to be calculated per segment on an appropriate time basis, or the LLD contours are used for representing each segmented behavior.

Both the ComParE set and the eGeMAPS have already successfully been applied for demonstrating the feasibility of an automatic preverbal vocalization-based differentiation between typically developing (TD) individuals and individuals with late detected developmental disorders (Pokorny et al. 2016a, 2017). Pokorny et al. (2016a) extracted the ComParE features from 4678 retrospectively collected preverbal vocalizations of four TD infants and four infants later diagnosed with RTT. A promising mean unweighted accuracy of 76.5% was achieved in the binary vocalization classification paradigm RTT versus TD using linear kernel support vector machines as classifier in a four-fold leave-one-speaker-pair-out cross-validation scheme. The study by Pokorny et al. (2017) built upon data collected at a participants' age of 10 months in a semi-standardized parent–child-interaction setting within the prospective ASD high-risk protocol EASE (Early Autism Sweden; http://www.earlyautism.se [as of 3 April 2019]). Both the eGeMAPS features and the LLDs of the ComParE set were extracted from 684 preverbal vocalizations of 10 TD individuals and 10 individuals later diagnosed with ASD. The eGeMAPS features were used as data representation for static classification by means of linear kernel support vector machines. In contrast, dynamic modeling was investigated by applying a neural network classifier on the basis of the LLDs of the ComParE set. Three-fold cross-validation was carried out for performance evaluation. Either approach led to 15 of 20 infants correctly assigned to group ASD or TD.

A popular data representation that builds upon extracted LLD or HLD contours is the Bag-of-Audio-Words (BoAW) representation (e.g., Lim et al. 2015; Pancoast and Akbacak 2014; Pokorny et al. 2015; Schmitt et al. 2016). It relies on the quantization of input feature trajectories according to a learned codebook. Finally, data are represented as histograms of the previously generated sequence of 'audio words'. A recently introduced open-source tool kit for the extraction of Bo(A)W from arbitrary, multidimensional input feature trajectories is openXBOW (https://github.com/openXBOW [as of 9 April 2019]; Schmitt and Schuller 2017). Since 2017, openXBOW has been used in addition to openSMILE as official baseline feature extractor of the annual INTERSPEECH ComParE Challenges (e.g., Schuller et al. 2017). Within the INTERSPEECH 2018 ComParE Challenge (Schuller et al. 2018), openXBOW was for the first time used to generate BoAW from preverbal behavior: In the Crying Sub-Challenge, vocalizations of the CRIED database had to be automatically told apart according to three mood-related vocalization classes, namely (1) neutral/positive

sounds, (2) crying sounds, and (3) fussing sounds, which can be regarded as transition behavior between (1) and (2) (Schuller et al. 2018).

Further feature-related data representations have been recently used for intelligent audio analysis applications, such as Bag-of-Context-Aware-Words (e.g., Han et al. 2018). However, to the best of our knowledge, such representations have played a minor role for the acoustic modeling of preverbal behavior in context of typical or atypical development so far.

To meet the requirements of some classifiers (Bishop 2006), a common feature post-processing step is feature normalization or standardization. This means that feature values are rescaled to be located within a defined interval, such as [0, 1], or to have zero mean and unit variance, respectively. Feature normalization/standardization can be carried out globally, i.e., based on all instances/vocalizations of a dataset. Alternatively, it can be done in separate for semantic instance sub-partitions, e.g., participant/infant-wisely.

Representation Learning

In contrast to feature-based modeling, representation learning seeks for automatically deriving a data representation that makes subsequent learning tasks easier (Goodfellow et al. 2016). Thereby, a trade-off between reaching beneficial properties for subsequent classification and keeping as much information about the input signal as possible, has to be found. Representation learning can be (a) supervised, i.e., based on class-labeled data, (b) unsupervised, i.e., based on unlabeled data, or (c) semi-supervised, i.e., based on a usually small amount of labeled and a usually large amount of unlabeled data (Goodfellow et al. 2016). Following general directions in machine learning, especially deep representation learning has received increasing attention (Bengio et al. 2013) and proven powerful in a number of intelligent audio analysis scenarios in recent years. An open-source tool kit for deep unsupervised representation learning in the audio domain was introduced by Freitag et al. (2017): AUDEEP is a Python tool based on the widely used open-source machine learning library TENSORFLOW™ (https://www.tensorflow.org [as of 9 April 2019]; Abadi et al. 2016). It allows for learning data representations from audio time series by means of a recurrent sequence-to-sequence autoencoder approach (Freitag et al. 2017). Complementing the brute-force feature extraction tools openSMILE and openXBOW, in 2018 AUDEEP was elected as open-source representation learning tool kit for official baseline evaluation within the ongoing series of INTERSPEECH ComParE Challenges (Schuller et al. 2018, 2019). Thus, in that year, AUDEEP was initially applied to learn representations from preverbal data. Within the Crying Sub-Challenge of the 2018 INTERSPEECH ComParE Challenge (Schuller et al. 2018), representations were automatically derived from Mel-scale spectrograms generated for the vocalizations of the CRIED database. Similarly, based on input spectrograms, Cummins et al. (2017) proposed a deep spectrum feature representation for emotional speech recognition in children. Here, spectrograms are passed through a deep image classification convolutional neural network. The intended representation is then derived from the activations of the last fully-connected layer of the network (Cummins et al. 2017). However, empirical studies using this approach for modeling preverbal data are still outstanding. (Deep) representation learning in general, still describes a very young methodology for processing preverbal data of typical and atypical development.

A method in which the optimal data representation and the subsequent classifier are learned simultaneously from the audio signal is end-to-end learning. End-to-end learning

models were also applied to the CRIED database in the framework of the 2018 INTERSPEECH ComParE Challenge (Schuller et al. 2018). However, this method is not further treated here, as in end-to-end learning representation learning can not be regarded independently from the subsequent classification algorithm.

Discussion

Preverbal human development has been a popular research field for almost 40 years (e.g., Nathani et al. 2006; Oller 1980, 2000; Papoušek 1994; Stark 1980). Researchers and clinicians from different disciplines, such as linguists, neurologists, pediatricians, psychologists, physiologists, speech-language therapists, and engineers, have synergetically characterized preverbal phenomena in typical and atypical development over the years. Objective empirical investigations of preverbal behavior have, however, always required the audio recording of preverbal behavior as well as the recorded behavior's meaningful representation for subsequent analyses. A number of strategies for the collection and representation of preverbal data exist. Some of these strategies have even been successfully applied in intelligent audio analysis paradigms on infants with late detected developmental disorders testing automatic earlier identification (e.g., Oller et al. 2010; Orlandi et al. 2012; Pokorny et al. 2016a, 2017; Xu et al. 2009). The question of what makes collection and representation of preverbal data efficient, has to be answered in the context of the specific learning task and its target area of application.

Regarding data collection, efficiency might be quantified as the proportion between (a) quality and quantity of collected preverbal data, and (b) collection efforts in time and money. In this context, data quality refers to recording quality on the one hand, and to influences on the recorded behavior of interest on the other hand. Quantity refers to preverbal behavior of interest within a recording, not to the overall recording duration. Both quality and quantity depend on recording setting and environment. The general demand of recording participants within their natural environment in order not to get their behavior influenced by the experimental setup should be considered as a function of age. Human vocal behavior has been discussed to be endogenously generated in early infancy by specific neural networks in the brain stem (Barlow et al. 2010), so-called central pattern generators (e.g., Barlow and Estep 2006; Fénelon et al. 1998). We thus might assume that an artificial laboratory environment will hardly influence the preverbal behavior of newborns or participants in the early postnatal period. When trying to collect representative data of participants' natural behavior during late infancy, the recording environment may then play a crucial role. However, apart from age-related aspects and the limitation that specific data collection strategies are not suitable for specific participant groups, e.g., prospective approaches for participants with rare late detected developmental disorders, each data collection strategy has its strengths and weaknesses with respect to the above discussed efficiency criteria (Fig. 1). A careful hypotheses-oriented a priori study design planning under consideration of available time and budget may guarantee maximum efficiency in data collection.

'Efficient' in representing preverbal data may mean that collected data are best possibly reduced to specific information that is needed to reach high data interpretability in subsequent analyses or high performance in subsequent learning tasks. Of course, required processing power and computing time have to be taken into account, especially in the context of potential real-time applications with devices of limited computing capacities, such as

smartphones. For example, when using a large brute-force feature set and a classifier that is not prone to overfitting, a feature selection procedure can be left out. However, the choice of efficient data representation not only depends on a subsequent analysis or learning task. It also depends on the nature of collected data and, thus, on the applied data collection strategy. For example, audio normalization as a representation pre-processing step might not make sense for recordings that include background noise events exceeding the level of preverbal behaviors.

In Table 1, we present recognition results that were achieved in a binary preverbal vocalization classification paradigm RTT versus TD. Data for this experiment were collected 'in the past'. 3502 preverbal vocalizations were extracted from home video recordings of 3 TD individuals and 3 individuals later diagnosed with RTT. The recordings were made in the individuals' second half year of life, respectively. Some alternatives with regard to data representation were tried out, namely (1) segment-wise audio normalization yes versus no, (2) ComParE set versus eGeMAPS, and (3) infant-wise versus global feature normalization. Then, we trained, optimized, and evaluated linear kernel support vector machines in a three-fold leave-one-speaker-pair-out cross-validation scheme. Finally, we stored the predictions of each iteration and calculated the unweighted average recall (UAR) for the whole dataset. The best performance—a UAR of .879—was achieved when (1) processing normalized audio segments, (2) using the eGeMAPS, and (3) applying infant-wise feature normalization. Generally, there were only marginal differences between the scenarios with and without audio normalization. However, on average the scenarios without audio normalization reached slightly better results. This might be due to the nature of data used for this experiment—home video data involving everyday background noise. Segment-wise audio normalization of background noise-prone recordings may have caused disadvantageous level imbalances between segments with background noise exceeding the level of the preverbal behavior of interest and segments without background noise. The greatest effect in this experiment was related to the choice of feature set. The eGeMAPS clearly outperformed the ComParE set here. Finally, infant-wise feature normalization turned out more beneficial compared to global feature normalization.

Table 1 Comparison of system configurations regarding audio normalization, used feature set, and feature normalization strategy (infant-wise: normalization in separate for all vocalizations of an infant, respectively; global: normalization over all instances of the dataset; see also last paragraph of "Feature-based representation" section) by means of the unweighted average recall (UAR) achieved in a binary vocalization classification paradigm RTT versus TD

Audio normalization	Feature set	Feature normalization	$UAR_{RTT\ vs.\ TD}$
–	ComParE	infant-wise	.356
–	ComParE	global	.399
–	eGeMAPS	infant-wise	.832
–	eGeMAPS	global	.674
✓	ComParE	infant-wise	.372
✓	ComParE	global	.281
✓	eGeMAPS	infant-wise	.879
✓	eGeMAPS	global	.535

UAR values are rounded to three decimal places

ComParE Computational Paralinguistics ChallengE (feature set), *eGeMAPS* extended Geneva Minimalistic Acoustic Parameter Set, *RTT* Rett syndrome, *TD* typical development, ✓ applied, – not applied

For answering the question of better to apply a brute-force feature set, such as the eGeMAPS, or features automatically learned from data, it should be considered that deep representation learning methods usually require very large datasets (Bengio et al. 2013) and that automatically learned features are hardly acoustically or voice-physiologically interpretable anymore. The latter aspect might be relevant if clinical conclusions shall be drawn from the data representation.

In this paper, we provided an overview and discussed current strategies for collecting and acoustically representing preverbal data for intelligent audio analysis paradigms on typical and atypical early development. A special focus was given to methodological requirements and limitations for studies on the preverbal behavior of individuals with late detected developmental disorders. With the rising age of deep learning, significant advancements have been made in intelligent audio analysis in recent years. However, the application of state-of-the-art audio processing and machine learning methods for preverbal behavior analysis, especially in a clinical context of automatically differentiating orthology and pathology, has only just started. Progress in the acoustic modeling of preverbal data over the coming years is thus warranted. In addition, future collection of preverbal data will most probably be influenced by industrial and social trends, such as by the combination of smartphone development and people's increasing affinity for self-tracking during daily life (e.g., Lupton and Smith 2018). Consumers, thus also parents, get more and more equipped with affordable powerful tools facilitating an extensive documentation of their own, but also of their children's lives. Masses of new data will thereby be generated. In a couple of years, these data will be available for being collected 'in the past'.

Acknowledgements Open access funding provided by the Austrian Science Fund (FWF). The authors want to express their gratitude to all families who participated in prospective studies on early development or provided video material for retrospective scientific analysis. Without their support, research on preverbal behavior would not have been possible. Special thanks go to Maximilian Schmitt for assistance in data analysis. Parts of this work were funded by the Austrian National Bank (OeNB; P16430), the Austrian Science Fund (FWF; P25241), and BioTechMed-Graz, Austria.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings OSDI* (pp. 265–283).

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 and Adolescent Psychiatry, 32*(3), 617–626.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. Washington, DC: American Psychiatric Pub.

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, encoding methyl-CpG-binding protein 2. *Nature Genetics, 23*(2), 185–188.

Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., et al. (2018). Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network. *MMWR Surveillance Summaries, 67*(6), 1–23.

Barlow, S. M., & Estep, M. (2006). Central pattern generation and the motor infrastructure for suck, respiration, and speech. *Journal of Communication Disorders, 39*(5), 366–380.

Barlow, S. M., Radder, J. P. L., Radder, M. E., & Radder, A. K. (2010). Central pattern generators for orofacial movements and speech. In S. M. Brudzynski (Ed.), *Handbook of behavioral neuroscience* (pp. 351–369). London: Academic Press.

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*(10), 3133–3138.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828.

Bhat, S., Acharya, U. R., Adeli, H., Bairy, G. M., & Adeli, A. (2014). Autism: Cause factors, early diagnosis and therapies. *Reviews in the Neurosciences, 25*(6), 841–850.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bloom, K., Moore-Schoenmakers, K., & Masataka, N. (1999). Nasality of infant vocalizations determines gender bias in adult favorability ratings. *Journal of Nonverbal Behavior, 23*(3), 219–236.

Bölte, S. (2014). Is autism curable? *Developmental Medicine and Child Neurology, 56*(10), 927–931.

Bölte, S., Girdler, S., & Marschik, P. B. (2019). The contribution of environmental exposure to the etiology of autism spectrum disorder. *Cellular and Molecular Life Sciences, 76*(7), 1275–1297.

Bölte, S., Marschik, P. B., Falck-Ytter, T., Charman, T., Roeyers, H., & Elsabbagh, M. (2013). Infants at risk for autism: A European perspective on current status, challenges and opportunities. *European Child and Adolescent Psychiatry, 22*(6), 341–348.

Boterberg, S., Charman, T., Marschik, P. B., Bölte, S., & Roeyers, H. (2019). Regression in autism spectrum disorder: A critical overview of retrospective findings and recommendations for future research. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2019.03.013.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing, 300,* 70–79.

Chang, C.-Y., Hsiao, Y.-C., & Chen, S.-T. (2015). Application of incremental SVM learning for infant cries recognition. In *Proceedings NBiS* (pp. 607–610).

Chang, C.-Y., & Li, J.-J. (2016). Application of deep learning for recognizing infant cries. In *Proceedings ICCE-TW* (pp. 1–2).

Chericoni, N., de Brito Wanderley, D., Costanzo, V., Diniz-Goncalves, A., Leitgel Gille, M., Parlato, E., et al. (2016). Pre-linguistic vocal trajectories at 6–18 months of age as early markers of autism. *Frontiers in Psychology, 7,* 1595.

Christensen, D. L., Baio, J., Braun, K. V. N., Bilder, D., Charles, J., Constantino, J. N., et al. (2016). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network. *MMWR Surveillance Summaries, 65*(3), 1–23.

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.

Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., & Schuller, B. W. (2017). An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings ACM-MM* (pp. 478–484).

Dehaene-Lambertz, G. (2017). The human infant brain: A neural architecture able to learn language. *Psychonomic Bulletin & Review, 24*(1), 48–55.

Deller, J. R., Hansen, J. H., & Proakis, J. G. (1993). *Discrete-time processing of speech signals*. New York: Macmillan Pub. Co.

Deng, J., Xu, X., Zhang, Z., Frühholz, S., Grandjean, D., & Schuller, B. (2017). Fisher kernels on phase-based features for speech emotion recognition. In K. Jokinen & G. Wilcock (Eds.), *Dialogues with social robots: Enablements, analyses, and evaluation* (pp. 159–203). Singapore: Springer.

Einspieler, C., & Marschik, P. B. (2019). Regression in Rett syndrome: Developmental pathways to its onset. *Neuroscience and Biobehavioral Reviews, 98,* 320–332.

Esposito, G., Nakazawa, J., Venuti, P., & Bornstein, M. H. (2013). Componential deconstruction of infant distress vocalizations via tree-based models: A study of cry in autism spectrum disorder and typical development. *Research in Developmental Disabilities, 34*(9), 2717–2724.

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*(8), 1093–1102.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., et al. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7*(2), 190–202.

Eyben, F., Weninger, F., Gro, F., & Schuller, B. (2013a). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings ACM-MM* (pp. 835–838).

Eyben, F., Weninger, F., & Schuller, B. (2013b). Affect recognition in real-life acoustic conditions—A new perspective on feature selection. In *Proceedings INTERSPEECH* (pp. 2044–2048).

Eyben, F., Weninger, F., Squartini, S., & Schuller, B (2013c). Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies. In *Proceedings ICASSP* (pp. 483–487).

Eyben, F., Wollmer, M., & Schuller, B. (2010). openSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proceedings ACM-MM* (pp. 1459–1462).

Fénelon, V. S., Casasnovas, B., Simmers, J., & Meyrand, P. (1998). Development of rhythmic pattern generators. *Current Opinion in Neurobiology, 8*(6), 705–709.

Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., & Schuller, B. (2017). auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research, 18*(1), 6340–6344.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.

Han, J., Zhang, Z., Schmitt, M., Ren, Z., Ringeval, F., & Schuller, B. (2018). Bags in bag: Generating context-aware bags for tracking emotions from speech. In *Proceedings INTERSPEECH* (pp. 3082–3086).

Harold, M. P., & Barlow, S. M. (2013). Effects of environmental stimulation on infant vocalizations and orofacial dynamics at the onset of canonical babbling. *Infant Behavior and Development, 36*(1), 84–93.

Holzki, J., Brown, K. A., Carroll, R. G., & Coté, C. J. (2018). The anatomy of the pediatric airway: Has our knowledge changed in 120 years? A review of historic and recent investigations of the anatomy of the pediatric larynx. *Pediatric Anesthesia, 28*(1), 13–22.

Kauschke, C. (2000). *Der Erwerb des frühkindlichen Lexikons: Eine empirische Studie zur Entwicklung des Wortschatzes im Deutschen*. Tübingen: Narr.

Keating, P., & Buhr, R. (1978). Fundamental frequency in the speech of infants and children. *Journal of the Acoustical Society of America, 63*(2), 567–571.

Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *Journal of the Acoustical Society of America, 72*(2), 353–365.

Lang, S., Bartl-Pokorny, K. D., Pokorny, F. B., Garrido, D., Mani, N., Fox-Boyer, A. V., et al. (2019). Canonical babbling: A marker for earlier identification of late detected developmental disorders? *Current Developmental Disorders Reports, 6*(3), 111–118.

Laurvick, C. L., Klerk, N. D., Bower, C., Christodoulou, J., Ravine, D., Ellaway, C., et al. (2006). Rett syndrome in Australia: A review of the epidemiology. *Journal of Pediatrics, 148*(3), 347–352.

Lavner, Y., Cohen, R., Ruinskiy, D., & IJzerman, H. (2016). Baby cry detection in domestic environment using deep learning. In *Proceedings ICSEE* (pp. 1–5).

Lim, H., Kim, M. J., & Kim, H. (2015). Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation. In *Proceedings INTERSPEECH* (pp. 3325–3329).

Locke, J. L. (1995). *The child's path to spoken language*. Cambridge: Harvard University Press.

Lohmander, A., Holm, K., Eriksson, S., & Lieberman, M. (2017). Observation method identifies that a lack of canonical babbling can indicate future speech and language problems. *Acta Paediatrica, 106*(6), 935–943.

Lupton, D., & Smith, G. J. (2018). 'A much better person': The agential capacities of self-tracking practices. In B. Ajana (Ed.), *Metric culture: Ontologies of self-tracking practices* (pp. 57–75). Bingley: Emerald Publishing Limited.

Lynch, M. P., Oller, D. K., Steffens, M. L., & Buder, E. H. (1995). Phrasing in prelinguistic vocalizations. *Developmental Psychobiology, 28*(1), 3–25.

Marschik, P. B. (2014). The pivotal role of parents in documenting early development. *North American Journal of Medical Sciences, 6*(1), 48–49.

Marschik, P. B., Bartl-Pokorny, K. D., Sigafoos, J., Urlesberger, L., Pokorny, F., Didden, R., et al. (2014a). Development of socio-communicative skills in 9- to 12-month-old individuals with fragile X syndrome. *Research in Developmental Disabilities, 35*(3), 597–602.

Marschik, P. B., Bartl-Pokorny, K. D., Tager-Flusberg, H., Kaufmann, W. E., Pokorny, F., Grossmann, T., et al. (2014b). Three different profiles: Early socio-communicative capacities in typical Rett syndrome, the preserved speech variant and normal development. *Developmental Neurorehabilitation, 17*(1), 34–38.

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*(6), 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*(6), 958–961.

Marschik, P. B., Einspieler, C., & Sigafoos, J. (2012a). Contributing to the early detection of Rett syndrome: The potential role of auditory Gestalt perception. *Research in Developmental Disabilities, 33*(2), 461–466.

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*(4), 1236–1239.

Marschik, P. B., Pini, G., Bartl-Pokorny, K. D., Duckworth, M., Gugatschka, M., Vollmann, R., et al. (2012b). Early speech-language development in females with Rett syndrome: Focusing on the preserved speech variant. *Developmental Medicine and Child Neurology, 54*(5), 451–456.

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*(5), 43.

Marschik, P. B., Sigafoos, J., Einspieler, C., Enzinger, C., & Bölte, S. (2016). The interdisciplinary quest for behavioral biomarkers pinpointing developmental disorders. *Developmental Neurorehabilitation, 19*(2), 73.

Nathani, S., Ertmer, D. J., & Stark, R. E. (2006). Assessing vocal development in infants and toddlers. *Clinical Linguistics & Phonetics, 20*(5), 351–369.

Nathani, S., & Oller, D. K. (2001). Beyond ba-ba and gu-gu: Challenges and strategies in coding infant vocalizations. *Behavior Research Methods, Instruments, & Computers, 33*(3), 321–330.

Neul, J. L., Kaufmann, W. E., Glaze, D. G., Christodoulou, J., Clarke, A. J., Bahi-Buisson, N., et al. (2010). Rett syndrome: Revised diagnostic criteria and nomenclature. *Annals of Neurology, 68*(6), 944–950.

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child phonology: Vol. 1. Production* (pp. 93–112). New York: Academic Press.

Oller, D. K. (2000). *The emergence of the speech capacity*. Mahwah: Lawrence Erlbaum Associates.

Oller, D. K., Eilers, R. E., Neal, A. R., & Cobo-Lewis, A. B. (1998). Late onset canonical babbling: A possible early marker of abnormal development. *American Journal on Mental Retardation, 103*(3), 249–263.

Oller, D. K., & Lynch, M. P. (1992). Infant vocalizations and innovations in infraphonology: Toward a broader theory of development and disorders. In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 509–536). Parkton: York Press.

Oller, D. K., Niyogi, P., Gray, S., Richards, J., Gilkerson, J., Xu, D., et al. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences, 107*(30), 13354–13359.

Orlandi, S., Manfredi, C., Bocchi, L., & Scattoni, M. (2012). Automatic newborn cry analysis: A non-invasive tool to help autism early diagnosis. In *Proceedings EMBC* (pp. 2953–2956).

Ozonoff, S., Iosif, A. M., Young, G. S., Hepburn, S., Thompson, M., Colombi, C., et al. (2011a). Onset patterns in autism: Correspondence between home video and parent report. *Journal of the American Academy of Child and Adolescent Psychiatry, 50*(8), 796–806.

Ozonoff, S., Young, G. S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., et al. (2011b). Recurrence risk for autism spectrum disorders: A baby siblings research consortium study. *Pediatrics, 128*(3), e488.

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

Pancoast, S., & Akbacak, M. (2014). Softening quantization in bag-of-audio-words. In *Proceedings ICASSP* (pp. 1370–1374).

Papoušek, M. (1994). *Vom ersten Schrei zum ersten Wort: Anfänge der Sprachentwicklung in der vorsprachlichen Kommunikation*. Bern: Hans Huber.

Patten, E., Belardi, K., Baranek, G. T., Watson, L. R., Labban, J. D., & Oller, D. K. (2014). Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency. *Journal of Autism and Developmental Disorders, 44*(10), 2413–2428.

Paul, R., Fuerst, Y., Ramsay, G., Chawarska, K., & Klin, A. (2011). Out of the mouths of babes: Vocal production in infant siblings of children with ASD. *Journal of Child Psychology and Psychiatry, 52*(5), 588–598.

Petroni, M., Malowany, A. S., Johnston, C. C., & Stevens, B. J. (1994). A new, robust vocal fundamental frequency (F0) determination method for the analysis of infant cries. In *Proceedings CBMS* (pp. 223–228).

Pokorny, F. B., Bartl-Pokorny, K. D., Einspieler, C., Zhang, D., Vollmann, R., Bölte, S., et al. (2018). Typical vs. atypical: Combining auditory Gestalt perception and acoustic analysis of early vocalisations in Rett syndrome. *Research in Developmental Disabilities, 82,* 109–119.

Pokorny, F., Graf, F., Pernkopf, F., & Schuller, B. (2015). Detection of negative emotions in speech signals using bags-of-audio-words. In *Proceedings WASA/ACII* (pp. 879–884).

Pokorny, F. B., Marschik, P. B., Einspieler, C., & Schuller, B. W. (2016a). Does she speak RTT? Towards an earlier identification of Rett syndrome through intelligent pre-linguistic vocalisation analysis. In *Proceedings INTERSPEECH* (pp. 1953–1957).

Pokorny, F. B., Peharz, R., Roth, W., Zöhrer, M., Pernkopf, F., Marschik, P. B., & Schuller, B. W. (2016b). Manual versus automated: The challenging routine of infant vocalisation segmentation in home videos to study neuro (mal) development. In *Proceedings INTERSPEECH* (pp. 2997–3001).

Pokorny, F. B., Schuller, B. W., Marschik, P. B., Brueckner, R., Nyström, P., N. Cummins, S., et al. (2017). Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach. In *Proceedings INTERSPEECH* (pp. 309–313).

Robb, M. P., Saxman, J. H., & Grant, A. A. (1989). Vocal fundamental frequency characteristics during the first two years of life. *Journal of the Acoustical Society of America, 85*(4), 1708–1717.

Roche, L., Zhang, D., Bartl-Pokorny, K. D., Pokorny, F. B., Schuller, B. W., Esposito, G., et al. (2018). Early vocal development in autism spectrum disorder, Rett syndrome, and fragile X syndrome: Insights from studies using retrospective video analysis. *Advances in Neurodevelopmental Disorders, 2*(1), 49–61.

Rodriguez, R. L., & Caluya, S. S. (2017). Waah: Infants cry classification of physiological state based on audio features. In *Proceedings ICSIIT* (pp. 7–10).

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.

Schmitt, M., Janott, C., Pandit, V., Qian, K., Heiser, C., Hemmert, W., & Schuller, B. (2016). A bag-of-audio-words approach for snore sounds' excitation localisation. In *Proceedings ITG Speech Communication* (pp. 230–234).

Schmitt, M., & Schuller, B. (2017). openXBOW: Introducing the Passau opensource crossmodal bag-of-words toolkit. *Journal of Machine Learning Research, 18*(96), 1–5.

Schuller, B. (2013). *Intelligent audio analysis*. Berlin: Springer.

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

Schuller, B. W., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., et al. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity. In *Proceedings INTERSPEECH* (pp. 2378–2382).

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., et al. (2017). The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, cold & snoring. In *Proceedings INTERSPEECH* (pp. 3442–3446).

Schuller, B. W., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., et al. (2016). The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, sincerity & native language. In *Proceedings INTERSPEECH* (pp. 2001–2005).

Schuller, B. W., Steidl, S., Batliner, A., Marschik, P. B., Baumeister, H., Dong, F., et al. (2018). The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & self-assessed affect, crying & heart beats. In *Proceedings INTERSPEECH* (pp. 122–126).

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. In *Proceedings INTERSPEECH* (pp. 148–152).

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*(5), 331–339.

Sicherman, N., Loewenstein, G., Tavassoli, T., & Buxbaum, J. D. (2018). Grandma knows best: Family structure and age of diagnosis of autism spectrum disorder. *Autism, 22*(3), 368–376.

Stark, R. E. (1980). Stages of speech development in the first year of life. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child phonology: Vol. 1. Production* (pp. 73–92). New York: Academic Press.

Stark, R. E. (1981). Infant vocalization: A comprehensive view. *Infant Mental Health Journal, 2*(2), 118–128.

Stark, R. E., Bernstein, L. E., & Demorest, M. E. (1993). Vocal communication in the first 18 months of life. *Journal of Speech, Language, and Hearing Research, 36*(3), 548–558.

Swanson, M. R., Shen, M. D., Wolff, J. J., Boyd, B., Clements, M., Rehg, J., et al. (2018). Naturalistic language recordings reveal "hypervocal" infants at high familial risk for autism. *Child Development, 89*(2), e60–e73.

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*(6), 585–591.

Tokaji, N., Ito, H., Kohmoto, T., Naruto, T., Takahashi, R., Goji, A., et al. (2018). A rare male patient with classic Rett syndrome caused by MeCP2_e1 mutation. *American Journal of Medical Genetics, 176*(3), 699–702.

World Health Organization. (2019). *International classification of diseases—Eleventh revision*. Geneva: World Health Organization.

Xu, D., Gilkerson, J., Richards, J., Yapanel, U., & Gray, S. (2009). Child vocalization composition as discriminant information for automatic autism detection. In *Proceedings EMBC* (pp. 2518–2522).

Xu, D., Yapanel, U., Gray, S., & Baer, C. T. (2008a). *The LENA language environment analysis system: The interpreted time segments (ITS) file*. Boulder: Infoture Inc.

Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., & Hansen, J. (2008b). Signal processing for young child speech language development. In *Proceedings of the 1st Workshop on Child, Computer and Interaction.*

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*(2), 114–122.

Affiliations

Florian B. Pokorny[1,2] · Katrin D. Bartl-Pokorny[1] · Dajie Zhang[1,3,4] · Peter B. Marschik[1,3,4,5] · Dagmar Schuller[6] · Björn W. Schuller[6,7,8]

[1] iDN – interdisciplinary Developmental Neuroscience, Division of Phoniatrics, Medical University of Graz, Graz, Austria

[2] Machine Intelligence & Signal Processing group (MISP), Chair of Human–Machine Communication, Technical University of Munich, Munich, Germany

[3] Department of Child and Adolescent Psychiatry and Psychotherapy, University Medical Center Göttingen, Göttingen, Germany

[4] Leibniz ScienceCampus Primate Cognition, Göttingen, Germany

[5] Center of Neurodevelopmental Disorders (KIND), Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

[6] audEERING GmbH, Gilching, Germany

[7] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany

[8] GLAM – Group on Language, Audio & Music, Department of Computing, Imperial College London, London, UK