TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Mensch-Maschine-Kommunikation

Towards an Automatic Earlier Recognition of Autism Spectrum Disorder, Fragile X Syndrome, and Rett Syndrome through Intelligent Pre-linguistic Vocalisation Analysis

Dipl.-Ing. Florian Pokorny

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr.sc.techn. Gerhard Kramer
Prüfer der Dissertation:	1. Prof. Dr.-Ing. habil. Björn W. Schuller
	2. Prof. Dr.-Ing. Werner Hemmert

Die Dissertation wurde am 13.11.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 07.06.2019 angenommen.

Abstract

Developmental disorders present with a variety of phenotypes characterised by deficits in different developmental domains that often become manifest right after birth. There are, however, developmental disorders without apparent early signs, such as autism spectrum disorder (ASD), fragile X syndrome (FXS), and Rett syndrome (RTT). The late onset of clinically detectable features in ASD, FXS, and RTT usually leads to an accurate diagnosis beyond infancy. Even though an earlier treatment would certainly be beneficial for affected individuals, research on early markers of these disorders is still underrepresented. Some studies reported on delays and early peculiarities in the socio-communicative domain – a key domain for the clinical diagnosis of ASD, FXS, and RTT. However, little to nothing is known about acoustic characteristics of disorder-related pre-linguistic vocalisations and their potential role for an automatic earlier identification of affected individuals.

The objective of this thesis was to evaluate the basic feasibility of an 'intelligent' pre-linguistic vocalisation-based approach for enabling an earlier recognition of ASD, FXS, and RTT. Therefore, early audio-video recordings of individuals later diagnosed with ASD, FXS, or RTT, and typically developing controls, were segmented for pre-linguistic vocalisations. Then, the differentiability of class-related vocalisations was analysed on the basis of an extended set of acoustic signal-level features extracted from the vocalisations. Finally, the automatic feature-based recognition of ASD, FXS, and RTT by means of machine learning methodology was investigated.

A high number of acoustic features was identified to significantly differ between vocalisations of individuals with different developmental outcomes. Moreover, promising recognition results were achieved for the automatic vocalisation-based identification of individuals with ASD, FXS, or RTT.

The generated results raise evidence for acoustic information in pre-linguistic vocalisations to reflect the integrity of the developing young nervous system. Thereby, this thesis may essentially contribute to a reliable earlier recognition of currently 'late diagnosed' developmental disorders and, ultimately, facilitate an earlier intervention.

Zusammenfassung

Die frühkindliche Entwicklung ist hinsichtlich zahlreicher Entwicklungsaspekte intensiv beforscht und verstanden. Ein Wissensdefizit – besonders in Bezug auf das erste Lebensjahr – besteht jedoch bei Entwicklungsstörungen, die zurzeit nach wie vor erst im Kleinkindalter diagnostiziert werden. Dazu zählen beispielsweise die Autismus-Spektrum-Störung (ASS), das Fragile-X-Syndrom (FXS) und das Rett-Syndrom (RTT). Wenngleich für diese Störungen bereits Verzögerungen und Auffälligkeiten im Zusammenhang mit der Sprachentwicklung attestiert werden konnten, so sind akustische Parameter prälinguistischer Lautäußerungen und deren Relevanz für die Früherkennung bislang noch weitgehend unerforscht.

Ziel der vorliegenden Arbeit war es, diese Wissenslücke zu schließen. So erfolgte die Segmentierung von prälinguistischen Lautäußerungen in Videoaufzeichnungen von Säuglingen, die später mit ASS, FXS oder RTT diagnostiziert wurden, sowie von sich normal entwickelnden Kontrollkindern. Aus diesen Lautäußerungen wurde anschließend eine Vielzahl an akustischen Signalparametern abgeleitet. Die Parameter wurden zum einen zur detaillierten Untersuchung gruppenspezifischer Lautunterschiede auf Signalebene, zum anderen zur automatischen Klassifikation von Lautäußerungen hinsichtlich der späteren Diagnose der Säuglinge basierend auf Methoden des Maschinellen Lernens verwendet.

Eine Reihe von Signaleigenschaften, die sich als valide für die Unterscheidung zwischen frühen Lautäußerungen von Säuglingen mit ASS, FXS oder RTT, sowie von sich normal entwickelnden Säuglingen herausgestellt hat, sowie vielversprechende Ergebnisse für die automatische lautbasierte Erkennung dieser Störungsbilder, sprechen dafür, dass sich anhand akustischer Informationen in prälinguistischen Lautäußerungen Aussagen über die Integrität des sich entwicklenden Nervensystems ableiten lassen.

Die in dieser Arbeit gewonnenen Erkenntnisse könnten einen wichtigen Beitrag für eine zuverlässige Früherkennung von derzeit erst im Kleinkindalter diagnostizierten Entwicklungsstörungen leisten und in weiterer Folge betroffenen Kindern künftig eine frühere Intervention ermöglichen.

Acknowledgement

In summer 2012 when I was still a master student, I had the pleasure and honour of working closely with Professor Dr. Björn Schuller (at that time) from the Technical University of Munich, Germany, for a few months. During that period, he not only gave me valuable advice on my master's thesis, but he also made me familiar with some important rules of the 'game of science' in general. In the end, he offered me to be my supervisor in case I would ever decide to write a doctoral thesis. In 2014 the time had come. I got a university assistant position at the Medical University of Graz, Austria, and could start with my highly interdisciplinary research endeavour as an external doctoral candidate at the Technical University of Munich under Professor Schuller's supervision. At this point, I would like to express my sincere gratitude to him for creating the opportunity of my doctorate, for his technical guidance, and his 24-hours-a-day, 7-days-a-week support over the last years primarily in form of Skype meetings and e-mails usually across many degrees of longitude, latitude, and time zones.

Moreover, I would like to thank my colleagues from Professor Schuller's team, especially Maximilian Schmitt from Professor Schuller's lab at the University of Augsburg, Germany, and Raymond Brueckner, who is another of Professor Schuller's external doctoral candidates at the Technical University of Munich, for cooperation in some experiments carried out in the framework of this thesis.

Next, I want to express my deep gratefulness to Professor Dr. Christa Einspieler and Associate Professor DDr. Peter Marschik for enabling me to get my current position at the Medical University of Graz and, thereby, making me a member of their Research Unit iDN – interdisciplinary Developmental Neuroscience, for giving me the generous opportunity to do an external doctorate on the basis of iDN's core data and research carried out in their lab, for always being open-minded for novel technical approaches to be integrated in iDN's future research direction, for introducing me into the 'scientific world', and for their unlimited efforts in teaching me the fundamentals of scientific working from the bottom up.

Many many thanks, at this point, also to the entire team of iDN for support, everlasting fruitful scientific exchange, and friendship over the last years. iDN has given me an understanding of how to join forces to achieve sustainable progress in science and medicine.

Another thank is due to Associate Professor Dr. Markus Gugatschka for housing our Research Unit iDN in his Divison of Phoniatrics and assuring a friendly research environment and a high-standard working infrastructure.

Then, I want to express my utmost appreciation to all the parents who provided our research unit with footage of their children as having been an essential basis for my approach to study early typical and atypical developmental trajectories. Moreover, I want to thank parent organisations that supported us with the collection of respective home video recordings, especially the Austrian Rett Syndrome Association and the *Interessensgemeinschaft Fragiles-X e.V.*, as well as our clinical and scientific cooperation partners who shared thesis-relevant audio-video material with us for analysis purposes. In this regard, special appreciation is due to Professor Michele Zappella from the Foundation for Autism Research, New York, USA, Dr. Alison Kerr from the University of Glasgow, UK, Professor Dr. Sven Bölte from the Karolinska Institutet, Stockholm, Sweden, Associate Professor Dr. Terje Falck-Ytter, also from the Karolinska Institutet as well as from the Uppsala University, Sweden, and Dr. Pär Nyström, also from the Uppsala University.

From a financial point of view, I want to acknowledge that parts of the research and short-term scientific missions carried out in the framework of this thesis were funded by the Austrian National Bank (*OeNB*; P16430), the Austrian Science Fund (*FWF*; P25241), the European Cooperation in Science and Technology (COST; Action BM1004), and BioTechMed-Graz, Austria.

Last but not least, I want to dedicate my deepest thankfulness to my whole family for having embedded my scientific ambitions into a solid and warm-hearted private backing structure. In particular, I want to thank my mother Irene, my stepfather August, my brother Christoph, my grandfather Leo, and my parents-in-law, Sonja and Mike, for always believing in my academic career, while at the same time providing an understanding of family to be the most valuable thing in life.

Nevertheless, demonstrating that professional and family/private lives can be synergetically combined when questing for a joint vision, my most heartfelt gratitude is due to Katrin who is not only my lovely wife and mother of our beautiful children, but also a highly respected colleague and long lasting precious scientific member of our Research Unit iDN. A thousand thanks to her for always being there for me and our kids, for her valuable advice in both private and professional matters, for often preventing me from frustrations by serving as a natural counterpart to my unlimited optimism, for exciting scientific discussions, as well as for many many hours between

10 p.m. and 2 a.m. over the last months sitting vis-à-vis at our dining table, drinking green tea, and proceeding with our doctoral theses. Katrin's support and the way she enriched my life over the last years is beyond words.

My final appreciation goes to my beloved children Noah and Flora not only for cheering me up after a hard day in the office, but also for teaching me lessons on early human development you can not find in papers or books and, thereby, further boosting my intrinsic motivation for unravelling some of the mysteries of the 'wonder' of early human life.

Graz, 24 September 2018 Florian Pokorny

Contents

Abstract i

Acknowledgement v

0 Preface 1

I Introduction 3

1 Early human development 5
1.1 Interdisciplinary developmental neuroscience 6
1.1.1 Machine learning in developmental neuroscience 7

2 Developmental disorders with a late clinical manifestation 9
2.1 Autism spectrum disorder 10
2.2 Fragile X syndrome 11
2.3 Rett syndrome 12

3 The speech-language domain 13
3.1 Typical pre-linguistic verbal development 14
3.2 Deviant pre-linguistic verbal development 15
3.3 Early vocalisation engineering 17

4 A technical tool for infant healthcare assistance 19
4.1 Motivation 19
4.2 Aims 20
4.3 Outline 22

II Methods **25**

5 Approach **27**
5.1 Retrospective audio-video analysis 28

6 Database **31**
6.1 GUARDIAN 32
6.2 Vocalisation segmentation 36
6.3 Vocalisation annotation 38

7 Intelligent vocalisation analysis **41**
7.1 Feature extraction 43
7.2 Feature processing 47
7.2.1 Feature analysis 47
7.3 Classification 48
7.3.1 Validation 52

III Experiments **57**

8 Infant voice activity detection **59**

9 Developmental disorder recognition **67**
9.1 Recognition of autism spectrum disorder 69
9.2 Recognition of fragile X syndrome 75
9.3 Recognition of Rett syndrome 81
9.3.1 How atypical does Rett syndrome sound? 91
9.4 Cross-syndrome recognition 96

IV Conclusion **109**

10 Feasibility of early recognition **111**
10.1 Limitations 114

11 Implications **117**
11.1 Technical perspectives 117
11.2 Clinical perspectives 118
11.3 Multi-domain fingerprint modelling 120

List of abbreviations **125**

List of symbols **131**

List of equations **133**

List of figures **135**

List of tables **137**

Bibliography **139**

0

Preface

The first year of human life – a period of significant neurofunctional adaptations and changes – represents one of the most striking periods for parents, but also for clinicians and researchers. Phenomena such as the transition from the first spontaneous movements to the first step, the emergence of social smiling, or the acquisition of verbal abilities from the first cry to the production of the first meaningful word, characterise the 'wonder' of early human development. However, there are parents, who cannot rejoice in their children to achieve specific developmental milestones as developmental processes appear delayed or even deviant. Starting to realise that something might be wrong, the fascinating period of early development often becomes a period of concerns and frustration and a starting point for diagnostic odysseys with uncertain outcomes.

This thesis is dedicated to all families – parents and children – who are affected by a developmental disorder with a current mean age of diagnosis beyond infancy, especially by autism spectrum disorder, fragile X syndrome, or Rett syndrome. By presenting a novel, interdisciplinary approach, I hope this work will contribute to a reliable earlier identification of children with 'late diagnosed' developmental disorders, facilitating an earlier entry into intervention...

Part I

Introduction

1

Early human development

Early human development is related to developmental phenomena within the broad continuum between foetal and early postnatal life. Science in context of early (human) development may be referred to as developmental (neuro)science. When dealing with developmental neuroscience, the ontogenetic adaptation paradigm postulated by the Austrian physician, zoologist, anthropologist, and founder of developmental neurology, Professor Dr. Heinz Friedrich Rudolf Prechtl, should be considered as a fundamental explanatory model. The concept of ontogenetic adaptation acknowledges a recurring sequence of developing neural structure realising itself through function that – for survival – has to meet the requirements of both the organism and its environment (e.g., in utero vs. ex utero). This, in turn, leads to the generation of adapted neural structure, which is prerequisite for differentiated neurofunction, and so forth. [1] For example, the formation of a foetus' diaphragm sets in at 8 weeks of gestation (postmenstrual age[1]) [2] and represents the anatomical precondition (structure) for the production of hiccup (function). First breathing movements usually follow 2–4 weeks later [3]. From around 30 weeks of gestation onwards, foetal breathing movements are the predominant type of diaphragmatic movements. By being required for the accomplishment of the morphological differentiation between pneumocytes of type I and type II [4], foetal breathing movements appear – amongst other reasons – to be essential for a normal foetal lung development allowing for functional postnatal respiration [5] and vocalisation. As a direct consequence of the onotogenetic adaptation paradigm, Professor Prechtl further emphasised that"[...] at different ages we are dealing with qualitatively different nervous systems. [...] If various developmental stages are studied in their own right, it is evident that an 'immature' nervous system does not exist." [1, p. 837]. This means that "[...] each foetus, infant, and child has a biologically different brain at different ages." [1, p. 837]

[1] Postmenstrual age refers to the time elapsed from the first day of the mother's last menstruation period.

The thesis at hand focusses on postnatal development in the first year of life, thus in reference to a term birth, on aspects related to brain development from the 41^{st} to the end of the 92^{nd}week postmenstrual age. Important developmental processes and milestones during this period are – amongst others – the development of the circadian rhythm [6], reaching [7], grasping [7], eye-hand coordination [8], social smiling [9], binocular vision [10], directional hearing [11], solid food intake [12], rolling over [13], unaided sitting [13], crawling [13], pulling up [13], unaided standing [13], walking [13], the use of gestures [14], behaviour imitation [15], and the emergence of the speech capacity from the onset of cooing around the third month post-term age via the onset of canonical babbling around the eighth month post-term age, to the production of the first meaningful words around the end of the first year of life [16, 17, 18, 19, 20, 21, 22]. Different developmental domains, such as the cognitive domain, the motor domain, or the socio-communicative/speech-language domain, allow different 'views' into the developing brain. The integrity of the young nervous system can be assessed by analysing domain-related parameters of interest (POI), such as the anticipation of coming events or object permanence in the cognitive domain, midline movements or laterality in the motor domain, or reaction to name calling or vocalisation type ratios in the socio-communicative/speech-language domain, to name but a few. Deviations from the typical structure-function-structure sequence of early development can manifest in a delayed, reduced, or qualitatively and/or quantitatively deviant functional repertoire in any subsequent developmental stage. The comprehensive understanding of typical and deviant developmental phenomena in the young human organism considering cross-domain and multi-POI relations, requires a complex and highly interdisciplinary scientific approach.

1.1 Interdisciplinary developmental neuroscience

Aiming for a multi-domain analysis of phenomena in typical and deviant early human development also envisaging potential healthcare applications, interdisciplinary developmental neuroscience is characterised by a close collaboration between different professional groups, namely between scientists of diverse expertise, physicians, and/or engineers that regard developmental phenomena from different angles, but communicate in a 'common language'. As illustrated in Figure 1.1, interdisciplinary developmental neuroscience can be understood as a synergy of various disciplines of basic research and applied clinical disciplines, such as gynaecology and obstetrics, neonatology, paediatrics, neurology, physiology, psychology, psychiatry, pathology, genetics, linguistics, speech-language pathology, biomedical engineering, sensor technology, and signal processing. Recent studies in this relatively young field, e.g., focussed on the effect of maternal Zika virus infection on foetal development [23], the association between the postnatal motor repertoire and language development [24], or on socio-communicative deficits in 9–24 months old children as markers for

developmental disabilities [25, 26, 27, 28]. Over the last years, more and more technical approaches have gained increasing importance for developmental neuroscience, especially approaches building on 'intelligent' analytics implying the utilisation of machine learning methodology.

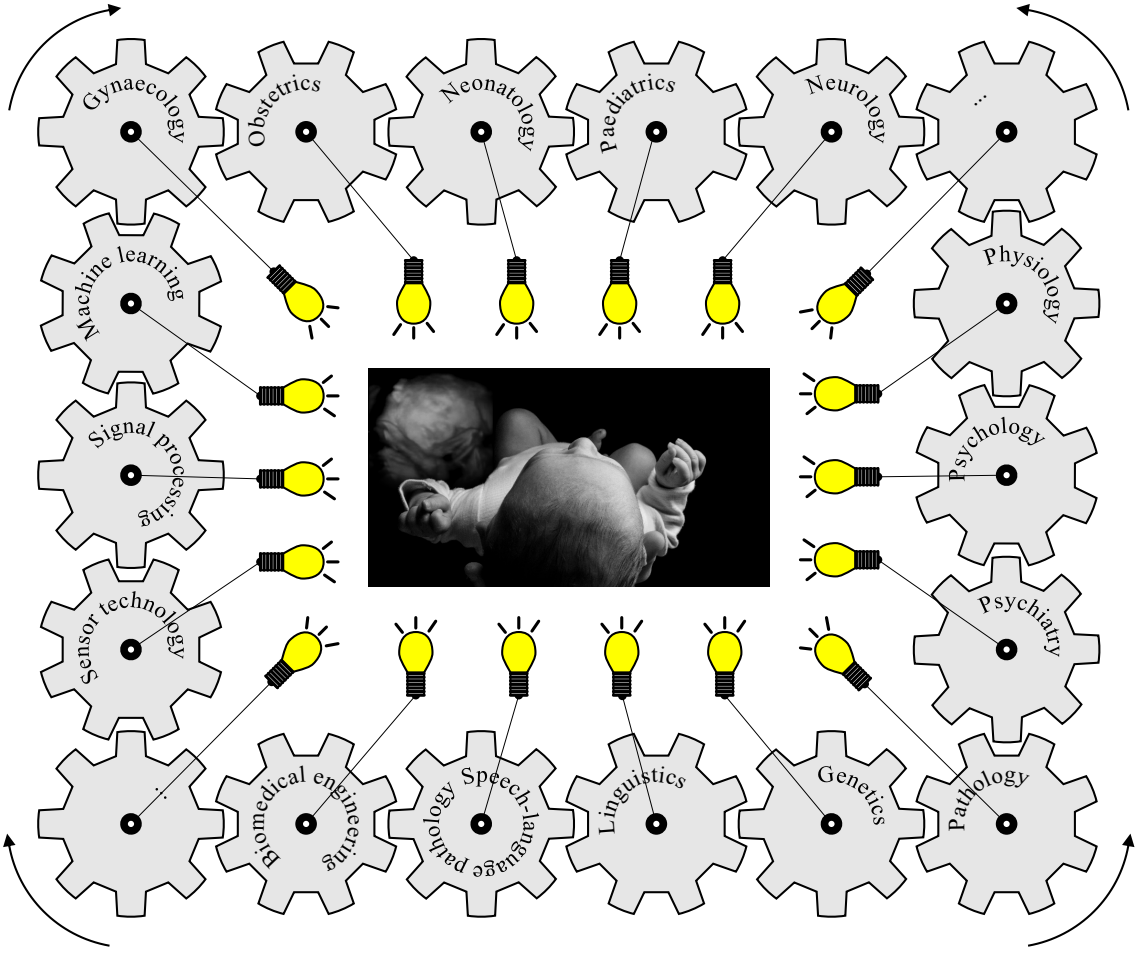


Figure 1.1: Interdisciplinary developmental neuroscience as synergy of various disciplines shedding light on currently unknown aspects of early human development. Picture Stangl 2014

1.1.1 Machine learning in developmental neuroscience

Building on fundamental mathematics such as linear algebra but also on probability and information theory [29], machine learning can be regarded as a form of data mining by means of applied statistics [30, 29]. 'Learning' in this context implies that knowledge is acquired from data [31], which have to be available in an electronic representation. The sought knowledge is about patterns within these data [30]. As

the quantity and/or complexity of the data exceed the analytical competences of a human being, a computer, a machine, is used to automatically perform this learning process. Then, based on the learned, future data or data-related outcomes of interest can be automatically predicted [30].

Also applying to other medical-related fields, machine learning methodology has found its way into developmental neuroscience in connection with applications for which human performance or the performance of conventional/deterministic signal-based approaches has been unsatisfactory and the probability value generally involved by machine learning practice has been arguable. Machine learning in developmental neuroscience allows for an automated analysis, classification, and/or detection of foetal or infant neurofunctions and/or (medical) states. Apart from its role in basic research, machine learning methodology currently has an essential impact on applied clinical research, especially on the development of assistive tools to support and automatise obstetric and paediatric diagnostic (pre-)processes.

Recent studies, e.g., focussed on the automated localisation of foetal body parts in cine-MRI scans [32], the automated diagnosis of hip dysplasia in infants from 2D ultrasound images [33], or on newborn face recognition via distance metric learning [34]. Li and colleagues [35] compared different machine learning approaches for an automated identification of infants, who were small for gestational age. Moreover, a number of studies built on machine learning technology for the automatic detection and categorisation of infant cries (e.g., [36, 37, 38, 39]). Another recent application for machine learning in developmental neuroscience is the automated recognition of infant sleep states, e.g., based on respiration [40] or EEG data [41]. Aiming to improve the quality of neonatal resuscitation procedures, respective video data were automatically analysed by Guo and colleagues [42]. An automated prediction of apnea of prematurity, a breathing disorder prevalent in preterm infants, from data collected during the first days of life was focussed on by Mago and colleagues [43]. Machine learning-based infant MRI analyses were applied, e.g., to segment infant brain tissue into white matter, grey matter, and cerebrospinal fluid [44], to characterise functional connectivity in the preterm brain [45], or to predict infant brain maturity providing a means for estimating an infant's neurodevelopmental outcome [46]. Obviously, machine learning represents an emerging methodological tool in developmental neuroscience not only for the classification of neurotypical behaviour, but also for the recognition or prediction of pathological conditions.

2

Developmental disorders with a late clinical manifestation

Developmental disorders subsume medical conditions that are characterised by deficits in any developmental domain and typically become manifest early in development [47]. A number of these disorders are apparent from the day of birth onwards or even earlier in the prenatal period due to behavioural peculiarities and/or physical atypicalities/dysmorphic features. For example, foetuses with chromosomal anomalies, e.g., present in Down syndrome (trisomy 21), can be identified with a certain probability between 10 and 14 week of gestation by means of a nuchal translucency ultrasound scan [48]. Characteristic postnatal features of individuals with Down syndrome are low set simple ears, an upward eye slant, epicanthic folds, a flat occiput, a short neck, a single palmar crease, and a gap between the first and the second toe [49]. Further examples of postnatal physical or behavioural signs for developmental disorders are a low muscle tone in combination with facial abnormalities including full lips, a wide mouth, a short upturned nose, a flat nasal bridge, and a broad brow in Williams syndrome [50, 51, 52], or high-pitched cat-like crying in Cri-du-chat syndrome [53].

However, there are developmental disorders without apparent early signs (see Figure 2.1). The late onset of clinically relevant features in individuals with such a disorder currently leads to an accurate diagnosis in toddlerhood, at preschool age, or even later. In addition, diagnosis is often hampered by (i) a wide phenotypical variation within one and the same disorder, (ii) different disorders sharing the same or similar phenotypical signs, and (iii) comorbidities of multiple developmental disorders. Examples[1] of 'late recognised' developmental disorders are attention deficit hyperactivity disorder (ADHD), Angelman syndrome, autism spectrum disorder (ASD), Cohen syndrome, fragile X syndrome (FXS), Pitt-Hopkins syndrome, Prader-Willi syndrome, Rett syndrome (RTT), Smith-Magenis syndrome,

[1] http://www.orpha.net, http://compbio.charite.de/phenomizer (as of 31 March 2018)

and Tourette syndrome, to name but a few. Some of these disorders are associated with genetic causes, such as FXS or RTT, but respective genetic testing is currently not standard for screening in newborn healthcare. In other developmental disorders with a late clinical manifestation, the exact cause is still unknown, as for example in ASD (e.g., [47, 54]). Some 'late recognised' developmental disorders are relatively prevalent, such as ADHD (1 of 20 individuals affected [47]) or ASD, others are rather rare, such as FXS, RTT, or the Pitt-Hopkins syndrome with only a few hundred confirmed cases worldwide [55].

Figure 2.1: Unawareness of the presence of a developmental disorder due to the lack of early behavioural and/or physical evidence. ASD = autism spectrum disorder; FXS = fragile X syndrome; RTT = Rett syndrome; TD = typical development. Picture Loske 2016; www.artigebilder.at

Due to the availability of data of both quantity and quality sufficient for meaningful scientific-technical analyses and for evaluating the potential of future clinical implications, experimentation in the framework of this thesis was based on individuals with ASD, FXS, and RTT, and on typically developing (TD) controls. Thereby, this work covers a relatively prevalent disorder of still unknown cause (ASD) as well as two rare genetic disorders (FXS and RTT) necessitating to consider different medical condition-related requirements. Henceforth, ASD, FXS, and RTT are also referred to as conditions of interest (COI).

2.1 Autism spectrum disorder

ASD is a neurodevelopmental disorder, which was first described by the Austro-American psychiatrist Leo Kanner and the Austrian paediatrician Hans Asperger, in the 1940s [56, 57]. Today, more than 70 years later, the exact cause of ASD is sill unknown (e.g., [47, 54]). However, a variety of environmental factors, such as low birth weight or advanced parental age, and genetic configurations were found to be associated with an increased risk for ASD [47]. According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; [47]), an ASD diagnosis is based on the criteria of persistent deficits in socio-communicative functioning as well as restricted, repetitive behavioural patterns. Symptoms need to become manifest early in development causing impairments of clinical relevance. Above all, the possibility of an intellectual developmental disorder or a global developmental delay to be a more appropriate exclusive classification of the condition must be excluded, even

though intellectual disability and ASD often co-occur. [47] Recent estimates from the ADDM network (Autism and Developmental Disabilities Monitoring network) funded by the Centers for Disease Control and Prevention, indicated an ASD prevalence of 1 in 59 children and an increased occurrence rate in males than in females by factor 3.2–4.9 across the USA [58]. Moreover, ASD was found to have a recurrence risk of up to 18% for (younger) siblings of individuals with an existing ASD diagnosis [59, 60]. This allows for longitudinal studies gathering prospective data from high risk cohorts (e.g., Early Autism Sweden[2]). Currently, there is no cure for ASD (e.g., [54, 61]), but the benefit of early treatment for affected children has been repeatedly demonstrated (e.g., [62, 63, 64]). However, even though research has been intensively focussing on early ASD screening over the last years (e.g., [65]), children with ASD are usually not identified before toddlerhood [66, 58, 67].

2.2 Fragile X syndrome

FXS is a condition that was first described as a 'mental defect' by the British human geneticists James Purdon Martin and Julia Bell in 1943 [68]. Based on the cytogenetic identification of a marker X chromosome with a secondary constriction close to the end of the long arm by Herbert Lubs in 1969 [69], FXS was found to result from a mutation in the X-linked Fragile X Mental Retardation 1 (*FMR1*) gene and classified as trinucleotide repeat disorder in the 1990s [70, 71]. FXS is the most prevalent inherited form of intellectual disability [72]. As usual for X-linked diseases, clinical manifestation in affected females tends to be milder than in males due to the presence of a second, 'normal' X chromosome [73]. The incidence of FXS is about 1.4 of 10 000 in males (1:7 143) and 0.9 of 10 000 in females (1:11 111) [74]. The disease is characterised by a general developmental delay, cognitive impairment, prominent stereotypies, ADHD symptoms, and anxious behaviour [75, 73]. Moreover, individuals with FXS often display autistic features, such as gaze aversion or repetitive speech patterns [76, 73]. FXS with a comorbid diagnosis of ASD is common [77, 78, 79]. Physical features related to FXS are, e.g., a large head, a long and narrow face, large and anteverted ears, a large jaw, and a seizure predisposition [80, 81]. These features, however, become prominent only later during development. On average, males with FXS are diagnosed at around 36 months of age, females at around 42 months [82]. Besides facilitating an earlier intervention for affected individuals, an earlier recognition of FXS could open parents the opportunity of timely genetic counselling with respect to further family planning.

[2]http://www.earlyautism.se (as of 27 March 2018)

2.3 Rett syndrome

RTT is a profound progressive developmental disorder that was first described by the Austrian neuropaediatrician Andreas Rett in 1966 [83, 84]. In 1999, the main cause of the disease could be identified to be a spontaneous mutation in the X chromosome-linked gene encoding Methyl-CpG-binding Protein 2 (*MECP2*) [85]. Therefore, RTT almost exclusively occurs in females[3] at a prevalence of about 1 in 10 000 live female births [87]. Nevertheless, there are both individuals with *MECP2* mutation without showing the RTT phenotype and patients with the RTT phenotype without *MECP2* mutation [88, 89]. Consequently, the diagnosis of RTT and its variants still remains a clinical matter based on four main consensus criteria [88]: (1) loss of already acquired purposeful hand skills, (2) loss of already acquired spoken language, (3) gait abnormalities, (4) and hand stereotypies. For a more fine-grained classification of the clinical picture of RTT, there are eleven supportive criteria [88], e.g., growth retardation, scoliosis/kyphosis, muscle tone abnormalities, breathing irregularities, bruxism during waking state, diminished pain response, and small cold hands and feet. The early development of individuals with RTT usually appears inconspicuous [90] for caregivers as well as for healthcare professionals being in charge of early paediatric standard examinations. In most cases, first suspicion is raised due to a regression of already acquired functions. Developmental regression is characteristic for the progression of RTT and typically sets in between 6 and 18 months of age [91]. Currently, RTT can only be treated symptomatically, but a reliable earlier recognition of the disease could add to improve life quality of affected individuals and to prevent diagnostic odysseys for parents from the day of their first concerns to the day of their child's definite RTT diagnosis. However, even though a number of studies have raised evidence for RTT signs already in the pre-regression period, such as atypicalities in early spontaneous movement patterns (e.g., [92]) or early vocalisation peculiarities (e.g., [93, 94, 95, 25, 26]), the mean age of a RTT diagnosis still is 2.7 years [96].

[3] As being X-linked, *MECP2* mutation in males usually leads to miscarriage, stillbirth, or a very low life expectancy [86].

3

The speech-language domain

Communication plays a central role in human life [97]. The intentional transmission and exchange of information including wants and need, or emotions is an essential human capacity and a requirement for individuals living together in a civilisation [98]. Defined by the use of a conventional (abstract) symbolic code [98], language is a fundamental means of human communication [99]. Spoken language represents a basic mode of language use [99].

From a technical perspective, spoken language, i.e., speech, can be regarded as a recordable signal with specific characteristics. These characteristics become evident when understanding and modelling the process of speech production. For engineering purposes, modelling the speech production process is not to exactly rebuild anatomy and (neuro-)physiology of the real speech production system, but to obtain a mathematical approximation fulfilling application-specific requirements, such as intelligibility in speech synthesis or efficiency in speech transmission [99]. A popular speech production model is the source-filter model, which approximates the effects of the lung and the vocal folds as the excitation source, and the influence of the vocal tract as a time-varying filter [99].

An infant's pre-linguistic period, i.e., the time before the production of linguistic utterances such as target language-related proto-words or words [100, 19], can be regarded as a phase in which both source and filter are fundamentally tuned due to underlying anatomical and physiological transformations and processes of infant brain development. For almost 40 years, studies have documented typical pre-linguistic verbal development (e.g., [101, 17, 18, 21, 22]). Besides aiming to categorise early vocalisations (e.g., [16, 17, 19, 20, 21]), research has focussed on defining milestones of speech-language development. Delays in achieving particular milestones have been discussed as potential early indicators of developmental disabilities (e.g., [102]).

3.1 Typical pre-linguistic verbal development

The first weeks of an infant's life are mainly characterised by the production of vegetative and discomfort sounds, such as fussy and crying vocalisations [20]. A relatively small proportion of vocalisations in this very early period is already composed of non-distress sounds. These are typically brief, voiced, low-pitched, grunt-like sounds with limited resonance referred to as quasi-resonant nuclei [17, 16]. The vocal quality of quasi-resonant nuclei is far from the quality of target language-related vowels [16]. Very early vocal behaviour has been discussed to be endogenously produced by specific neural networks located in the brain stem, the so-called central pattern generators [103, 104].

Still not transcribable as target language-related vowels according to the International Phonetic Alphabet (IPA), fully resonant nuclei to a small proportion occur along quasi-resonant nuclei during the first two months of life. Then, they become more and more prominent [17, 16]. Fully resonant nuclei are vowel-like vocalisations longer than quasi-resonant nuclei with energy over a broader frequency range. Nevertheless, fully resonant nuclei may exhibit a poor, e.g., harsh or high-pitched vocal quality [16].

Shortly after infants have achieved a certain level of cortical control over sound production, a more complex, distinctive type of vocalisation emerges around the third month of life, namely the cooing vocalisation [17, 20]. A cooing vocalisation consists of a usually velar consonant-like element, such as a voiced fricative, which is optionally combined with a vowel-like element [17, 16], and frequently exhibits a distinct melodic contour [19]. Even though the cooing pattern considerably varies in structure and quality, cooing vocalisations are the first vocalisations in the course of early verbal development exhibiting discernible tongue movements [18] and can be regarded as a very first form of syllables [101]. The cooing pattern typically develops alongside the pattern of social smiling and is frequently generated in face-to-face interaction settings with a caregiver [20, 18].

The period around the fourth and fifth month of age is characterised by an infant's exploration of his or her vocal apparatus' full potential. The playful use of vocal behaviours, such as squealing, yelling, growling, or raspberry sounds is typical for this phase. [17, 20, 19] Furthermore, the early vocal repertoire is expanded by vowel-like vocalisations more and more turning into vowels of target language-related quality and by vowel glides, i.e., vowel-like elements or vowels with slow variations in vowel-quality, such as in pitch or loudness, without an audible gap [16].

Around the end of the first half year of life, infants typically start to produce first sequences of consonant(-like)-vowel(-like) sound combinations with yet prolonged or shaky formant transitions referred to as marginal babbling [17, 19, 16].

From the beginning of the second half year of life onwards, an infant's so far acquired vocal competence becomes manifest in the transition from marginal babbling to the onset of canonical syllables – consonant-vowel combinations with target

language-related quality and timing – and series of canonical syllables referred to as canonical babbling [17, 20, 19, 16]. Canonical babbling is typically well in place not later than 10 months of age representing a key milestone in pre-linguistic verbal development [17, 20, 105]. At first, babbling consists of repetitions of the same consonant-vowel sequence (reduplicated babbling). Then, infants more and more systematically produce vocalisations with differing consonant-vowel combinations referred to as variegated (non-reduplicated) babbling [17, 20, 19, 16].

Over the last months of the first year of life, the advanced level of phonation control and the perpetual competence of acquiring new (vocal) structures by combining already acquired structures involve the generation of even more complex syllabic vocalisation types, such as gibberish or jargon. These are patterns of target-language-like intonation but still without lexical meaning [17, 20, 19, 16].

The reached phonation competence in combination with the influence of an infant's language environment, finally, leads to the production of first target language-related proto-words – phonetically consistent word-like vocalisations with consistent lexical meaning – and words around the end of the first year of life [20, 19]. This marks the end of the pre-linguistic period of speech-language development [100, 19].

3.2 Deviant pre-linguistic verbal development

Based on our knowledge on typical pre-linguistic verbal development (cf. Section 3.1), deviant verbal behaviour – especially the delayed achievement or even the non-achievement of specific speech-language milestones, e.g., canonical babbling, has been discussed as potential early indicator of developmental disorders (e.g., [102]). As this thesis focusses on a selection of three developmental disorders with a late clinical manifestation, this section will report on pre-linguistic verbal peculiarities related to these three disorders, namely to ASD, FXS, and RTT. Studies dealing with the early vocal development in these disorders have been recently reviewed by Roche and colleagues [106].

A substantial number of studies on pre-linguistic verbal specificities in ASD have exclusively focussed on crying vocalisations (e.g., [107, 108, 109]). Reasons for this may be that crying is a frequent vocalisation pattern present right after birth. Moreover, the identification and segmentation of crying sequences in audio recordings for scientific purposes is comparably easy. Sheinkopf and colleagues [108] investigated a group of 6-months-old individuals at heightened risk for ASD, i.e., each participant had an older sibling with an ASD diagnosis. Pain-related cries in this sample exhibited a higher fundamental frequency (F0) as well as a higher F0 range compared to a group of low-risk controls. Those high-risk infants diagnosed with ASD at 36 months of age produced among the highest F0 values. In addition, their crying was more poorly phonated than the crying of infants without ASD [108].

Esposito and Venuti [107] documented an unchanged F0 in cries between the first and the second year of life in ASD, as opposed to a decrease in the respective F0 trajectory in TD infants. Furthermore, Esposito and colleagues [109] found out that ASD-related crying is more negatively perceived by adults than crying related to typical development (TD). In this study, listeners' perceptions of negativity were more determined by the length of pauses than by, e.g., the F0 [109].

Beside investigations of crying, much research on pre-linguistic verbal behaviour in ASD has focussed on babbling. Patten and colleagues [110] reported on significantly lower canonical babbling ratios (# canonical syllables divided by # all syllables) in ASD infants between 9 and 18 months of age compared to TD controls. This was also shown by Paul and colleagues [111] for a group of infants at risk for ASD, particularly at 9 months. In contrast, Chericioni and colleagues [112] did not find ASD-related atypicalities concerning babbling frequency (rate per minute), but significantly less first words produced by ASD participants than by controls between 12 and 18 months of age. They further attested a decrease in vocalising in infants with ASD subsequent to the first half year of life and significantly less vocalisations compared to TD controls during the second half year of life [112]. A lower volubility in ASD infants compared to TD controls was also reported by Patten and colleagues [110]. Other ASD-related peculiarities concerning pre-linguistic verbal behaviour reported in different studies are (i) an absence of cooing [113], (ii) hardly modulated unspecific vocalisations [113], (iii) less speech-like vocalisations in favour of non-speech vocalisation [111], (iv) significantly less consonant types [111], (v) a reduced number of vowel sounds produced per minute [114], (vi) significantly less complex vocalisation contours in terms of melodic modulation [115], and (vii) significantly less vocalising to people in both the first and second half year of life [116, 117].

Similar to ASD, also FXS was found to be associated with a significantly lower volubility and a reduced likeliness for affected infants to achieve the canonical babbling milestone by 9 to 12 months of age [118]. Marschik and colleagues [26] reported on the dominance of non-verbal behaviours over pre-linguistic communication forms in connection with FXS.

The pre-linguistic, pre-regressional verbal development of individuals with RTT has been repeatedly characterised by an intermittent occurrence of apparently typical and atypical vocalisations (e.g., [95, 119]). A listening experiment by Marschik and colleagues [94] revealed that listeners were able to differentiate between a set of pre-selected atypical vocalisations of individuals with RTT and vocalisations of TD infants. Described prominent features of RTT-associated early vocalisation atypicality are (i) inspiratory, (ii) pressed, and (iii) high-pitched crying-like phonation (e.g., [120, 95, 119]). Moreover, some infants with RTT were found not to achieve certain pre-linguistic speech-language milestones, such as cooing, babbling, and proto-words ([25, 119, 27]). In a case study by Marschik and colleagues [120], a girl with the

preserved speech variant (PSV) of RTT, a relatively milder variant of RTT, e.g., in terms of a better recovery of speech-language capabilities, was documented to produce repetitive, unmodulated vocalisations at 6 months of age [120]. She further produced typical babbling sequences interspersed with atypical babbling patterns [120]. Finally, in home video recordings of a sample of six female infants with RTT, Bartl-Pokorny and colleagues [25] could not observe any pre-linguistic vocalisations in use for communicative purposes, such as for answering, imitating, or requesting an action, an object, or information (also see the Inventory of Potential Communicative Acts; IPCA [121, 122]).

3.3 Early vocalisation engineering

Technical investigations of early verbal behaviour have been conducted for more than 40 years. Early studies already dealt with specific infant vocalisation types, e.g., with fricative and trill vocalisations [123], and with acoustic features and methodological aspects of the extraction of acoustic features from infant vocalisations. Acoustic analyses of infant vocalic utterances conducted by Kent and Murray [124] were based on features such as vocalisation duration, formant frequencies, or vocal tract source excitation variation. Among acoustic vocalisation features, especially the F0 and its robust determination have been frequently focussed on in context of vocalisation analysis – also for infant vocalisation analysis (e.g., [125, 124, 126]).

As acoustic measures for the characterisation of infant vocal behaviour have been usually taken from traditional acoustic feature sets for the analysis of the significantly different vocal behaviour of adults, Warlaumont and colleagues [127] proposed a neural network approach for a data-driven derivation of features for infant vocalisation analysis.

A popular and highly researched field of engineering activity is constituted by infant crying and its computer-based processing. This may be motivated by similar reasons why crying vocalisations are frequently focussed on in investigations of deviant early verbal behaviour (crying as easily segmentable, frequent behaviour present from the day of birth; cf. Section 3.2). Moreover, infant crying research opens up the opportunity for several useful healthcare-related applications and assistive applications for parents. As already indicated in Section 1.1.1, early crying processing primarily targets the automatic detection and categorisation of infant crying patterns (e.g., [36, 37, 38, 39]). Moreover, Aucouturier and colleagues [128], for instance, examined the segmentation of expiratory and inspiratory vocalisation phases in infant cries using hidden Markov models. Current work in the field also aimed for the identification of the cause of crying [129, 130]. Myakala and colleagues [131, 132] proposed intelligent systems for infant monitoring and crying detection in real time. Furthermore, infant crying even built the central topic in a sub-task of this year's (2018) Interspeech Computational Paralinguistics and Emotion (ComParE)

Challenge [133]. In this sub-task, participants had to compete in the automatic classification of more than 5000 infant vocalisations constituting the CRIED (Cry Recognition In Early Development) database, which was recorded in the framework of a study on neuro-physiological changes in early infancy at the Medical University of Graz, Austria [134]. In particular, participants had to differentiate between three mood-related classes of vocalisations, namely neutral/positive vocalisations, fussy vocalisations, and crying vocalisations, labelled by professionals in the field of early speech-language development. In context of this challenge, the idea of an 'intelligent' babyphone was mentioned. Such a babyphone would not only report an increased sound level in a baby's bedroom, but also indicate if a baby was 'only' vocalising, fussy, or already crying.

Dependent on the specific research question and the intended target application, the availability of suitable data, i.e. in this context, the availability of audio recordings of infant vocal behaviour, plays a crucial and often limiting role in vocalisation processing. On this, the setting in which infant recordings are performed, represents an essential factor. On the one hand, the recording of infants under laboratory conditions following a pre-defined recording protocol, best possibly allows for a high recording quality and (semi-)standardised analyses, but infant behaviour might be influenced by the artificial environment. On the other hand, recording in an infant's natural environment, e.g., at home, poses some methodological challenges. Issues in connection with infant home recordings for early vocalisation engineering purposes are, e.g., related to (i) the recording device and its positioning, (ii) the presence of everyday background noise events or noise caused by the infant manipulating objects, such as toys, or (iii) the selection of a representative recording time span. For versatile use in context of early communication development, the LENA® (Language Environment Analysis) system[1] provides (i) a 24-hours audio recording of an infant in its natural environment by means of a child-safe recording device fixed in a vest, as well as (ii) an automated (pre-)analysis of the recorded infant's developmental status and language environment on the basis of different calculated metrics [135]. The continuously growing 'LENA® community' emphasises the relevance of naturalistic recordings for early speech-language-related research and healthcare applications, and the potential of the link between the fields of speech-language acquisition and vocalisation engineering for an earlier identification of speech-language-related (developmental) disabilities (e.g., [136]).

[1] https://www.lena.org (as of 25 April 2018)

4

A technical tool for infant healthcare assistance

By combining knowledge and methods from technical and medical-related disciplines, this thesis can be regarded as a highly interdisciplinary endeavour exploring the technical feasibility of particular modules of a tool for pathology recognition. This proposed tool is intended as an assistive pre-diagnostic screening tool, which shall contribute to a future improvement of infant healthcare.

4.1 Motivation

In consideration of the beneficial role of an earlier identification of individuals with developmental disorders such as ASD, FXS, or RTT (cf. Sections 2.1–2.3), research on the prodromal period of these disorders appears to be underrepresented. While to date there are more than 95 000 COI-related articles indexed in Web of Science, of which significantly less deal with the relatively rare genetic disorders FXS and RTT compared to ASD (see Table 4.1), in total not even 2% of these articles address an earlier recognition of these diseases. The socio-communicative domain represents a key domain for the clinical manifestation and diagnosis of COI. Accordingly, almost the half of articles on an earlier recognition of COI focussed on vocal behaviour. However, only two articles were published on an automated approach for the earlier identification of individuals with ASD. In one of these articles, Xu and colleagues [137] report on a fully automatic machine learning-based autism detection mechanism exploiting vocalisation composition data as an extension of the LENA® system (cf. Section 3.3). In the other article, Orlandi and colleagues [138] describe an automatic infant cry recording system aiming for an acoustic differentiation between newborns at heightened risk for ASD and a group of control individuals. To the best of my knowledge, there are no empirical studies on an automated early identification of individuals with FXS or RTT.

Table 4.1: Number of articles indexed in Web of Science (all databases including Web of Science Core Collection, BIOSIS Citation Index, BIOSIS Previews, KCI-Korean Journal Database, MEDLINE®, Russian Science Citation Index, and SciELO Citation Index; as of 21 March 2018) on autism spectrum disorder (ASD), fragile X syndrome (FXS), and Rett syndrome (RTT) in total, and filtered for relevance for this thesis (early recognition of a condition of interest (i) based on vocalisations from the first year of life (ii) using an automated approach (iii)). For search terms/strategy including boolean operators see footnotes[1–6]. No additional relevant articles were found in the IEEE *Xplore*® Digital Library (as of 21 March 2018).

	ASD[1]	FXS[2]	RTT[3]
Total	75 571	13 546	6 528
(i) Early recognition[4]	1 630	155	84
(ii) Vocalisation-based early recognition[5]	755	49	35
(iii) Automated vocalisation-based early recognition[6]	2	0	0

In summary, a number of studies raised the evidence for early speech-language-related deviances in COI (cf. Section 3.2) and at least a small proportion of articles on COI focussed on an earlier COI recognition based on verbal capacities (see Table 4.1). However, little is currently known about COI-specific acoustic characteristics within the complete range of early vocalisations on signal level. The 'intelligent' acoustic analysis of pre-linguistic vocalisations for an early recognition of COI almost seems to be a blind spot in studying developmental disorders.

4.2 Aims

The objective of this work was to bridge the gap of 'intelligent' pre-linguistic vocalisation engineering for an earlier recognition of 'late recognised' developmental disorders, especially of ASD, FXS, or RTT. Accordingly, in the framework of this thesis the answers to the following main research questions (MQs) were aimed to be found:

[1] autis* OR Asperger OR Kanner

[2] 'fragile X' OR 'Martin-Bell'

[3] Rett

[4] 'earl* detect*' OR 'earl* diagnos*' OR 'earl* identif*' OR 'earl* recogni*' OR 'earl* screen*'

[5] Search terms from footnote[4] AND babbl* OR communicat* OR cooing OR cry OR crying OR language OR linguistic* OR verbal* OR vocal* OR voice OR oral* OR sound* OR speech OR utter* OR word*

[6] Search terms from footnote[5] AND automat* OR comput* OR 'deep learning' OR engineer* OR machine* OR 'pattern recognition' OR signal* OR technical* OR technologi*; subsequent manual review for relevance

(MQ1) Do individuals with ASD and TD individuals differ in terms of acoustic signal-level characteristics of pre-linguistic vocalisations?

(MQ2) Do individuals with FXS and TD individuals differ in terms of acoustic signal-level characteristics of pre-linguistic vocalisations?

(MQ3) Do individuals with RTT and TD individuals differ in terms of acoustic signal-level characteristics of pre-linguistic vocalisations?

(MQ4) Are individuals with different COI and TD individuals distinguishable in terms of acoustic signal-level characteristics of pre-linguistic vocalisations?

(MQ5) Can individuals with ASD be automatically recognised vs. TD individuals on the basis of pre-linguistic vocalisations?

(MQ6) Can individuals with FXS be automatically recognised vs. TD individuals on the basis of pre-linguistic vocalisations?

(MQ7) Can individuals with RTT be automatically recognised vs. TD individuals on the basis of pre-linguistic vocalisations?

(MQ8) Can individuals with different COI and TD individuals be automatically differentiated on the basis of pre-linguistic vocalisations?

Apart from these questions on basic feasibility of an early vocalisation-based identification of individuals with COI, the following additional research questions (AQs) were aimed to be answered:

(AQ1) Which acoustic features allow best to distinguish individuals with different COI and TD individuals?

(AQ2) How do COI-related auditory atypicalities manifest in the acoustic signal domain?

(AQ3) Can infant vocalisations be automatically segmented in audio material recorded 'in the wild', i.e., in infants' natural environments?

By answering to the above listed MQs and AQs, the feasibility, practicability, and impact of a fully automatic, vocalisation-based early COI recognition tool should have been evaluated. Figure 4.1 reveals the overall system block diagram of the intended tool. It consists of two main modules, namely (i) an infant voice activity detector, and (ii) the COI recogniser. Even though the first module certainly describes an interesting research topic addressed in this thesis, the main focus of this work was the second module.

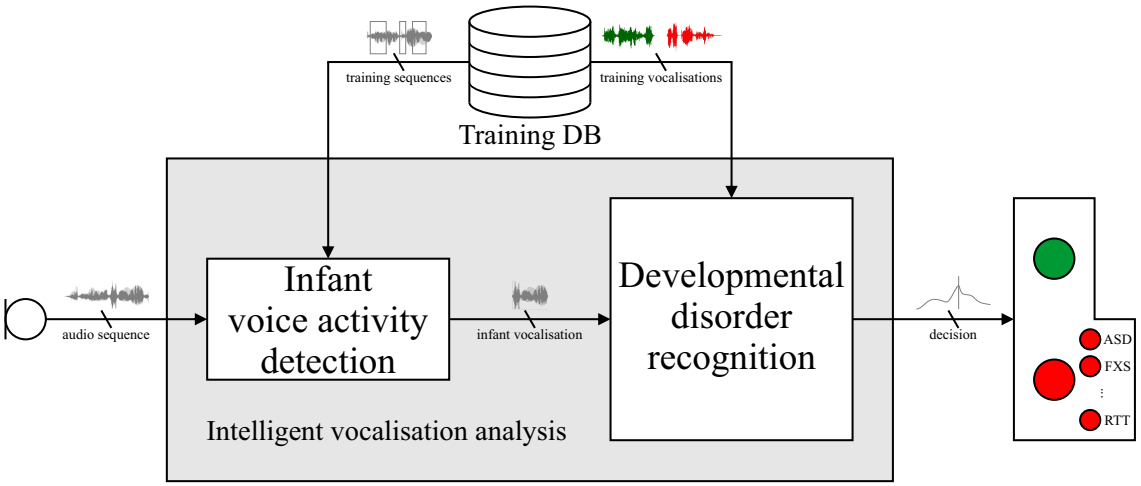


Figure 4.1: Block diagram for the automated segmentation of pre-linguistic vocalisations from an audio input sequence and the subsequent pre-linguistic vocalisation-based recognition of the potential presence of a developmental disorder, such as ASD, FXS, or RTT, in the infant, who produced the respective vocalisations. ASD = autism spectrum disorder; DB = database; FXS = fragile X syndrome; RTT = Rett syndrome

To the best of my knowledge, this thesis is among the first empirical studies dealing with an automatic identification of individuals with either ASD, FXS, or RTT – three developmental disorders with a late clinical manifestation – exclusively based on acoustic characteristics of pre-linguistic vocalisations. Thereby, this work might have potential to break new ground in future paediatric diagnostics, which may contribute to pave the way for a transition from a 'wait-and-see' to a 'find-early-to-intervene-early' strategy in COI-related healthcare.

4.3 Outline

The remainder of this thesis is structured as follows: In Part II, first, the chosen approach for the acoustic analysis of pre-linguistic vocalisations of infants later diagnosed with a developmental disorder is introduced in general (Chapter 5). Then, the corpus building the basis for analyses in the framework of this thesis is described (Chapter 6). Finally, the applied method of 'intelligent' vocalisation analysis is specified (Chapter 7). Part III reports on the individual experiments carried out in the framework of this thesis in order to find the answers to the posed research questions (cf. Section 4.2), i.e. basically, to evaluate the feasibility of an automated early COI recognition tool as depicted in Figure 4.1. Here, Chapter 8 deals with the proposed infant voice activity detector and Chapter 9 with the automatic COI

recogniser. Finally, this thesis concludes in Part IV by summarising the generated results in context of the thesis' aims (Chapter 10) and addressing approach-related limitations (Section 10.1), as well as by discussing the thesis' implications (Chapter 11) from both a technical and a clinical perspective (Sections 11.1 and 11.2) and presenting a potentially more powerful future early COI recognition tool based on speech-language-related parameters combined with parameters derived from additional domains, such as the motor domain (Section 11.3).

Part II

Methods

5

Approach

In the framework of this thesis, a relatively young field of engineering – 'intelligent' audio analysis (cf. [139]) – was applied to more or less unique behavioural data of infants with a 'late recognised' developmental disorder, namely with either ASD, FXS, or RTT, and to data of TD controls. Methodological procedures were approved by the Institutional Review Board of the Medical University of Graz (MUG), Austria (27-388 ex 14/15), where experimentation was carried out under compliance with the respective Standards of Good Scientific Practice[1].

When aiming to investigate early behavioural phenomena in developmental disorders with a late clinical manifestation, appropriate approaches are limited. This is due to the fact that analyses have to build on data from an infant's period, in which his or her developmental outcome, therefore, his or her eligibility to be included in a study on a specific 'late recognised' developmental disorder is not yet known. Thus, both the collection of infant data has to be managed and the respective infant's developmental outcome has to be found out. This allows for two approaches with respected to participant recruiting strategy and study timeline.

On the one hand, infants fulfilling specific study-relevant a priori criteria, such as to be younger siblings of children with an ASD diagnosis and, therefore, at heightened risk for ASD themselves, can be systematically recruited and prospectively studied in the framework of a longitudinal study design including outcome evaluations. Such a procedure is well-suited for investigating developmental disorders with a high prevalence and/or recurrence risk, but also for investigating typical early development. The prospective collection of behavioural data on rare, inheritable developmental disorders is hardly possible.

On the other hand, participants can be recruited at a time at which their developmental outcome is already known, i.e., the diagnosis of a 'late recognised'

[1] https://www.medunigraz.at/fileadmin/forschen/gsp/GSP_Standards_engl.pdf
(as of 11 April 2018)

developmental disorder has already been made. Families affected by specific developmental disorders can, for example, be searched out via parents' associations. Then, early data of an included participant/data from an included participant's prodromal/pre-diagnostic period have to be collected retrospectively. For example, a considerable part of the current knowledge on socio-communicative and speech-language development in individuals with rare developmental disorders with a late clinical manifestation, such as FXS, builds on retrospective family/parent interviews and questionnaires (e.g., [140, 141, 82, 142, 143]). More objective, signal-based retrospective analyses require the collection of early recordings of a participant, such as audio-video recordings. Being aware of methodological limitations (cf. Part IV), experimentation in the framework of this thesis was based on a retrospective signal-analytical approach building on both prospectively and retrospectively collected audio-video data of TD individuals and individuals with ASD, FXS, or RTT.

5.1 Retrospective audio-video analysis

Retrospective analysis of audio-video data in context of this thesis means that audio-video recordings, i.e., video recordings including audio tracks as made by a standard video camera, showing individuals in a period before a potential later diagnosis with a COI were analysed at a time the diagnosis of a COI had already been made, or the presence of a COI or a comparable other developmental disorder could had been excluded and the individual considered as TD. This procedure is illustrated in Figure 5.1.

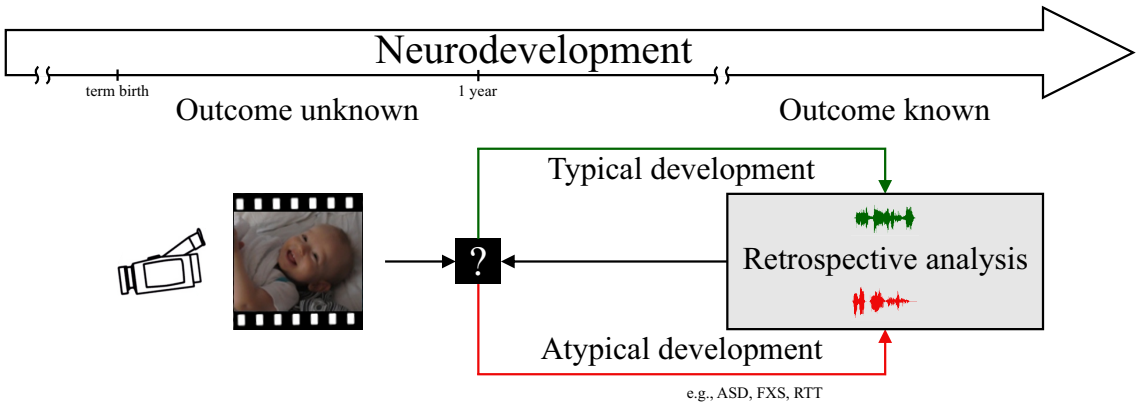


Figure 5.1: Procedure of retrospective audio-video analysis for the identification of early acoustic atypicalities in recorded infant vocalisations that could have already predicted an infant's later developmental outcome with regard to the presence of a developmental disorder with a late clinical manifestation vs. typical development, known at the time of analysis. ASD = autism spectrum disorder; FXS = fragile X syndrome; RTT = Rett syndrome

Currently, retrospective audio-video analysis represents one of the most promising available means to 'look' and/or 'listen back' for studying the prodromal/pre-diagnostic period in developmental disorders with a late clinical manifestation (e.g., [144, 145, 146, 147, 148, 27, 26]).

The goal of retrospective audio-video analysis as used in context of this thesis was to find out whether infants' developmental outcomes with respect to the presence of a COI vs. TD would have already been deducible on the basis of information within the infants' early audio-video recordings. More precisely, the approach was intended for the identification and delineation of early behavioural markers within a given set of early infant recordings. In particular, the aim was to identify and delineate acoustic atypicalities in recorded pre-linguistic vocalisations, usable to reliably predict an infant's later diagnosis of a COI, or his or her TD. By following an 'intelligent' audio analysis approach (cf. Chapter 7), the acoustic characterisation of vocalisations and the vocalisation-based recognition of infants' later outcomes were carried out by means of signal processing and machine learning methodology.

In principle, for 'intelligent' audio analysis as applied here the collection of audio recordings only (as against audio-video recordings) would have been enough, but (i) audio-video recordings are usually more commonly available and, therefore, easier to acquire, and (ii) video information is essential for a number of data pre-processing steps, such as for the vocalisation segmentation process (cf. Chapter 6).

Experimentation in the framework of this thesis was based on audio-video material of TD individuals and individuals later diagnosed with either ASD, FXS, or RTT. For data availability reasons, the focus of this work was set to the second half year of life, a period typically including the onset of canonical babbling as an important milestone in pre-linguistic speech-language development. Material of individuals with ASD as well as material of a matched group of TD controls was prospectively recorded in the framework of a longitudinal ASD high-risk study and included for experimentation in this thesis only after the individuals' developmental outcomes were assessed (cf. Section 9.1). Material of individuals with the rare genetic disorders, FXS and RTT, again respective material of matched TD controls, as well as material of another individuals with ASD was provided by parents/families for the purpose of retrospective scientific analysis at a time they already knew about their children's outcomes (cf. Chapter 8 and Sections 9.2–9.4).

6

Database

The data constituting the basis for this work were brought together from two different sources (see Figure 6.1).

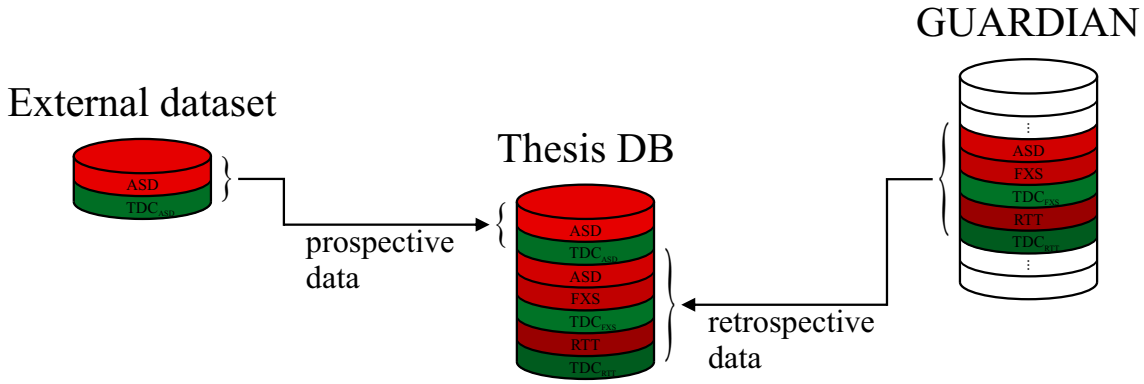


Figure 6.1: Database used for experimentation in the framework of this thesis as a composition of an external dataset and a sub-dataset of GUARDIAN. ASD = autism spectrum disorder; DB = database; FXS = fragile X syndrome; GUARDIAN = Graz University Audiovisual Research Database for the Interdisciplinary Analysis of Neurodevelopment; RTT = Rett syndrome; $TDC_{ASD/FXS/RTT}$ = typically developing control dataset matched for ASD/FXS/RTT dataset

First, data of individuals with ASD as well as respective data of a matched group of TD controls (cf. Section 9.1) were provided by cooperation partners from the Karolinska Institutet, Stockholm, Sweden, and the Uppsala University, Sweden. These data were collected within the project EASE (Early Autism Sweden[1]), a longitudinal study on a Swedish population of individuals at heightened risk for ASD as being younger siblings of individuals with an ASD diagnosis, and low-risk controls. Developmental outcome assessments regarding ASD were carried

[1]http://www.earlyautism.se (as of 27 March 2018)

out at a participant age of 3 years. The provided dataset contained audio-video sequences of participants at 10 months of age, each recorded in a parent-child interaction setting. This setting was designed as follows: Parents were instructed to playfully interact with their children in a closed room on a mat with different toys same as they would do in everyday life. Multi-view audio-video recording was guaranteed by a four-angle recording system of Panasonic HC-V700 camcorders fixed in the four room corners at the ceiling. Apart from the participant and his or her parent(s), nobody/no study personnel was present in the room during the recording. Therefore, potential distractions of the participant were prevented and, from an acoustic point of view, toy manipulation sounds and parental voice were the only background noise events contained in the recordings. For analyses in the framework of this thesis, synchronised 4-angle video information as well as the audio information recorded by one of the four cameras were provided. Audio information was provided in 2-channel AAC format and converted to single-channel format at 44.1 kHz and 16 bits, for further acoustic analysis.

Second, retrospectively collected recordings of TD individuals and individuals with ASD, FXS, or RTT (cf. Chapter 8 and Sections 9.2–9.4) were taken from GUARDIAN (see Section 6.1), the core database of the Research Unit iDN – interdisciplinary Developmental Neuroscience, at the MUG, Austria, where the research for this thesis was carried out.

6.1 GUARDIAN

GURADIAN is the acronym for 'Graz University Audiovisual Research Database for the Interdisciplinary Analysis of Neurodevelopment'. Even though the database's name was first introduced in 2014, active data collection already started much earlier, in the 1980s. In the meanwhile, GUARDIAN has grown to an extensive and unique digital archive of both early typical and atypical human development in form of prospectively and retrospectively collected data (mostly but not exclusively audio-video data), physically stored on a secure network-attached storage (NAS) server at the MUG, Austria, and administrated and employed by the MUG's Research Unit iDN. GUARDIAN's main datasets are introduced and characterised in Table 6.1.

Over the last years, GUARDIAN was substantially enlarged and enhanced with regard to data administration and data (pre-)processing automation. Especially, Dataset 3 (cf. Table 6.1), a prospective collection of multi-domain recordings of TD infants at 4–16 weeks, was built up and prepared for multivariate analyses between 2015 and 2018 in the course of this work. This dataset will be essential for the further development of the approach delineated in this thesis (cf. Section 11.3). However, for experimentation in this work, appropriate parts of GUARDIAN's Dataset 4 (cf. Table 6.1) were used.

Table 6.1: Overview (non-exhaustive) of GUARDIAN with respect to its main sets of prospectively and retrospectively recorded data. ASD = autism spectrum disorder; CP = cerebral palsy, d = day(s), DGS = DiGeorge syndrome, DB = database; FXS = fragile X syndrome; GUARDIAN = Graz University Audiovisual Research Database for the Interdisciplinary Analysis of Neurodevelopment; MND = minor neurological dysfunction; mo = month(s); PHS = Pitt-Hopkins syndrome; RTT = Rett syndrome; TD = typical development/typically developing; w = week(s); y = year(s); ZIKV = Zika virus, * = post-term; ** = ongoing data collection

	GUARDIAN			
	Prospective data			Retrospective data
	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Participant age* at time of recording	birth–5mo, 1d–365d	1mo–2y, 2mo–13y	4w–16w + follow-up at 10mo–12mo	0mo–24mo
Participant diagnoses/ inclusion conditions	e.g., TD, preterm birth, MND, brain lesion, CP, maternal ZIKV infection	apnoea, TD	TD	e.g., ASD, DGS, FXS, PHS, RTT, TD
Participant family languages	e.g., English, German, Portuguese	German, Spanish-German	German, German-Persian, German-Romanian, German-Russian	e.g., English, German, Italian, Portuguese; Spanish-German
Recording periods	1986–**	1986–1990, 1998–2012	2015–2018	1964–**
Recording settings	(semi-)standardised, participant (awake) in supine position	(semi-)standardised under laboratory conditions, participant completing tasks	(semi-)standardised, participant (awake) in supine position, participant (awake) seated in infant carrier	non-standardised home recording showing participant during daily routines and special family events
Recording types	(audio-)video	(audio-)video, gaze (eye tracking)	audio, audio-video, depth image, inertial acceleration, pressure distribution, gaze (eye tracking)	(audio-)video
Applicable methods	motor analysis, vocalisation analysis	e.g., motor analysis, assessment of communication development, neurological examination, gaze analysis	e.g., motor analysis, assessment of communication development, vocalisation analysis, sensor fusion, 3D motion modelling, gaze analysis	motor analysis, analysis of communication development, vocalisation analysis

Dataset 4 of GUARDIAN contains retrospectively collected home video material of TD individuals and individuals with 'late recognised' developmental disorders. The individuals come from around the world and, therefore, from families of different family languages. The material was recorded by the individuals' parents within the first 2 years of their children's life – a period in which the parents of individuals with developmental disorders were not yet aware of their children's medical condition. In the material, the children are shown in everyday situations, such as playing situations, during typical family routines, such as feeding situations, bathing situations, and changing situations, or during special (family) events, such as birthday parties or Christmas eve (see Figure 6.2). Consequently, the material is non-standardised as being characterised by inhomogeneous recording settings, e.g., with regard to the camera position or the number of persons present in a scene, and recording locations, e.g., ranging from a cot in a children's room to an outdoor swimming pool. These factors involve a large acoustic variability within the dataset a priori posing a challenge/representing a limitation for acoustic vocalisation analyses (cf. Section 10.1). All videos of GUARDIAN's Dataset 4 were provided by parents or in the meanwhile grown-up participants themselves for the purpose of retrospective scientific analysis[2]. The data collection is still ongoing, but new material is not included until a participant's developmental outcome is known/the diagnosis of a developmental disorder was already made. The earliest recordings contained in GUARDIAN's Dataset 4 were shot in 1964, the most recently included material stems from 2016. Dependent on the variety of used recording devices and their compatible recording media from analogue standards such as Super 8, VHS, or S-VHS, to digital standards such as DV tape or DVD, the original audio-video codecs vary dramatically within the dataset.

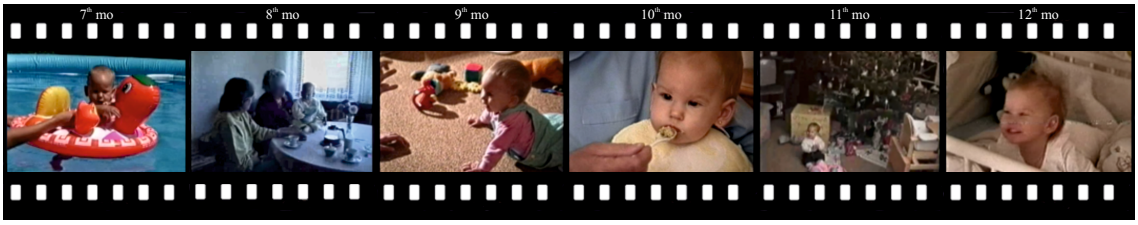


Figure 6.2: Sample frames of a typically developing female over the second half year of life demonstrating inhomogeneity within GUARDIAN's retrospective home video dataset in terms of recording setting and recording location. mo = month (of life)

In consideration of the high recording-related variability within the raw material of GUARDIAN's Dataset 4, each recording included for the purpose of vocalisation analysis was manually pre-checked for suitability (e.g., regarding continuous existence/quality of audio-video information, or information on participant age or

[2]Families/data were either recruited by the Research Unit iDN or by cooperation partners.

developmental outcome) and then passed a 5-stage data pre-processing flow (see Figure 6.3) to guarantee a best possible degree of comparability within the dataset.

In the first stage (I), original input material (as provided by the families) was converted to a standard format/audio-video codec. This standard format is defined and periodically updated by the Research Unit iDN. Since 2015, it has been MPEG-4. Analogue input media were AD-converted in an initial step prior to the MPEG-4 conversion. For the purpose of subsequent vocalisation analyses on signal level as carried out in the framework of this thesis, audio streams were additionally exported in single-channel PCM format paradigm-dependently either at 44.1 kHz or 16 kHz, and 16 bits.

In the second stage (II), the converted audio-video material was screened for scenes of interest, i.e., scenes in which the participant was present (not necessarily visible, e.g., only present in a room but not within the camera's shooting angle) and in a condition he or she was potentially able to vocalise. Thus, e.g., sequences of participants sleeping were discarded. Furthermore, sequences showing more than one infant or toddler present in a scene (e.g., the participant plus his or her younger and/or older sibling(s) in everyday situations, or the participant plus his or her playmate(s) at special events such as birthday parties) were discarded, if vocalisations could not be determined with absolute certainty to have been produced by the participant. In connection with the process of scene selection, another – sometimes quite challenging task – was performed. As early human behaviour has to be always regarded in context of developmental processes associated with brain development, the participant's (term corrected) age in months of life was validated for each included scene. Date stamps often displayed for a few seconds at the beginning of a recording served as an important indication for the age validation process.

The third stage (III) comprised the annotation of selected scenes continuously over time according to pre-defined categories, such as (a) 'bathtime', 'mealtime', or 'play event' with the additional, mutually exclusive modifiers 'indoor' and 'outdoor' for a general scene description, (b) 'alone' vs. 'interactive' with additional interaction type modifiers for a description of the communicative setting in a scene, or (c) 'no', 'low', and 'high' to specify a participant's physical restriction in the scene (e.g., infant crawling on the floor vs. seated in a high chair or being held by a caregiver). Scene annotations can be regarded as information required for a detailed dataset description or for generating scene-dependent or scene-matched (independent) subdatasets for specific analyses. However, detailed scene information was not used for experimentation in the framework of this thesis.

The fourth (IV) and the fifth stage (V) comprised the processes of vocalisation segmentation and vocalisation annotation. As these processes were not only applied to material of GUARDIAN's Dataset 4, but also to material from the external, prospective ASD dataset provided for experimentation in this work, vocalisation segmentation and vocalisation annotation are treated in detail in the following sections of this thesis (cf. Sections 6.2–6.3).

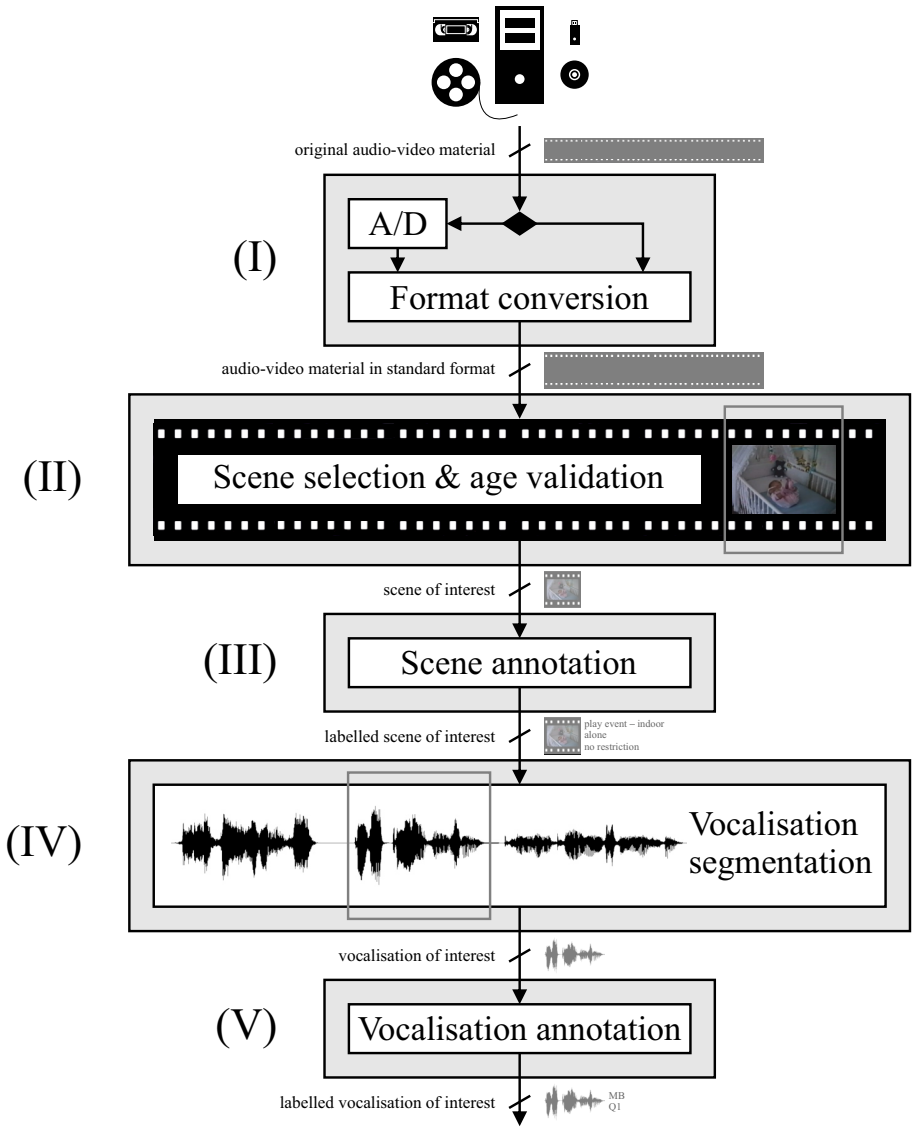


Figure 6.3: GUARDIAN's 5-stage (I–V) data pre-processing routine applied to retrospectively collected raw audio-video data for the purpose of subsequent vocalisation analysis. A/D = analogue-digital converter; MB = marginal babbling; Q = quality

6.2 Vocalisation segmentation

The process of segmenting infant vocalisations in audio-video material basically consists of two tasks. First, the presence of infant voice has to be identified. Second, meaningful segment boundaries have to be set around the identified voice episodes to generate semantic units alongside the continuous audio-video stream. 'Meaningful' in this context highly depends on the intended further processing/analysis steps and

the research question or target application behind. Thus, in some cases segmentation into single phones might make more sense than, for example, segmentation into phrases.

In the context of this thesis, a vocalisation was regarded as an utterance that generally lies in between the duration of a phone and a phrase (e.g., [149, 150]). There are two established strategies for utterance segmentation. The first strategy is based on setting segment boundaries at vocalisation pauses (episodes of silence/without vocalisation) exceeding a defined pause duration (e.g., [150]). This strategy might have advantages, e.g, regarding automatic vocalisation segmentation, however, per definition it does not follow the physiological process of voice production as being linked to the pulmonic air stream. Therefore, in the framework of this thesis the second common strategy was applied. It relies on the criterion that a vocalisation has to be assigned to a distinct vocal breathing group (e.g., [151, 150]). Thus, a segment boundary has to coincide with audible ingressive breathing or with a vocalisation pause at which an experienced judge would assume ingressive breathing (e.g., [151, 150]). Especially when segmenting home video material characterised by occasionally low audio recording quality and background noise (partly) overlaying vocalisations of interest, infant breathing activity is often not auditorily perceivable. However, in theses cases the available video information, e.g., showing mouth opening and closing, turned out to be very helpful to identify the beginning and the end of a vocalisation. On the one hand, segmentation according to breathing groups occasionally led to long vocalisation pauses within a segment, namely in case the infant started with an utterance, then held his or her breath, and finally continued with the utterance. In such a case, segmentation according to vocalisation pauses would have led to two separate segments. On the other hand, segmentation on the basis of underlying breathing activity sometimes necessitated to separate semantic patterns of very short consecutive vocal sounds into series of isolated segments. Vocalised ingressive breaths/ingressive sounds were treated as (part of) a vocalisation and not as the marker for a segment boundary. Furthermore, no segment boundary was set, if an ingressive breath was extremely shortened and a distinct intonation curve started before the breath and continued after it.

Figure 6.4 illustrates the segmentation of infant vocalisations on the basis of vocal breathing groups and strategy-related challenges.

Segments of isolated vegetative sounds such as breathing sounds, hiccups, smacking sounds, or burp sounds were excluded from experimentation in the framework of this thesis as they are not primarily meaningful when studying an infant's pre-linguistic speech-language development. Moreover, infant vocalisation segments completely overlayed with high level background noise were also excluded.

Basically, vocalisation segmentation in the framework of this thesis was done manually using both audio and video information. All segment boundaries were

verified by at least one second, independent coder. However, different automatic audio-based infant voice activity detection approaches were investigated and evaluated on the basis of a set of manually segmented data (cf. Chapter 8).

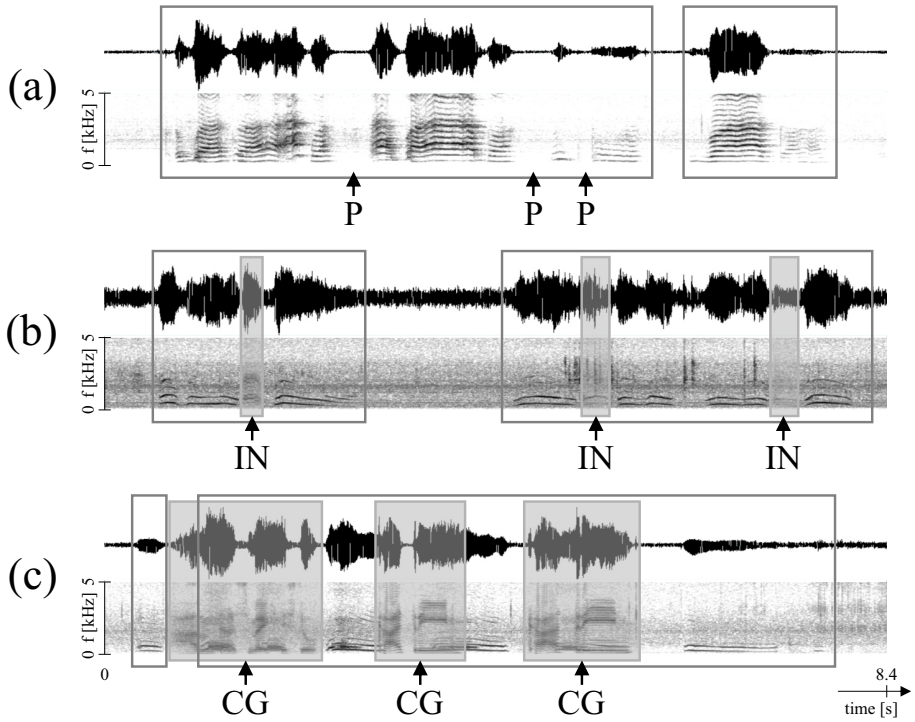


Figure 6.4: Examples of infant vocalisation segmentation on the basis of vocal breathing groups in form of spectrograms and corresponding waveforms: (a) Infant vocalisation including vocalisation pauses (without ingessive breaths); (b) Ingressive breaths/sounds as components of infant vocalisations; (c) Infant vocalisation intermittently overlayed with caregiver vocalisations. Spectrogram settings: window length = 0.025 s, window shape = Gaussian, time step = 0.00625 s (window length/4), dynamic range = 50 dB; dark grey frames = (target) vocalisation segments of a typically developing female in the 7^{th} month of life; light grey frames with filling = ingressive breaths/sounds in (b) and caregiver vocalisation segments in (c); CG = caregiver; f = frequency; IN = inspiration; P = vocalisation pause

6.3 Vocalisation annotation

Manually created segments of infant vocalisations built the basis for each experiment conducted in the framework of this thesis. The most important label of each included vocalisation was the developmental outcome of the infant, who produced the vocalisation, regarding the presence or absence of a COI. Moreover, the gender

as well as the family language of the infant, who vocalised, were registered for each segment. Even though vocalisations from all over the participants' second half year of life were pooled together for the different experiments, each vocalisation was processed as a function of the exact month of life at which it was produced. This information was considered in the experimental design of this work (cf. Section 7.2). Furthermore, the age label together with the gender and family language labels were essential for the compilation of balanced/matched sub-corpora from the available database(s) for the different experiments. Whereas age, gender, family language, and outcome actually represent vocalisation-independent variables linked to the participant only, vocalisation-dependent annotations of vocalisation type and vocalisation recording background quality were performed. However, these annotations were not taken into account for resolving the MQs of this thesis, but for considering experiment-related side aspects.

Vocalisation type annotation in the framework of this thesis was based on the Stark Assessment of Early Vocal Development-Revised (SAEVD-R [16]). Depending on the participant age and the particular research question, selected categories of the SAEVD-R were applied. The complete SAEVD-R comprises 23 mutually exclusive vocalisation type descriptors including, e.g., vowel (V), vowel glide (VG), marginal babbling (MB), canonical babbling (CB), but also crying/fussing (CR), divided into 5 levels of ascending complexity for the categorisation of pre-linguistic vocalisations and, thus, the assessment of the early speech-language development [16]. As useful with regard to the fact that vocalisations in the framework of this thesis were treated as utterances potentially containing a number of single vocalisation episodes, the SAEVD-R allowed for annotating series of vocalisation type descriptors per vocalisation segment with an overall vocalisation level always corresponding to the highest level associated with an involved vocalisation type descriptor [16]. Vocalisation type annotation was basically done by at least two independent raters. In cases of disagreement the raters discussed until consensus was achieved.

Vocalisation recording background quality was annotated according to the following four mutually exclusive classes in descending order from best to worst quality from an acoustic point of view: (Q1) No background noise present; (Q2) Stationary background noise (partly) overlaying the vocalisation, such as vehicle noises or the sound of a hair blower; (Q3) Transient background noise (partly) overlaying the vocalisation, such as parental/caregiver voice or a turned-on radio; (Q4) Both stationary and transient background noise (partly) overlaying the vocalisation simultaneously. Vocalisations recorded with a markedly low signal-to-noise ratio were assigned to Q2, as a recording's noise floor at increased sound level equals a stationary background noise.

Finally, external factors of potential influences on vocalisation acoustics were documented, e.g., food, a toy or a pacifier in the infant's mouth while vocalising.

Vocalisation annotation as well as vocalisation segmentation and scene annotation were carried out by means of the video coding system The Observer® XT[3] by Noldus, which allowed for the individual set-up of a coding scheme of behaviour groups (e.g., vocalisation: start/stop) and related modifiers (e.g., vocalisation recording background quality: Q1/Q2/Q3/Q4) in different hierarchical layers. Figure 6.5 shows the tool's coding surface. For all audio-video coding processes (segmentation and annotation) AKG K240 studio headphones were used.

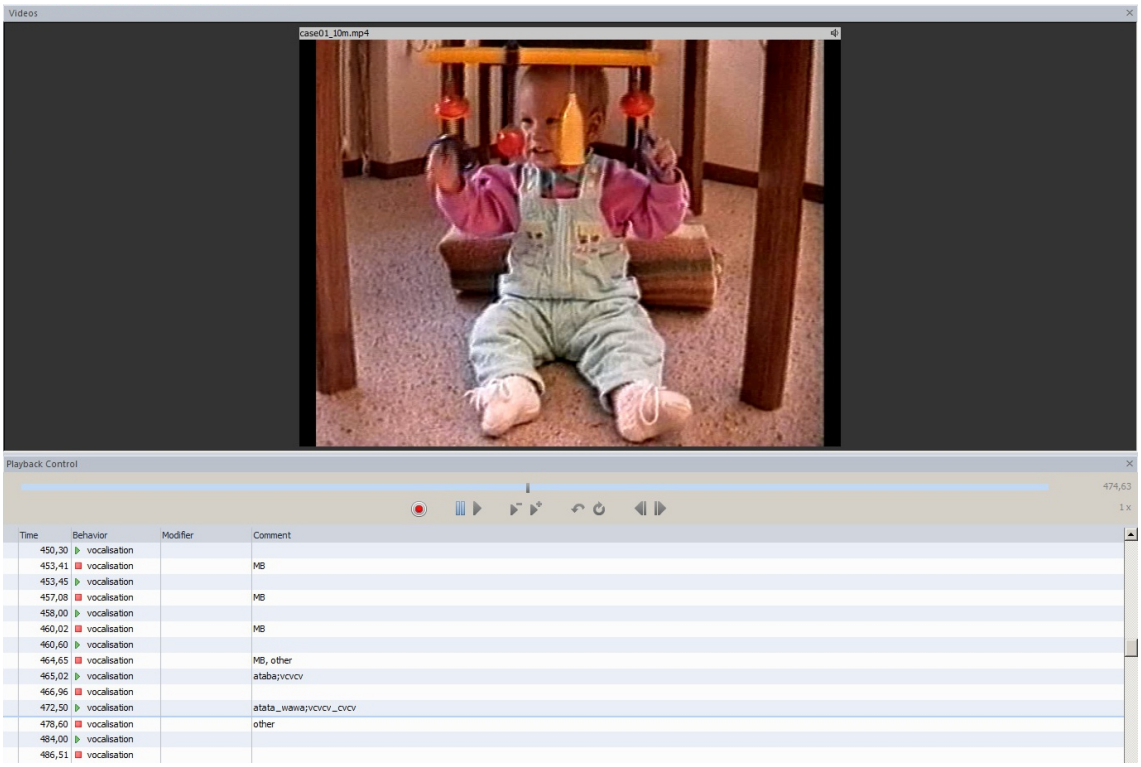


Figure 6.5: Coding surface of Noldus' The Observer® XT showing a typically developing female individual in the 10th month of life in an indoor playing situation. Examples of categorical and orthographic vocalisation type annotations can be found in the Comment column on the basis of previously created vocalisation segments with the respective start and stop times in the Time column. c = consonant(-like) sound; MB = marginal babbling; v = vowel(-like) sound; _ = juncture

[3]https://www.noldus.com/human-behavior-research/products/the-observer-xt (as of 25 May 2018)

7

Intelligent vocalisation analysis

The method of intelligent vocalisation analysis can be regarded as a special application of the field of intelligent audio analysis [139]. Whereas intelligent audio analysis deals with all types of audio signals such as speech, music, or sounds in general [139], intelligent vocalisation analysis exclusively focusses on the analysis of recorded vocal behaviour. In the framework of this thesis, vocal behaviour was further limited to pre-linguistic vocalisations of human infants. Intelligent vocalisation analysis aims for the automatic recognition of vocalisation-related (meta-)information, such as semantic behavioural patterns/states (e.g., vocalisation vs. non-vocalisation, vocalisation types, or affective states) or attributes/traits associated with the individual who produced the vocalisation (e.g., age, gender, or the medical condition of the speaker). In the framework of this thesis, the method was used for (i) the automatic detection of infant vocal behaviour within continuous audio sequences[1] (Chapter 8), and (ii) the automatic vocalisation-wise recognition of the (neuro)developmental outcome of the individual, who produced the vocalisation (Chapter 9). In principle, intelligent vocalisation analysis is related to the fields of speech and language processing (cf. [152]) and computational paralinguistics (cf. [153]). However, 'intelligent' analysis in this context means that conventional signal processing techniques for acoustic information retrieval are complemented by the involvement of artificial/computational intelligence strategies in form of machine learning algorithms [154]. Research questions in the framework of this thesis were basically associated with supervised learning paradigms. Thus, in unknown data, information of interest had to be deduced on the basis of knowledge previously learned from training data, in which the information of interest was known, e.g., as acquired or generated in a data pre-processing procedure. In this work, information used for system training was represented by (a) the posi-

[1] In context of this thesis, the automatic audio-based detection of infant vocal behaviour is ranked among applications of intelligent 'vocalisation' analysis, even though, strictly speaking, types of audio other than vocalisations, such as everyday background noise events, were also included in the analysed material (cf. Chapter 6).

tions in time of infant vocalisation segment boundaries alongside continuous audio sequences, and (b) the outcome labels of the individuals, who vocalised, respectively.

In the framework of this thesis, the method of intelligent vocalisation analysis was basically implemented – according to the general chain of audio processing in intelligent audio analysis systems [155, 139] – as a sequence of three steps, namely a feature extraction step, a feature processing step, and a classification step. In the feature extraction step, acoustic signal-level features were extracted from input vocalisations, which can be understood as the generation of each vocalisation's individual acoustic fingerprint. In the feature processing step, extracted features were either analysed or prepared in a way to obtain an advantageous/optimal representation for the subsequent classification step. Here, the actual class decisions on the input vocalisations were generated. This sequence of steps is illustrated in Figure 7.1.

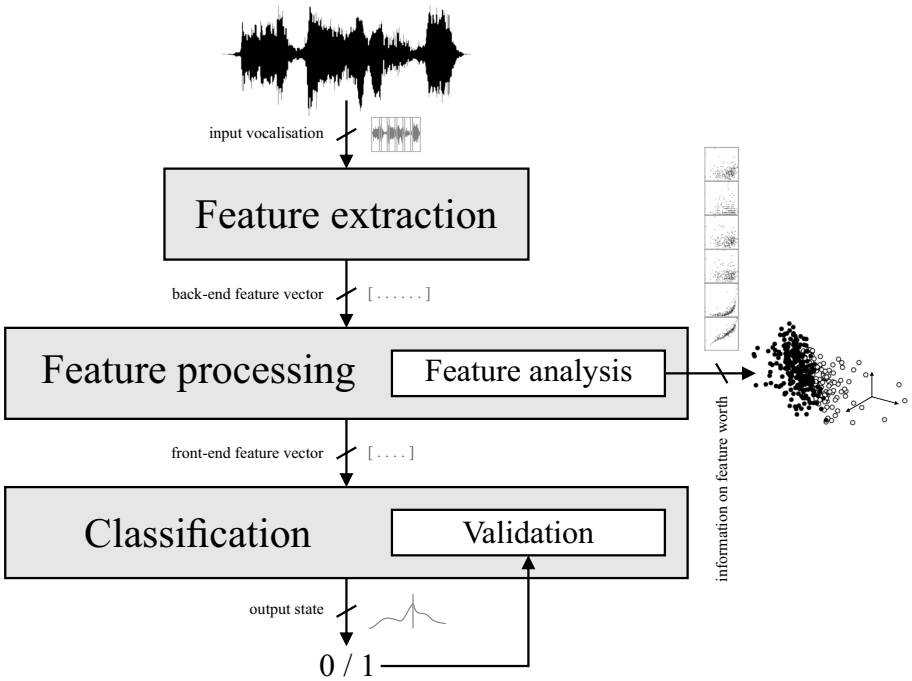


Figure 7.1: Applied procedure of intelligent vocalisation analysis for (i) the generation of output states on the basis of input vocalisations including performance validation by means feature extraction, feature processing and classification (cf. MQ5–MQ8 in Section 4.2), and (ii) the analysis of acoustic features with regard to their informative value for a specific differentiation task as part of the feature processing step (cf. MQ1–MQ4 and AQ1 in Section 4.2).

In this work, a feature-based learning approach was preferred to a representation learning approach (e.g., [156]), as in addition to the automatic recognition of an

infant's later developmental outcome with regard to the presence or absence of a COI exclusively based on pre-linguistic vocal behaviour, also potential specific relations between a COI and acoustic vocalisation characteristics were aimed to be studied (cf. MQ1–MQ4 and AQ1 in Section 4.2). On the one hand, this was because an early acoustic characterisation of a 'late recognised' developmental disorder on the basis of features can constitute the groundwork for a future approach to increase the sensitivity of healthcare professionals and, especially of caregivers/parents for early vocalisation atypicalities and, therefore, an auditory earlier recognition of deviant developmental. On the other hand, acoustic vocalisation features allow for drawing conclusions on the underlying speech production processes and, to some extent, on the involved neural processes behind. Consequently, feature analyses in the framework of this thesis can be regarded as a contribution to a better understanding of developmental disorders themselves, in case of ASD, potentially even of the cause of the disease.

In the following, the steps of feature extraction, feature processing, and classification as used for experimentation in this work are described in detail. Prior to the extraction of features, audio normalisation was tested by, first, shifting each input sequence's mean to zero and, then, setting the maximum amplitude of the corrected sequence to -3 dB. Depending on the nature of the used set of audio recordings regarding variability of vocalisation energy and the presence of background noise events, audio normalisation is a common pre-processing procedure in studies using intelligent vocalisation analysis (e.g., [157, 158, 159]). However, neither audio normalisation nor all subsequently described feature extraction, feature processing, and classification methods were applied in each experiment. For reasons of reproducibility, technical computing in the framework of this thesis was either carried out using open-source software or Matlab[2] by MathWorks®, which is widely used in the technical community.

7.1 Feature extraction

The extraction of acoustic features from audio recordings represented the basic step for further analyses in the framework of this thesis. Acoustic features are mathematical descriptors defined with the intention to characterise audio signals in a compact, informative, but preferably non-redundant way, that is meaningful/optimal for subsequent processing/analysis steps or tasks/applications of interest [139]. Thus, optimal acoustic features for speech applications (cf. [153]) might differ from optimal acoustic features for applications of music information retrieval (cf. [160]). Anyway, the transformation of an audio signal into a meaningful feature representation basically implies a reduction of information [139].

[2]https://de.mathworks.com/products/matlab.html (as of 5 June 2018)

Natural audio signals such as speech signals are usually time variant, i.e., they change over time [161]. This also holds true for infant vocal behaviour. Therefore, the extraction of acoustic features is usually carried out on the basis of short, window-function-weighted, overlapping time frames, within which audio information can be considered as stationary [161]. Descriptors calculated at this first level of signal sub-sampling are denominated as low-level descriptors (LLDs). By extracting further information, such as statistical functionals from frames alongside LLD contours, so called higher-level descriptors or supra-segmental features can be generated [139, 153].

In the framework of this thesis, experiments on two basically different intelligent vocalisation analysis paradigms were carried out. On the one hand, different algorithms for the automatic detection of infant vocalisations within continuous audio recordings were implemented and evaluated. On the other hand, the basic feasibility of a vocalisation-based recognition of an infant's developmental outcome in a COI vs. TD design was investigated. Whereas in the first paradigm, infant voice had to be acoustically differentiated from other recorded sound sources, in the second paradigm, meta-information-related discrimination between different vocalisations had to be managed. According to these different paradigm-individual requirements, different sets of acoustic features were employed throughout the different experiments of this work.

Experiments on infant voice activity detection were based on two different sets of features extracted for overlapping frames of 25 ms with a step size of 10 ms between consecutive frames.

The first set contained 100 features including, on the one hand, 40 established features for speech analysis applications [162, 163, 152], namely pitch, 13 Mel-frequency cepstral coefficients (MFCCs), 13 perceptual linear predictive coding coefficients (PLPCCs), and 13 relative spectral transform PLPCCs (RASTA-PLPCCs), and on the other hand, 60 features suggested by Temko and Nadeu [164] for general acoustic event detection purposes and used in the 2006/2007 CLEAR (Classification of Events, Activities and Relationships) Challenges [165, 166, 167]. Extraction of these 100 features was performed by means of Matlab.

The second feature set consisted of line spectral frequencies, Mel spectra, and energy, and constituted the set of a standard rule-based voice activity detection implementation, which was used for baseline evaluations only (cf. Chapter 8). Baseline infant voice activity detection as well as the extraction of features for all experiments on the paradigm of developmental disorder recognition were carried out by means of openSMILE[3] by audEERING™ *GmbH*.

[3]https://audeering.com/technology/opensmile/ (as of 5 June 2018)

openSMILE is an open-source, C++-based, real-time capable feature extraction tool kit [168, 169]. Originally, openSMILE was designed for both speech and music applications, which explains 'SMILE' to be the acronym for 'Speech and Music Interpretation by Large-space Extraction'. The development of openSMILE began in 2008 at the Technical University of Munich, Germany. Hosted by audEERING™ *GmbH* since 2014, openSMILE is still continuously updated, enhanced, and provided free of charge for non-commercial research use. Consequently, the tool is widely used and allows for comparability and reproducibility of feature-based experiments throughout the research community.

For investigating the feasibility of automatic vocalisation-based developmental disorder recognition, again two different feature sets were employed in the framework of this thesis. Ensuring reproducibility, suitability for audio/speech analysis applications, and currentness, both used feature sets are standard sets included in the recent release of openSMILE (version 2.3).

The first set is denominated 'ComParE set' as it represented the official baseline feature set of the 2013–2018 Computational Paralinguistics and Emotion Challenges [154, 170, 171, 159, 172, 133] that have been carried out in connection with the Annual Conferences of the International Speech Communication Association (Interspeech conferences) since 2009. Comprising 6 373 acoustic features, the ComParE set is still the most comprehensive standardised feature set contained in openSMILE. Due to its extensiveness, it is a popular 'out-of-the-box'/'brute-force' choice for baseline/feasibility experiments and a powerful basis for subsequent feature selection/analysis steps. The ComParE set consists of higher-order features in form of statistical functionals that were computed for the trajectories of a wide range of acoustic time-, energy-, and/or spectral/cepstral-based LLDs as well as their first order derivatives (see Table 7.1) [154].

As an alternative to the comprehensive ComParE set, openSMILE's most current feature set was applied: The so-called 'extended Geneva Minimalistic Acoustic Parameter Set' (eGeMAPS) was launched and published in 2016 by Eyben and colleagues [173] and represents a comparatively small set of 88 acoustic higher-order features (see Table 7.1) that were selected based on their theoretical and practical relevance for applications of automatic voice analysis also including clinical applications as well as their proven value in previous studies in field.

Keeping the standard settings, LLD trajectory extraction for both the ComParE set and the eGeMAPS was based on overlapping time frames of 60 ms alongside an input vocalisation at a frame step size of 10 ms. Extracted LLD contours were smoothed using a symmetric moving average filter over an interval of three frames (except for unvoiced–voiced transitions in pitch, jitter, and shimmer contours). Finally, statistical functionals were calculated for the smoothed LLD contours over the individual length of each vocalisation (variable frame size) resulting in exactly one vector of 6 373 or 88 higher-order feature values per vocalisation, respectively.

Table 7.1: Overview (in alphabetical order) of LLDs and statistical functionals calculated for LLD contours contained in the ComParE set [154] and in the eGeMAPS [173] (not all functionals are applied to all LLDs; in the eGeMAPS functionals are basically calculated for voiced regions, i.e., regions of non-zero F0 only, however, functionals of specific LLDs are exclusively or additionally calculated for all regions and/or unvoiced regions; for details see [173]), as well as additional features contained in the eGeMAPS. A3 = amplitude of harmonic closest to third formant; ComParE = Computational Paralinguistics and Emotion; eGeMAPS = extended Geneva Minimalistic Acoustic Parameter Set; F0 = fundamental frequency; H1 = amplitude of first harmonic; H2 = amplitude of second harmonic; LLD = low-level descriptor; MFCC = Mel-frequency cepstral coefficient; LP = linear prediction; RASTA = relative spectral transform; RMS = root mean square; SD = standard deviation; # = number of; * = used in the ComParE set only; ** = used in the eGeMAPS only; neither* nor **: used in both the ComParE set and in the eGeMAPS

LLDs	Functionals
Alpha ratio**	Amplitude mean of minima*
Formant 1**, 2**, 3** bandwidth	Amplitude mean of peaks*
Formant 1**, 2**, 3** frequency	Amplitude range of peaks*
Formant 1**, 2**, 3** relative energy	Arithmetic mean
F0	Arithmetic mean value of peaks*
Hammarberg index**	Coefficient of variation**
Harmonic difference H1–A3**	Contour centroid*
Harmonic difference H1–H2**	Contour flatness*
Harmonics-to-noise ratio	Interquartile ranges (3)*
Jitter	Kurtosis*
Loudness (sum of auditory spectrum)	LP coefficient 1–5*
MFCC 1–4, 5–14*	LP gain*
Psychoacoustic harmonicity*	Linear regression offset*
Psychoacoustic sharpness*	Linear regression quadratic error*
RASTA-filtered auditory spectral band 1–26*	Linear regression slope*
RASTA-filtered auditory spectrum sum*	Maximum of segment length*
RMS energy*	Mean of inter-peak distance*
Shimmer	Mean of falling slopes
Spectral energy 250–650 Hz*, 1–4 kHz*	Mean of rising slopes
Spectral centroid*	Mean of segment length*
Spectral entropy*	Mean value of peaks*
Spectral flux	Minimum of segment length*
Spectral kurtosis*	Percentage of non-zero frames*
Spectral roll-off point 0.25*, 0.5*, 0.75*. 0.9*	Percentile 1*, 20**, 50**, 80**, 99*
Spectral skewness*	Percentile range 1–99*, 20–80**
Spectral slope*, 0–500 Hz**, 500–1500 Hz**	Position of minimum*
Spectral variance*	Position of maximum*
Voicing probability*	Quadratic regression a*
Zero-crossing rate*	Quadratic regression b*
	Quadratic regression offset*
Additional features	Quadratic regression quadratic error*
Equivalent sound level**	Range (maximum – minimum)*
Mean length of voiced regions**	Rel. dur. > 25%*, 50%*, 75%*, 90%* range
Mean length of unvoiced regions**	Relative duration positive curvature*
# continuous voiced regions per second**	Relative duration rising*
Rate of loudness peaks**	Root quadratic mean*
SD of length of voiced regions**	SD*
SD of length of unvoiced regions**	SD of inter peak distance*
	SD of falling slopes
	SD of rising slopes
	SD of segment length*
	Skewness*
	Quartile 1–3*

7.2 Feature processing

In the framework of this thesis, feature processing subsumes strategies (i) to adapt the representation of previously extracted features to the intended classification procedure in order to optimise classification performance, or (ii) to analytically deduce further information from the raw features in order to systematically gain insights into acoustic characteristics underlying a specific classification task.

As this thesis aimed at investigating the basic feasibility of a novel approach, one common, minor feature adaptation step and standard feature analysis methods were applied only. Thereby, the number of degrees of freedom in the single experiments could be kept as small as possible. Feature processing/analysis was basically carried out by means of Matlab, or by means of Weka[4], which is a widely used Java-based open-source data mining and machine learning tool kit provided by the University of Waikato, New Zealand [174, 175].

Considering potential dataset-related influences on feature distributions with respect to absolute feature values (e.g., speaker/infant-dependent specificities and age-dependent variations regarding physiological/anatomical changes of the vocal folds and in the vocal tract or regarding varying predominant vocalisation types over the second half year of life), feature normalisation/standardisation was tried out according to different strategies. On the one hand, features were normalised to a defined interval, such as $[0, 1]$. On the other hand, features were standardised to have a mean of zero and unit variance. Feature normalisation/standardisation was carried out either speaker/infant-dependently, i.e., in separate for each infant's instances, (infant-)age-dependently, i.e., in separate for instances of each available month of age of an infant, or globally, i.e., on the basis of all instances, contained in a respective set of data.

7.2.1 Feature analysis

In order to investigate the value of each individual acoustic feature in context of a vocalisation-based differentiation between COI and TD or between different COI (cf. MQ1–MQ4 and AQ1 in Section 4.2), statistical tests of difference were applied. As feature values extracted throughout this work were basically not always normally distributed, non-parametric tests were used, namely the Mann-Whitney U-test in two-class paradigms, e.g., when examining differences between a COI and TD (cf. Sections 9.1–9.3), and the Kruskal-Wallis test in paradigms of more than two classes, e.g., when examining differences between multiple COI and TD (cf. Section 9.4). The significance level was set to $\alpha = 0.05$. For each feature, these tests were used to evaluate the null hypothesis that feature values extracted from vocalisations of

[4] https://www.cs.waikato.ac.nz/ml/weka/ (as of 13 June 2018)

individuals with a COI, in case of a paradigm of more than two classes, feature values extracted from vocalisations of individuals with an other COI, and feature values extracted from vocalisations of TD individuals, were likely to derive from the same population [176]. Features were then ranked according to the effect size estimate r (see Equation 7.1 [177]).

$$r = \frac{z\text{-}value}{\sqrt{\#\ samples}} \tag{7.1}$$

7.3 Classification

The final step of the applied intelligent vocalisation analysis procedure, classification, is understood as a task in which an audio sequence/a vocalisation 'under test' has to be assigned to a specific category/class [29], such as a specific medical condition related to the individual who produced the vocalisation. In this work, assignment decisions were generated by classifiers on the basis of previously extracted and optionally processed acoustic features. In this context, a classifier refers to the entity that performs a classification tasks, i.e., a computational decision making unit operating on the basis of a specific classification algorithm. As already mentioned at the beginning of Chapter 7, the classifiers applied in the framework of this thesis built on supervised machine learning algorithms, i.e., the intended assignment of a sequence of audio/a vocalisation to one of different discrete classes was made on the basis of knowledge/a set of rules the classifier had previously acquired/learned from training data in which the assignments between the contained audio sequences/vocalisations and the possible class labels were given [178]. In the field of intelligent audio analysis, classifiers can be roughly divided into two groups or 'learners', namely static learners and dynamic (or sequential) learners [139, 153] both designable to operate on features extracted from an audio sequence/a vocalisation. While static learners exclusively handle acoustic information within single audio frames, e.g., one feature vector for one audio sequence/vocalisation, dynamic learners operate on the acoustic content's time trajectory, i.e., the sequence of acoustic information within consecutive frames over time [139].

The two scenarios dealt with in the framework of this thesis, namely (i) the scenario of detecting infant vocalisations in recorded audio according to the first module of the proposed, fully automatic developmental disorder recognition tool (cf. Figure 4.1 in Section 4.2), and (ii) the scenario of the actual vocalisation-based recognition of an infant's developmental outcome with respect to the presence or absence of a COI according to the proposed tool's second module, posed different

requirements for the application of a learning algorithm and its performance validation. However, due to their widespread use for feasibility/baseline evaluations (e.g., [154, 170, 171, 159, 172, 133]), experimentation on either scenario was primarily carried out using support vector machines (SVMs), which are a representative of static learners.

The first scenario, i.e., the identification of sequences of infant voice alongside a recorded audio sequence actually does not represent a classification task, but a detection task. In contrast to the assignment of a frame of audio to a specific class, in a detection task, the presence or absence of specific states has to be evaluated, e.g., frame by frame, alongside a usually continuous audio sequence. Apart from SVMs, for the detection of infant voice in audio recordings, another frequently used static learner and a popular dynamic learner were tested. The additional static learner was a random forest (RF) classifier, the dynamic learner was a hidden Markov model (HMM). Furthermore, baseline evaluations were generated using a standard rule-based approach, i.e., an algorithm does not performing on the basis of previously learned knowledge, but making deterministic decisions as a function of implemented rules relying on expert knowledge [153].

For the second scenario, i.e., the classification of vocalisations according to the developmental outcomes of the individuals who produced the vocalisations, in one of the experiments, one of the most popular, currently used learning approaches was tested in addition to the SVM approach, namely an artificial neural network (ANN). In particular, a bidirectional long short-term memory neural network (BLSTMNN) was employed, which is a special implementation of an ANN capable of modelling temporal context.

In the following, all learning algorithms applied in the framework of this thesis are briefly described. Classifier implementation/design and validation were basically carried out using Matlab, Weka, and/or TensorFlow™[5], which is an open-source library widely used in machine learning practice across different scientific domains as it allows for high performance numerical computation on the basis of a flexible architecture [179].

Introduced in 1995 by Cortes and Vapnik [180], SVMs are binary decision making units, i.e., they are capable of discriminating between two classes by directly providing class identities instead of posterior probability outputs. Multiple-class paradigms can be processed through strategies of combining binary SVM decisions. The principle of SVMs relies on statistical learning and optimisation. The basic idea is a non-linear mapping of input feature vectors to a higher dimensional decision space in which an optimal hyperplane of high generalisation ability is determined on the basis of training instances in order to achieve a best possible linear discrimination between the two classes. For mapping into the higher dimensional decision

[5]https://www.tensorflow.org/ (as of 19 June 2018)

space, a kernel function is applied. This is called the 'kernel trick'. The margins of the largest separation between the two classes are defined by the so-called 'support vectors'. These are determined by solving a quadratic optimisation problem. Based on the support vectors, classification only builds on a subset of learning instances. Thereby, SVMs can handle from large feature spaces to features being zero most of the time while the risk of overfitting to the learning instances is reduced. [180, 181, 139, 153, 29]

In the framework of this thesis, both linear kernels and a Gaussian kernel were employed. The kernel complexity parameter C, which is a factor regulating the trade-off between training error and generalisation ability of the classifier, and in case of the Gaussian kernel, additionally the kernel width γ, were optimised on respective data(sub)sets. The sequential minimal optimisation (SMO) algorithm [182] was selected for SVM training.

Competitive to the SVM approach, the RF classifier is another supervised learning and decision making algorithm. It is based on the concept of decision trees. Decision trees generate output states for input observations, e.g., feature vectors, on the basis of a sequence of rules that are, for example, implemented in the form of comparisons to constants alongside the tree's branches. In contrast to a standard rule-base classification approach, in which rules are base on expert knowledge, here the optimal sequence of rules is automatically derived from training data. [139, 153]. Finally, RFs are ensembles of multiple decision trees constructed by random feature and training set sub-sampling as defined by ensemble learning methodology [183, 184, 185, 186, 187, 139, 153]. Thereby, the sensitivity of decision trees for overfitting to training data is corrected [188].

The number of trees, the maximum tree depth, or the minimum number of samples per leaf represent design parameters of RFs. Experimentation in the framework of this thesis was designed to determine these parameters in an optimisation task.

A state-of-the-art alternative to SVMs or RFs are ANNs, especially ANNs with specific architecture-related properties. Over the last years, for example, BLSTMNNs have proven to be powerful classification models for audio applications, also including speech-related tasks (e.g., [189, 190, 191, 192, 193]). Inspired by neuro-physiological structures in vertebrates, ANNs are generally based on neurons as main information processing units, which are connected to other neurons via synapses according to a specific network topology. Information is propagated across the network as a function of the properties related to neurons (mathematical capacities) and connections (e.g., weights). Thereby, ANNs allow for learning arbitrary, even non-linear functions. Typically, ANNs are structured into different layers of neurons potentially exhibiting different properties. Information propagates from the input layer via one or more hidden layers to the output layer and, thereby, gets

processed from observational input data, e.g., in form of feature vectors, to class predictions. [194, 181, 139, 153]

The architecture of BLSTMNNs is intended for the classification of time series as it builds on recurrent neural networks (RNNs). RNNs are characterised by both forward and backward (recurrent) connections between neurons of different layers, which yields a kind of memory. Consequently, not only a current input, but also the history of previous inputs can influence the current output and future outputs of the network. The implementation of a bidirectional RNN (BRNN) enables the accessibility of the full input sequence in forward as well as in backward direction at any time by using two independent hidden layers – one responsible for input sequence forward processing, the other for input sequence backward processing – both connected to the same previous/input and subsequent/output layers. Finally, a BLSTMNN is a BRNN equipped with long short-term memory (LSTM) layers of LSTM blocks instead of the original neurons. An LSTM block comprises at least one linear cell featuring an internal recurrent connection weighted with 1 ('constant error carousel'). In addition, the dynamic data flow including a potential storing of new content and an erasing of content is controlled by multiplicative units ('input', 'output' and 'forget gates'). While a BRNN is limited with regard to the range of manageable temporal context as dependent on the weights of recurrent connections (< 1 or ≥ 1) the input's influence decays or blows-up exponentially ('vanishing gradient problem' [195]), LSTM cells, yield a more functional, potentially permanent memory and allow for considering an optimal amount of contextual information for a specific learning paradigm. [29, 139, 153]

For BLSTMNN implementation in the framework of this thesis, the popular 'vanilla architecture' was used [196]. As a BLSTMNN allows for modelling time series, each vocalisation was input in the form of its extracted, smoothed LLD trajectories based on time steps of 10 ms (cf. Section 7.1). For training, the first-order gradient-based Adam optimisation algorithm [197] was employed. The optimum number of layers and cells was determined by performing a grid search.

In the end, HMMs are among the most frequently encountered learning models for audio sequence classification. Both Markov chains and HMMs can be regarded as extended finite state automata, i.e., they build on sets of states and sets of transitions between the states with associated probabilities that indicate how likely a respective transition is to be taken. At any time step, the state may be changed. In a Markov chain, a specific input sequence uniquely determines the sequence of states stepped through by the automaton. This is useful when needing to calculate the probability of a sequence of observable events. However, in HMMs the states are unobservable, latent attributes presumed to have a causal relation to observational data. An observation in form of a feature vector is modelled and generated as a function of the current state and a respective emission probability at each time step. Thus in a recognition task, the sequence of the most probable 'hidden' states is to be

derived from a given sequence of state-related observational data. Underlying model parameters have to be previously determined on the basis of training observations with the related states given. [198, 152, 181, 139, 153]

Regarding HMM implementation in the framework of this thesis, state prediction was performed by means of the Viterbi algorithm [199, 200]. As usual for intelligent audio analysis applications, the emission probabilities were modelled by Gaussian mixtures [139, 153].

7.3.1 Validation

To evaluate the performance of (learning) algorithms applied in the framework of this thesis for (i) the automatic detection of infant vocalisations within audio recordings, and (ii) the automatic classification of vocalisations according to the developmental outcomes of the infants, who produced the vocalisations, respectively, different evaluation strategies were pursued.

Infant voice activity detection validation was based on each audio recording's vocalisation segmentation reference information, i.e., the manually set start and stop markers for infant vocalisations alongside each recording. The set of available audio recordings was split into a partition for training and tuning the employed learning algorithms, and a test partition for performance evaluation. The detection output of each employed approach was given in form of a raw binary data stream alongside each recording of the test partition on a time basis defined by the feature extraction process, i.e. in this case, on a frame basis of 10 ms steps (cf. Section 7.1). In this binary data stream, the value '1' indicated that infant voice was detected, the value '0' indicated that no infant voice was detected in a respective frame. In order not to allow for unrealistic vocalisation detection outputs, a moving median filter was applied to the raw data streams prior to the validation process. Thereby, an isolated value or too short sequences of the same value surrounded by sequences of the other value were eliminated. Consequently, implausibly short detected vocalisation periods were precluded and too short interruptions between two detected vocalisation periods were prevented by generating one single vocalisation of extended duration. The length of the moving median filter was set to 15 frames according to the duration of the shortest reference vocalisation contained in the training set ($\sim$150 ms). By setting the filter length to an odd frame number, no further processing step was necessary to ensure that the filtered data stream still exclusively contained the values '1' and '0'. The filter length of 15 frames led to a duration threshold for sequences of isolated values in a row to be eliminated of half of the length of the shortest reference vocalisation within the training set. Accordingly, sequences of less than 8 same values in a row preceded and followed by sequences of at least 8 times the other value, respectively, were eliminated, i.e., set to the value of the surrounding sequences.

Infant voice activity detection performance was evaluated in two different ways, namely frame-based on the one hand, and vocalisation-based on the other hand.

In the frame-based evaluation approach, the filtered binary output stream of each employed detection algorithm was compared to the vocalisation segmentation reference information frame by frame. In doing so, true positives (TPs; value '1' in the detection output stream for a frame within the period of a reference vocalisation), false negatives (FNs; value '0' in the detection output stream for a frame within the period of a reference vocalisation), and false positives (FPs; value '1' in the detection output stream for a frame outside the period of a reference vocalisation) were counted.

In the vocalisation-based evaluation approach, the locations of detected vocalisation periods, i.e., sequences of the value '1' in the filtered detection output stream, were compared to the locations of reference vocalisations. According to Phan and colleagues [201], in this scenario a reference vocalisation was regarded as correctly identified (TP), either if the centre of a detected vocalisation was located within the boundaries of the reference vocalisation, or if the centre of the reference vocalisation was located within the boundaries of a detected vocalisation. A FN was counted if a reference vocalisation was not correctly identified according to the afore-explained criterion concerning the centre of vocalisations. Finally, a FP was given if a detected vocalisation did not include the centre of a reference vocalisation or the detected vocalisation's centre was not included in a reference vocalisation. The assignment of TPs, FNs, and FPs for the vocalisation-based evaluation scenario is exemplified in Figure 7.2.

On the basis of the determined TPs, FNs, and FPs, for both the frame-based and the vocalisation-based validation scenario, four common measures for evaluating the performance of acoustic event detection algorithms were calculated according to Phan and colleagues [201] and following the 2006/2007 CLEAR challenges [165, 167], namely precision, recall, the acoustic event detection accuracy (AED-ACC $\equiv$ F-measure), and the acoustic event error rate (AEER; see Equations 7.2–7.5).

$$precision = \frac{\#\,TPs}{\#\,TPs + \#\,FPs} \tag{7.2}$$

$$recall = \frac{\#\,TPs}{\#\,TPs + \#\,FNs} \tag{7.3}$$

$$AED\text{-}ACC = \frac{2 * precision * recall}{precision + recall} \tag{7.4}$$

$$AEER = \frac{\#\,FNs + \#\,FPs}{\#\,TPs + \#\,FNs} \tag{7.5}$$

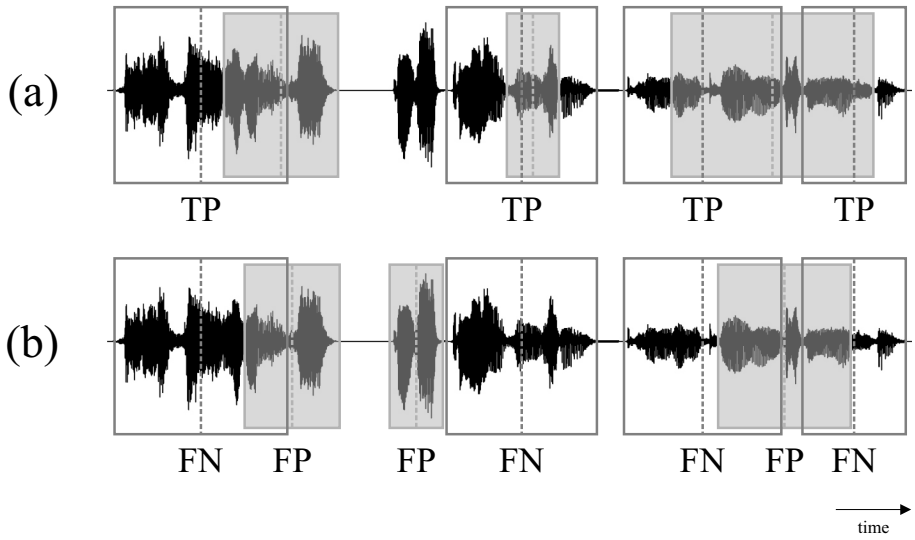


Figure 7.2: Exemplification of comparing reference vocalisations (dark grey frames) and detected vocalisations (light grey frames with filling) according to their relative positions of vocalisation centres (dark grey and light grey dashed lines, respectively) in order to assign (a) true positives (TPs), and (b) false negatives (FNs) and false positives (FPs) for the vocalisation-based evaluation scenario.

The frame-based evaluation delivered precise performance information within the scope of detecting fragments of infant vocal patterns in a continuous audio stream irrespective of their role as elements of larger entities, such as vocalisations, whereas the vocalisation-based evaluation referred to the detection of the rough locations of infant vocalisations alongside the audio stream. Both the frame-based evaluation approach and the vocalisation-based evaluation approach have their justification in context of an intended fully automatic infant vocalisation detection system. However, as in the vocalisation-based evaluation approach one detected vocalisation period of extended duration could lead to more than one reference vocalisations to be correctly identified (one single segment over a whole recording would lead to all reference vocalisations within the recording to be correctly detected without causing any FPs), the respective performance measures must not be interpreted without also carefully considering the results from the frame-based scenario, especially the precision value.

For evaluating the performance of the automatic classification of infant vocalisations according to the infants' developmental outcomes with respect to the presence or absence of a COI, cross-validation strategies were applied. For this purpose, the datasets of the individual experiments were split into, as far as possible, equally sized, speaker/infant independent partitions alternately used for classifier training, classifier optimisation if applicable, and classifier testing throughout the single iterations of the cross-validation scheme. In this context, 'infant independent' means that

vocalisations of one and the same infant were never split up over different partitions and, therefore, could not be used for both classifier training and testing during one iteration of the cross-validation procedure. In case of very small datasets available for individual experiments, leave-one-matched-speaker-group-out cross-validation was employed. This means that each created partition only contained vocalisations of a matched group of one infant per class, i.e. in case of a two-class scenario, one infant later diagnosed with a COI and one matched TD control. In case of highly imbalanced training partitions with regard to the number of instances per class, i.e., vocalisations associated with a specific COI and vocalisations associated with TD, the instances of the underrepresented class were upsampled.

Analogous to the evaluation of the infant voice activity detection performance, also the performance of vocalisation classification was carried out in two different ways, namely vocalisation-wisely and infant-wisely. In the first approach, the vocalisation-wise classification decisions were evaluated, whereas for the infant-wise approach a global decision on each infant was generated on the basis of the single decisions on all of his or her vocalisations, e.g., by applying majority vote[6]. Both the vocalisation-wise scenario and the infant-wise scenario make sense for evaluating the basic feasibility of an automatic early COI recognition tool.

In accordance with all so far held Interspeech[7] challenges (2009–2018 [202, 203, 204, 205, 154, 170, 171, 159, 172, 133]; since 2013 called 'ComParE challenges'), the unweighted average recall/unweighted accuracy (UAR/UA; average of class-specific recall values [139]; for recall see Equation 7.3) was selected to be the primary measure for vocalisation classification performance in the framework of this thesis. As compared to the (class) weighted average recall/weighted accuracy (WAR/WA; average of class-specific recall values weighted according to the distribution of instances per class in the test dataset $\equiv$ number of correctly identified instances divided by the total number of instances in the test dataset [139]), the UAR is more adequate for class imbalanced (test) datasets [139]. On the one hand, the UAR was calculated in separate for each iteration of the cross-validation procedure and then averaged over all iterations. On the other hand, the predictions of each iteration were stored and, finally, the UAR was calculated globally for the whole dataset.

[6]In the particular case of exactly the same number of class-wise predictions (representing the majority), the involved class with the smaller/smallest internal numerical identifier was arbitrarily defined to be selected.

[7]Annual Conference of the International Speech Communication Association (cf. Section 7.1)

Part III

Experiments

8

Infant voice activity detection

The first experiment conducted in the framework of this thesis dealt with the automatic detection/segmentation of infant vocalisations in audio recordings. Thereby, the experiment addressed AQ3 of this thesis (cf. Section 4.2) and refers to the first module of the proposed, fully automatic developmental disorder recognition tool (cf. Figure 4.1 in Section 4.2). Parts of the analysis of this experiment were carried out by colleagues, namely by Dr. Robert Peharz, currently affiliated with the Machine Learning Group at the University of Cambridge, UK (at the time of experimentation he was with the Research Unit iDN, at the MUG, Austria), and by Wolfgang Roth from the Signal Processing and Speech Communication Laboratory, at the Graz University of Technology, Austria. Experiment-related procedures and findings were published in 2016 by Pokorny and colleagues [206].

In order to investigate automatic infant voice activity detection in consideration of representing the audio input pre-selection step for a complex vocalisation classification system intended for application in clinical but also home scenarios, experimentation was carried out on the basis of audio-video material recorded under real-world conditions. For this experiment, more than 20 hours of home video material were extracted from GUARDIAN's Dataset 4 (cf. Section 6.1). The material comprised variable quality audio-video recordings of 32 participants[1] including both female and male TD individuals and individuals later diagnosed with a COI. Moreover, the material covered individuals of the four different nationalities Austria, Germany, Italy, or United Kingdom, each having one of the three different mother tongues/family languages German, Italian, and English. In total, the dataset used for the experiment on infant voice activity detection consisted of 4 903 vocalisations from the included individuals' pre-linguistic period, more precisely, from the individuals' respective second half year of life. A detailed overview of the dataset is given in Table 8.1.

[1]Participant identification codes (IDs) throughout this thesis are unique, i.e., one and the same ID, such as 'TD01', in different experiments implies one and the same participating individual.

Table 8.1: Number of included vocalisations on the basis of the available audio-video duration in format hh:mm:ss (two-digit hour number:two-digit minute number:two-digit second number; seconds rounded down to integer values; sums calculated on the basis of exact durations) as a function of age per participant for the experiment on infant voice activity detection. This table includes information of Table 1 published by Pokorny and colleagues [206] (DOI: 10.21437/Interspeech.2016-1341). ASD = autism spectrum disorder; AT = Austria; DE = Germany; Eng = English; FXS = fragile X syndrome; Gen = gender; Ger = German; ID = (unique participant) identification code; IT = Italy; Ita = Italian; L1 = mother tongue; Nat = nationality; RTT = Rett syndrome; TD = typical development/typically developing; UK = United Kingdom; ♀ = female; ♂ = male; - = no audio-video material available

ID	Gen	Nat/L1	Age [month of life]						Σ
			7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	
ASD01	♂	IT/Ita	-	-	-	3 00:04:13	32 00:04:47	13 00:02:56	48 00:11:56
ASD02	♂	IT/Ita	-	73 00:13:46	3 00:03:28	24 00:04:55	23 00:04:53	10 00:05:45	133 00:32:50
ASD03	♂	IT/Ita	-	-	-	-	-	41 00:05:08	41 00:05:08
ASD04	♂	IT/Ita	-	3 00:01:09	-	-	6 00:03:26	-	9 00:04:36
ASD05	♂	IT/Ita	-	-	-	-	-	3 00:07:39	3 00:07:39
ASD06	♂	IT/Ita	1 00:00:47	-	-	-	19 00:02:08	1 00:02:14	21 00:05:10
ASD07	♂	IT/Ita	-	-	3 00:00:48	-	-	-	3 00:00:48
ASD08	♂	IT/Ita	-	-	-	-	1 00:02:15	130 00:13:34	131 00:15:50
ASD09	♂	IT/Ita	-	-	-	33 00:03:10	2 00:03:02	-	35 00:06:13
ASD10	♂	IT/Ita	5 00:01:15	4 00:02:45	-	1 00:05:18	48 00:09:35	26 00:07:19	84 00:26:14
ASD11	♂	IT/Ita	2 00:03:18	12 00:07:38	80 00:24:49	11 00:03:26	-	4 00:01:51	109 00:41:03
ASD12	♂	IT/Ita	35 00:02:48	5 00:07:48	-	-	-	10 00:03:07	50 00:13:45
ASD13	♂	IT/Ita	24 00:05:27	-	-	-	-	5 00:03:07	29 00:08:34
FXS01	♂	AT/Ger	-	-	87 00:17:05	-	-	-	87 00:17:05
RTT01	♀	AT/Ger	-	-	-	150 00:13:46	-	-	150 00:13:46
RTT02	♀	AT/Ger	-	-	138 00:42:53	120 01:02:18	46 00:19:05	-	304 02:04:16
RTT03	♀	DE/Ger	-	363 00:57:45	273 00:44:25	323 01:23:21	180 00:57:42	243 00:50:36	1 382 04:53:51
RTT04	♀	DE/Ger	73 00:10:23	26 00:04:25	57 00:08:35	73 00:15:28	68 00:08:47	66 00:12:41	363 01:00:21
RTT05	♀	UK/Eng	-	3 00:00:29	-	1 00:00:31	10 00:02:36	1 00:00:31	15 00:04:09
RTT06	♀	UK/Eng	-	44 00:16:37	-	19 00:04:01	46 00:07:56	-	109 00:28:35
RTT07	♀	UK/Eng	8 00:05:16	-	-	-	-	-	8 00:05:16
RTT08	♀	UK/Eng	-	116 00:20:34	7 00:02:26	83 00:18:43	9 00:03:40	20 00:09:32	235 00:54:57
RTT09	♀	UK/Eng	7 00:07:09	-	1 00:02:36	4 00:12:15	-	7 00:13:28	19 00:35:30
TD01	♂	AT/Ger	-	22 00:05:44	-	-	-	-	22 00:05:44
TD02	♂	AT/Ger	9 00:06:36	-	-	4 00:03:36	21 00:06:14	5 00:02:59	39 00:19:27
TD03	♂	AT/Ger	32 00:19:07	20 00:09:39	11 00:11:55	-	4 00:03:39	46 00:22:03	113 01:06:26
TD04	♂	AT/Ger	2 00:01:04	2 00:00:27	-	-	-	5 00:03:22	9 00:04:55
TD05	♂	AT/Ger	-	148 00:24:25	-	160 00:17:40	-	-	308 00:42:06
TD06	♀	AT/Ger	-	3 00:03:43	-	3 00:01:48	-	-	6 00:05:32
TD07	♀	AT/Ger	12 00:06:01	-	1 00:03:19	-	-	2 00:00:51	15 00:10:12
TD08	♀	AT/Ger	231 01:11:58	97 00:32:04	89 00:16:02	80 00:16:02	78 00:15:07	106 00:12:47	681 02:44:01
TD09	♀	AT/Ger	9 00:07:35	48 00:28:54	84 00:18:22	102 00:30:41	51 00:11:46	48 00:18:18	342 01:55:40
		Σ	450 02:28:53	989 03:58:01	834 03:16:48	1 194 05:01:20	653 02:46:47	792 03:19:58	4 903 20:51:50

The mean vocalisation duration was 1.72 s (± 1.41 s standard deviation), the median duration was 1.33 s. The shortest vocalisation had a duration of 150 ms. The longest vocalisation had a duration of 21.31 s, which was an extended sequence of crying patterns by TD03 in the 9^{th} month of life. The total duration of infant vocalisations within the dataset was 2 hours, 20 minutes, and 19 seconds. Consequently, vocalisation periods made up 11.2% of the total audio-video duration.

The number of studies on an automatic detection of infant vocal behaviour (e.g., [207]) not exclusively focussing on specific infant vocalisation types, such as crying, is limited most likely due to the so far limited number of application areas for such a detection paradigm. However, following an obvious trend in the highly researched field of voice activity detection in general, namely the replacement of traditional rule-based voice activity detection (e.g., [208, 209, 210, 211, 212, 213]) by machine learning-based voice activity detection (e.g., [214, 215, 216, 217, 218, 219, 220, 221]) over the last years, in this experiment three different machine learning approaches, namely an SVM, an RF, and HMMs as already outlined in Section 7.3, were implemented and tested against a standard rule-based approach constituting the experiment's baseline. For separate detection model learning/optimisation and detection performance evaluation, the experiment's dataset was split into a training/development and a test partition. Partitioning was carried out participant-independently in order to best possibly assign the material of two-thirds of individuals per developmental outcome, gender, and mother tongue (e.g., three of the five female individuals later diagnosed with RTT stemming from English-speaking families) to the training/development partition (22 participants), and the remaining third to the test partition (10 participants). As there was only material of one individual with a later FXS diagnosis available for this experiment, FXS-associated vocalisations could not be used for both training/development and testing. In order not to evaluate on a COI that is unknown to the detector, the FXS-associated material was assigned to the training/development partition. Dataset partitioning for this experiment is shown in Table 8.2.

As already specified in Section 7.1 and Subsection 7.3.1, this voice activity detection experiment was based on frames of 25 ms at a step size of 10 ms alongside each included audio(-video) recording. Exactly the first half of audio frames of the training/development partition was designated to be the training partition, the second half to be the development partition.

Baseline infant voice activity detection was performed using the 'cVadV1' component of openSMILE (cf. Section 7.1), which provided a frame-wise decision on the presence or absence of voice. This decision was made on the basis of fuzzy scores derived from the current deviations from the mean long-term trajectories of three pre-selected acoustic LLDs (voice detected if all three scores unequal zero), namely energy, line spectral frequencies, and Mel spectra [222] (cf. Section 7.1).

Table 8.2: Assignment of participants to training/development or test partition for the experiment on infant voice activity detection. This table includes information of Table 2 published by Pokorny and colleagues [206] (DOI: 10.21437/Interspeech.2016-1341). ASD = autism spectrum disorder; AT = Austria; DE = Germany; Dur = available audio-video duration in format hh:mm:ss (two-digit hour number:two-digit minute number:two-digit second number; seconds rounded down to integer values; sums calculated on the basis of exact durations); Eng = English; FXS = fragile X syndrome; Gen = gender; Ger = German; ID = (unique participant) identification code; IT = Italy; Ita = Italian; L1 = mother tongue; Nat = nationality; RTT = Rett syndrome; TD = typical development/typically developing; UK = United Kingdom; # = number of; ♀ = female; ♂ = male

Partition	ID	Gen	Nat/L1	# Vocalisations Dur [hh:mm:ss]
Training/development	ASD01	♂	IT/Ita	48 00:11:56
	ASD02	♂	IT/Ita	133 00:32:50
	ASD03	♂	IT/Ita	41 00:05:08
	ASD04	♂	IT/Ita	9 00:04:36
	ASD05	♂	IT/Ita	3 00:07:39
	ASD06	♂	IT/Ita	21 00:05:10
	ASD07	♂	IT/Ita	3 00:00:48
	ASD08	♂	IT/Ita	131 00:15:50
	ASD09	♂	IT/Ita	35 00:06:13
	FXS01	♂	AT/Ger	87 00:17:05
	RTT01	♀	AT/Ger	150 00:13:46
	RTT02	♀	AT/Ger	304 02:04:16
	RTT03	♀	DE/Ger	1382 04:53:51
	RTT07	♀	UK/Eng	8 00:05:16
	RTT08	♀	UK/Eng	235 00:54:57
	RTT09	♀	UK/Eng	19 00:35:30
	TD01	♂	AT/Ger	22 00:05:44
	TD04	♂	AT/Ger	9 00:04:55
	TD05	♂	AT/Ger	308 00:42:06
	TD06	♀	AT/Ger	6 00:05:32
	TD07	♀	AT/Ger	15 00:10:12
	TD08	♀	AT/Ger	681 02:44:01
			Σ	3 650 14:27:32
Test	ASD10	♂	IT/Ita	84 00:26:14
	ASD11	♂	IT/Ita	109 00:41:03
	ASD12	♂	IT/Ita	50 00:13:45
	ASD13	♂	IT/Ita	29 00:08:34
	RTT04	♀	DE/Ger	363 01:00:21
	RTT05	♀	UK/Eng	15 00:04:09
	RTT06	♀	UK/Eng	109 00:28:35
	TD02	♂	AT/Ger	39 00:19:27
	TD03	♂	AT/Ger	113 01:06:26
	TD09	♀	AT/Ger	342 01:55:40
			Σ	1 253 06:24:17

All three investigated learning algorithms built on the same set of 100 acoustic features as describes in Section 7.1. The features were partition-wisely standardised to have zero-mean and unit variance.

For the SVM approach, a Gaussian kernel was employed. In order to decrease training time, the detection model was trained on a representative subset of 125 000 samples per class (no infant voice vs. infant voice). The kernel complexity parameter $C \in \{2^{-10}, \ldots, 2^{10}\}$ and the kernel width $\gamma \in \{2^{-10}, \ldots, 2^{10}\}$ were cross-validated with regard to optimise the AED-ACC (cf. Subsection 7.3.1) on the development partition.

Considering the same optimisation criterion as already applied for SVM parameter tuning (AED-ACC optimisation on the development partition), in the RF approach, the number of trees $T \in \{50, 100, 200, 300\}$, the maximal tree depth $D \in \{5, 10, 15, 20\}$, and the minimum number of samples per leaf $M \in \{1, 10, 100\}$ were cross-validated.

For the HMM implementation, two states were used. One state modelled the absence of infant voice, the other state modelled the presence of infant voice in a frame. The HMM was designed to have a uniform prior distribution, but a transition probability for both states of only 0.5% leading to a strong a priori state blocking (of 99.5%) that was also observed within the training data. As already mentioned in Section 7.3, the Viterbi algorithm was used to perform state prediction and Gaussian mixture models (GMMs) were trained as observational models by applying the expectation maximisation (EM) algorithm [152, 139, 153, 29] and by cross-validating the number of components. Training was performed based on a stratified subset of all (training) frames including infant voice and the same number of frames without infant voice. As the full set of 100 features yielded a poor detection performance (in terms of a low AED-ACC on the development set), greedy feature forward selection was carried out. In doing so, a model was trained on each individual feature. Then, the feature involved in the best performing model (again in terms of the best AED-ACC on the development set) was selected. This procedure was iterated while keeping the already selected features fixed. Thereby, eight features could be selected for the final model until the detection performance started to gradually degrade. As an alternative to the GMMs trained by means of EM, discriminative GMMs were applied as observational models by means of large-margin training [223] based on the reduced set of the eight selected features. In the following, the HMM using GMMs trained by means of EM is referred to as 'HMM_{gen}' (generative HMM). The HMM using discriminative GMMs is referred to as 'HMM_{dis}' (discriminative HMM).

The performance of the employed approaches in terms of precision, recall, AED-ACC, and the AEER (cf. Subsection 7.3.1) for both frame-based and vocalisation-based infant voice activity detection is revealed in Table 8.3.

Table 8.3: Frame-based and vocalisation-based infant voice activity detection performance of a standard rule-based voice activity detector as the baseline approach, a support vector machine (SVM) approach, a random forest (RF) approach, and both a generative and a discriminative hidden Markov model (HMM_{gen}; HMM_{dis}) approach in terms of precision, recall, acoustic event detection accuracy (AED-ACC), and the acoustic event error rate (AEER). This table is based on Table 3 published by Pokorny and colleagues [206] (DOI: 10.21437/Interspeech.2016-1341). Precision, recall, and AED-ACC values are rounded to two decimal places. The AEER is rounded to integers.

Measure		Approach				
		Baseline	SVM	RF	HMM_{gen}	HMM_{dis}
Precision	frame-based	0.15	0.19	0.11	0.18	0.20
	vocalisation-based	0.11	0.11	0.07	0.17	0.26
Recall	frame-based	0.59	0.56	0.79	0.66	0.60
	vocalisation-based	0.89	0.88	0.96	0.89	0.74
AED-ACC	frame-based	0.24	0.29	0.19	0.28	0.31
	vocalisation-based	0.19	0.20	0.13	0.29	0.38
AEER	frame-based	90 295	96 846	45 994	74 172	87 597
	vocalisation-based	151	159	67	143	319

The results of this thesis' experiment on automatic infant voice activity detection/segmentation in home video material yielded a very unclear picture of the performance of the evaluated approaches with respect to the applied performance measures. Obviously, the different employed approaches had different strengths and weaknesses, i.e., there was not one specific approach that clearly outperformed the other approaches. Whereas the RF approach achieved the highest recall as well as the lowest AEER in both the frame-based and the vocalisation-based evaluation scenario, respectively, the HMM_{dis} approach performed best in terms of reaching the highest precision as well as the highest AED-ACC again in both the frame-based and the vocalisation-based evaluation scenario, respectively. The RF approach caused a low number of FNs at the expense of a high number of FPs, the HMM_{dis} approach vice verca. Generally, all approaches tended to cause a relatively high number of FPs leading to precision values ranging from only 0.11–0.20 in the frame-based evaluation scenario and from 0.07–0.26 in the vocalisation-based evaluation scenario. In contrast, the highest recall value was even 0.96 achieved by the RF approach in the vocalisation-based evaluation scenario. However, the highest AED-ACC combining both precision and recall within one measure was only 0.38, achieved by the HMM_{dis} approach in the vocalisation-based evaluation scenario. A good trade-off between all four evaluation measures was obtained by the HMM_{gen} approach. Finally, it needs to be mentioned that the rule-based

detector representing the baseline for the experiment could absolutely keep up with the applied machine learning approaches, especially with the SVM approach.

Currently, none of the evaluated detection approaches seems to be already usable as a reliable input stage for a fully automatic infant vocalisation-based classification system, but for supporting/semi-automatising the time-consuming vocalisation segmentation process for infant vocalisation-based experiments. However, the rather low detection performance of any of the applied approaches might be attributed to the nature of the used audio data most of the time (88.2%) does not containing an infant (target) vocalisation, but comprising a number of background noise events including sequences of human voice, such as voice from television or radio switched on during the recording, parental voice, or even voice of children or infants other than the participating individual such as an older sibling, a cousin, and/or an other playmate of the participant also present in the recorded setting. This potential explanation for the poor overall detection performance is supported by the fact that the applied rule-based detector, which represents a tool for detecting human voice in general, thus, not explicitly for detecting infant voice, achieved results comparable to the results of the machine learning approaches. Another problematical aspect with respect to both the frame-based and the vocalisation-based evaluation scenario might be the applied vocalisation segmentation strategy relying on each vocalisation coming along with a vocal breathing group and, therefore, causing vocalisation pauses/sequences of silence occasionally being part of vocalisations (cf. Section 6.2). Finally, the AEER turned out to be a measure that can be hardly interpreted and compared between the different approaches, as its value range is not normalised to an upper limit of 1.

This experiment was carried out without a prior audio normalisation, as (i) a global normalisation, i.e., for example the normalisation of whole audio-video clips, would have led to mainly setting the gain according to contained, distinct noise events that usually had higher signal energy than infant (target) vocalisations, or (ii) a frame-wise normalisation would have led to setting background noise to the maximum gain for a large number of frames of 'silence'/frames without recorded acoustic events. However, dependent on the scenario of application of the proposed vocalisation-based infant developmental disorder recogniser, audio normalisation as a data pre-processing step might make sense, for example in a clinical setting with controllable recording conditions (cf. Section 11.2), and should be considered. Furthermore, from a technical point of view, the length of the applied moving median filter for raw detection data stream pre-processing (cf. Subsection 7.3.1) represents an important design parameter for further experiments.

9

Developmental disorder recognition

A number of experiments were conducted in order to evaluate the basic feasibility of an automatic vocalisation-based identification of individuals later diagnosed with a COI vs. TD individuals. These experiments are dealt with in the following sections. They each refer to the second module of the proposed developmental disorder recognition tool (cf. Figure 4.1 in Section 4.2) and, thereby, represent this thesis' core experiments. Table 9.1 provides the corresponding experimental overview or rather a preview of the upcoming sections.

For a better comparability of COI-related vocalisation atypicalities/specificities across the different experiments, acoustic feature analysis in the framework of this thesis (cf. Subsection 7.2.1) was decided to be carried out on the basis of a homogeneous vocalisation and feature (pre-)processing procedure, namely on the basis of the unnormalised/unstandardised eGeMAPS features (cf. Section 7.1) extracted from unnormalised vocalisation data only. This setup was chosen as both not applying audio normalisation to the vocalisation data in a pre-processing step and using the eGeMAPS (not necessarily concurrently) more often led to a better classification performance in terms of the applied performance measures (cf. Subsection 7.3.1), compared to applying audio normalisation to the vocalisation data and using the ComParE set, respectively, and (iii) the role/benefit of specific feature normalisation/standardisation strategies was heterogeneous throughout the conducted experiments (cf. Sections 9.1–9.4 and Chapter 10). Moreover, the eGeMAPS's reduced complexity in contrast to the ComParE set and its systematic composition based on relevance criteria also regarding clinical speech applications (cf. Section 7.1), makes most eGeMAPS features more transparent and descriptive regarding their relation to underlying (neuro-)physiological processes of speech production. Consequently, eGeMAPS features were expected to reveal a meaningful information basis for a clinically relevant understanding of vocalisation differences between individuals with different COI and TD individuals.

Table 9.1: Overview of experiments carried out on developmental disorder recognition paradigms with regard to the addressed research questions (RQs; cf. Section 4.2) as well as to the used data subset(s), acoustic feature set(s), and classification approach(es), respectively. AD = atypical developmental/atypically developing; AQ = additional (research) question; ASD = autism spectrum disorder; BLSTMNN = bidirectional long short-term memory neural network; ComParE = Computational Paralinguistics and Emotion (set); eGeMAPS = extended Geneva Minimalistic Acoustic Parameter Set; FXS = fragile X syndrome; LLD = low-level descriptor; MQ = main (research) question; RTT = Rett syndrome; SVM = support vector machine; TD = typical development/typically developing; # = number of

	Experimentation on developmental disorder recognition			
Sections	9.1	9.2	9.3	9.4
References	[224]	[225]	[225], [226]	[225]
Paradigms	ASD vs. TD	FXS vs. TD	RTT vs. TD	AD vs. TD, FXS vs. RTT vs. TD
Addressed RQs	MQ1 & MQ5 & AQ1	MQ2 & MQ6 & AQ1	MQ3 & MQ7 & AQ1	MQ4 & MQ8 & AQ1
Subset sizes [# Participants]	20 (10 vs. 10)	6 (3 vs. 3)	6 (3 vs. 3), 8 (4 vs. 4)	12 (6 vs. 6), 12 (3 vs. 3 vs. 6)
Subset sizes [# Vocalisations]	684 (259 vs. 425)	1 164 (942 vs. 222)	3 502 (2 011 vs. 1 491), 4 678 (2 199 vs. 2 479)	4 666 (2 953 vs. 1 713), 4 666 (942 vs. 2 011 vs. 1 713)
Feature sets	eGeMAPS, ComParE LLDs	ComParE & eGeMAPS	ComParE & eGeMAPS, ComParE	ComParE & eGeMAPS
Classifiers	SVM, BLSTMNN	SVM	SVM	SVM

9.1 Recognition of autism spectrum disorder

In this thesis' first feasibility experiment on an automatic, vocalisation-based COI recognition, the two-class paradigm ASD vs. TD was investigated. Thereby, the experiment focussed on the first of the three COI treated in this work and addressed MQ1, MQ5, and AQ1 of this thesis (cf. Section 4.2). As already mentioned in Chapter 6, the audio(-video) data constituting the basis for this experiment were provided by cooperation partners from Sweden, namely by (i) Professor Dr. Sven Bölte, director of the Center of Neurodevelopmental Disorders and head of the Neuropsychiatry Division at the Department of Women's and Children's Health of the Karolinska Institutet, Stockholm, by (ii) Associate Professor Dr. Terje Falck-Ytter, also from the Department of Women's and Children's Health of the Karolinska Institutet, and from the Department of Psychology at the Uppsala University, as well as by (iii) Dr. Pär Nyström, also from the Department of Psychology at the Uppsala University. Parts of the analysis in the framework of this experiment were carried out by Raymond Brueckner, a colleague affiliated with both the Machine Intelligence & Signal Processing group of the Technical University of Munich, Germany, and Nuance Communications, Ulm, Germany. Experiment-related procedures and findings were published in 2017 by Pokorny and colleagues [224].

The dataset of this experiment comprised a total of 240 minutes of audio(-video) material recorded in semi-standardised parent-child interaction settings under laboratory conditions (cf. Chapter 6). It contained recordings of 10 Swedish individuals in the 11^{th} month of life later diagnosed with ASD (5 females, 5 males) and recordings of 10 nationality-/family language-, age-, and gender-matched TD controls. Each recording had an exact duration of 12 minutes. Within the provided material a total of 684 vocalisations could be segmented. An overview of participants and participant-specific numbers of included vocalisations is given in Table 9.2.

The mean duration of the included vocalisations was 2.01 s at a standard deviation of 2.31 s. The median vocalisation duration was 1.39 s. Vocalisation durations ranged from 0.42 s to 36.27 s. The longest vocalisation was an extended sequence of crying patterns of participant TD13. The median duration of the vocalisations of individuals later diagnosed with ASD was 1.49 s, the median duration of vocalisations of TD participants was 1.3 s. At a value of $p = 0.019$ and a significance level of $\alpha = 0.05$, a significant difference in the vocalisation duration between the participants with ASD and the TD participants was found using the Mann-Whitney U-test (vocalisation durations were not normally distributed).

Due to the standardised recording time per participant across this dataset on the one hand, and due to the more standardised recording setting (infants in physically unrestricted position motivated by their parents for interaction/communication in a closed room without external influencing factors) as compared to the home video

setting underlying the data used in the other experiments on developmental disorder recognition (cf. Sections 9.2–9.4) on the other hand, the volubility/vocalisation rate (VR) was calculated in this experiment as an additional, potentially meaningful evaluation measure predictive for an infant's socio-communicative outcome and, thus, for a later diagnosis of a COI (e.g., [95, 110, 112, 118, 227]; cf. Section 3.2). However, no significant difference in volubility between participants with ASD and TD participants could be found using a paired t-test (vocalisation rates were normally distributed; $\alpha = 0.05$; $p = 0.0922$).

Table 9.2: Number of included vocalisations and vocalisation rate (VR) per participant, per group, and in total for the experiment on automatic autism recognition. This table is based on Table 1 published by Pokorny and colleagues [224] (DOI: 10.21437/Interspeech.2017-1007). The VR is rounded to two decimal places. ASD = autism spectrum disorder; Gen = gender; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; SE = Sweden; Swe = Swedish ; TD = typical development/typically developing; # = number of; ♀ = female; ♂ = male

ID	Gen	Nat/L1	# Vocalisations	VR [# vocalisations/minute]
ASD14	♀	SE/Swe	10	0.83
ASD15	♂	SE/Swe	38	3.17
ASD16	♀	SE/Swe	42	3.50
ASD17	♀	SE/Swe	26	2.17
ASD18	♂	SE/Swe	19	1.58
ASD19	♀	SE/Swe	31	2.58
ASD20	♂	SE/Swe	28	2.33
ASD21	♂	SE/Swe	9	0.75
ASD22	♂	SE/Swe	17	1.42
ASD23	♀	SE/Swe	39	3.25
Σ_{ASD}		$Mean_{ASD}$	259	2.16
TD12	♂	SE/Swe	18	1.50
TD13	♂	SE/Swe	45	3.75
TD14	♂	SE/Swe	35	2.92
TD15	♀	SE/Swe	15	1.25
TD16	♀	SE/Swe	98	8.17
TD17	♀	SE/Swe	29	2.42
TD18	♂	SE/Swe	59	4.92
TD19	♀	SE/Swe	52	4.33
TD20	♂	SE/Swe	37	3.08
TD21	♀	SE/Swe	37	3.08
Σ_{TD}		$Mean_{TD}$	425	3.54
$\Sigma_{ASD \cup TD}$		$Mean_{ASD \cup TD}$	684	2.85

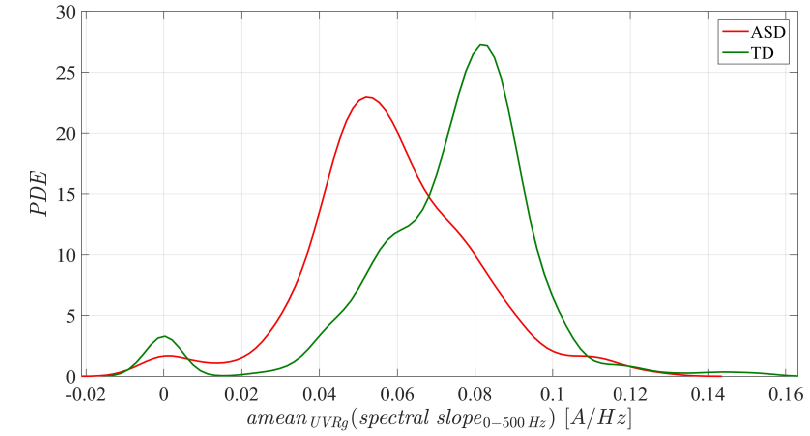
The analysis of the 88 eGeMAPS features revealed the value distributions of 54 acoustic feature to significantly differ between the vocalisations of individuals later diagnosed with ASD and vocalisations of TD individuals. The top ten eGeMAPS features according to the effect size estimate r (cf. Subsection 7.2.1) are given in Table 9.3. These top ten features are mainly based on spectral LLDs, namely the spectral slope, MFCCs, the Hammarberg index, and the alpha ratio. Class-specific distribution visualisations of the acoustic feature with the highest differentiation effect as well as of the three acoustic features with the highest differentiation effects are shown in Figure 9.1.

Table 9.3: Top ten acoustic features to differentiate between vocalisations of individuals later diagnosed with autism spectrum disorder and typically developing individuals, ranked according to the magnitude of the effect size estimate r, at given p-values of the underlying Mann-Whitney U-tests. This table is based on Table 2 published by Pokorny and colleagues [224] (DOI: 10.21437/Interspeech.2017-1007). r is rounded to three decimal places. Multiplicators of p-values are rounded to one decimal place. amean = arithmetic mean; coeffVar = coefficient of variation; F0 = fundamental frequency; $\text{Idx}_{\text{eGeMAPS}}$ = extended Geneva Minimalistic Acoustic Parameter Set-internal feature index [173]; MFCC = Mel-frequency cepstral coefficient; pctlR = percentile range; UVRg = unvoiced regions; VRg = voiced regions; neither UVRg nor VRg = all regions

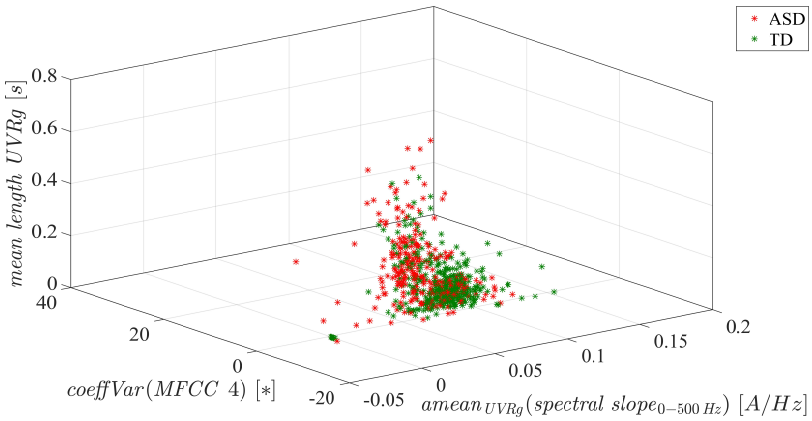
Rank	r	p	Feature	$\text{Idx}_{\text{eGeMAPS}}$
1	0.397	$3.3 \cdot 10^{-25}$	$amean_{UVRg}(spectral\ slope_{0–500\,Hz})$	79
2	0.340	$5.7 \cdot 10^{-19}$	$coeffVar(MFCC\ 4)$	30
3	-0.337	$1.2 \cdot 10^{-18}$	$mean\ length\ UVRg$	86
4	0.337	$1.2 \cdot 10^{-18}$	$amean_{UVRg}(Hammarberg\ index)$	78
5	-0.336	$1.7 \cdot 10^{-18}$	$amean_{UVRg}(spectral\ slope_{500–1500\,Hz})$	80
6	0.293	$1.7 \cdot 10^{-14}$	$amean_{VRg}(F0)$	02
7	-0.291	$2.7 \cdot 10^{-14}$	$amean_{UVRg}(\alpha\ ratio)$	77
8	0.255	$2.5 \cdot 10^{-11}$	$VRg\ per\ second$	83
9	0.254	$3.1 \cdot 10^{-11}$	$pctlR_{20–80,VRg}(F0)$	06
10	-0.250	$6.0 \cdot 10^{-11}$	$coeffVar(MFCC\ 1)$	24

As the material used for this experiment was recorded under laboratory conditions with one and the same recording device, it was decided not to apply audio normalisation to the vocalisation segments as a pre-processing step for experimentation on the automatic classification of vocalisations. Moreover, also feature normalisation/standardisation was not applied in this experiment.

In order to investigate this experiment's binary vocalisation classification paradigm ASD vs. TD, an infant independent three-fold cross-validation scheme was applied. Therefore, participants with ASD and TD participants were pair-



(a)



(b)

Figure 9.1: Comparison between vocalisations of individuals later diagnosed with ASD and vocalisations of TD individuals by means of (a) probability density estimates of the acoustic feature with the highest group differentiation effect (cf. Table 9.3), and by means of (b) vocalisation distributions within the three-dimensional space of the acoustic features with the three highest group differentiation effects (cf. Table 9.3). This figure includes information of Figure 1 published by Pokorny and colleagues [224] (DOI: 10.21437/Interspeech.2017-1007). A = amplitude; amean = arithmetic mean; ASD = autism spectrum disorder; coeffVar = coefficient of variation; MFCC = Mel-frequency cepstral coefficient; PDE = probability density estimate; TD = typical development/typically developing; UVRg = unvoiced regions; * = real measurement unit not existent as feature values refer to the amplitude of the digital audio signal

wisely split into three gender-matched and best possibly gender-balanced partitions with the requirement to obtain both a ratio between the number of vocalisations per class and an overall number of vocalisations as constant as possible across the three partitions. As specified in Table 9.4, the first partition contained one pair of male participants and two pairs of female participants, the second partition vice versa. Finally, the third partition contained two participant pairs per gender.

Table 9.4: Class-matched pairing of participants and assignment of participant pairs to partitions for the experiment on autism recognition. This table includes information of Table 1 published by Pokorny and colleagues [224] (DOI: 10.21437/Interspeech.2017-1007). ASD = autism spectrum disorder; Gen = gender; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; SE = Sweden; Swe = Swedish; TD = typical development/typically developing; Voc = vocalisations; # = number of; ♀ = female; ♂ = male

Partition	Participant pair								Σ
	$\text{Participant}_{\text{ASD}}$				$\text{Participant}_{\text{TD}}$				
	ID	Gen	Nat/L1	# Voc	ID	Gen	Nat/L1	# Voc	
1	ASD15	♂	SE/Swe	38	TD18	♂	SE/Swe	59	97
	ASD17	♀	SE/Swe	26	TD15	♀	SE/Swe	15	41
	ASD19	♀	SE/Swe	31	TD19	♀	SE/Swe	52	83
Σ_1				95				126	221
2	ASD16	♀	SE/Swe	42	TD16	♀	SE/Swe	98	140
	ASD20	♂	SE/Swe	28	TD20	♂	SE/Swe	37	65
	ASD21	♂	SE/Swe	9	TD14	♂	SE/Swe	35	44
Σ_2				79				170	249
3	ASD14	♀	SE/Swe	10	TD17	♀	SE/Swe	29	39
	ASD18	♂	SE/Swe	19	TD13	♂	SE/Swe	45	64
	ASD22	♂	SE/Swe	17	TD12	♂	SE/Swe	18	35
	ASD23	♀	SE/Swe	39	TD21	♀	SE/Swe	37	76
Σ_3				85				129	214
$\Sigma_{1\cup2\cup3}$				259				425	684

For each of the three validation runs, one partition was determined to be the training partition, another one the development partition, and the remaining one the test partition. Subsequent to parameter tuning/optimisation based on the development partition, the training and the development partitions were merged to an ultimate partition for training the classifier, which was then validated on the test partition. Throughout the three-fold cross-validation procedure, each of the three partitions was used as training partition, development partition, and test partition exactly one time.

For testing the automatic recognition of vocalisations of individuals later diagnosed with ASD vs. vocalisations of TD individuals, two basically different classification approaches were employed, namely linear kernel SVMs on the one hand, and BLSTMNNs as a topical alternative on the other hand. The SVM approach was based on the 88 eGeMAPS features calculated per vocalisation (over its individual vocalisation length), whereas for the BLSTMNN approach – an approach for modelling time series – LLD trajectories extracted on the basis of 10 ms time steps were used for each vocalisation as already mentioned in Section 7.3. As the eGeMAPS only comprises 25 LLDs [173], the BLSTMNN approach was decided to be based on the 130 LLDs/ΔLLDs of the ComParE set [170].

SVM optimisation comprised the determination of the kernel complexity parameter $C \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ to achieve the best UAR on the development partition. As already mentioned in Section 7.3, BLSTMNN training was performed using the first-order gradient-based Adam optimisation algorithm. Cross-entropy loss was found to work best when the posterior probabilities were averaged across the full vocalisations. Following a patience-based procedure, training was stopped when the UAR on the development partition did not improve for more than five epochs. Subsequently, the best model was selected. In a grid search, a single-layer network with eight cells was identified to be optimal for this experiment.

Both the SVM approach and the BLSTMNN approach were evaluated vocalisation-wisely and infant-wisely (cf. Subsection 7.3.1). Infant-wise evaluation was built upon the single classification decisions on the vocalisations of each infant. In other words, a global decision on each infant was made on the basis of the percentage of his or her vocalisations assigned to the ASD class against a defined decision threshold. In order to classify an infant as an infant with a later diagnosis of ASD, this threshold had to be exceeded. In this experiment, both (i) a threshold of 0.5 (majority voting), i.e., an individual was for example classified as an individual later diagnosed with ASD in case that more than 50% of his or her vocalisations were assigned to the class ASD, and (ii) decision threshold optimisation throughout the 3-fold cross-validation procedure were tested. As to the latter strategy, in each validation run the optimal decision threshold was determined on the basis of the vocalisation-wise classification decisions obtained for the merged training and development partitions and, subsequently, applied to the test partition. Optimisation was based on the criterion to maximise the distance between the mean number of vocalisations of individuals with ASD and the mean number of vocalisations of TD individuals assigned to the class ASD, respectively. For the SVM approach, the optimised thresholds for partitions 1–3 were 0.33, 0.42, and 0.31, for the BLSTMNN approach, 0.46, 0.44, and 0.39. Thus, on the basis of the data used for this experiment, the excess of an infant's percentage of vocalisations assigned to the class ASD somewhere between 31% and 46% seems to be optimal for recognising that an infant has ASD, while, at the same time, avoiding that a potential TD individual gets misclassified.

As can be seen in Table 9.5, in this experiment, the SVM approach and the BLSTMNN approach achieved similar recognition results. The highest UAR for vocalisation-wise classification was 0.645. It was achieved using SVMs. For infant-wise classification both approaches performed at an UAR of 0.75 (related to # individuals) when using the optimised decision thresholds. Eight of ten individuals later diagnosed with ASD and seven of ten TD individuals were correctly classified.

Finally it needs to be mentioned that the BLSTMNN approach might not have developed its full potential in this experiment due to the number of vocalisations/instances available for model training, which was – from a machine learning point of view – rather low.

Table 9.5: Cross-validation results of the experiment on autism recognition in form of class-specific numbers of vocalisations (in-)correctly classified as class ASD or TD (confusion matrices), and in form of mean and standard deviation (SD) of the unweighted average recall (UAR) for vocalisation-wise and infant-wise evaluation, respectively, for both the SVM and the BLSTMNN approach. This table is based on Tables 3 and 4 published by Pokorny and colleagues [224] (DOI: 10.21437/Interspeech.2017-1007). UAR values are rounded to three decimal places. ASD = autism spectrum disorder; BLSTMNN = bidirectional long short-term memory neural network; opt = optimised; SVM = support vector machine; TD = typical development/typically developing; Th = (decision) threshold

			Approach			
			SVM		BLSTMNN	
	classified as →		ASD	TD	ASD	TD
	ASD		131	128	155	104
	TD		83	342	141	284
Evaluation mode	Measure					
Vocalisation-wise	UAR	mean	0.645		0.629	
		SD	0.033		0.020	
			$Th_{0.5}$	Th_{opt}	$Th_{0.5}$	Th_{opt}
Infant-wise	UAR	mean	0.694	0.750	0.708	0.750
		SD	0.048	0.083	0.110	0.083

9.2 Recognition of fragile X syndrome

The second experiment on an automatic COI recognition conducted in the framework of this thesis dealt with the binary classification task FXS vs. TD. Thus, this experiment was intended to give answers to MQ2, MQ6, and AQ1 (cf. Section 4.2).

Parts of the analysis of this experiment were carried out by Maximilian Schmitt from the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. Procedures and findings related to this experiment are currently prepared for publication by Pokorny and colleagues [225].

The dataset used for this experiment consisted of retrospective home video recordings taken from GUARDIAN's Dataset 4 (cf. Section 6.1). As specified in Table 9.6, it comprised the material of three male participants later diagnosed with FXS and three male TD controls. All participants stem from German-speaking families. The recordings were made over the participants' respective second half year of life. Within the available video duration of almost 4 hours, 1 164 vocalisations, of which 942 were produced by participants with FXS, could be segmented.

Table 9.6: Number of included vocalisations on the basis of the available audio-video duration in format hh:mm:ss (two-digit hour number:two-digit minute number:two-digit second number; seconds rounded down to integer values; sums calculated on the basis of exact durations) as a function of age per participant for the experiment on FXS recognition. AT = Austria; DE = Germany; FXS = fragile X syndrome; Gen = gender; Ger = German; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; TD = typical development/typically developing; ♂ = male; - = no audio-video material available

ID	Gen	Nat/L1	Age [month of life] 7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	Σ
FXS01	♂	AT/Ger	85 00:16:01	62 00:05:30	88 00:17:05	45 00:05:45	127 00:15:02	257 00:30:07	664 01:29:32
FXS02	♂	DE/Ger	-	4 00:00:29	-	25 00:02:55	-	1 00:01:35	30 00:05:00
FXS03	♂	DE/Ger	-	7 00:01:24	117 00:10:49	20 00:03:10	83 00:15:16	21 00:02:45	248 00:33:25
		Σ_{FXS}	85 00:16:01	73 00:07:24	205 00:27:55	90 00:11:51	210 00:30:18	279 00:34:28	942 02:07:59
TD01	♂	AT/Ger	36 00:04:39	22 00:05:44	-	-	-	12 00:03:16	70 00:13:40
TD02	♂	AT/Ger	9 00:06:36	-	-	4 00:03:36	21 00:06:14	5 00:02:59	39 00:19:27
TD03	♂	AT/Ger	32 00:19:07	20 00:09:39	11 00:11:55	-	4 00:03:39	46 00:22:03	113 01:06:26
		Σ_{TD}	77 00:30:24	42 00:15:24	11 00:11:55	4 00:03:36	25 00:09:54	63 00:28:19	222 01:39:34
		$\Sigma_{FXS \cup TD}$	162 00:46:25	115 00:22:49	216 00:39:50	94 00:15:27	235 00:40:12	342 01:02:47	1 164 03:47:33

The vocalisations included in this experiment had a mean duration of 2.33 s with a standard deviation of 1.95 s, and a median duration of 1.76 s. The shortest

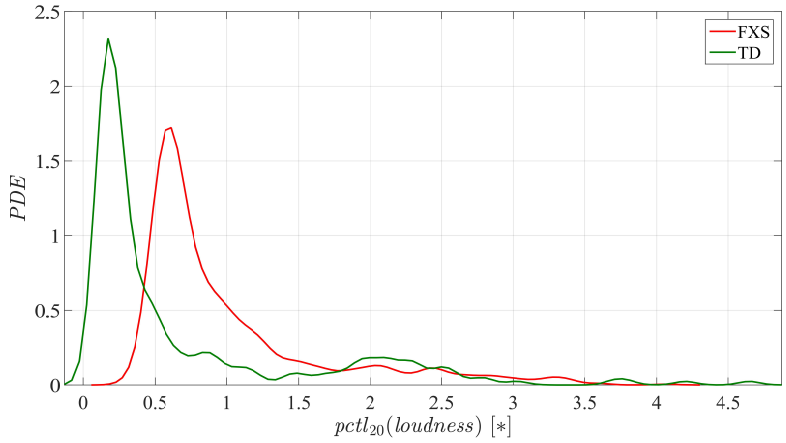
vocalisation had a duration of 232 ms. The longest vocalisation was an extended sequence of crying patterns produced by participant TD03 and had a duration of 21.32 s. Neither the durations of vocalisations produced by participants with FXS nor the durations of vocalisations produced by TD controls were normally distributed. The vocalisations of participants with FXS had a median duration of 1.9 s, the vocalisations of TD participants a duration of 1.21 s. At a value of $p = 7.1 \cdot 10^{-12}$, a significant difference in the duration of vocalisations between participants with FXS and TD participants was found using the Mann-Whitney U-test.

The class-specific distribution analysis of the 88 eGeMAPS features revealed the distributions of 59 features to significantly differ between the vocalisations of participants later diagnosed with FXS and the TD participants. Table 9.7 lists the top ten acoustic features according to the effect size estimate r. Seven of ten of these features are based on an energy/amplitude related LLD, namely on loudness. Figure 9.2 presents class-specific distribution visualisations of both the top and the top three acoustic features.

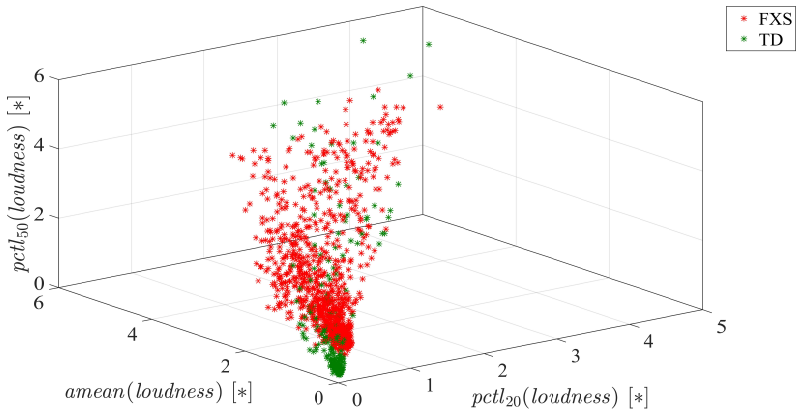
Table 9.7: Top ten acoustic features to differentiate between vocalisations of individuals later diagnosed with fragile X syndrome and typically developing individuals, ranked according to the magnitude of the effect size estimate r, at given p-values of the underlying Mann-Whitney U-tests. r is rounded to three decimal places. Multiplicators of p-values are rounded to one decimal place. amean = arithmetic mean; FS = falling slope; $\text{Idx}_{\text{eGeMAPS}}$ = extended Geneva Minimalistic Acoustic Parameter Set-internal feature index [173]; pctl = percentile; pctlR = percentile range; RS = rising slope; SD = standard deviation; UVRg = unvoiced regions; VRg = voiced regions; neither UVRg nor VRg = all regions

Rank	r	p	Feature	$\text{Idx}_{\text{eGeMAPS}}$
1	-0.348	$1.4 \cdot 10^{-32}$	$pctl_{20}(loudness)$	13
2	-0.335	$3.6 \cdot 10^{-30}$	$amean(loudness)$	11
3	-0.330	$1.8 \cdot 10^{-29}$	$pctl_{50}(loudness)$	14
4	-0.328	$4.5 \cdot 10^{-29}$	$pctl_{80}(loudness)$	15
5	-0.305	$2.4 \cdot 10^{-25}$	$pctlR_{20\text{–}80}(loudness)$	16
6	-0.301	$8.4 \cdot 10^{-25}$	*equivalent sound level*	88
7	-0.295	$7.5 \cdot 10^{-24}$	$SD_{RSs}(loudness)$	18
8	-0.293	$1.7 \cdot 10^{-23}$	$mean_{FSs}(loudness)$	19
9	-0.292	$2.1 \cdot 10^{-23}$	$amean_{UVRg}(spectral\ flux)$	81
10	-0.286	$1.8 \cdot 10^{-22}$	$amean_{VRg}(spectral\ flux)$	67

For the automatic recognition of vocalisations produced by participants with FXS vs. vocalisations produced by TD participants, amplitude normalisation of vocalisation segments as well as different strategies for normalising feature values to



(a)



(b)

Figure 9.2: Comparison between vocalisations of individuals later diagnosed with FXS and vocalisations of TD individuals by means of (a) probability density estimates of the acoustic feature with the highest group differentiation effect (cf. Table 9.7), and by means of (b) vocalisation distributions within the three-dimensional space of the acoustic features with the three highest group differentiation effects (cf. Table 9.7). amean = arithmetic mean; FXS = fragile X syndrome; pctl = percentile; PDE = probability density estimate; TD = typical development/typically developing; * = real measurement unit not existent as feature values refer to the amplitude of the digital audio signal

the interval $[0, 1]$, namely infant-dependent normalisation, age-dependent normalisation, and global normalisation (cf. Section 7.2), were tested. In this experiment, the features of both the ComparE set and the eGeMAPS were employed.

Due to the small number of only three participants later diagnosed with FXS and three gender-matched TD participants available for this classification experiment, leave-one-speaker-pair-out cross-validation was applied without using a training split as development partition, i.e., in each of three validation runs the classifier was trained on the basis of vocalisations/instances of two participants with FXS and two TD participants and, subsequently, optimised and tested on the instances of the remaining participant pair. As specified in Table 9.8, pairing of participants with FXS and TD participants was carried out according to the number of vocalisations available per participant, namely in a way to pair the participant with FXS with the highest number of vocalisations with the TD participant with the highest number of vocalisations, the participant with FXS with the second highest number of vocalisations with the TD participant with the second highest number of vocalisations, and finally, the participant with FXS with the lowest number of vocalisations with the TD participant with the lowest number of vocalisations. Throughout the cross-validation procedure, each participant pair was used as test partition exactly one time.

Table 9.8: Assignment of class-matched participant pairs to partitions for the experiment on FXS recognition. AT = Austria; DE = Germany; FXS = fragile X syndrome; Gen = gender; Ger = German; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; TD = typical development/typically developing; Voc = vocalisations; # = number of; ♂ = male

Partition	Participant pair								Σ
	$\text{Participant}_{\text{FXS}}$				$\text{Participant}_{\text{TD}}$				
	ID	Gen	Nat/L1	# Voc	ID	Gen	Nat/L1	# Voc	
1	FXS01	♂	AT/Ger	664	TD03	♂	AT/Ger	113	777
2	FXS02	♂	DE/Ger	30	TD02	♂	AT/Ger	39	69
3	FXS03	♂	DE/Ger	248	TD01	♂	AT/Ger	70	318
Σ				942				222	1 164

In order to obtain a best possibly balanced number of instanced per participant within the training partition for each validation run, upsampling (integer multiplication) of instances according to the number of instances of the participant with the highest number of instances within the respective training partition was tested. Partitioning did not play a role for both infant-dependent and age-dependent feature normalisation. In contrast, global normalisation was meant to be applied to all instances, thus, also to the instances of the test partition, according to the calculated range (respective minimum and maximum) of feature values within the training partition of each validation run. In this experiment, linear kernel SVMs were employed. The complexity parameter C was optimised $\in \{10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ to achieve the best UAR on the test partition.

Table 9.9: Cross-validation results of the experiment on fragile X syndrome recognition in form of the mean unweighted average recall (UAR) of the three validation runs and the globally calculated UAR based on the gathered predictions of the whole dataset for both vocalisation-wise and infant-wise evaluation and different system configurations regarding audio normalisation, used feature set, feature normalisation strategy (infant-dependent, age-dependent, or normalisation over all instances), and upsampling of the training set. UAR values are rounded to three decimal places. ComParE = Computational Paralinguistics and Emotion (set); eGeMAPS = extended Geneva Minimalistic Acoustic Parameter Set; ✓ = applied; ✗ = not applied

Evaluation mode	Audio normalisation	Feature set	Feature normalisation	Upsampling	Measure	
					UAR_{mean}	UAR_{global}
Vocalisation-wise	✗	ComParE	infant	✗	0.668	0.725
	✗	ComParE	infant	✓	0.672	0.758
	✗	ComParE	age	✗	0.664	0.671
	✗	ComParE	age	✓	0.674	0.700
	✗	ComParE	all	✗	0.496	0.469
	✗	ComParE	all	✓	0.482	0.488
	✗	eGeMAPS	infant	✗	0.506	0.507
	✗	eGeMAPS	infant	✓	0.449	0.561
	✗	eGeMAPS	age	✗	0.604	0.623
	✗	eGeMAPS	age	✓	0.606	0.623
	✗	eGeMAPS	all	✗	0.500	0.499
	✗	eGeMAPS	all	✓	0.412	0.430
	✓	ComParE	infant	✗	0.625	0.676
	✓	ComParE	infant	✓	0.625	0.697
	✓	ComParE	age	✗	0.645	0.652
	✓	ComParE	age	✓	0.662	0.709
	✓	ComParE	all	✗	0.492	0.480
	✓	ComParE	all	✓	0.441	0.492
	✓	eGeMAPS	infant	✗	0.539	0.559
	✓	eGeMAPS	infant	✓	0.569	0.684
	✓	eGeMAPS	age	✗	0.647	0.682
	✓	eGeMAPS	age	✓	0.666	0.698
	✓	eGeMAPS	all	✗	0.500	0.499
	✓	eGeMAPS	all	✓	0.463	0.538
Infant-wise	✗	ComParE	infant	✗	0.667	0.588
	✗	ComParE	infant	✓	0.667	0.730
	✗	ComParE	age	✗	0.667	0.588
	✗	ComParE	age	✓	0.833	0.745
	✗	ComParE	all	✗	0.500	0.500
	✗	ComParE	all	✓	0.667	0.781
	✗	eGeMAPS	infant	✗	0.500	0.500
	✗	eGeMAPS	infant	✓	0.500	0.765
	✗	eGeMAPS	age	✗	0.667	0.781
	✗	eGeMAPS	age	✓	0.667	0.781
	✗	eGeMAPS	all	✗	0.500	0.500
	✗	eGeMAPS	all	✓	0.333	0.368
	✓	ComParE	infant	✗	0.500	0.572
	✓	ComParE	infant	✓	0.667	0.826
	✓	ComParE	age	✗	0.667	0.588
	✓	ComParE	age	✓	0.667	0.588
	✓	ComParE	all	✗	0.500	0.500
	✓	ComParE	all	✓	0.500	0.695
	✓	eGeMAPS	infant	✗	0.667	0.560
	✓	eGeMAPS	infant	✓	0.500	0.739
	✓	eGeMAPS	age	✗	1.000	1.000
	✓	eGeMAPS	age	✓	1.000	1.000
	✓	eGeMAPS	all	✗	0.500	0.500
	✓	eGeMAPS	all	✓	0.500	0.572

Classification performance was evaluated vocalisation-wisely, and infant-wisely (related to # vocalisations) by applying majority voting over the participant-specific vocalisation-wise class predictions (cf. Subsection 7.3.1). On the one hand, the mean UAR of the three validation runs, on the other hand, the global UAR based on the predictions gathered over all three validation runs, were calculated. Table 9.9 reveals the achieved results for different configurations of the classification system.

The best UAR achieved for vocalisation-wise FXS vs. TD recognition was a globally calculated UAR of 75.8% for the classification system (i) processing unnormalised audio sequences, (ii) using the infant-dependently normalised ComParE features, and (iii) applying training partition upsampling. The mean UAR of the three validation runs for the same configuration was 67.2%, representing the second highest mean UAR achieved for vocalisation-wise classification. For infant-wise recognition, both a mean and a globally calculated UAR of even 100% were achieved for the classification system (i) processing normalised audio sequences, (ii) using the age-dependently normalised eGeMAPS features, and (iii) applying as well as not applying training partition upsampling. Thus, based on the strategy to assign each participant to the class, which was predicted for the majority of his or her vocalisations, all six participants were correctly assigned. In this classification experiment, neither system configuration options on audio normalisation and training partition upsampling, nor the decision for a specific feature set in combination with the choice for infant-dependent or age-dependent feature normalisation turned out to systematically influence the classification performance. However, feature normalisation over all instances led to the classifier performing around chance level (UAR = 50%) for most configurations.

9.3 Recognition of Rett syndrome

In the framework of this thesis, feasibility investigations on an automatic recognition of vocalisations produced by individuals later diagnosed with RTT vs. vocalisations produced by TD individuals were dealt with in two sub-experiments. Addressing MQ3, MQ7, and AQ1 (cf. Section 4.2), both of these sub-experiments are treated in this section. Moreover, in order to give answer to AQ2, a closer look at auditory atypicalities of RTT-associated early vocalisations and their manifestation in the acoustic signal domain is taken in Subsection 9.3.1. Analyses in the framework of the first sub-experiment on automatic RTT recognition were partly conducted by Maximilian Schmitt from the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. Procedures and findings related to this first sub-experiment on RTT recognition are currently prepared for publication by Pokorny and colleagues [225]. The second sub-experiment on RTT recognition was published in 2016 by Pokorny and colleagues [226].

Same as the previously dealt with experiment on automatic FXS recognition (cf. Section 9.2), also the experiments of this thesis focussing on early vocalisations in RTT were based on material taken from GUARDIAN's Dataset 4.

Table 9.10 overviews the dataset used for the first sub-experiment on RTT recognition. It comprised home video recordings from the respective second half year of life of 3 female participants from German-speaking families, later diagnosed with RTT, and 3 gender- and family language-matched TD controls, with a total audio-video duration of more than 14 hours. Altogether, 3 502 segmented vocalisations, of which 2 011 stem from the participants with RTT, were included in the experiment.

Table 9.10: Number of included vocalisations on the basis of the available audio-video duration in format hh:mm:ss (two-digit hour number:two-digit minute number:two-digit second number; seconds rounded down to integer values; sums calculated on the basis of exact durations) as a function of age per participant for the first sub-experiment on RTT recognition. AT = Austria; DE = Germany; Gen = gender; Ger = German; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; RTT = Rett syndrome; TD = typical development/typically developing; ♀ = female; - = no audio-video material available

ID	Gen	Nat/L1	Age [month of life] 7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	Σ
RTT01	♀	AT/Ger	-	-	-	149 00:13:46	-	-	149 00:13:46
RTT02	♀	AT/Ger	21 00:14:48	-	138 42:53	120 01:02:18	46 00:19:05	-	325 02:19:04
RTT03	♀	DE/Ger	142 00:41:35	370 00:57:45	276 00:44:25	325 01:23:21	180 00:57:42	244 00:50:36	1 537 05:35:27
		Σ_{RTT}	163 00:56:23	370 00:57:45	414 01:27:18	594 02:39:25	226 01:16:48	244 00:50:36	2 011 08:08:18
TD08	♀	AT/Ger	232 01:11:58	97 00:32:04	90 00:16:02	80 00:16:02	82 00:15:07	108 00:12:47	689 02:44:01
TD09	♀	AT/Ger	9 00:07:35	48 00:28:54	84 00:18:22	101 00:30:41	51 00:11:46	48 00:18:18	341 01:55:40
TD10	♀	AT/Ger	78 00:07:09	183 00:29:15	4 00:05:30	26 00:08:13	84 00:14:59	86 00:09:31	461 01:14:40
		Σ_{TD}	319 01:26:43	328 01:30:14	178 00:39:55	207 00:54:57	217 00:41:53	242 00:40:38	1 491 05:54:22
		$\Sigma_{RTT \cup TD}$	482 02:23:07	698 02:27:59	592 02:07:13	801 03:34:22	443 01:58:42	486 01:31:14	3 502 14:02:40

The vocalisations of this first sub-experiment on RTT recognition had a mean duration of 1.72 s with a standard deviation of 1.32 s. The median vocalisation duration was 1.35 s. Vocalisation durations ranged from 186 ms to 16.86 s with the longest vocalisation being an extended sequence of canonical babbling realisations

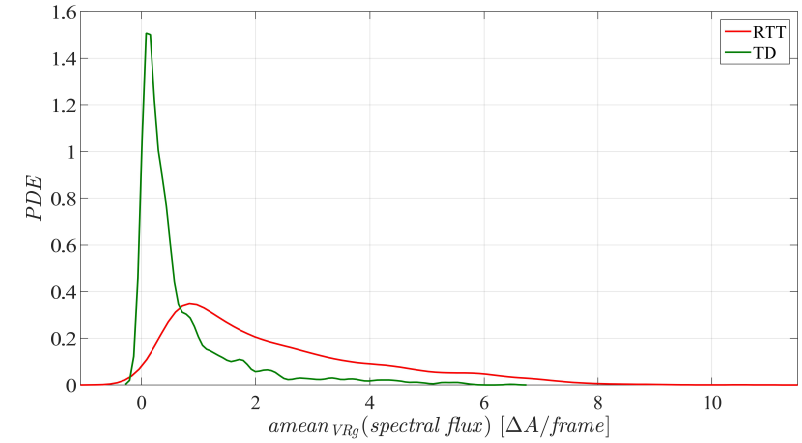
produced by participant TD08. Class-specific vocalisation durations were not normally distributed. The median durations of vocalisations produced by participants with RTT and TD participants were 1.35 s and 1.39 s, respectively. By applying the Mann-Whitney U-test, vocalisation durations were found to differ significantly between the class RTT and the class TD ($\alpha = 0.05$; $p = 0.0068$).

By means of feature analysis based on the eGeMAPS, 79 of 88 acoustic features were identified to significantly differ in value distributions between vocalisations of participants with RTT and vocalisations of TD participants. The top ten eGeMAPS features according to the effect size estimate r are listed in Table 9.11, distribution visualisations of the feature with the highest class differentiation effect, i.e., the arithmetic mean over voiced regions of the spectral flux, and of the three features with the highest class differentiation effects are presented in Figure 9.3. The top two features were related to the LLD spectral flux. Six of the top ten features were related to the energy/amplitude-based LLD loudness.

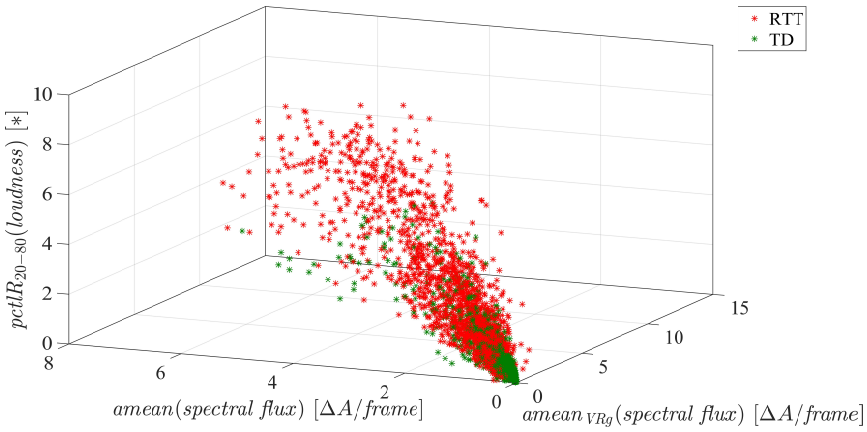
Table 9.11: Top ten acoustic features to differentiate between vocalisations of individuals later diagnosed with Rett syndrome and typically developing individuals, ranked according to the magnitude of the effect size estimate r, at given p-values of the underlying Mann-Whitney U-tests. r is rounded to three decimal places. Multiplicators of p-values are rounded to one decimal place. amean = arithmetic mean; FS = falling slope; $\text{Idx}_{\text{eGeMAPS}}$ = extended Geneva Minimalistic Acoustic Parameter Set-internal feature index [173]; pctl = percentile; pctlR = percentile range; RS = rising slope; SD = standard deviation; UVRg = unvoiced regions; VRg = voiced regions; neither UVRg nor VRg = all regions

Rank	r	p	Feature	$\text{Idx}_{\text{eGeMAPS}}$
1	-0.604	$5.2 \cdot 10^{-280}$	$amean_{VRg}(spectral\ flux)$	67
2	-0.596	$1.6 \cdot 10^{-272}$	$amean(spectral\ flux)$	21
3	-0.566	$2.2 \cdot 10^{-246}$	$pctlR_{20\text{–}80}(loudness)$	16
4	-0.562	$7.9 \cdot 10^{-243}$	$mean_{RSs}(loudness)$	17
5	-0.559	$1.4 \cdot 10^{-239}$	$pctl_{80}(loudness)$	15
6	-0.553	$1.2 \cdot 10^{-234}$	$mean_{FSs}(loudness)$	19
7	-0.544	$1.2 \cdot 10^{-227}$	$amean(loudness)$	11
8	-0.536	$1.5 \cdot 10^{-220}$	$equivalent\ sound\ level$	88
9	-0.513	$1.5 \cdot 10^{-202}$	$SD_{RSs}(loudness)$	18
10	0.437	$2.4 \cdot 10^{-147}$	$amean_{UVRg}(Hammarberg\ index)$	78

The procedure for testing the automatic vocalisation classification paradigm RTT vs. TD in this sub-experiment was chosen the same as in the experiment on FXS recognition (cf. Section 9.2), i.e., (i) the dataset was divided into three partitions of class-matched participant pairs grouped according to the class-wisely ranked



(a)



(b)

Figure 9.3: Comparison between vocalisations of individuals later diagnosed with RTT and vocalisations of TD individuals by means of (a) probability density estimates of the acoustic feature with the highest group differentiation effect (cf. Table 9.11), and by means of (b) vocalisation distributions within the three-dimensional space of the acoustic features with the three highest group differentiation effects (cf. Table 9.11). A = amplitude; amean = arithmetic mean; pctlR = percentile range; PDE = probability density estimate; RTT = Rett syndrome; TD = typical development/typically developing; VRg = voiced regions; * = real measurement unit not existent as feature values refer to the amplitude of the digital audio signal

numbers of available vocalisations per participant (see Table 9.12), (ii) leave-one-speaker-pair-out cross-validation was applied, (iii) audio normalisation was tested, (iv) both the ComParE features and the eGeMAPS features were employed, (v)

three different strategies to normalise features to the interval $[0, 1]$ were tried out, namely infant-dependent normalisation, age-dependent normalisation, and normalisation over all instances according to the determined feature value ranges within the training partition, (vi) training set upsampling was tested, and (vii) linear kernel SVMs were employed as classifier and optimised for the complexity parameter $C \in \{10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ to achieve the best UAR on the test partition.

Table 9.12: Assignment of class-matched participant pairs to partitions for the first sub-experiment on RTT recognition. AT = Austria; DE = Germany; Gen = gender; Ger = German; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; RTT = Rett syndrome; TD = typical development/typically developing; Voc = vocalisations; # = number of; ♀ = female

Partition	Participant pair								Σ
	$\text{Participant}_{\text{RTT}}$				$\text{Participant}_{\text{TD}}$				
	ID	Gen	Nat/L1	# Voc	ID	Gen	Nat/L1	# Voc	
1	RTT01	♀	AT/Ger	149	TD09	♀	AT/Ger	341	490
2	RTT02	♀	AT/Ger	325	TD10	♀	AT/Ger	461	786
3	RTT03	♀	DE/Ger	1 537	TD08	♀	AT/Ger	689	2 226
Σ				2 011				1 491	3 502

Same as in the experiment on FXS recognition, the performance of the classification system was evaluated vocalisation-wisely, and infant-wisely in terms of a class majority voting over the infant-specific vocalisation-wise predictions. As performance measures, again the mean UAR of the three validation runs and the globally determined UAR, which is based on the predictions of the whole dataset gathered throughout the cross-validation process, were calculated. The respective results for different classification system configurations are given in Table 9.13.

For vocalisation-wise classification, the best mean UAR of 86.1% as well as the best global UAR of 87.9% were achieved for (i) applying audio normalisation, (ii) using the eGeMAPS, (iii) normalising the feature values infant-dependently, and (iv) not upsampling the training partitions in each validation run. Keeping this configuration with the only difference not to apply training set upsampling, led to the respective second best mean and global UAR. The respective third and fourth best results were achieved when using the just described configurations that led to the top two mean and global UAR values, but leaving out audio normalisation as a pre-processing step. The best mean UAR achieved for infant-wise classification was 83.3%, the best global UAR for infant-wise classification was 84.5%. These results were obtained for diverse configurations, namely the same configurations that also led to the top three vocalisation-wise results and when setting the system not to apply audio normalisation, to use the eGeMAPS, and to normalise features across all test instances of a validation run.

Table 9.13: Cross-validation results of the first sub-experiment on Rett syndrome recognition in form of the mean unweighted average recall (UAR) of the three validation runs and the globally calculated UAR based on the gathered predictions of the whole dataset for both vocalisation-wise and infant-wise evaluation and different system configurations regarding audio normalisation, used feature set, feature normalisation strategy (infant-dependent, age-dependent, or normalisation over all instances), and upsampling of the training set. UAR values are rounded to three decimal places. ComParE = Computational Paralinguistics and Emotion (set); eGeMAPS = extended Geneva Minimalistic Acoustic Parameter Set; ✓ = applied; ✗ = not applied

Evaluation mode	Audio normalisation	Feature set	Feature normalisation	Upsampling	Measure	
					UAR_{mean}	UAR_{global}
Vocalisation-wise	✗	ComParE	infant	✗	0.328	0.356
	✗	ComParE	infant	✓	0.278	0.293
	✗	ComParE	age	✗	0.437	0.389
	✗	ComParE	age	✓	0.419	0.421
	✗	ComParE	all	✗	0.388	0.399
	✗	ComParE	all	✓	0.389	0.425
	✗	eGeMAPS	infant	✗	0.770	0.832
	✗	eGeMAPS	infant	✓	0.705	0.814
	✗	eGeMAPS	age	✗	0.500	0.539
	✗	eGeMAPS	age	✓	0.404	0.521
	✗	eGeMAPS	all	✗	0.700	0.674
	✗	eGeMAPS	all	✓	0.699	0.665
	✓	ComParE	infant	✗	0.329	0.372
	✓	ComParE	infant	✓	0.279	0.304
	✓	ComParE	age	✗	0.424	0.380
	✓	ComParE	age	✓	0.401	0.402
	✓	ComParE	all	✗	0.310	0.281
	✓	ComParE	all	✓	0.296	0.340
	✓	eGeMAPS	infant	✗	0.861	0.879
	✓	eGeMAPS	infant	✓	0.843	0.863
	✓	eGeMAPS	age	✗	0.494	0.347
	✓	eGeMAPS	age	✓	0.339	0.372
	✓	eGeMAPS	all	✗	0.565	0.535
	✓	eGeMAPS	all	✓	0.590	0.564
Infant-wise	✗	ComParE	infant	✗	0.333	0.386
	✗	ComParE	infant	✓	0.333	0.386
	✗	ComParE	age	✗	0.500	0.423
	✗	ComParE	age	✓	0.500	0.423
	✗	ComParE	all	✗	0.333	0.419
	✗	ComParE	all	✓	0.333	0.537
	✗	eGeMAPS	infant	✗	0.833	0.845
	✗	eGeMAPS	infant	✓	0.667	0.808
	✗	eGeMAPS	age	✗	0.500	0.728
	✗	eGeMAPS	age	✓	0.500	0.728
	✗	eGeMAPS	all	✗	0.833	0.845
	✗	eGeMAPS	all	✓	0.833	0.845
	✓	ComParE	infant	✗	0.333	0.386
	✓	ComParE	infant	✓	0.333	0.386
	✓	ComParE	age	✗	0.500	0.423
	✓	ComParE	age	✓	0.500	0.423
	✓	ComParE	all	✗	0.167	0.155
	✓	ComParE	all	✓	0.167	0.155
	✓	eGeMAPS	infant	✗	0.833	0.845
	✓	eGeMAPS	infant	✓	0.833	0.845
	✓	eGeMAPS	age	✗	0.500	0.349
	✓	eGeMAPS	age	✓	0.167	0.231
	✓	eGeMAPS	all	✗	0.667	0.688
	✓	eGeMAPS	all	✓	0.667	0.614

Generally, in this first sub-experiment on the automatic recognition of RTT vs. TD, the best classification performances were achieved when using the eGeMAPS features infant-dependently normalised, irrespective of the settings for audio normalisation and training set upsampling.

For this thesis' second sub-experiment on the automatic recognition of RTT, the dataset of the first sub-experiment on RTT recognition was extended for material from the respective second half year of life of another participant pair, namely of another female individual with RTT from a German-speaking family and another gender- and family language-matched TD control, taken from GUARDIAN's Dataset 4. Thus, the second sub-experiment on RTT recognition was based on home video recordings of four participants with RTT and four TD controls with a total audio-video duration of almost 17 hours. As specified in Table 9.14, within the available material, 4 678 target vocalisations could be segmented, of which 2 199 vocalisations were produced by participants with RTT. Compared to the dataset used for the first sub-experiment on RTT recognition, here the number of vocalisations per class was more balanced. However, it needs to be mentioned that the newly added participant RTT04 did not receive a later diagnosis of typical RTT, but of a special variant of RTT, namely the relatively milder PSV/Zappella variant of RTT that is, amongst other variations from the course of typical RTT, characterised by a comparably better recovery of speech-language capabilities after regression [88].

The vocalisations building the basis for this second sub-experiment on RTT recognition had a mean duration of 1.66 s at a standard deviation of 1.31 s, and a median duration of 1.32 s. Vocalisation durations ranged from a minimum of 300 ms to a maximum of 26.62 s. The vocalisation with the maximum duration was a sequence of isolated nasal sounds in different modulations by participant TD11. Class-specific vocalisation durations were not normally distributed. Therefore, the Mann-Whitney U-test was applied to check the RTT- and TD-associated vocalisation duration distributions for a potential difference. The median duration of vocalisations produced by participants with RTT was 1.35 s, the median duration of vocalisations produced by TD participants was 1.3 s. The class-specific duration distributions differed significantly at a value of $p = 5.7 \cdot 10^{-4}$.

In this second sub-experiment on RTT recognition, again linear kernel SVMs were employed. However, compared to the first RTT recognition experiment, this time the system configuration was more straightforward, namely (i) no audio normalisation was applied to the vocalisation segments, (ii) the comprehensive ComParE set was used only, and (iii) feature normalisation/standardisation was only tested infant-dependently. Feature values were optionally, on the one hand, normalised to the interval $[0, 1]$, on the other hand, standardised to have a mean of zero and unit variance.

Table 9.14: Number of included vocalisations on the basis of the available audio-video duration in format hh:mm:ss (two-digit hour number:two-digit minute number:two-digit second number; seconds rounded down to integer values; sums calculated on the basis of exact durations) as a function of age per participant for the second sub-experiment on RTT recognition. This table is based on Table 1 published by Pokorny and colleagues [226] (DOI: 10.21437/Interspeech.2016-520). AT = Austria; DE = Germany; Gen = gender; Ger = German; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; RTT = Rett syndrome; TD = typical development/typically developing; ♀ = female; - = no audio-video material available

ID	Gen	Nat/L1	Age [month of life]						Σ
			7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	
RTT01	♀	AT/Ger	-	-	-	150 00:13:46	-	-	150 00:13:46
RTT02	♀	AT/Ger	-	-	138 42:53	120 01:02:18	46 00:19:05	-	304 02:04:16
RTT03	♀	DE/Ger	-	363 00:57:45	273 00:44:25	323 01:23:21	180 00:57:42	243 00:50:36	1 382 04:53:51
RTT04	♀	DE/Ger	73 00:10:23	26 00:04:25	57 00:08:35	73 00:15:28	68 00:08:47	66 00:12:41	363 01:00:21
		Σ_{RTT}	73 00:10:23	389 01:02:10	468 01:35:53	666 02:54:54	294 01:25:36	309 01:03:17	2 199 08:12:15
TD08	♀	AT/Ger	231 01:11:58	97 00:32:04	89 00:16:02	80 00:16:02	78 00:15:07	106 00:12:47	681 02:44:01
TD09	♀	AT/Ger	9 00:07:35	48 00:28:54	84 00:18:22	102 00:30:41	51 00:11:46	48 00:18:18	342 01:55:40
TD10	♀	AT/Ger	78 00:07:09	183 00:29:15	4 00:05:30	26 00:08:13	84 00:14:59	86 00:09:31	461 01:14:40
TD11	♀	AT/Ger	35 00:17:00	-	-	137 00:27:33	109 00:19:16	714 01:42:00	995 02:45:51
		Σ_{TD}	353 01:43:43	328 01:30:14	177 00:39:55	345 01:22:30	322 01:01:10	954 02:22:38	2 479 08:40:13
		$\Sigma_{RTT \cup TD}$	426 01:54:06	717 02:32:24	645 02:15:48	1 011 04:17:25	616 02:26:46	1 263 03:25:56	4 678 16:52:28

Adapted to the dataset of this sub-experiment comprising vocalisations of four participants with RTT and four TD participants, a four-fold leave-one-speaker-pair-out cross-validation procedure was carried out. Therefore, first of all, each participant with RTT was paired with a TD participant according to the class-wisely ranked numbers of infant-specific vocalisations, i.e., the participant with RTT with the highest number of vocalisations was paired with the TD participant with the highest number of vocalisations, the participant with RTT with the second highest number of vocalisations was paired with the TD participant with the second highest number of vocalisations, and so forth. Then, for each of the four validation runs,

the dataset was split into a training partition of vocalisations of two participant pairs, a development partition containing the vocalisations of another participant pair, and a test partition containing the vocalisations of the remaining participant pair. Throughout the cross-validation procedure, the vocalisations of each participant pair were used for testing exactly one time. Participant pairing and dataset partitioning are itemised in Table 9.15.

Table 9.15: Class-matched pairing of participants and assignment of participants to training, development, and test partitions at given class proportions of the respective numbers of vocalisations (RTT/TD) for the four-fold leave-one-speaker-pair-out cross-validation procedure applied in the second sub-experiment on RTT recognition. This table is based on Table 2 published by Pokorny and colleagues [226] (DOI: 10.21437/Interspeech.2016-520). AT = Austria; DE = Germany; Gen = gender; Ger = German; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; RTT = Rett syndrome; TD = typical development/typically developing; Voc = vocalisations # = number of; ♀ = female

Partition	Participant pair								Σ
	$\text{Participant}_{\text{RTT}}$				$\text{Participant}_{\text{TD}}$				
	ID	Gen	Nat/L1	# Voc	ID	Gen	Nat/L1	# Voc	
1	RTT01	♀	AT/Ger	150	TD09	♀	AT/Ger	342	492
2	RTT02	♀	AT/Ger	304	TD10	♀	AT/Ger	461	765
3	RTT03	♀	DE/Ger	1 382	TD11	♀	AT/Ger	995	2 377
4	RTT04	♀	DE/Ger	363	TD08	♀	AT/Ger	681	1 044
Σ				2 199				2 479	4 678

Run	Training	Development	Test
1	RTT01, RTT03, TD10, TD11 1 532/1 456	RTT02, TD09 304/342	partition 4 363/681
2	RTT03, RTT04, TD09, TD11 1 745/1 337	RTT01, TD08 150/681	partition 2 304/461
3	RTT01, RTT02, TD08, TD09 454/1 023	RTT04, TD10 363/461	partition 3 1 382/995
4	RTT02, RTT04, TD08, TD10 667/1 142	RTT03, TD11 1 382/995	partition 1 150/342

In a first step, in each validation run, SVM training was carried out on the basis of the respective training partition. Then, SVM optimisation was done by determining the complexity parameter $C \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ that led to the best UAR on the respective development partition. Subsequently, the training partition and the development partition were merged to obtain an ultimate partition

for model training using the previously determined C. This model's classification performance was, finally, evaluated on the basis of vocalisations in the respective test partition. Training partitions of class-imbalanced numbers of vocalisations exceeding a ratio of 3 : 2, were upsampled using Weka's implementation of the synthetic minority oversampling technique (SMOTE [228]). This procedure was necessary for the training partitions used in the third and the fourth validation run as well as for the merged training and development partition used in the third validation run.

Classification performance in this second sub-experiment on RTT recognition was evaluated vocalisation-wisely and based on the mean UAR of the single validation runs only. Respective results are given in Table 9.16.

Table 9.16: Cross-validation results of the second sub-experiment on RTT recognition in form of class-specific numbers of vocalisations (in-)correctly classified as class RTT or TD (confusion matrices), and in form of mean and standard deviation (SD) of the unweighted average recall (UAR) for different feature normalisation/standardisation options. This table is based on Tables 3 published by Pokorny and colleagues [226] (DOI: 10.21437/Interspeech.2016-520). UAR values are rounded to three decimal places. RTT = Rett syndrome; TD = typical development/typically developing

		No feature normalisation/ standardisation		Feature normalisation		Feature standardisation	
classified as →		RTT	TD	RTT	TD	RTT	TD
RTT		1 650	549	1 586	613	1 392	807
TD		1 239	1 240	615	1 864	1 608	871
Measure							
UAR	mean	0.594		0.765		0.498	
	SD	0.114		0.234		0.009	

According to the mean UAR, the three tested classifier configurations, namely (i) classification without prior feature normalisation/standardisation, (ii) classification on the basis of normalised feature values, and (iii) classification on the basis of standardised feature values, were identified via one-sided z-tests to lead to significantly different performances at a significance level of $\alpha = 0.001$. Yielding a mean UAR of 76.5% at a standard deviation of 23.4%, the best (vocalisation-wise) classification performance in this second sub-experiment on RTT recognition was achieved when applying feature normalisation. In contrast, when applying feature standardisation, the classifier did not perform better than chance. Using the feature standardisation option, about two-thirds of vocalisations produced by TD participants were

incorrectly assigned to the class RTT. Classification without applying neither feature normalisation nor feature standardisation, led to a mean UAR of 59.4% at a standard deviation of 11.4%. In this configuration, vocalisations produced by TD participants were about equally assigned to the TD and the RTT class.

9.3.1 How atypical does Rett syndrome sound?

In addition to this thesis' feasibility investigations on an automatic pre-linguistic vocalisation-based recognition of RTT, early RTT-associated vocalisation atypicalities were studied by examining auditory Gestalt perception of RTT-associated pre-linguistic vocalisations in relation to the vocalisations' acoustic signal-level representations. Referring to AQ2, the procedure treated in this subsection was intended to contribute to an objectification of the repeatedly described intermittent character of typical vs. atypical early vocalisations in RTT (cf. Section 3.2, and, e.g., [95, 119]). Thereby, the used approach may, on the one hand, help to systematically characterise auditory vocalisation atypicalities associated with RTT in order to increase the sensitivity of healthcare professionals but also of caregivers for specific patterns in early vocalisations pointing to potential adverse developmental trajectories. On the other hand, by underpinning the feasibility of an automatic vocalisation-based RTT recognition, an approach like the one dealt with in this subsection, may pave the way for the generation of an objective acoustic model of early vocalisation atypicality, as a start, associated with RTT, but then also associated with other COI. Study-related procedures and findings were published in 2018 by Pokorny and colleagues [229].

Taken from GUARDIAN's Dataset 4, the corpus of this study comprised 88 clips of home video recordings showing participant RTT04 within the second half year of life. The same clips were also included in the dataset for the second sub-experiment on automatic RTT recognition. As already specified, RTT04, a female individual from a German-speaking family, received a later diagnosis of the PSV of RTT. By including only the material of one participant (exclusively recorded with a single recording device), the homogeneity of recording quality as well as of intrinsic voice parameters throughout the whole dataset was guaranteed. The decision for building the study on a participant with the PSV of RTT instead of a participant with typical RTT was motivated by the a priori expectation of a higher number and a broader range of different types of pre-linguistic vocalisations produced by an individual with a normally milder course of the disease (also) regarding speech-language capacities. As depicted in Table 9.17, within the available audio-video duration of 1 hour and 21 seconds, a number of 363 target vocalisations of participant RTT04 were segmented. The amplitudes of vocalisation segments were not normalised in this study.

The 363 included vocalisations had a mean duration of 1.82 s at a standard deviation of 1.33 s. The median vocalisation duration was 1.44 s. Vocalisation durations ranged from 450 ms to 12 s. The longest vocalisation comprised a pattern of vowel-like sounds and vowel glides.

The annotation of the included vocalisations according to recording background quality (cf. Section 6.3) yielded 58% of vocalisation (210) related to the best quality class Q1, 3% of vocalisations (10) related to the second best class Q2, 34% of vocalisations (125) related to the second worst class Q3, and 5% of vocalisations (18) related to the worst quality class Q4.

Table 9.17: Number of included vocalisations on the basis of the available audio-video duration in format hh:mm:ss (two-digit hour number:two-digit minute number:two-digit second number; seconds rounded down to integer values; sums calculated on the basis of exact durations) as a function of infant age for the study on auditory atypicalities in RTT-associated vocalisations. This table includes information published by Pokorny and colleagues [229] (DOI: 10.1016/j.ridd.2018.02.019). DE = Germany; Gen = gender; Ger = German; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; RTT = Rett syndrome; ♀ = female

ID	Gen	Nat/L1	Age [month of life]						Σ
			7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	
RTT04	♀	DE/Ger	73 00:10:23	26 00:04:25	57 00:08:35	73 00:15:28	68 00:08:47	66 00:12:41	363 01:00:21

In order to acoustically characterise potential auditory patterns of vocalisation atypicality, on the one hand, the 6 373 features of the comprehensive ComParE set were extracted from each vocalisation. No feature normalisation/standardisation strategy was applied in this study. On the other hand, in the framework of a listening experiment, the vocalisations were independently rated for atypicality by five professionals in the fields of speech-language development and/or developmental psychology/physiology. In detail, the listeners had to rate whether they perceived a presented vocalisation as (i) typical, (ii) atypical, or whether they (iii) did not know. In case a vocalisation was rated 'atypical' by a listener, he or she was further asked to specify whether the perceived atypicality was predominantly associated with the vocalisation's (a) rhythm, (b) timbre, and/or (c) pitch, and/or with (d) an other, not listed auditory vocalisation attribute (multiple selections possible). The listeners were allowed to replay a vocalisation as often as they needed to come to a decision. For later intra-rater reliability evaluations, 10% of randomly selected vocalisations were presented twice throughout the listening experiment. The presentation order of vocalisation duplicates was randomised fulfilling the only constraint not to allow for one and the same vocalisation to be presented directly after one another.

In this study, auditory Gestalt perception of atypicality in pre-linguistic RTT-associated vocalisations was evaluated by calculating an information factor and an atypicality factor for each vocalisation as a function of the ratings for 'do not know' and 'atypical' of the five listeners (l) on the respective vocalisation according to Equations 9.1 and 9.2. Whereas the information factor was intended to reflect the listeners' level of certainty to give an answer on a vocalisation unequal to 'do not know' (an information factor of 0 means that all five listeners rated for 'do not know', an information factor of 1 means that all five listeners rated either for 'typical' or 'atypical'), the atypicality factor was used to indicate a kind of degree of atypicality based on the proportion of ratings on a vocalisation for 'atypical' (an atypicality factor of 0 means that none of the five listeners rated for 'atypical', an atypicality factor of 1 means that all five listeners rated for 'atypical').

$$\textit{information factor} = 1 - \left(\frac{1}{5} \sum_{l=1}^{5} \textit{'do not know'}_l \right) \tag{9.1}$$

$$\textit{atypicality factor} = \frac{1}{5} \sum_{l=1}^{5} \textit{'atypical'}_l \tag{9.2}$$

To begin with, the intra-rater reliability for the listening experiment was reflected by an average Cohen's kappa of 0.66 considering the three main rating options ('typical', 'atypical', and 'do not know'). In 83.2% of cases, vocalisation duplicates were rated consistently. In contrast, a very low inter-rater reliability for the experiment reflected by an overall Fleiss' kappa of 0.2 indicates that even professionals in the relevant field did not share a common concept of (a)typicality in pre-linguistic vocalisations. This finding, once again, motivates the utilisation of objective, computer-assisted pre-linguistic vocalisation information retrieval approaches – like the one primarily followed throughout this thesis – for an automatic and intelligent classification/recognition of early COI-related phenomena in the speech-language domain.

As can be seen in Figure 9.4a, 211 of the 363 vocalisations were rated unequal to 'do not know' by all 5 listener, i.e., they had an information factor of 1. Among these 211 vocalisations, 93 vocalisations were consentaneously rated as 'typical' (*atypicality factor* = 0; see Figure 9.4c). In contrast, only 9 vocalisations were consentaneously rated as 'atypical' (*atypicality factor* = 1; see Figure 9.4b). No vocalisation had an information factor of 0. However, for three vocalisations only one listener came to a decision for either 'typical' or 'atypical'. More than half of the vocalisations (186) were rated 'atypical' by at least one listener (*atypicality factor* > 0; see Figure 9.4b).

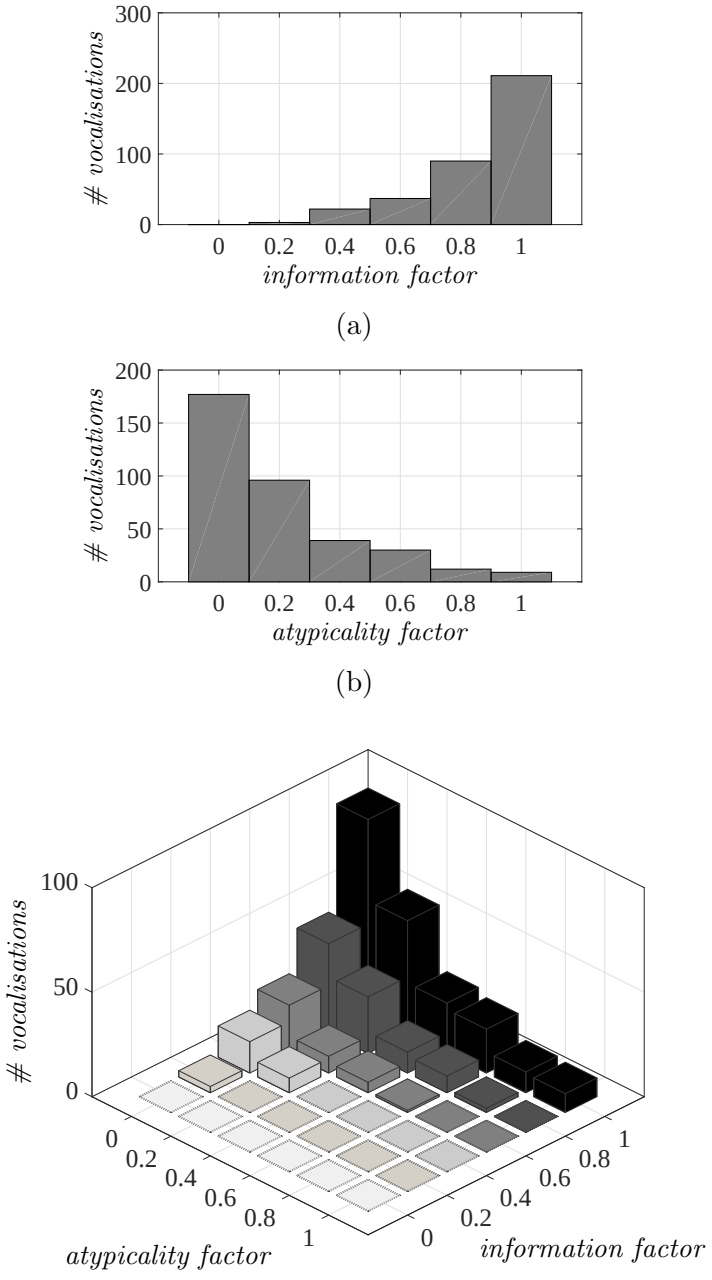


Figure 9.4: Histograms giving absolute numbers of vocalisations as a function of (a) the information factor, (b) the atypicality factor, and (c) the information factor and the atypicality factor in combination, reflecting the rating results of the listening experiment on atypicality in Rett syndrome-associated vocalisations. This figure is based on Figure 1 published by Pokorny and colleagues [229] (DOI: 10.1016/j.ridd.2018.02.019). # = number of

Both the information factor and the atypicality factor were identified to have fair [230] positive correlations with the vocalisation duration. The Spearman's rank correlation coefficients (r_S) were 0.38 and 0.36, respectively. These correlations might indicate that (i) pre-linguistic vocalisations need to have a certain duration in order to contain enough auditory information for a listener to give a definite rating, and (ii) RTT-associated vocalisation atypicality comes out more clearly in specific vocalisation types associated with a longer duration, such as crying sequences or more complex, syllabic vocalisation types. Furthermore, also the specification of auditory attributes to characterise vocalisation atypicality generally depends on the vocalisation duration, as for example, a vocalisation's rhythm can not be perceived in too short vocalisations.

No relation was found between the vocalisations' recording background quality and the information factor ($r_S = -0.16$) as well as between the recording background quality and the atypicality factor ($r_S = -0.04$).

Auditorily perceived atypicality in RTT-associated vocalisations reflected by the atypicality factor was, on the one hand, analysed for relations with the specified auditory vocalisation attributes 'rhythm', 'timbre', 'pitch', and 'other', and, on the other hand, with the acoustic vocalisation characteristics in form of the extracted signal-level features. Therefore, correlations between the atypicality factor, i.e., the proportion of listeners that rated for 'atypical' on a vocalisation, and the respective proportion of listeners that rated for a specific attribute[1], as well as between the atypicality factor and the values of each of the 6 373 acoustic features were calculated. The respective Spearman's rank correlation coefficients with respect to the four auditory attributes as well as to the ten acoustic features with the highest correlation with the atypicality factor are given in Table 9.18.

At a correlation of $r_S = 0.876$, vocalisation atypicality was predominantly specified by the listeners to be related to the auditory attribute 'timbre'. The fact that the proportion of listeners that rated for the option 'other' yielded the second highest correlation with the atypicality factor might suggest that the pre-determined categories were not sufficient for characterising the full range of RTT-associated vocalisation atypicality. In the acoustic domain, vocalisation atypicality was predominantly related to features based on spectral LLDs, namely on RASTA-filtered auditory spectral bands on the one hand, which might reflect the predominance in the listeners' selection for the auditory attribute 'timbre' to characterise perceived atypicality, and based on the energy-related LLD loudness on the other hand, which potentially reflects the high proportion of selections for an 'other', not listed auditory attribute to characterise atypicality.

[1]Positive correlations between the atypicality factor and the proportion of listeners that rated for the particular auditory attributes were expected a priori as – defined by the experimental design of the listening experiment – the specification of an attribute was only required in case of rating a vocalisation as 'atypical'.

Table 9.18: Auditory attributes and top ten acoustic features in correlation with Rett syndrome-associated vocalisation atypicality as given by the atypicality factor ranked according to the magnitude of Spearman's rank correlation coefficients r_S. r_S-values are rounded to three decimal places. audSpecB = auditory spectral band; F0 = fundamental frequency; LregQerr = linear regression quadratic error; pctl = percentile; MFCC = Mel-frequency cepstral coefficient; QregQerr = quadratic regression quadratic error; Rfilt = relative spectral transform (RASTA)-filtered; SD = standard deviation

Rank	r_S	Auditory attribute
1	0.876	timbre
2	0.715	other
3	0.585	rhythm
4	0.535	pitch
Rank	r_S	Acoustic feature
1	-0.522	$pctl_1(audSpecB_{Rfilt}\ 12)$
2	0.517	$pctl_{99}(F0)$
3	-0.509	$pctl_1(MFCC\ 1)$
4	0.506	$QregQerr(loudness)$
5	0.506	$SD(loudness)$
6	-0.504	$pctl_1(audSpecB_{Rfilt}\ 11)$
7	0.502	$range(loudness)$
8	-0.500	$pctl_1(audSpecB_{Rfilt}\ 20)$
9	0.499	$LregQerr(loudness)$
10	-0.493	$pctl_1(audSpecB_{Rfilt}\ 19)$

Finally, it needs to be mentioned that features with potential specific non-linear relations with the atypicality factor might have not been identified on the basis of Spearman's rank correlation analysis.

9.4 Cross-syndrome recognition

Addressing the remaining, not so far treated main research questions (MQ)4 and (MQ)8, and one more time the additional research question (AQ)1 (cf. Section 4.2), this thesis' final experiment was split into two sub-experiments both differing from the previous experiments on developmental disorder recognition as they considered scenarios more relevant for a potential real-world/clinical application, namely scenarios processing more than one COI in one and the same classification paradigm. In the first sub-experiment, a classifier was trained and evaluated on the two-class

task to recognise whether a vocalisation/vocalisations was/were produced by a TD individual, or by an atypically developing (AD) individual, i.e., an individual later diagnosed with one of different possible COI. In the second sub-experiment, the classifier was not only employed to differentiate between TD and AD, but to differentiate between TD and more than one specific developmental disorders in a multi-class paradigm. Parts of the analyses of these two sub-experiments were carried out by Maximilian Schmitt from the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. Experiment-related procedures and findings are currently prepared for publication by Pokorny and colleagues [225].

The dataset used for this thesis' sub-experiments on cross-syndrome recognition was composed from the data already used for the experiment on FXS recognition (cf. Section 9.2) and the first sub-experiment on RTT recognition (cf. Section 9.3). Thus, it contained vocalisations from the respective second half year of life of three male individuals later diagnosed with FXS, three female individuals later diagnosed with RTT, and six gender-matched TD controls, all of which coming from German-speaking families. Altogether, cross-syndrome recognition was based on almost 18 hours of retrospective home video material containing 4 666 target vocalisations. A dataset overview is given in Table 9.19. For the first sub-experiment on cross-syndrome recognition, the participants with FXS and the participants with RTT were pooled together to represent a group of AD individuals as to be classified vs. the six TD individuals. Retaining the TD class represented by the group of three female and three male TD controls, in the second sub-experiment on cross-syndrome recognition, the individuals with FXS and the individuals with RTT were kept separate in order to investigate the three-class paradigm FXS vs. RTT vs. TD. The reason for this thesis' cross-syndrome experimentation only reusing data from the previous experiments on the genetically-caused COI FXS and RTT, and not from the experiment on automatic ASD recognition was that, the ASD dataset differed from the datasets of the other experiments on developmental disorder recognition in essential aspects, namely audio-video material was collected prospectively under (semi-)standardised conditions, and the participants came from Swedish-speaking families.

The mean duration of the 4 666 vocalisations included in the sub-experiments on cross-syndrome recognition was 1.87 s at a standard deviation of 1.52 s. The median vocalisation duration was 1.44 s. Vocalisation durations within the dataset ranged from 186 ms to 21.32 s. The dataset's longest vocalisation was the extended sequence of crying patterns produced by participant TD03, which had already been the longest vocalisation included in the experiment on FXS recognition. Neither the durations of the vocalisations of the participants with FXS, the participants with RTT, and the TD participants, nor the durations of the vocalisations of the

Table 9.19: Number of included vocalisations on the basis of the available audio-video duration in format hh:mm:ss (two-digit hour number:two-digit minute number:two-digit second number; seconds rounded down to integer values; sums calculated on the basis of exact durations) as a function of age per participant for cross-syndrome recognition. AD = atypical development/atypically developing; AT = Austria; DE = Germany; FXS = fragile X syndrome; Gen = gender; Ger = German; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; RTT = Rett syndrome; TD = typical development/typically developing; ♀ = female; ♂ = male; - = no audio-video material available

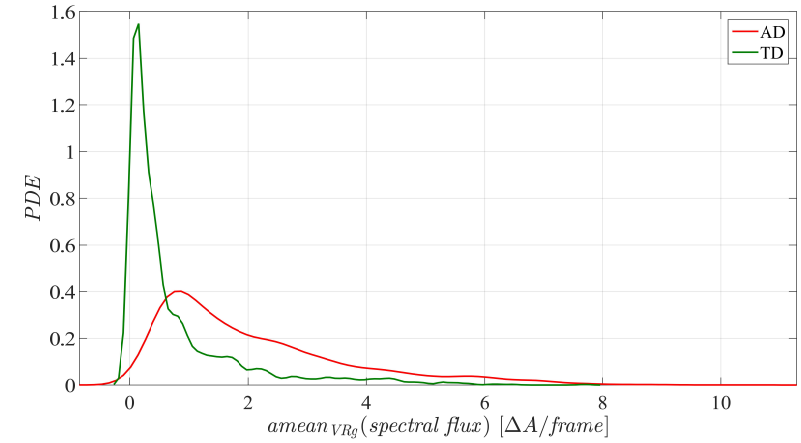
ID	Gen	Nat/L1	Age [month of life] 7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	Σ
FXS01	♂	AT/Ger	85 00:16:01	62 00:05:30	88 00:17:05	45 00:05:45	127 00:15:02	257 00:30:07	664 01:29:32
FXS02	♂	DE/Ger	-	4 00:00:29	-	25 00:02:55	-	1 00:01:35	30 00:05:00
FXS03	♂	DE/Ger	-	7 00:01:24	117 00:10:49	20 00:03:10	83 00:15:16	21 00:02:45	248 00:33:25
		Σ_{FXS}	85 00:16:01	73 00:07:24	205 00:27:55	90 00:11:51	210 00:30:18	279 00:34:28	942 02:07:59
RTT01	♀	AT/Ger	-	-	-	149 00:13:46	-	-	149 00:13:46
RTT02	♀	AT/Ger	21 00:14:48	-	138 42:53	120 01:02:18	46 00:19:05	-	325 02:19:04
RTT03	♀	DE/Ger	142 00:41:35	370 00:57:45	276 00:44:25	325 01:23:21	180 00:57:42	244 00:50:36	1 537 05:35:27
		Σ_{RTT}	163 00:56:23	370 00:57:45	414 01:27:18	594 02:39:25	226 01:16:48	244 00:50:36	2 011 08:08:18
		Σ_{AD}	248 01:12:24	443 01:05:10	619 01:55:13	684 02:51:16	436 01:47:06	523 01:25:05	2 953 10:16:17
TD01	♂	AT/Ger	36 00:04:39	22 00:05:44	-	-	-	12 00:03:16	70 00:13:40
TD02	♂	AT/Ger	9 00:06:36	-	-	4 00:03:36	21 00:06:14	5 00:02:59	39 00:19:27
TD03	♂	AT/Ger	32 00:19:07	20 00:09:39	11 00:11:55	-	4 00:03:39	46 00:22:03	113 01:06:26
TD08	♀	AT/Ger	232 01:11:58	97 00:32:04	90 00:16:02	80 00:16:02	82 00:15:07	108 00:12:47	689 02:44:01
TD09	♀	AT/Ger	9 00:07:35	48 00:28:54	84 00:18:22	101 00:30:41	51 00:11:46	48 00:18:18	341 01:55:40
TD10	♀	AT/Ger	78 00:07:09	183 00:29:15	4 00:05:30	26 00:08:13	84 00:14:59	86 00:09:31	461 01:14:40
		Σ_{TD}	396 01:57:08	370 01:45:38	189 00:51:51	211 00:58:33	242 00:51:47	305 01:08:57	1 713 07:33:56
		$\Sigma_{AD \cup TD}$	644 03:09:32	813 02:50:48	808 02:47:04	895 03:49:50	678 02:38:54	828 02:34:02	4 666 17:50:13

participants with FXS and the participants with RTT merged together to represent the group of AD participants, were normally distributed. The median durations of vocalisations produced by the group of AD participants vs. the group of TD participants were 1.44 s vs. 1.39 s. The participants with FXS produced vocalisations with a median duration of 1.9 s, the participants with RTT of 1.35 s. Using the Mann-Whitney U-test, duration distributions of the vocalisations produced by AD participants and the vocalisations produced by TD participants were identified to differed significantly at $p = 0.0026$. Moreover, also the class-specific duration distributions in the three-class scenario FXS vs. RTT vs. TD were found to bear significant differences at $p = 1.8 \cdot 10^{-41}$ via the Kruskal-Wallis test. Post-testing by means of the Mann-Whitney U-test revealed significant differences between all of the three class-specific vocalisation duration distributions, namely between FXS and TD at $p = 6.2 \cdot 10^{-26}$, between RTT and TD at $p = 0.0456$, and between FXS and RTT at $p = 9.9 \cdot 10^{-44}$.

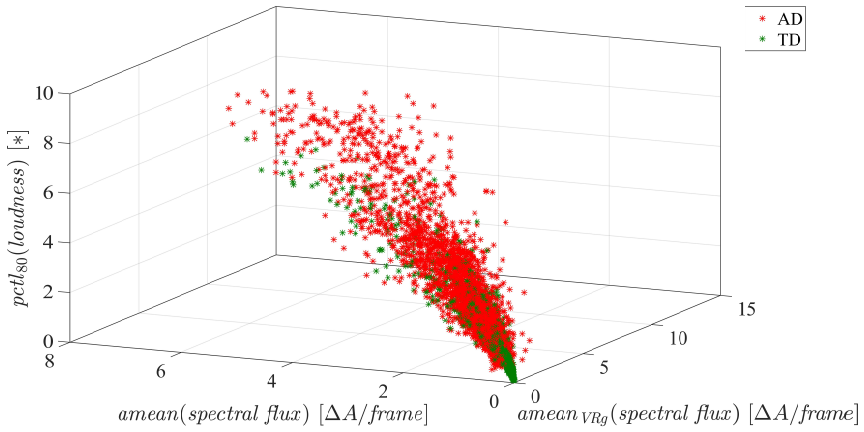
Table 9.20: Top ten acoustic features to differentiate between vocalisations of individuals later diagnosed with a developmental disorder and typically developing individuals, ranked according to the magnitude of the effect size estimate r, at given p-values of the underlying Mann-Whitney U-tests. r is rounded to three decimal places. Multiplicators of p-values are rounded to one decimal place. amean = arithmetic mean; FS = falling slope; $\text{Idx}_{\text{eGeMAPS}}$ = extended Geneva Minimalistic Acoustic Parameter Set-internal feature index [173]; pctl = percentile; pctlR = percentile range; RS = rising slope; SD = standard deviation; VRg = voiced regions; neither UVRg (= unvoiced regions) nor VRg = all regions

Rank	r	p	Feature	$\text{Idx}_{\text{eGeMAPS}}$
1	-0.545	$1.1 \cdot 10^{-303}$	$amean_{VRg}(spectral\ flux)$	67
2	-0.542	$3.4 \cdot 10^{-300}$	$amean(spectral\ flux)$	21
3	-0.523	$6.4 \cdot 10^{-280}$	$pctl_{80}(loudness)$	15
4	-0.519	$1.7 \cdot 10^{-275}$	$pctlR_{20-80}(loudness)$	16
5	-0.515	$6.7 \cdot 10^{-271}$	$amean(loudness)$	11
6	-0.512	$4.7 \cdot 10^{-268}$	$mean_{RSs}(loudness)$	17
7	-0.512	$6.8 \cdot 10^{-268}$	$mean_{FSs}(loudness)$	19
8	-0.494	$8.6 \cdot 10^{-250}$	*equivalent sound level*	88
9	-0.484	$8.3 \cdot 10^{-240}$	$SD_{RSs}(loudness)$	18
10	-0.432	$1.5 \cdot 10^{-191}$	$pctl_{50}(loudness)$	14

In the first sub-experiment on cross-syndrome recognition, i.e., the experiment on the differentiation of vocalisations produced by AD individuals and vocalisations produced by TD individuals, the analysis of the 88 eGeMAPS features via the Mann-Whitney U-test revealed the feature value distributions of 81 features to



(a)



(b)

Figure 9.5: Comparison between vocalisations of individuals later diagnosed with a developmental disorder and vocalisations of TD individuals by means of (a) probability density estimates of the acoustic feature with the highest group differentiation effect (cf. Table 9.20), and by means of (b) vocalisation distributions within the three-dimensional space of the acoustic features with the three highest group differentiation effects (cf. Table 9.20). A = amplitude; AD = atypical development/atypically developing; amean = arithmetic mean; pctl = percentile; PDE = probability density estimate; TD = typical development/typically developing; VRg = voiced regions; * = real measurement unit not existent as feature values refer to the amplitude of the digital audio signal

significantly differ between the two classes. The top ten acoustic features according to the effect size estimate r are given in Table 9.20. Moreover, class-specific

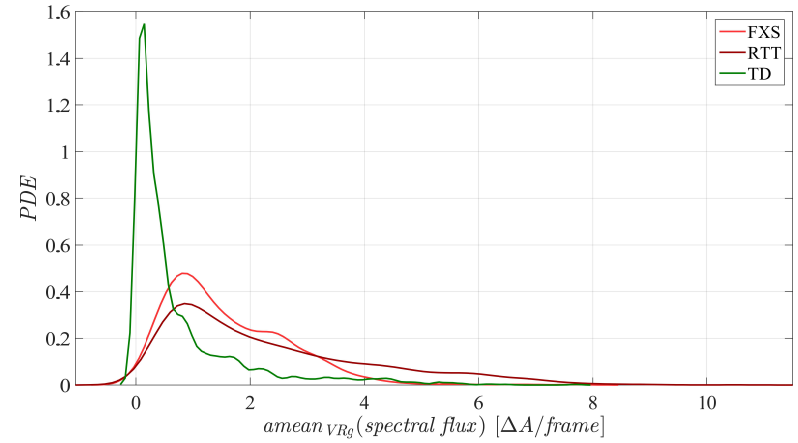
distribution visualisations of the feature with the highest differentiation effect as well as of the top three features as spanning a three-dimensional feature space, are shown in Figure 9.5. Same as in the experiment on RTT recognition, the top two features were based on the LLD spectral flux. Furthermore, similar to both the experiment on RTT recognition and the experiment on FXS recognition, a high proportion of the top ten features – here seven of ten – were related to the LLD loudness. However, this is not very surprising as the dataset used for this sub-experiment on cross-syndrome recognition was composed from the datasets used for the FXS recognition experiment and the RTT recognition experiment.

A similar picture was registered for the second sub-experiment on cross-syndrome recognition, in which the three-class paradigm FXS vs. RTT vs. TD was dealt with. The Kruskal-Wallis test in combination with subsequent Mann-Whitney U post-testing revealed that the distributions of all of the 88 eGeMAPS features significantly differed between at least 2 class pairs. The distributions of 57 features significantly differed between all three class constellations. Table 9.21 lists the ten acoustic features with the highest mean feature ranks according to the effect size estimates r of all three class pair specific distribution differences. Again, class-specific distributions of the top feature as well as of the top three features in the multi-dimensional feature space are shown in Figure 9.6. Same as in the experiment on RTT recognition and in the first sub-experiment on cross-syndrome recognition, the top two features were related to the LLD spectral flux. Additionally, here the sixth highest ranked feature was spectral flux-related too. Furthermore, again a high proportion of the top ten features – here four of ten – was based on loudness. Two top ten features were related to MFCCs. The eighth rank was achieved by the equivalent sound level – a feature which has already been among the top ten features in all previous (sub-)experiments on developmental disorder recognition based on retrospective home video material taken from GUARDIAN's Dataset 4.

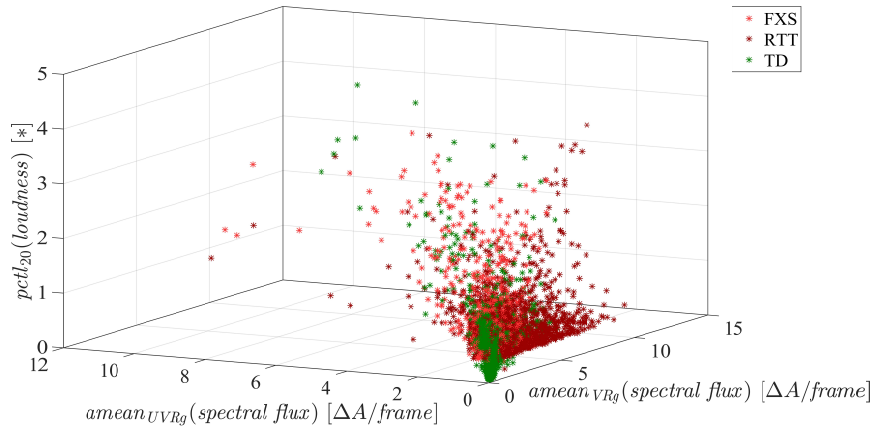
Both the automatic vocalisation-based recognition of atypical vs. typical development and the three-class paradigm FXS vs. RTT vs. TD were investigated similarly to the procedure that has already been used for the experiment on FXS recognition (cf. Section 9.2) and the first sub-experiment on RTT recognition (cf. Section 9.3). Thus, (i) for each of the two sub-experiments on cross-syndrome recognition the dataset was split into partitions for subsequent leave-one-speaker-group-out cross-validation. In case of the two-class paradigm AD vs. TD, six partitions were created each containing one class-matched participant pair (see Table 9.22). The three-class paradigm FXS vs. RTT vs. TD was investigated on the basis of three partitions each containing a participant quadruple of one male participant with FXS, one female participant with RTT, and one male and one female TD participant (see Table 9.23). Grouping in either paradigm was carried out gender-matched according to the class-wisely ranked numbers of available vocalisations per participant. In both

Table 9.21: Top ten acoustic features to differentiate between vocalisations of individuals later diagnosed with FXS, individuals later diagnosed with RTT, and TD individuals, in ascending order according to their mean feature rank over the three class pair specific Mann-Whitney U-tests on the basis of the effect size estimate r, at given p-values of the underlying global three-class Kruskal-Wallis tests as well as of the individual two-class Mann-Whitney U post-tests. The mean feature rank is rounded to integers. Multiplicators of p-values are rounded to one decimal place. amean = arithmetic mean; coeffVar = coefficient of variation; FXS = fragile X syndrome; $\text{Idx}_{\text{eGeMAPS}}$ = extended Geneva Minimalistic Acoustic Parameter Set-internal feature index [173]; MFCC = Mel-frequency cepstral coefficient; pctl = percentile; pctlR = percentile range; RTT = Rett syndrome; TD = typical development/typically developing; UVRg = unvoiced regions; VRg = voiced regions; neither UVRg nor VRg = all regions

Rank	Mean rank	p FXS vs. RTT vs. TD	p FXS vs. TD	p RTT vs. TD	p FXS vs. RTT	Feature	$\text{Idx}_{\text{eGeMAPS}}$
1	9	$2.9 \cdot 10^{-314}$	$4.0 \cdot 10^{-139}$	$5.3 \cdot 10^{-284}$	$3.8 \cdot 10^{-21}$	$amean_{VRg}(spectral\ flux)$	67
2	11	$3.2 \cdot 10^{-171}$	$1.0 \cdot 10^{-128}$	$1.0 \cdot 10^{-104}$	$6.5 \cdot 10^{-30}$	$amean_{UVRg}(spectral\ flux)$	81
3	13	$1.2 \cdot 10^{-140}$	$5.3 \cdot 10^{-105}$	$3.7 \cdot 10^{-69}$	$4.2 \cdot 10^{-50}$	$pctl_{20}(loudness)$	13
4	13	$2.3 \cdot 10^{-200}$	$3.7 \cdot 10^{-143}$	$2.0 \cdot 10^{-141}$	$1.5 \cdot 10^{-15}$	$pctl_{50}(loudness)$	14
5	14	$1.2 \cdot 10^{-279}$	$4.1 \cdot 10^{-132}$	$6.2 \cdot 10^{-253}$	$2.6 \cdot 10^{-10}$	$pctlR_{20\text{–}80}(loudness)$	16
6	15	$2.5 \cdot 10^{-302}$	$1.5 \cdot 10^{-150}$	$2.6 \cdot 10^{-270}$	$1.9 \cdot 10^{-07}$	$amean(spectral\ flux)$	21
7	16	$4.6 \cdot 10^{-120}$	$2.5 \cdot 10^{-93}$	$1.1 \cdot 10^{-65}$	$7.1 \cdot 10^{-30}$	$amean(MFCC\ 4)$	29
8	18	$1.4 \cdot 10^{-253}$	$2.8 \cdot 10^{-121}$	$3.2 \cdot 10^{-228}$	$3.6 \cdot 10^{-10}$	*equivalent sound level*	88
9	20	$1.4 \cdot 10^{-89}$	$6.4 \cdot 10^{-83}$	$1.7 \cdot 10^{-36}$	$4.2 \cdot 10^{-26}$	$amean_{VRg}(MFCC\ 4)$	75
10	22	$1.4 \cdot 10^{-171}$	$6.7 \cdot 10^{-09}$	$1.5 \cdot 10^{-151}$	$8.8 \cdot 10^{-78}$	$coeffVar(loudness)$	12



(a)



(b)

Figure 9.6: Comparison between vocalisations of individuals later diagnosed with FXS, individuals later diagnosed with RTT, and TD individuals by means of (a) probability density estimates of the acoustic feature with the best mean rank according to its differentiation effect for all three group constellations (FXS–TD, RTT–TD, FXS–RTT; cf. Table 9.21), and by means of (b) vocalisation distributions within the three-dimensional space of the acoustic features with the three best mean ranks according to their differentiation effects for all three group constellations (cf. Table 9.21). A = amplitude; amean = arithmetic mean; FXS = fragile X syndrome; pctl = percentile; PDE = probability density estimate; RTT = Rett syndrome; TD = typical development/typically developing; UVRg = unvoiced regions; VRg = voiced regions; * = real measurement unit not existent as feature values refer to the amplitude of the digital audio signal

scenarios (ii) audio normalisation was tested, (iii) the ComParE and the eGeMAPS features were tried out, (iv) feature normalisation to the interval $[0, 1]$ was carried out infant-dependently, age-dependently, and globally as a function of the feature value distributions within the respective training partition, (v) upsampling of the training set was tested, and (vi) linear kernel SVMs were used as classifier with the complexity parameter C optimised $\in \{10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ to achieve the best UAR on the test partition. As SVMs actually represent a two-class discrimination approach, for the three-class paradigm, final decisions were generated by combining single binary one-vs.-rest decisions (cf. multi-class classification in Python's machine learning library scikit-learn[2] [231]).

Table 9.22: Assignment of class-matched participant pairs to partitions for the experiment on the recognition of AD. AD = atypical development/atypically developing; AT = Austria; DE = Germany; FXS = fragile X syndrome; Gen = gender; Ger = German; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; RTT = Rett syndrome; TD = typical development/typically developing; Voc = vocalisations; # = number of; ♀ = female; ♂ = male

Partition	Participant pair								Σ
	$\text{Participant}_{\text{AD}}$				$\text{Participant}_{\text{TD}}$				
	ID	Gen	Nat/L1	# Voc	ID	Gen	Nat/L1	# Voc	
1	FXS01	♂	AT/Ger	664	TD03	♂	AT/Ger	113	777
2	FXS02	♂	DE/Ger	30	TD02	♂	AT/Ger	39	69
3	FXS03	♂	DE/Ger	248	TD01	♂	AT/Ger	70	318
4	RTT01	♀	AT/Ger	149	TD09	♀	AT/Ger	341	490
5	RTT02	♀	AT/Ger	325	TD10	♀	AT/Ger	461	786
6	RTT03	♀	DE/Ger	1 537	TD08	♀	AT/Ger	689	2 226
Σ				2 953				1 713	4 666

Same as already in previous experiments, also the classification performance evaluation of the cross-syndrome recognition tasks was carried out vocalisation-wisely, and infant-wisely based on class majority voting over the infant-specific vocalisation-wise predictions. Performance measures were again the mean UAR of all validation runs and the global UAR determined at once from the entirety of all predictions gathered throughout the cross-validation process. The results for the two-class paradigm AD vs. TD and the three-class paradigm FXS vs. RTT vs. TD for different classification system configurations are given in Table 9.24 and Table 9.25, respectively.

[2]http://scikit-learn.org/ (as of 11 October 2018)

Table 9.23: Assignment of class- and gender-matched participant quadruples to partitions for the multi-class experiment on FXS and RTT recognition. AT = Austria; DE = Germany; FXS = fragile X syndrome; Gen = gender; Ger = German; ID = (unique participant) identification code; L1 = mother tongue; Nat = nationality; RTT = Rett syndrome; TD = typical development/typically developing; Voc = vocalisations; # = number of; ♀= female; ♂ = male

Partition	Participant quadruple												Σ
	$\text{Participant}_{\text{FXS}}$				$\text{Participant}_{\text{RTT}}$				$\text{Participants}_{\text{TD}}$				
	ID	Gen	Nat/L1	# Voc	ID	Gen	Nat/L1	# Voc	ID	Gen	Nat/L1	# Voc	
1	FXS01	♂	AT/Ger	664	RTT03	♀	DE/Ger	1 537	TD03	♂	AT/Ger	113	3 003
									TD08	♀	AT/Ger	689	
2	FXS02	♂	DE/Ger	30	RTT01	♀	AT/Ger	149	TD02	♂	AT/Ger	39	559
									TD09	♀	AT/Ger	341	
3	FXS03	♂	DE/Ger	248	RTT02	♀	AT/Ger	325	TD01	♂	AT/Ger	70	1 104
									TD10	♀	AT/Ger	461	
Σ				942				2 011				1713	4 666

Table 9.24: Cross-validation results of the experiment on recognising atypical development in form of the mean unweighted average recall (UAR) of the three validation runs and the globally calculated UAR based on the gathered predictions of the whole dataset for both vocalisation-wise and infant-wise evaluation and different system configurations regarding audio normalisation, used feature set, feature normalisation strategy (infant-dependent, age-dependent, or normalisation over all instances), and upsampling of the training set. UAR values are rounded to three decimal places. ComParE = Computational Paralinguistics and Emotion (set); eGeMAPS = extended Geneva Minimalistic Acoustic Parameter Set; ✓ = applied; ✗ = not applied

Evaluation mode	Audio normalisation	Feature set	Feature normalisation	Upsampling	Measure	
					UAR_{mean}	UAR_{global}
Vocalisation-wise	✗	ComParE	infant	✗	0.660	0.521
	✗	ComParE	infant	✓	0.660	0.534
	✗	ComParE	age	✗	0.627	0.590
	✗	ComParE	age	✓	0.634	0.606
	✗	ComParE	all	✗	0.584	0.605
	✗	ComParE	all	✓	0.611	0.623
	✗	eGeMAPS	infant	✗	0.465	0.478
	✗	eGeMAPS	infant	✓	0.460	0.475
	✗	eGeMAPS	age	✗	0.550	0.552
	✗	eGeMAPS	age	✓	0.578	0.593
	✗	eGeMAPS	all	✗	0.657	0.726
	✗	eGeMAPS	all	✓	0.616	0.668
	✓	ComParE	infant	✗	0.564	0.416
	✓	ComParE	infant	✓	0.604	0.481
	✓	ComParE	age	✗	0.625	0.586
	✓	ComParE	age	✓	0.626	0.603
	✓	ComParE	all	✗	0.482	0.494
	✓	ComParE	all	✓	0.539	0.529
	✓	eGeMAPS	infant	✗	0.488	0.527
	✓	eGeMAPS	infant	✓	0.460	0.495
	✓	eGeMAPS	age	✗	0.505	0.497
	✓	eGeMAPS	age	✓	0.564	0.495
	✓	eGeMAPS	all	✗	0.545	0.609
	✓	eGeMAPS	all	✓	0.530	0.573
Infant-wise	✗	ComParE	infant	✗	0.750	0.550
	✗	ComParE	infant	✓	0.750	0.550
	✗	ComParE	age	✗	0.667	0.604
	✗	ComParE	age	✓	0.667	0.604
	✗	ComParE	all	✗	0.667	0.712
	✗	ComParE	all	✓	0.667	0.690
	✗	eGeMAPS	infant	✗	0.500	0.500
	✗	eGeMAPS	infant	✓	0.500	0.488
	✗	eGeMAPS	age	✗	0.667	0.695
	✗	eGeMAPS	age	✓	0.667	0.807
	✗	eGeMAPS	all	✗	0.750	0.834
	✗	eGeMAPS	all	✓	0.667	0.801
	✓	ComParE	infant	✗	0.583	0.505
	✓	ComParE	infant	✓	0.667	0.384
	✓	ComParE	age	✗	0.583	0.505
	✓	ComParE	age	✓	0.667	0.604
	✓	ComParE	all	✗	0.417	0.436
	✓	ComParE	all	✓	0.667	0.681
	✓	eGeMAPS	infant	✗	0.583	0.613
	✓	eGeMAPS	infant	✓	0.500	0.513
	✓	eGeMAPS	age	✗	0.500	0.500
	✓	eGeMAPS	age	✓	0.583	0.468
	✓	eGeMAPS	all	✗	0.667	0.827
	✓	eGeMAPS	all	✓	0.583	0.759

Table 9.25: Cross-validation results of the multi-class experiment on fragile X and Rett syndrome recognition in form of the mean unweighted average recall (UAR) of the three validation runs and the globally calculated UAR based on the gathered predictions of the whole dataset for both vocalisation-wise and infant-wise evaluation and different system configurations regarding audio normalisation, used feature set, feature normalisation strategy (infant-dependent, age-dependent, or normalisation over all instances), and upsampling of the training set. UAR values are rounded to three decimal places. ComParE = Computational Paralinguistics and Emotion (set); eGeMAPS = extended Geneva Minimalistic Acoustic Parameter Set; ✓ = applied; ✗ = not applied

Evaluation mode	Audio normalisation	Feature set	Feature normalisation	Upsampling	Measure	
					UAR_{mean}	UAR_{global}
Vocalisation-wise	✗	ComParE	infant	✗	0.278	0.291
	✗	ComParE	infant	✓	0.287	0.299
	✗	ComParE	age	✗	0.312	0.306
	✗	ComParE	age	✓	0.307	0.317
	✗	ComParE	all	✗	0.277	0.240
	✗	ComParE	all	✓	0.305	0.270
	✗	eGeMAPS	infant	✗	0.419	0.521
	✗	eGeMAPS	infant	✓	0.355	0.519
	✗	eGeMAPS	age	✗	0.336	0.381
	✗	eGeMAPS	age	✓	0.310	0.373
	✗	eGeMAPS	all	✗	0.373	0.351
	✗	eGeMAPS	all	✓	0.314	0.310
	✓	ComParE	infant	✗	0.298	0.310
	✓	ComParE	infant	✓	0.308	0.316
	✓	ComParE	age	✗	0.308	0.309
	✓	ComParE	age	✓	0.309	0.320
	✓	ComParE	all	✗	0.223	0.204
	✓	ComParE	all	✓	0.250	0.238
	✓	eGeMAPS	infant	✗	0.432	0.462
	✓	eGeMAPS	infant	✓	0.262	0.404
	✓	eGeMAPS	age	✗	0.319	0.294
	✓	eGeMAPS	age	✓	0.300	0.279
	✓	eGeMAPS	all	✗	0.333	0.262
	✓	eGeMAPS	all	✓	0.234	0.252
Infant-wise	✗	ComParE	infant	✗	0.333	0.333
	✗	ComParE	infant	✓	0.326	0.333
	✗	ComParE	age	✗	0.444	0.358
	✗	ComParE	age	✓	0.333	0.333
	✗	ComParE	all	✗	0.301	0.315
	✗	ComParE	all	✓	0.314	0.327
	✗	eGeMAPS	infant	✗	0.433	0.570
	✗	eGeMAPS	infant	✓	0.412	0.570
	✗	eGeMAPS	age	✗	0.425	0.581
	✗	eGeMAPS	age	✓	0.347	0.510
	✗	eGeMAPS	all	✗	0.544	0.570
	✗	eGeMAPS	all	✓	0.322	0.261
	✓	ComParE	infant	✗	0.333	0.333
	✓	ComParE	infant	✓	0.333	0.333
	✓	ComParE	age	✗	0.444	0.358
	✓	ComParE	age	✓	0.333	0.333
	✓	ComParE	all	✗	0.338	0.257
	✓	ComParE	all	✓	0.338	0.333
	✓	eGeMAPS	infant	✗	0.433	0.512
	✓	eGeMAPS	infant	✓	0.226	0.374
	✓	eGeMAPS	age	✗	0.336	0.273
	✓	eGeMAPS	age	✓	0.336	0.333
	✓	eGeMAPS	all	✗	0.417	0.296
	✓	eGeMAPS	all	✓	0.222	0.243

For the automatic vocalisation-wise recognition of AD vs. TD, a best mean UAR of 66% was achieved when (i) not normalising the amplitudes of the vocalisation segments, (ii) using the ComParE feature set, and (iii) applying infant-dependent feature normalisation. If training set upsamping was applied or not did not make a difference. The same configuration also led to the best mean UAR for infant-wise classification, which was 75%. However, (i) again not applying audio normalisation, but (ii) using the eGeMAPS, (iii) normalising features over all instances according to the feature distributions within the respective training set, and (iv) not applying training set upsampling, also yielded in a mean UAR of 75% for infant-wise classification and in a second best mean UAR for vocalisation-wise classification of 65.7%. This configuration led to both the best global UAR of 72.6% for vocalisation-wise classification and the best global UAR of 83.4% for infant-wise classification.

The classification performance for the three-class paradigm FXS vs. RTT vs. TD, i.e., a classification scenario with the chance level at UAR = $33.\overline{3}$% as compared to 50% for the two-class paradigms, was as follows: Vocalisation-wisely evaluated, the best mean UAR was 43.2%. The underlying classification system was configured to (i) apply audio normalisation, (ii) use the eGeMAPS, (iii) infant-dependently normalise the acoustic features, and (iv) not apply training set upsampling. The same configuration with the only difference not to apply audio normalisation, led to the second best mean UAR of 41.9% and to the best global UAR of even 52.1% for vocalisation-wise classification. This system configuration also represented the first of the three configurations that yielded the second best global UAR for infant-wise classification, namley 57%. The other two configurations that also achieved a global UAR of 57% only differed from the first configuration in one setting each, namely, on the one hand, in the option on training set upsampling to be changed from 'not applied' to 'applied', and, on the other hand, in the feature normalisation strategy to be changed from the infant-dependent mode to the normalisation over all instances according to the feature distributions within the respective training set. This latter configuration, i.e., the system (i) not applying audio normalisation, (ii) using the eGeMAPS, (iii) normalising features over all instances, and (iv) not applying training set upsampling, also achieved the best mean UAR at 54.4% in the infant-wise evaluation scenario.

Part IV

Conclusion

10

Feasibility of early recognition

Employing the method of intelligent vocalisation analysis (Chapter 7) in context of a novel, highly interdisciplinary approach (Chapter 5), this thesis aimed to evaluate the feasibility of a fully automatic pre-linguistic vocalisation-based tool to enhance an earlier identification of infants with a currently 'late recognised' developmental disorder, in particular, with ASD, FXS, or RTT (see Figure 4.1). For this purpose, eight MQs and three AQs were addressed (cf. Section 4.2) in eight (sub-)experiments (Part III). On the one hand, differences between vocalisations of individuals with different developmental outcomes in terms of acoustic signal-level characteristics were investigated (MQ1–MQ4). Moreover, acoustic features with the highest class differentiation effects were identified (AQ1). On the other hand, the automatic classification of vocalisations according to the developmental outcomes of the infants that produced the vocalisations was tested for different COI-vs.-TD paradigms and a COI-vs.-COI-vs.-TD paradigm using different classification system configurations (MQ5–MQ8). An overview of both the LLDs related to the top ten acoustic features per paradigm and the best paradigm-specific recognition performances for vocalisation-wise and infant-wise classification in terms of either the highest mean or globally determined UAR is given in Table 10.1.

Feature analyses in the framework of this thesis revealed that high differentiation effects between vocalisations produced by individuals with different developmental outcomes were obtained via features related to spectral LLDs, such as spectral flux, spectral slope, MFCCs, the Hammarberg index, or the alpha ratio. However, the respective highest number of identified top ten acoustic features for (i) the differentiation paradigm ASD vs. TD and for (ii) all other investigated differentiation paradigms were not spectral-related, but based on (i) F0, and (ii) the energy/amplitude-related LLD loudness. Moreover, another energy-related feature – the equivalent sound level – was also among the ten features with the highest differentiation effects for each paradigm except for the paradigm ASD vs. TD. Loudness, spectral flux, and the equivalent sound level were identified to represent top LLDs

Table 10.1: Overview of top feature analysis and classification results from this thesis' experiments on developmental disorder recognition in form of the low-level descriptors (LLDs) related to the top ten acoustic features, and the highest (mean* or global**) unweighted average recall (UAR) for vocalisation-wise and infant-wise classification per paradigm. LLDs are ranked according to the number of occurrence within the top ten acoustic features (per LLD given in brackets). In case of an equal number of occurrence, LLDs are ranked according to the magnitude of the effect size estimate r. UAR values are rounded to three decimal places. AD = atypical development/atypically developing; ASD = autism spectrum disorder; FXS = fragile X syndrome; MFCC = Mel-frequency cepstral coefficient; RTT = Rett syndrome; TD = typical development/typically developing; UVRg = unvoiced regions; VRg = voiced regions; # = number of; $^{\diamond}$ = related to # individuals

Paradigm	Top 10 feature LLDs	Top UAR	
		Vocalisation-wise	Infant-wise
ASD vs. TD	F0 (2) Spectral slope 0–500 Hz (1) MFCC 4 (1) Mean length of UVRg (1) Hammarberg index (1) Spectral slope 500–1500 Hz (1) Alpha ratio (1) # continuous VRg per second (1) MFCC 1 (1)	0.645*	0.750*,$^{\diamond}$
FXS vs. TD	Loudness (7) Spectral flux (2) Equivalent sound level (1)	0.758**	1.000*,**
RTT vs. TD	Loudness (6) Spectral flux (2) Equivalent sound level (1) Hammarberg index (1)	0.879**	0.845**
AD vs. TD	Loudness (7) Spectral flux (2) Equivalent sound level (1)	0.726**	0.834**
FXS vs. RTT vs. TD	Loudness (4) Spectral flux (3) MFCC 4 (2) Equivalent sound level (1)	0.521**	0.581**

for vocalisation differentiation in the paradigms FXS. vs. TD, RTT vs. TD, AD vs. TD, and FXS vs. RTT vs. TD, but not in the paradigm ASD vs. TD. This might, on the one hand, indicate that pre-linguistic vocalisations of individuals with ASD differ from pre-linguistic vocalisations of TD individuals in an other way than pre-linguistic vocalisations of individuals with FXS or RTT do in terms of acoustic signal-level characteristics, or, on the other hand, reflect the fact that investigations

on the paradigm ASD vs. TD were based on an autonomous set of semi-standardised laboratory recordings of participants from Swedish-speaking families and investigations on the other paradigms were based on a shared pool of home video recordings of participants from German-speaking families.

Basically, the signal-level features identified in the framework of this thesis to best allow for a differentiation between individuals with different developmental outcomes, seem to acoustically underpin previous findings on vocalisation atypicalities in connection with COI (cf. Section 3.2), such as F0 peculiarities and, generally, peculiarities in the (melodic) modulation contour of early ASD-related vocalisations (e.g., [107, 108, 115, 113]), or such as inspiratory, pressed, and/or high-pitched crying-like early phonation patters in early vocalisations of individuals with RTT (e.g., [120, 95, 119]).

Best recognition performances in terms of either the mean or the globally determined UAR for the different two-class paradigms, i.e., paradigms with the chance level at UAR = 50%, ranged from 64.5% (ASD vs. TD) to 87.9% (RTT vs. TD) in the vocalisation-wise evaluation scenario and from 75.0% (ASD vs. TD) to 100% (FXS vs. TD) in the infant-wise evaluation scenario. For the three-class paradigm FXS vs. RTT vs. TD, thus, a paradigm with the chance level at UAR = $33.\overline{3}$%, the highest UAR in the vocalisation-wise evaluation scenario was 52.1%, in the infant-wise evaluation scenario 58.1%.

Most of the top recognition performances were obtained when using the eGeMAPS features as against the ComParE features (cf. Section 7.1). The audio pre-processing strategy of amplitude normalisation did not yield an advantage in many cases. Certainly, the usefulness of audio normalisation highly depends on the audio quality of the underlying dataset. Distinct performance differences were registered as a function of the different applied feature normalisation strategies (cf. Section 7.2). However, no feature normalisation strategy clearly outperformed the others. Only the testing of feature standardisation in the second sub-experiment on RTT recognition caused the classification system not to perform better than chance. At last, not surprisingly for SVM-based classification systems, the configuration option whether to apply or not to apply training set upsampling was without effect in many cases.

In the course of experimentation of this thesis building on different datasets of COI- and TD-associated pre-linguistic vocalisations, an additional vocalisation parameter of potential predictive value for an infant's developmental outcome was found, namely the vocalisation duration. Considering the datasets used for the experiment on ASD recognition (cf. Section 9.1), the experiment on FXS recognition (cf. Section 9.2), and the second sub-experiment on RTT recognition (cf. Section 9.3), pre-linguistic vocalisations of TD individuals turned out to be significantly shorter than pre-linguistic vocalisations of individuals with ASD, FXS,

or RTT. In contrast, dataset statistics relating to the first sub-experiment on RTT recognition yielded RTT-associated vocalisations to be significantly shorter than TD-associated vocalisations. However, vocalisations of individuals with RTT and the significantly longer vocalisations of individuals with FXS pooled together to represent a set of vocalisations of AD individuals were in turn significantly longer than vocalisations of TD individuals according to the dataset used for the cross-syndrome recognition sub-experiments (cf. Section 9.4).

In conclusion, this thesis could raise evidence for acoustic information in pre-linguistic vocalisations bearing potential to reflect the integrity of the young nervous system and to predict an individual's developmental trajectory. The automatic acoustic feature-based classification of pre-linguistic vocalisations according to the developmental outcomes of the infants that produced the vocalisations with respect to the presence or absence of a COI, was successfully[1] tested. Therefore, the basic feasibility of a fully automatic developmental disorder recognition tool as proposed in Figure 4.1 could be demonstrated. Responding to AQ3, fair achievements were made for the automatic detection of infant vocalisations alongside variable quality real-world audio recordings as needed for a vocalisation-based classification system's audio input pre-selection stage (cf. Chapter 8).

Further motivation for the proposed automatic recognition tool is provided by the AQ2-related finding that there seem to be acoustically manifesting auditory atypicalities within at least a proportion of pre-linguistic vocalisations of an individual with a 'late recognised' developmental disorder such as RTT, but human listeners – even if they are professionals in the fields of speech-language development and/or developmental psychology/physiology – are hardly able to consistently recognise these atypicalities (cf. Subsection 9.3.1).

10.1 Limitations

Even though knowledge gained in the framework of this thesis appears to be highly promising for the implementation and practical application of a fully automatic, pre-linguistic vocalisation-based developmental disorder recognition tool, the generated results have to be interpreted carefully due to a number of methodological shortcomings.

One of the main drawbacks of this thesis reducing the generalisability of its experimental findings is the critically small number of available participants for the different (sub-)experiments, especially for the (sub-)experiments including participants with FXS or RTT. Certainly, this is not surprising as FXS and RTT are rare diseases with a current mean age of diagnosis not before toddlerhood and

[1] Classification performance was above chance level.

the acquisition of audio(-video) recordings – usually, in the best case, home video recordings – from an affected individual's prodromal period poses a challenging task. Furthermore, this thesis' claim to minimise additional factors potentially influencing early verbal behaviour and, therefore, to exclusively use gender- and family language-matched datasets for developmental disorder recognition experiments, did not make the recruiting of participants easier. Anyway, the used datasets were not only characterised by small numbers of participants, but also by unbalances among the numbers of available vocalisations per participant and unbalanced numbers of instances per class. These imbalances were most probably due to the different amount of material/durations of audio-video recordings available per participant, but might to a certain extent also be phenotype-related. For all (sub-)experiments except for the experiment on ASD recognition, the unbalanced number of available vocalisations per month of participant age, necessitated to pool vocalisations from all over the participants' respective second half year of life together considering the interval from the seventh to the end of the twelfth month of life as one discrete age point. Thereby, a broad range of vocalisation types was contained and mixed up in the individual datasets reflecting different processes and phases of early vocalisation development (cf. Section 3.1). Certainly, a sufficient number of realisations provided, the isolated analysis of particular, age-specific vocalisation types, such as canonical babbling, would have been of additional interest. Anyway, the factors 'dataset size' and 'dataset balance' significantly determined a priori considerations on suitable classification system configuration options as well as on classifier evaluation strategies, and limited the proper use of specific, state-of-the-art machine learning approaches, such as deep learning algorithms currently representing a highly popular and powerful methodology for classification in larger sample sizes.

Another dataset-related limitation in connection with all experiments of this thesis except for the experiment on ASD recognition, is the use of home video material as the basis for retrospective acoustic analyses at the signal level (cf. Section 5.1). Central issues with respect to home video material are that (i) recordings are non-standardised regarding both recording setting and recording quality, (ii) a broad range of acoustic background noise events has to be expected, (iii) recordings were not made for the purpose of later scientific analyses and primarily show infants in interactive, non-distress situations[2], therefore, the absence of a specific behaviour of interest within the material does not necessarily mean that this behaviour is absent at all, and (iv) reliable investigations on the rate of occurrence of specific behavioural patterns, such as volubility measures, are impossible.

[2]For example, parents usually tend to stop recording when their children fall into a negative mood and/or start to behave somehow 'peculiarly'. Consequently, specific atypical behavioural patterns representing potential predictors for developmental trajectories might be underrepresented in home videos datasets.

Nevertheless, the scientific analysis of home video material not only provides a unique opportunity to retrospectively 'observe' or 'listen to' infants in their natural environments; for now, it still represents one of the best available approaches for studying early behaviour in individuals with 'late recognised' developmental disorders (e.g., [144, 145, 146, 147, 148, 27, 26]).

Generally, the results generated in the framework of this thesis can hardly be discussed in context of previous work as comparable (detection/recognition) studies focussing on signal-level characteristics of pre-linguistic verbal behaviour in ASD, FXS, or RTT to the best of my knowledge do not exist. Moreover, the thesis-internal comparability of results – especially of feature analysis results – from the different (sub-)experiments is limited due to (i) the experiment on ASD recognition having been based on an external dataset of semi-standardised laboratory recordings (cf. Chapter 6) vs. the remaining experiments having been based on GUARDIAN's (home video) Dataset 4 (cf. Section 6.1), and (ii) the reuse of one and the same recordings of participants in different (sub-)experiments.

As a final remark, it needs to be mentioned that conclusions from the reported absolute numbers of features significantly differing between the value distributions of class-specific vocalisations per paradigm have to be drawn with caution as no alpha error correction strategy [176] was applied for multiple feature testing in the framework of this thesis.

11

Implications

Although methodological limitations suggest to interpret the generated experimental results with caution (cf. Section 10.1), this thesis outlines the potential of a novel approach based on pre-linguistic vocalisation acoustics and machine learning to enable an earlier identification of individuals with currently 'late diagnosed' developmental disorders facilitating an earlier treatment. Thereby, this work may build the basis for future research to increase the reliability of the proposed automatic developmental disorder recognition tool and to fit its requirements to realistic healthcare-relevant application scenarios.

11.1 Technical perspectives

From an engineering point of view, the collection of more data is an essentially required next step. So far, the basic technical feasibility of pre-linguistic vocalisation-based COI recognition was demonstrated on the basis of small datasets mainly using linear kernel SVMs. Follow-up experiments should be carried out on more comprehensive datasets in order to (i) generate results with higher reliability, (ii) allow for generalisations with regard to infants' different family language/cultural backgrounds, (iii) extend the approach to other 'late recognised' developmental disorders in the context of multi-class paradigms, and to (iv) allow for alternative, more powerful machine learning tools, such as sophisticated deep learning algorithms [29].

Apart from the careful selection and configuration of an algorithm for the classification step, also the preceding steps of the intelligent vocalisation analysis procedure (cf. Chapter 7), i.e., the feature extraction step and the feature processing step, should be more intensively focussed on in future work. The capabilities of feature sets other than the two applied openSMILE standard sets, the ComParE set and the eGeMAPS, should be evaluated. Alternatively, a special feature set optimised for the acoustic differentiation of developmental outcome-specific vocalisation patterns could be compiled on the basis of the findings of this

thesis' feature analysis efforts. The appliance of promising feature processing techniques, such as the transformation of extracted acoustic features into the bag-of-words representation using the open-source tool kit openXBOW[1] [232], could be tested. The bag-of-words approach was originally applied for text document information retrieval [233], but has proven its value in intelligent vocalisation analysis applications, such as speech emotion recognition (e.g., [234]), over the last years.

Finally, as constituting the crucial audio input pre-selection procedure for the proposed fully automatic developmental disorder recognition tool, intensive engineering efforts need to be made to attain an adequate performance for robust infant voice activity detection in real-world settings.

11.2 Clinical perspectives

Provided that essential technical improvements as suggested in the previous Section (11.1) were made, there are two major healthcare-related application scenarios in which the proposed pre-linguistic vocalisation-based developmental disorder recognition tool could be implemented, namely a home scenario and a clinical scenario. In matters of the underlying training and evaluation datasets, both scenarios were experimentally addressed in the framework of this thesis, though a greater focus was placed on the home scenario due to the accessibility of larger datasets of home video material as compared to laboratory recordings (more or less related to the medical specificities of the studied disorders). Certainly, an implementation in either scenario depends on the healthcare system as well as on the clinical and technological infrastructure available. Automatic developmental outcome prediction in both the home scenario and the clinical scenario could be provided area-wide as an early screening procedure, or, specifically, for infants at heightened risk for adverse developmental outcomes, such as preterms or individuals from families with predispositions for specific diseases.

The home scenario requires parents/caregivers to prospectively collect audio or audio-video data from their children while being awake in natural settings. Data could be recorded, for example, using a standard video camera, an audio recorder, or even a smartphone. Then, these data could be brought along to the physician's office in the course of the next paediatric routine examination. In a first step, the physician would need to transfer the data to his or her computer, e.g., via a USB port, a memory card reader, or Bluetooth. In a second step, the included audio tracks would be automatically segmented for infant target vocalisations. The final step would then comprise the automatic analysis of the set of detected/pre-selected target vocalisations in order to recognise the presence or absence of a (specific)

[1] https://github.com/openXBOW/ (as of 17 September 2018)

developmental disorder in the recorded infant. On the basis of the recognition tool's output decision, the physician could subsequently inform the parents/caregivers about a suspicion of a certain developmental disorder, advice them correspondingly, and potentially initiate a diagnostic cascade.

Alternatively, the fully automatic infant vocalisation segmentation and developmental disorder recognition tool could be directly integrated in an audio recording device that is commercially available or provided for parents by a healthcare institution for home use. As exemplified in Figure 11.1, the tool could potentially be even implemented as a smartphone application – a so-called 'BabbleApp' – that can be downloaded and installed by interested parents or by parents specifically worried about their children's developmental outcomes. Besides providing the core functions of audio recording, vocalisation segmentation, and vocalisation analysis, the 'BabbleApp' could offer some additional useful features for parents/caregivers, such as (i) reminders to record their children during specific age periods of special interest from a speech-language developmental point of view or to perform long-term recordings from time to time, (ii) instructions to record or not to record their children in specific settings, or (iii) warnings not to start a recording in case of a detected insufficient audio background quality. In this scenario of an 'intelligent' home recording device, the parents/caregivers would have the opportunity to seek for a specialist as soon as the tool would advice them to do so.

Figure 11.1: Demonstration of using a hypothetical smartphone application – the 'BabbleApp' – for a fully automatic infant voice activity detection alongside recorded audio input information and the classification of the detected vocalisations according to the expectable developmental outcome of the infant that produced the vocalisations in order to alert parents/caregivers in case of the suspicion of a developmental disorder. FXS = fragile X syndrome

In the clinical scenario, both data acquisition again in the form of audio or audio-video recording and the automatic processing of the recorded material could be automatically carried out at the physician's office or a special clinical laboratory, for example during an ambulant paediatric routine examination or in the framework of an examination procedure to specifically assess individuals at heightened risk for developmental disorders. An inpatient setting would allow for a (24-hour) long-term recording. As compared to the home scenario, the clinical scenario would have the advantage of controllable audio recording conditions and a (semi-)standardisable recording setting, such as a physician-parent-child interaction setting. The disadvantage would be the artificial environment most probably more and more influencing (spontaneous) infant vocalisation behaviour with increasing age and potentially causing a higher proportion of fussy or crying vocalisations.

In both the home scenario and the clinical scenario the proposed developmental disorder recognition tool would generate its final decision on the basis of a set of vocalisations available from an individual. This approach corresponds to the infant-wise system operation mode as evaluated in the framework of this thesis in addition to the vocalisation-wise operation mode. However, in many cases, especially in the clinical scenario, the recognition tool would have to handle very small sets of only a few single available vocalisations.

Finally, it should be emphasised that the suspicion of an infant to have a developmental disorder represents a highly sensitive issue for parents/caregivers. Therefore, healthcare professionals and parents/caregivers that might use a tool like the one proposed in this thesis in future need to keep in mind that the underlying classification system is based on a probabilistic model allowing for misclassification (FPs and FNs). A potential earlier recognition of a developmental disorder would, however, be beneficial in many cases due to specific therapeutic programs, or at least avoid exhausting diagnostic odysseys and enable an earlier symptomatic treatment. Nevertheless, assuming that the methodology presented in this thesis will someday lead to a reliable earlier identification of individuals with a currently 'late recognised' developmental disorder, the need for adapted or novel intervention strategies comes with it.

11.3 Multi-domain fingerprint modelling

By reason of an automatic early recognition of currently 'late diagnosed' developmental disorders exclusively based on pre-linguistic vocalisations to be basically feasible, disorder-specific acoustic characteristics in infant vocal behaviour may be considered as speech-language domain-related neurofunctional markers to (at least partly) reflect the integrity of the developing nervous system. However, also other de-

velopmental domains have proven their value to capture disorder-related behavioural patterns usable for the prediction of atypical developmental trajectories, such as the motor domain (e.g., [235, 236, 92, 237, 238]). Consequently, a fully automatic infant behaviour classification model that simultaneously processes input information from different developmental domains might very likely outperform models that consider information from a single domain.

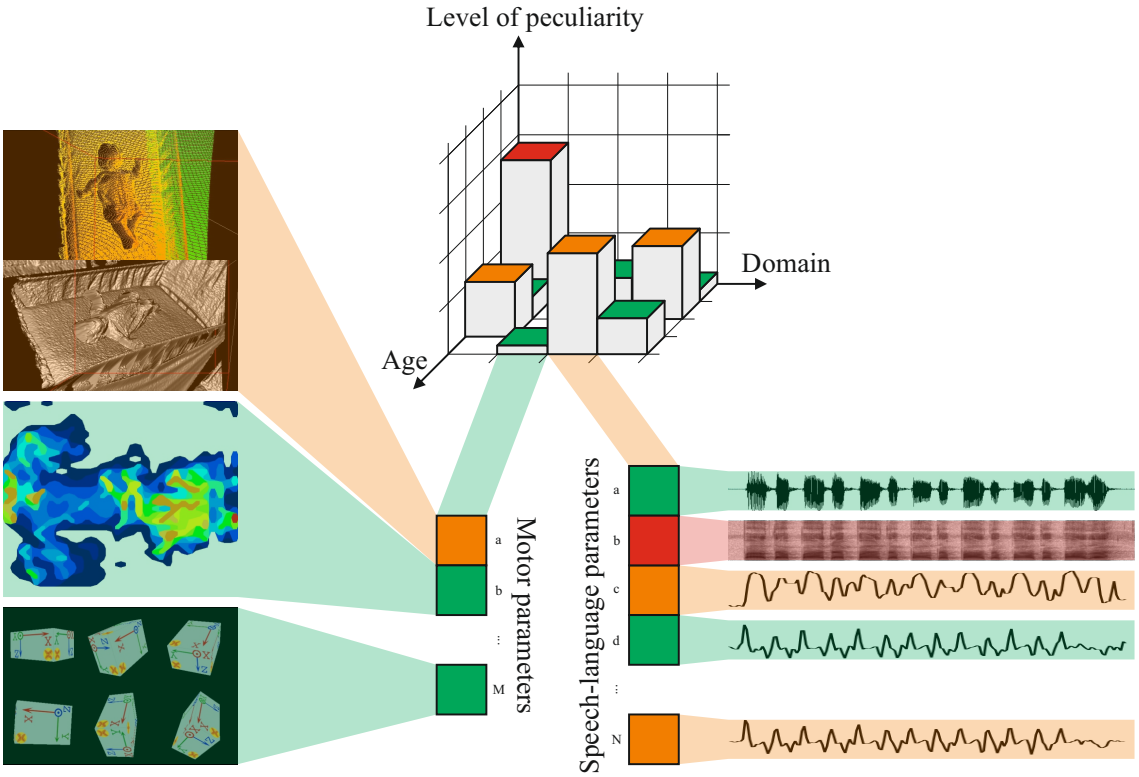


Figure 11.2: Proposed fingerprint model for an automatic recognition of developmental disorders on the basis of age-specific peculiarity distributions over different developmental domains, such as the motor domain and the speech-language domain, captured by domain-specific, recording-derived parameters. Motor parameters are exemplarily derived from a depth channel recording, a pressure distribution recording, and a body-part acceleration recording of an infant in supine position. Speech-language parameters are exemplarily derived from acoustic descriptors of recorded pre-linguistic vocalisations. The basic concept of this figure is based on Figure 1 published by Marschik and colleagues [134] (DOI: https://doi.org/10.1007/s11910-017-0748-8). Colour code: green = optimal/marginally peculiar; orange = moderately peculiar; red = highly peculiar. a, b, c, ... = parameter indices; M = total number of motor parameters; N = total number of speech-language parameters

A potential concept of a respective model is illustrated in Figure 11.2. It can be denominated as a multi-domain fingerprint model as it builds upon the assumption

that each developmental disorder has its characteristic fingerprint in terms of an age-specific peculiarity/symptom distribution over different developmental domains of interest. Respective fingerprint information could be modelled as a kind of level of peculiarity calculated for age-domain constellations on the basis of domain-specific, recording-derived parameters.

Once the fingerprints of different developmental disorders would have been identified, i.e., the model would have been initialised with fingerprint information derived from a set of respective training data, decisions on new data could be made by means of probabilistic fingerprint comparisons. From an application perspective, the suggested model should ideally manage to cope with missing data as in clinical real-world settings not all domain-related parameters might necessarily be available at each age (cf. [134]). Furthermore, the model, so far described to process recording-derived 'online' knowledge only, could be combined with comprehensive collections of additional, non-recording-derived 'offline' knowledge about (e.g., genomic) specificities of different developmental disorders, similar as provided by the Research Domain Criteria (RDoC) Initiative[2] of the National Institute of Mental Health (NIMH). In a clinical setting, the fingerprint model would then have to be set up with metadata of an individual, such as age, gender, weight, pregnancy- and birth-related information, or information about the medical history of the individual's family, before the actual collection of 'online' data, i.e., the multi-domain recording was started.

From 2015 to 2018, GUARDIAN's Dataset 3 (see Table 6.1) was generated by the Research Unit iDN – interdisciplinary Developmental Neuroscience, at the MUG, Austria, with the intention to enable the initialisation of a fingerprint model as described before [134]. It contains both comprehensive metadata and multi-domain recordings of neurotypical participants. Therefore, the dataset allows to define the fingerprint of TD that includes information on each derived parameter's range of normality reflecting neurotypical variability, and represents an optimal training set to initialise a model for atypicality recognition. In order to capture neuro-behavioural adaptations and neuro-functional changes that typically occur around the third month of postnatal life [239, 240], the participants were recorded (7 times) in a bi-weekly interval with the first recording at 4 weeks and the last recording at 16 weeks of post-term age, each time for 5 minutes while lying awake and unstimulated in supine position in a cot. Spontaneous vocal behaviour was recorded with a microphone. Spontaneous motor behaviour was recorded using (i) standard high-definition (HD) camcorders, (ii) Microsoft Kinect sensors, (iii) a pressure-sensitive mat placed between mattress and sheet, and (iv) motion sensors attached to different body parts. Subsequently, all recordings were synchronised among each other to allow for sensor fusion-derived model parameters.

[2] https://www.nimh.nih.gov/research-priorities/rdoc/index.shtml (as of 20 September 2018)

Currently, the dataset is about to be extended at the University Medical Center Göttingen, Germany, for participants at heightened risk for ASD.

In the scientific community, there is an ever growing belief that a multi-domain fingerprint model will substantially enhance the automatic earlier identification of individuals with currently 'late recognised' developmental disorders in order to fulfil the requirements for operating as a screening tool. Building upon this thesis and current trends in technological development such a solution might be close and realistically achievable in the near future.

List of abbreviations

A ..... Amplitude
AAC ..... Advanced Audio Coding
AD ..... Atypical development/atypically developing
A(/)D ..... Analogue-digital (converter)
ADDM ..... Autism and Developmental Disabilities Monitoring
ADHD ..... Attention deficit hyperactivity disorder
AED-ACC ..... Acoustic event detection accuracy
AEER ..... Acoustic event error rate
amean ..... Arithmetic mean
ANN ..... Artificial neural network
AQ ..... Additional (research) question
ASD ..... Autism spectrum disorder
ASS ..... *Autismus-Spektrum-Störung*
AT ..... Austria
audSpecB ..... Auditory spectral band
BIOSIS ..... BioSciences Information Service
BLSTMNN ..... Bidirectional long short-term memory neural network
BRNN ..... Bidirectional recurrent neural network
c ..... Consonant(-like) sound
CB ..... Canonical babbling
CG ..... Caregiver

CLEAR ..... Classification of Events, Activities and Relationships
coeffVar ..... Coefficient of variation
COI ..... Condition(s) of interest
COST ..... European Cooperation in Science and Technology
CP ..... Cerebral palsy
CRIED ..... Cry Recognition In Early Development
ComParE ..... Computational Paralinguistics and Emotion
d ..... Day(s)
DB ..... Database
DE ..... Germany
DGS ..... DiGeorge syndrome
DOI ..... Digital object identifier
DSM ..... Diagnostic and Statistical Manual of Mental Disorders
DV ..... Digital video
DVD ..... Digital versatile disc
EASE ..... Early Autism Sweden
EEG ..... Electroencephalography
eGeMAPS ..... extended Geneva Minimalistic Acoustic Parameter Set
EM ..... Expectation maximisation
Eng ..... English
e.V. ..... *Eingetragener Verein*
f ..... Frequency
F0 ..... Fundamental frequency
FMR1 ..... Fragile X Mental Retardation 1 (gene)
FN ..... False negative
FP ..... False positive
FS ..... Falling slope
FWF ..... *Fonds zur Förderung der wissenschaftlichen Forschung*
FXS ..... Fragile X syndrome/*Fragiles-X-Syndrom*
GmbH ..... *Gesellschaft mit beschränkter Haftung*

Ger ..... German

Gen ..... Gender

GMM ..... Gaussian mixture model

GUARDIAN ... Graz University Audiovisual Research Database for the Interdisciplinary Analysis of Neurodevelopment

H ..... Harmonic

HD ..... High-definition

hh ..... Two-digit hour number

HMM ..... Hidden Markov model

HMM_{dis} ..... Discriminative hidden Markov model

HMM_{gen} ..... Generative hidden Markov model

ID ..... Identification code

iDN ..... Interdisciplinary Developmental Neuroscience

Idx ..... Index

IN ..... Inspiration/inspiratory/ingressive

IPA ..... International Phonetic Alphabet

IPCA ..... Inventory of Potential Communicative Acts

IT ..... Italy

Ita ..... Italian

KCI ..... Korea Citation Index

l ..... Listener

L1 ..... Mother tongue

LENA ..... Language Environment Analysis

LLD ..... Low-level descriptor

LP ..... Linear prediction

LregQerr ..... Linear regression quadratic error

LSTM ..... Long short-term memory

MB ..... Marginal babbling

MECP2 ..... Methyl-CpG-binding Protein 2 (gene)

MFCC ..... Mel-frequency cepstral coefficient

mm ..... Two-digit minute number
MND ..... Minor neurological dysfunction
mo ..... Month(s)
MPEG ..... Moving Picture Experts Group
MQ ..... Main (research) question
MRI ..... Magnetic resonance image
MUG ..... Medical University of Graz
NAS ..... Network-attached storage
Nat ..... Nationality
NIMH ..... National Institute of Mental Health
OeNB ..... *Österreichische Nationalbank*
opt ..... Optimised
P ..... Pause
pctl ..... Percentile
pctlR ..... Percentile range
PDE ..... Probability density estimate
PHS ..... Pitt-Hopkins syndrome
PLPCC ..... Perceptual linear predictive coding coefficient
POI ..... Parameter of interest
PSV ..... Preserved speech variant
Q ..... Quality
QregQerr ..... Quadratic regression quadratic error
RASTA ..... Relative spectral transform
RDoC ..... Research Domain Criteria
RF ..... Random forest
Rfilt ..... Relative spectral transform (RASTA-)filtered
RMS ..... Root mean square
RNN ..... Recurrent neural network
RS ..... Rising slope
RTT ..... Rett syndrome/*Rett-Syndrom*

RQ . . . . . Research question
SAEVD-R . . . . . Stark Assessment of Early Vocal Development-Revised
SciELO . . . . . Scientific Electronic Library Online
SD . . . . . Standard deviation
SE . . . . . Sweden
SMILE . . . . . Speech and Music Interpretation by Large-space Extraction
SMO . . . . . Sequential minimal optimisation
SMOTE . . . . . Synthetic minority oversampling technique
ss . . . . . Two-digit second number
S-VHS . . . . . Super Video Home System
SVM . . . . . Support vector machine
Swe . . . . . Swedish
TD . . . . . Typical development/typically developing
Th . . . . . Threshold
TP . . . . . True positive
UA . . . . . Unweighted accuracy
UAR . . . . . Unweighted average recall
UK . . . . . United Kingdom
USA . . . . . United States of America
USB . . . . . Universal Serial Bus
UVRg . . . . . Unvoiced region(s)
v . . . . . Vowel(-like) sound
V . . . . . Vowel
Vg . . . . . Vowel glide
VHS . . . . . Video Home System
Voc . . . . . Vocalisation(s)
VR . . . . . Vocalisation rate
VRg . . . . . Voiced region(s)
w . . . . . Week(s)
WA . . . . . Weighted accuracy

WAR . . . . . Weighted average recall

y . . . . . Year(s)

ZIKV . . . . . Zika virus

List of symbols

Δ . . . . . Change

♀ . . . . . Female

$\equiv$ . . . . . Identical to

_ . . . . . Juncture

♂ . . . . . Male

\- . . . . . Not available

\# . . . . . Number of

Σ . . . . . Sum

$\cup$ . . . . . Union

List of equations

7.1 Effect size 48
7.2 Precision 53
7.3 Recall 53
7.4 Acoustic event detection accuracy 53
7.4 Acoustic event error rate 53
9.1 Information factor 93
9.2 Atypicality factor 93

List of figures

1.1 Interdisciplinary developmental neuroscience 7

2.1 Lack of early signs 10

4.1 Early recognition tool 22

5.1 Retrospective audio-video analysis 28

6.1 Database composition 31
6.2 Inhomogeneity within retrospective home video material 34
6.3 Data pre-processing routine 36
6.4 Vocalisation segmentation 38
6.5 Vocalisation annotation 40

7.1 Intelligent vocalisation analysis procedure 42
7.2 Vocalisation-based infant voice activity detection validation 54

9.1 Top acoustic feature(s) of autism spectrum disorder 72
9.2 Top acoustic feature(s) of fragile X syndrome 78
9.3 Top acoustic feature(s) of Rett syndrome 84
9.4 Rating results of the study on auditory atypicalities in Rett syndrome-associated vocalisations 94
9.5 Top acoustic feature(s) of atypical development 100
9.6 Top acoustic feature(s) of fragile X syndrome and Rett syndrome . . 103

11.1 BabbleApp 119
11.2 Multi-domain fingerprint model 121

List of tables

4.1 Current state of literature 20

6.1 GUARDIAN 33

7.1 Feature overview 46

8.1 Dataset for the experiment on infant voice activity detection 60
8.2 Partitioning for the experiment on infant voice activity detection . . . 62
8.3 Results of the experiment on infant voice activity detection 64

9.1 Overview of experiments on developmental disorder recognition . . . 68
9.2 Dataset for the experiment on autism recognition 70
9.3 Top acoustic features of autism spectrum disorder 71
9.4 Partitioning for the experiment on autism recognition 73
9.5 Results of the experiment on autism recognition 75
9.6 Dataset for the experiment on fragile X syndrome recognition 76
9.7 Top acoustic features of fragile X syndrome 77
9.8 Partitioning for the experiment on fragile X syndrome recognition . . 79
9.9 Results of the experiment on fragile X syndrome recognition 80
9.10 Dataset for the 1st sub-experiment on Rett syndrome recognition . . . 82
9.11 Top acoustic features of Rett syndrome 83
9.12 Partitioning for the 1st sub-experiment on Rett syndrome recognition 85
9.13 Results of the 1st sub-experiment on Rett syndrome recognition . . . 86
9.14 Dataset for the 2nd sub-experiment on Rett syndrome recognition . . 88
9.15 Partitioning for the 2nd sub-experiment on Rett syndrome recognition 89
9.16 Results of the 2nd sub-experiment on Rett syndrome recognition . . . 90
9.17 Dataset for the study on auditory atypicalities in Rett syndrome-associated vocalisations 92
9.18 Correlation results of the study on auditory atypicalities in Rett syndrome-associated vocalisations 96

9.19 Dataset for the sub-experiments on cross-syndrome recognition 98
9.20 Top acoustic features of atypical development 99
9.21 Top acoustic features of fragile X syndrome and Rett syndrome . . . 102
9.22 Partitioning for the experiment on atypical development recognition . 104
9.23 Partitioning for the experiment on multi-syndrome recognition 105
9.24 Results of the experiment on atypical development recognition 106
9.25 Results of the experiment on multi-syndrome recognition 107

10.1 Overview of top results on developmental disorder recognition 112

Bibliography

[1] H. F. Prechtl, "General movement assessment as a method of developmental neurology: New paradigms and their consequences. The 1999 Ronnie MacKeith Lecture," *Developmental Medicine and Child Neurology*, vol. 43, no. 12, pp. 836–842, 2001.

[2] L. J. Wells, "Development of the human diaphragm and pleural sacs," *Contributions to Embryology*, vol. 35, pp. 107–134, 1954.

[3] J. I. de Vries, G. H. Visser, and H. F. Prechtl, "The emergence of fetal behaviour. I. Qualitative aspects," *Early Human Development*, vol. 7, no. 4, pp. 301–322, 1982.

[4] M.-R. Inanlou, M. Baguma-Nibasheka, and B. Kablar, "The role of fetal breathing-like movements in lung organogenesis," *Histology and Histopathology*, vol. 20, no. 4, pp. 1261–1266, 2005.

[5] C. Einspieler, D. Prayer, and H. F. Prechtl, "Fetal behaviour: A neurodevelopmental approach," *Clinics in Developmental Medicine*, vol. 189, no. 1, pp. 1–200, 2012.

[6] K. McGraw, R. Hoffmann, C. Harker, and J. H. Herman, "The development of circadian rhythms in a human infant," *Sleep*, vol. 22, no. 3, pp. 303–310, 1999.

[7] G. Young, S. J. Segalowitz, C. M. Corter, and S. E. Trehub, *Manual Specialization and the Developing Brain*. New York, NY, USA: Academic Press, 1983.

[8] P. Rochat, "Self-sitting and reaching in 5-to 8-month-old infants: The impact of posture and its development on early eye-hand coordination," *Journal of Motor Behavior*, vol. 24, no. 2, pp. 210–220, 1992.

[9] D. Messinger and A. Fogel, "The interactive development of social smiling," *Advances in Child Development and Behaviour*, vol. 35, pp. 328–366, 2007.

[10] D. Brémond-Gignac, H. Copin, A. Lapillonne, and S. Milazzo, "Visual development in infants: Physiological and pathological mechanisms," *Current Opinion in Ophthalmology*, vol. 22, pp. S1–S8, 2011.

[11] R. Y. Litovsky, D. H. Ashmead, R. Gilkey, and T. Anderson, "Development of binaural and spatial hearing in infants and children," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds. New York, NY, USA: Psychology Press, 2014, ch. 27, pp. 571–592.

[12] L. M. Grummer-Strawn, K. S. Scanlon, and S. B. Fein, "Infant feeding and feeding transitions during the first year of life," *Pediatrics*, vol. 122, no. Supplement 2, pp. S36–S42, 2008.

[13] J. P. Piek, *Infant Motor Development*. Champaign, IL, USA: Human Kinetics, 2006, vol. 10.

[14] N. C. Capone and K. K. McGregor, "Gesture development: A review for clinical and research practices," *Journal of Speech, Language, and Hearing Research*, vol. 47, no. 1, pp. 173–186, 2004.

[15] S. S. Jones, "The development of imitation in infancy," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1528, pp. 2325–2335, 2009.

[16] S. Nathani, D. J. Ertmer, and R. E. Stark, "Assessing vocal development in infants and toddlers," *Clinical Linguistics & Phonetics*, vol. 20, no. 5, pp. 351–369, 2006.

[17] D. K. Oller, "The emergence of the sounds of speech in infancy," in *Child Phonology: Vol. 1. Production*, G. Yeni-Komshian, J. Kavanagh, and C. Ferguson, Eds. New York, NY, USA: Academic Press, 1980, pp. 93–112.

[18] ——, *The Emergence of the Speech Capacity*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2000.

[19] M. Papoušek, *Vom ersten Schrei zum ersten Wort: Anfänge der Sprachentwicklung in der vorsprachlichen Kommunikation*. Bern, Switzerland: Verlag Hans Huber, 1994.

[20] R. E. Stark, "Stages of speech development in the first year of life," in *Child Phonology: Vol. 1. Production*, G. Yeni-Komshian, J. Kavanagh, and C. Ferguson, Eds. New York, NY, USA: Academic Press, 1980, pp. 73–92.

[21] ——, "Infant vocalization: A comprehensive view," *Infant Mental Health Journal*, vol. 2, no. 2, pp. 118–128, 1981.

[22] R. E. Stark, L. E. Bernstein, and M. E. Demorest, "Vocal communication in the first 18 months of life," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 3, pp. 548–558, 1993.

[23] P. Brasil, J. P. P. Jr., M. E. Moreira, R. M. R. Nogueira, L. Damasceno, M. Wakimoto, R. S. Rabello, S. G. Valderramos, U.-A. Halai, T. S. Salles *et al.*, "Zika virus infection in pregnant women in Rio de Janeiro," *New England Journal of Medicine*, vol. 375, no. 24, pp. 2321–2334, 2016.

[24] S. Salavati, C. Einspieler, G. Vagelli, D. Zhang, J. Pansy, J. G. Burgerhof, P. B. Marschik, and A. F. Bos, "The association between the early motor repertoire and language development in term children born after normal pregnancy," *Early Human Development*, vol. 111, pp. 30–35, 2017.

[25] K. D. Bartl-Pokorny, P. B. Marschik, J. Sigafoos, H. Tager-Flusberg, W. E. Kaufmann, T. Grossmann, and C. Einspieler, "Early socio-communicative forms and functions in typical Rett syndrome," *Research in Developmental Disabilities*, vol. 34, no. 10, pp. 3133–3138, 2013.

[26] P. B. Marschik, K. D. Bartl-Pokorny, J. Sigafoos, L. Urlesberger, F. Pokorny, R. Didden, C. Einspieler, and W. E. Kaufmann, "Development of socio-communicative skills in 9-to 12-month-old individuals with fragile X syndrome," *Research in Developmental Disabilities*, vol. 35, no. 3, pp. 597–602, 2014.

[27] P. B. Marschik, K. D. Bartl-Pokorny, H. Tager-Flusberg, W. E. Kaufmann, F. Pokorny, T. Grossmann, C. Windpassinger, E. Petek, and C. Einspieler, "Three different profiles: Early socio-communicative capacities in typical Rett syndrome, the preserved speech variant and normal development," *Developmental Neurorehabilitation*, vol. 17, no. 1, pp. 34–38, 2014.

[28] G. S. Townend, K. D. Bartl-Pokorny, J. Sigafoos, L. M. C. Sven, Bölte, L. Poustka, C. Einspieler, and P. B. Marschik, "Comparing social reciprocity in preserved speech variant and typical Rett syndrome during the early years of life," *Research in Developmental Disabilities*, vol. 43, pp. 80–86, 2015.

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: The MIT Press, 2016.

[30] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: The MIT Press, 2012.

[31] J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, "An overview of machine learning," in *Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Eds. Berlin Heidelberg: Springer, 2013, ch. 1, pp. 3–24.

[32] C. Bowles, N. C. Nowlan, T. T. Hayat, C. Malamateniou, M. Rutherford, J. V. Hajnal, D. Rueckert, and B. Kainz, "Machine learning for the automatic localisation of foetal body parts in cine-MRI scans," in *Medical Imaging 2015: Image Processing*, vol. 9413. Orlando, FL, USA: SPIE, February 2015, p. 94130N.

[33] A. R. Hareendranathan, D. Zonoobi, M. Mabee, D. Cobzas, K. Punithakumar, M. Noga, and J. L. Jaremko, "Toward automatic diagnosis of hip dysplasia from 2D ultrasound," in *Proceedings of the 14th IEEE International Symposium on Biomedical Imaging, ISBI 2017*. Melbourne, Australia: IEEE, April 2017, pp. 982–985.

[34] S. Bharadwaj, H. S. Bhatt, M. Vatsa, and R. Singh, "Domain specific learning for newborn face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 7, pp. 1630–1641, 2016.

[35] J. Li, L. Liu, J. Sun, H. Mo, J.-J. Yang, S. Chen, H. Liu, Q. Wang, and H. Pan, "Comparison of different machine learning approaches to predict small for gestational age infants," *IEEE Transactions on Big Data*, 2016.

[36] C.-Y. Chang, Y.-C. Hsiao, and S.-T. Chen, "Application of incremental SVM learning for infant cries recognition," in *Proceedings of the 18th IEEE International Conference on Network-Based Information Systems, NBiS 2015*. Taipei, Taiwan: IEEE, September 2015, pp. 607–610.

[37] C.-Y. Chang and J.-J. Li, "Application of deep learning for recognizing infant cries," in *Proceedings of the IEEE International Conference on Consumer Electronics-Taiwan, ICCE-TW 2016*. Nantou County, Taiwan: IEEE, May 2016, pp. 1–2.

[38] Y. Lavner, R. Cohen, D. Ruinskiy, and H. IJzerman, "Baby cry detection in domestic environment using deep learning," in *Proceedings of the IEEE International Conference on the Science of Electrical Engineering, ICSEE 2016*. Eilat, Israel: IEEE, November 2016, pp. 1–5.

[39] R. L. Rodriguez and S. S. Caluya, "Waah: Infants cry classification of physiological state based on audio features," in *Proceedings of the IEEE International Conference on Soft Computing, Intelligent System and Information Technology, ICSIIT 2017*. Denpasar, Indonesia: IEEE, September 2017, pp. 7–10.

[40] J. R. Isler, T. Thai, M. M. Myers, and W. P. Fifer, "An automated method for coding sleep states in human infants based on respiratory rate variability," *Developmental psychobiology*, vol. 58, no. 8, pp. 1108–1115, 2016.

[41] L. Fraiwan and K. Lweesy, "Neonatal sleep state identification using deep learning autoencoders," in *Proceedings of the 13th IEEE International Colloquium on Signal Processing & its Applications, CSPA 2017*. Penang, Malaysia: IEEE, March 2017, pp. 228–231.

[42] Y. Guo, J. Wrammert, K. Singh, K. Ashish, K. Bradford, and A. Krishnamurthy, "Automatic analysis of neonatal video data to evaluate resuscitation performance," in *Proceedings of the 6th IEEE International Conference on Computational Advances in Bio and Medical Sciences, ICCABS 2016*. Atlanta, GA, USA: IEEE, October 2016, pp. 1–6.

[43] N. Mago, S. Srivastava, R. D. Shirwaikar, U. D. Acharya, L. E. Lewis, and M. Shivakumar, "Prediction of Apnea of Prematurity in neonates using support vector machines and random forests," in *Proceedings of the 2nd IEEE International Conference on Contemporary Computing and Informatics, IC3I 2016*. Greater Noida, India: IEEE, December 2016, pp. 693–697.

[44] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.

[45] G. Ball, P. Aljabar, T. Arichi, N. Tusor, D. Cox, N. Merchant, P. Nongena, J. V. Hajnal, A. D. Edwards, and S. J. Counsell, "Machine-learning to characterise neonatal functional connectivity in the preterm brain," *NeuroImage*, vol. 124, pp. 267–275, 2016.

[46] C. D. Smyser, N. U. Dosenbach, T. A. Smyser, A. Z. Snyder, C. E. Rogers, T. E. Inder, B. L. Schlaggar, and J. J. Neil, "Prediction of brain maturity in infants using machine-learning algorithms," *NeuroImage*, vol. 136, pp. 1–9, 2016.

[47] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. Arlington, VA, USA: American Psychiatric Association, 2013.

[48] K. H. Nicolaides, G. Azar, D. L. Byrne, C. A. Mansur, and K. Marks, "Fetal nuchal translucency: ultrasound screening for chromosomal defects in first trimester of pregnancy," *BMJ*, vol. 304, no. 6831, pp. 867–869, 1992.

[49] T. Lissauer, A. A. Fanaroff, L. Miall, and J. Fanaroff, *Neonatology at a Glance*. West Sussex, UK: Wiley, 2015.

[50] C. A. Morris and C. B. Mervis, "Williams syndrome," in *Handbook of Neurodevelopmental and Genetic Disorders in Children*, S. Goldstein and C. R. Reynolds, Eds. New York, NY, USA: The Guilford Press, 1999, ch. 24, pp. 555–591.

[51] M. A. Martens, S. J. Wilson, and D. C. Reutens, "Research review: Williams syndrome: a critical review of the cognitive, behavioral, and neuroanatomical phenotype," *Journal of Child Psychology and Psychiatry*, vol. 49, no. 6, pp. 576–608, 2008.

[52] C. B. Mervis and A. E. John, "Williams syndrome: Psychological characteristics," in *Neurogenetic Syndromes: Behavioral Issues and Their Treatment*, B. K. Shapiro and P. J. Accardo, Eds. Baltimore, MD, USA: Paul H. Brookes Publishing Co., 2010, ch. 6, pp. 81–98.

[53] P. C. Mainardi, "Cri du Chat syndrome," *Orphanet Journal of Rare Diseases*, vol. 1, no. 1, p. 33, 2006.

[54] S. Bhat, U. R. Acharya, H. Adeli, G. M. Bairy, and A. Adeli, "Autism: Cause factors, early diagnosis and therapies," *Reviews in the Neurosciences*, vol. 25, no. 6, pp. 841–850, 2014.

[55] K. Goodspeed, C. Newsom, M. A. Morris, C. Powell, P. Evans, and S. Golla, "Pitt-Hopkins syndrome: A review of current literature, clinical approach, and 23-patient case series," *Journal of Child Neurology*, pp. 1–12, DOI: 10.1177/0 883 073 817 750 490, 2018.

[56] L. Kanner, "Early infantile autism," *Journal of Pediatrics*, 1944.

[57] H. Asperger, "Die Autistischen Psychopathen im Kindesalter," *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 117, no. 1, pp. 76–136, 1944.

[58] J. Baio, L. Wiggins, D. L. Christensen, M. J. Maenner, J. Daniels, Z. Warren, M. Kurzius-Spencer, W. Zahorodny, C. R. Rosenberg, T. White, M. S. Durkin, P. Imm, L. Nikolaou, M. Yeargin-Allsopp, L.-C. Lee, R. Harrington, M. Lopez, R. T. Fitzgerald, A. Hewitt, S. Pettygrove, J. N. Constantino, A. Vehorn, J. Shenouda, J. Hall-Lande, K. V. N. Braun, and N. F. Dowling, "Prevalence of autism spectrum disorder among children aged 8 years – Autism and Developmental Disabilities Monitoring network," *MMWR. Surveillance Summaries*, vol. 67, no. 6, pp. 1–23, 2018.

[59] S. Sumi, H. Taniai, T. Miyachi, and M. Tanemura, "Sibling risk of pervasive developmental disorder estimated by means of an epidemiologic survey in Nagoya, Japan," *Journal of Human Genetics*, vol. 51, no. 6, pp. 518–522, 2006.

[60] S. Ozonoff, G. S. Young, A. Carter, D. Messinger, N. Yirmiya, L. Zwaigenbaum, S. Bryson, L. J. Carver, J. N. Constantino, K. Karen Dobkins, T. Hutman, J. M. Iverson, R. Landa, S. J. Rogers, M. Sigman, and W. L. Stone, "Recurrence risk for autism spectrum disorders: A Baby Siblings Research Consortium study," *Pediatrics*, vol. 128, no. 3, pp. e488–e495, 2011.

[61] S. Bölte, "Is autism curable?" *Developmental Medicine & Child Neurology*, vol. 56, no. 10, pp. 927–931, 2014.

[62] L. Zwaigenbaum, M. L. Bauman, R. Choueiri, D. Fein, C. Kasari, K. Pierce, W. L. Stone, N. Yirmiya, A. Estes, R. L. Hansen, J. C. McPartland, M. R. Natowicz, T. Buie, A. Carter, P. A. Davis, D. Granpeesheh, Z. Mailloux, C. Newschaffer, D. Robins, S. S. Roley, S. Wagner, and A. Wetherby, "Early identification and interventions for autism spectrum disorder: Executive summary," *Pediatrics*, vol. 136, no. Supplement 1, pp. S1–S9, 2015.

[63] L. Zwaigenbaum, M. L. Bauman, R. Choueiri, C. Kasari, A. Carter, D. Granpeesheh, Z. Mailloux, S. S. Roley, S. Wagner, D. Fein, K. Pierce, T. Buie, P. A. Davis, C. Newschaffer, D. Robins, A. Wetherby, W. L. Stone, N. Yirmiya, A. Estes, R. L. Hansen, J. C. McPartland, and M. R. Natowicz, "Early intervention for children with autism spectrum disorder under 3 years of age: Recommendations for practice and research," *Pediatrics*, vol. 136, no. Supplement 1, pp. S60–S81, 2015.

[64] J. H. Elder, C. M. Kreider, S. N. Brasher, and M. Ansell, "Clinical impact of early diagnosis of autism on the prognosis and parent-child relationships," *Psychology Research and Behavior Management*, vol. 10, p. 283, 2017.

[65] L. Zwaigenbaum, M. L. Bauman, D. Fein, K. Pierce, T. Buie, P. A. Davis, C. Newschaffer, D. L. Robins, A. Wetherby, R. Choueiri, C. Kasari, W. L. Stone, N. Yirmiya, A. Estes, R. L. Hansen, J. C. McPartland, M. R. Natowicz, A. Carter, D. Granpeesheh, Z. Mailloux, S. S. Roley, and S. Wagner, "Early screening of autism spectrum disorder: Recommendations for practice and research," *Pediatrics*, vol. 136, no. Supplement 1, pp. S41–S59, 2015.

[66] D. L. Christensen, J. Baio, K. V. N. Braun, D. Bilder, J. Charles, J. N. Constantino, J. Daniels, M. S. Durkin, R. T. Fitzgerald, M. Kurzius-Spencer, L.-C. Lee, S. Pettygrove, C. Robinson, E. Schulz, C. Wells, M. S. Wingate, W. Zahorodny, and M. Yeargin-Allsopp, "Prevalence and characteristics of autism spectrum disorder among children aged 8 years – Autism and Developmental Disabilities Monitoring network," *MMWR. Surveillance Summaries*, vol. 65, no. 3, pp. 1–23, 2016.

[67] N. Sicherman, G. Loewenstein, T. Tavassoli, and J. D. Buxbaum, "Grandma knows best: Family structure and age of diagnosis of autism spectrum disorder," *Autism*, vol. 22, no. 3, pp. 368–376, 2018.

[68] James P. Martin and Julia Bell, "A pedigree of mental defect showing sex-linkage," *Journal of Neurology and Psychiatry*, vol. 6, no. 3-4, p. 154, 1943.

[69] H. A. Lubs, "A marker X chromosome," *American Journal of Human Genetics*, vol. 21, no. 3, p. 231, 1969.

[70] A. J. Verkerk, M. Pieretti, J. S. Sutcliffe, Y.-H. Fu, D. P. Kuhl, A. Pizzuti, O. Reiner, S. Richards, M. F. Victoria, F. Zhang, B. E. Eussen, G.-J. B. van Ommen, L. A. Blonden, G. J. Riggins, J. L. Chastain, C. B. Kunst, H. Galjaard, C. T. Caskey, D. L. Nelson, B. A. Oostra, and S. T. Warren, "Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome," *Cell*, vol. 65, no. 5, pp. 905–914, 1991.

[71] W. E. Kaufmann and A. L. Reiss, "Molecular and cellular genetics of fragile X syndrome," *American Journal of Medical Genetics*, vol. 88, no. 1, pp. 11–24, 1999.

[72] D. Cohen, N. Pichard, S. Tordjman, C. Baumann, L. Burglen, E. Excoffier, G. Lazar, P. Mazet, C. Pinquier, A. Verloes, and D. Hron, "Specific genetic disorders and autism: Clinical contribution towards their identification," *Journal of Autism and Developmental Disorders*, vol. 35, no. 1, pp. 103–116, 2005.

[73] G. S. Fisch, "Developmental influences on psychological phenotypes," in *Neurogenetic Syndromes: Behavioral Issues and Their Treatment*, B. K. Shapiro and P. J. Accardo, Eds. Baltimore, MD, USA: Paul H. Brookes Publishing Co., 2010, ch. 7, pp. 99–114.

[74] J. Hunter, O. Rivero-Arias, A. Angelov, E. Kim, I. Fotheringham, and J. Leal, "Epidemiology of fragile X syndrome: A systematic review and meta-analysis," *American Journal of Medical Genetics*, vol. 164, no. 7, pp. 1648–1658, 2014.

[75] L. Boyle and W. E. Kaufmann, "The behavioral phenotype of FMR1 mutations," *American Journal of Medical Genetics*, vol. 154C, no. 4, pp. 469–476, 2010.

[76] R. J. Hagerman, "The physical and behavioral phenotype," in *Fragile X Syndrome: Diagnosis, Treatment, and Research*, R. J. Hagerman and P. J. Hagerman, Eds. Baltimore, MD, USA: The Johns Hopkins University Press, 2002, ch. 1, pp. 3–109.

[77] W. E. Kaufmann, R. Cortell, A. S. Kau, I. Bukelis, E. Tierney, R. M. Gray, C. Cox, G. T. Capone, and P. Stanard, "Autism spectrum disorder in fragile X syndrome: Communication, social interaction, and specific behaviors," *American Journal of Medical Genetics*, vol. 129, no. 3, pp. 225–234, 2004.

[78] D. Z. Loesch, Q. M. Bui, C. Dissanayake, S. Clifford, E. Gould, D. Bulhak-Paterson, F. Tassone, A. K. Taylor, D. Hessl, R. Hagerman, and R. M. Huggins, "Molecular and cognitive predictors of the continuum of autistic behaviours in fragile X," *Neuroscience & Biobehavioral Reviews*, vol. 31, no. 3, pp. 315–326, 2007.

[79] S. W. Harris, D. Hessl, B. Goodlin-Jones, J. Ferranti, S. Bacalman, I. Barbato, F. Tassone, P. J. Hagerman, K. Herman, and R. J. Hagerman, "Autism profiles of males with fragile X syndrome," *American Journal on Mental Retardation*, vol. 113, no. 6, pp. 427–438, 2008.

[80] J. M. Opitz and G. R. Sutherland, "International workshop on the fragile X and X-linked mental retardation," *American Journal of Medical Genetics*, vol. 17, no. 1, pp. 5–94, 1984.

[81] B. A. Oostra and D. J. Halley, "Complex behavior of simple repeats: The fragile X syndrome," *Pediatric Research*, vol. 38, no. 5, p. 629, 1995.

[82] D. B. Bailey, M. Raspa, E. Bishop, and D. Holiday, "No change in the age of diagnosis for fragile X syndrome: Findings from a national parent survey," *Pediatrics*, vol. 124, no. 2, pp. 527–533, 2009.

[83] A. Rett, "Über ein zerebral-atrophisches Syndrom bei Hyperammonämie," *Wiener Medizinische Wochenschrift*, vol. 116, pp. 723–726, 1966.

[84] ——, "On a remarkable syndrome of cerebral atrophy associated with hyperammonaemia in childhood," *Wiener Medizinische Wochenschrift*, vol. 166, no. 11-12, pp. 322–324, 2016.

[85] R. E. Amir, I. B. van den Veyver, M. Wan, C. Q. Tran, U. Francke, and H. Y. Zoghbi, "Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2," *Nature Genetics*, vol. 23, no. 2, pp. 185–188, 1999.

[86] N. Tokaji, H. Ito, T. Kohmoto, T. Naruto, R. Takahashi, A. Goji, T. Mori, Y. Toda, M. Saito, S. Tange, K. Masuda, S. Kagami, and I. Imoto, "A rare male patient with classic Rett syndrome caused by MeCP2_e1 mutation," *American Journal of Medical Genetics*, vol. 176, no. 3, pp. 699–702, 2018.

[87] C. L. Laurvick, N. D. Klerk, C. Bower, J. Christodoulou, D. Ravine, C. Ellaway, S. Williamson, and H. Leonard, "Rett syndrome in Australia: A review of the epidemiology," *Journal of Pediatrics*, vol. 148, no. 3, pp. 347–352, 2006.

[88] J. L. Neul, W. E. Kaufmann, D. G. Glaze, J. Christodoulou, A. J. Clarke, N. Bahi-Buisson, H. Leonard, M. E. S. Bailey, N. C. Schanen, M. Zappella, A. Renieri, P. Huppke, and A. K. Percy, "Rett syndrome: Revised diagnostic criteria and nomenclature," *Annals of Neurology*, vol. 68, no. 6, pp. 944–950, 2010.

[89] B. Suter, D. Treadwell-Deering, H. Y. Zoghbi, D. G. Glaze, and J. L. Neul, "Brief report: MECP2 mutations in people without Rett syndrome," *Journal of Autism and Developmental Disorders*, vol. 44, no. 3, pp. 703–711, 2014.

[90] L. Burd and G. G. Gascon, "Rett syndrome: Review and discussion of current diagnostic criteria," *Journal of Child Neurology*, vol. 3, no. 4, pp. 263–268, 1988.

[91] J. L. Neul, J. B. Lane, H.-S. Lee, S. Geerts, J. O. Barrish, F. Annese, L. M. Baggett, K. Barnes, S. A. Skinner, K. J. Motil, D. G. Glaze, W. E. Kaufmann, and A. K. Percy, "Developmental delay in Rett syndrome: Data from the natural history study," *Journal of Neurodevelopmental Disorders*, vol. 6, no. 1, p. 20, 2014.

[92] C. Einspieler, J. Sigafoos, K. D. Bartl-Pokorny, R. Landa, P. B. Marschik, and S. Bölte, "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*, vol. 8, no. 3, pp. 286–291, 2014.

[93] S. Tams-Little and G. Holdgrafer, "Early communication development in children with Rett syndrome," *Brain and Development*, vol. 18, no. 5, pp. 376–378, 1996.

[94] P. B. Marschik, C. Einspieler, and J. Sigafoos, "Contributing to the early detection of Rett syndrome: The potential role of auditory Gestalt perception," *Research in Developmental Disabilities*, vol. 33, no. 2, pp. 461–466, 2012.

[95] P. B. Marschik, G. Pini, K. D. Bartl-Pokorny, M. Duckworth, M. Gugatschka, R. Vollmann, M. Zappella, and C. Einspieler, "Early speech-language development in females with Rett syndrome: Focusing on the preserved speech variant," *Developmental Medicine & Child Neurology*, vol. 54, no. 5, pp. 451–456, 2012.

[96] D. C. Tarquinio, W. Hou, J. L. Neul, J. B. Lane, K. V. Barnes, H. M. O'Leary, N. M. Bruck, W. E. Kaufmann, K. J. Motil, D. G. Glaze, S. A. Skinner,

F. Annese, L. Baggett, J. O. Barrish, S. P. Geerts, and A. K. Percy, "Age of diagnosis in Rett syndrome: Patterns of recognition among diagnosticians and risk factors for late diagnosis," *Pediatric Neurology*, vol. 52, no. 6, pp. 585–591, 2015.

[97] S. W. Littlejohn and K. A. Foss, *Theories of Human Communication*. Belmont, CA, USA: Thomson Learning, Inc., 2010.

[98] M. Tomasello, *Origins of Human Communication*. Cambridge, MA, USA: The MIT Press, 2010.

[99] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. West Sussex, UK: Wiley, 2006.

[100] E. Bates, *Language and Context: The Acquisition of Pragmatics*. New York, NY, USA: Academic Press, 1976.

[101] J. L. Locke, *The Child's Path to Spoken Language*. Cambridge, MA, USA: Harvard University Press, 1995.

[102] D. K. Oller, R. E. Eilers, A. R. Neal, and A. B. Cobo-Lewis, "Late onset canonical babbling: A possible early marker of abnormal development," *American Journal on Mental Retardation*, vol. 103, no. 3, pp. 249–263, 1998.

[103] S. M. Barlow and M. Estep, "Central pattern generation and the motor infrastructure for suck, respiration, and speech," *Journal of Communication Disorders*, vol. 39, no. 5, pp. 366–380, 2006.

[104] S. M. Barlow, J. P. L. Radder, M. E. Radder, and A. K. Radder, "Central pattern generators for orofacial movements and speech," in *Handbook of Mammalian Vocalization: An Integrative Neuroscience Approach*, S. M. Brudzynski, Ed. London, UK: Academic Press, 2010, ch. 8.4, pp. 351–369.

[105] F. J. Koopmans-Van Beinum and J. M. van der Stelt, "Early stages in the development of speech movements," in *Precursors of Early Speech*, B. Lindblom and R. Zetterström, Eds. London, UK: Palgrave Macmillan, 1986, ch. 4, pp. 37–50.

[106] L. Roche, D. Zhang, K. D. Bartl-Pokorny, F. B. Pokorny, B. W. Schuller, G. Esposito, S. Bölte, H. Roeyers, L. Poustka, M. Gugatschka, H. Waddington, R. Vollmann, and P. B. Einspieler, Christa Marschik, "Early vocal development in autism spectrum disorder, Rett syndrome, and fragile X syndrome: Insights from studies using retrospective video analysis," *Advances in Neurodevelopmental Disorders*, pp. 49–61, 2018.

[107] G. Esposito and P. Venuti, "Developmental changes in the fundamental frequency (f0) of infants' cries: A study of children with autism spectrum disorder," *Early Child Development and Care*, vol. 180, no. 8, pp. 1093–1102, 2010.

[108] S. J. Sheinkopf, J. M. Iverson, M. L. Rinaldi, and B. M. Lester, "Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder," *Autism Research*, vol. 5, no. 5, pp. 331–339, 2012.

[109] G. Esposito, J. Nakazawa, P. Venuti, and M. H. Bornstein, "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*, vol. 34, no. 9, pp. 2717–2724, 2013.

[110] E. Patten, K. Belardi, G. T. Baranek, L. R. Watson, J. D. Labban, and D. K. Oller, "Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency," *Journal of Autism and Developmental Disorders*, vol. 44, no. 10, pp. 2413–2428, 2014.

[111] R. Paul, Y. Fuerst, G. Ramsay, K. Chawarska, and A. Klin, "Out of the mouths of babes: Vocal production in infant siblings of children with ASD," *Journal of Child Psychology and Psychiatry*, vol. 52, no. 5, pp. 588–598, 2011.

[112] N. Chericoni, D. de Brito Wanderley, V. Costanzo, A. Diniz-Gonçalves, M. Leitgel Gille, E. Parlato, D. Cohen, F. Apicella, S. Calderoni, and F. Muratori, "Pre-linguistic vocal trajectories at 6–18 months of age as early markers of autism," *Frontiers in Psychology*, vol. 7, p. 1595, 2016.

[113] M. Zappella, C. Einspieler, K. D. Bartl-Pokorny, M. Krieber, M. Coleman, S. Bölte, and P. B. Marschik, "What do home videos tell us about early motor and socio-communicative behaviours in children with autistic features during the second year of life – An exploratory study," *Early Human Development*, vol. 91, no. 10, pp. 569–575, 2015.

[114] E. Werner, G. Dawson, J. Osterling, and N. Dinno, "Brief report: Recognition of autism spectrum disorder before one year of age: A retrospective study based on home videotapes," *Journal of Autism and Developmental Disorders*, vol. 30, no. 2, pp. 157–162, 2000.

[115] J. Brisson, K. Martel, J. Serres, S. Sirois, and J.-L. Adrien, "Acoustic analysis of oral productions of infants later diagnosed with autism and their mother," *Infant Mental Health Journal*, vol. 35, no. 3, pp. 285–295, 2014.

[116] S. Maestro, F. Muratori, M. C. Cavallaro, F. Pei, D. Stern, B. Golse, and F. Palacio-Espasa, "Attentional skills during the first 6 months of age in autism

spectrum disorder," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 41, no. 10, pp. 1239–1245, 2002.

[117] S. Maestro, F. Muratori, M. C. Cavallaro, C. Pecini, A. Cesari, A. Paziente, D. Stern, B. Golse, and F. Palacio-Espasa, "How young children treat objects and people: An empirical study of the first year of life in autism," *Child psychiatry and human development*, vol. 35, no. 4, pp. 383–396, 2005.

[118] K. Belardi, L. R. Watson, R. A. Faldowski, H. Hazlett, E. Crais, G. T. Baranek, C. McComish, E. Patten, and D. K. Oller, "A retrospective video analysis of canonical babbling and volubility in infants with fragile X syndrome at 9–12 months of age," *Journal of Autism and Developmental Disorders*, vol. 47, no. 4, pp. 1193–1206, 2017.

[119] P. B. Marschik, W. E. Kaufmann, J. Sigafoos, T. Wolin, D. Zhang, K. D. Bartl-Pokorny, G. Pini, M. Zappella, H. Tager-Flusberg, C. Einspieler, and M. V. Johnston, "Changing the perspective on early development of Rett syndrome," *Research in Developmental Disabilities*, vol. 34, no. 4, pp. 1236–1239, 2013.

[120] P. B. Marschik, C. Einspieler, A. Oberle, F. Laccone, and H. F. Prechtl, "Case report: Retracing atypical development: A preserved speech variant of Rett syndrome," *Journal of Autism and Developmental Disorders*, vol. 39, no. 6, pp. 958–961, 2009.

[121] J. Sigafoos, G. Woodyatt, D. Keen, K. Tait, M. Tucker, D. Roberts-Pennell, and N. Pittendreigh, "Identifying potential communicative acts in children with developmental and physical disabilities," *Communication Disorders Quarterly*, vol. 21, no. 2, pp. 77–86, 2000.

[122] J. Sigafoos, M. Arthur-Kelly, and N. Butterfield, *Enhancing Everyday Communication for Children with Disabilities*. Baltimore, MD, USA: Paul H Brooks Publishing Company, 2006.

[123] H. R. Bauer and R. D. Kent, "Acoustic analyses of infant fricative and trill vocalizations," *Journal of the Acoustical Society of America*, vol. 81, no. 2, pp. 505–511, 1987.

[124] R. D. Kent and A. D. Murray, "Acoustic features of infant vocalic utterances at 3, 6, and 9 months," *Journal of the Acoustical Society of America*, vol. 72, no. 2, pp. 353–365, 1982.

[125] P. Keating and R. Buhr, "Fundamental frequency in the speech of infants and children," *Journal of the Acoustical Society of America*, vol. 63, no. 2, pp. 567–571, 1978.

[126] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens, "A new, robust vocal fundamental frequency (F0) determination method for the analysis of infant cries," in *Proceedings of the 7th IEEE Symposium on Computer-Based Medical Systems, CBMS 1994*. Winston-Salem, NC, USA: IEEE, June 1994, pp. 223–228.

[127] A. S. Warlaumont, D. K. Oller, E. H. Buder, R. Dale, and R. Kozma, "Data-driven automated acoustic analysis of human infant vocalizations using neural network tools," *Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2563–2577, 2010.

[128] J.-J. Aucouturier, Y. Nonaka, K. Katahira, and K. Okanoya, "Segmentation of expiratory and inspiratory sounds in baby cry audio recordings using hidden Markov models," *Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2969–2977, 2011.

[129] S. Sharma, P. R. Myakala, R. Nalumachu, S. V. Gangashetty, and V. Mittal, "Acoustic analysis of infant cry signal towards automatic detection of the cause of crying," in *Proceedings of the 7th IEEE International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2017*. San Antonio, TX, USA: IEEE, October 2017, pp. 117–122.

[130] S. Sharma and V. K. Mittal, "A qualitative assessment of different sound types of an infant cry," in *Proceedings of the 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics, UPCON 2017*. Mathura, India: IEEE, October 2017, pp. 532–537.

[131] P. R. Myakala, R. Nalumachu, S. Sharma, and V. Mittal, "An intelligent system for infant cry detection and information in real time," in *Proceedings of the 7th IEEE International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2017*. San Antonio, TX, USA: IEEE, October 2017, pp. 141–146.

[132] ——, "A low cost intelligent smart system for real time infant monitoring and cry detection," in *Proceedings of the IEEE Region 10 Conference, TENCON 2017*. Penang, Malaysia: IEEE, November 2017, pp. 2795–2800.

[133] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *Proceedings of the 19th Annual Conference of the International Speech Communication*

Association, Interspeech 2018. Hyderabad, India: ISCA, September 2018, pp. 122–126.

[134] P. B. Marschik, F. B. Pokorny, R. Peharz, D. Zhang, J. O'Muircheartaigh, H. Roeyers, S. Blte, A. J. Spittle, B. Urlesberger, B. Schuller, L. Poustka, S. Ozonoff, F. Pernkopf, T. Pock, K. Tammimies, C. Enzinger, M. Krieber, I. Tomantschger, K. D. Bartl-Pokorny, J. Sigafoos, L. Roche, G. Esposito, M. Gugatschka, K. Nielsen-Saines, C. Einspieler, W. E. Kaufmann, and the BEE-PRI Study Group, "A novel way to measure and predict development: A heuristic approach to facilitate the early detection of neurodevelopmental disorders," *Current Neurology and Neuroscience Reports*, vol. 17, no. 5, p. 43, 2017.

[135] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. Hansen, "Signal processing for young child speech language development," in *Proceedings of the 1st Workshop on Child, Computer and Interaction*. Crete, Greece: ISCA, October 2008.

[136] D. K. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13 354–13 359, 2010.

[137] D. Xu, J. Gilkerson, J. Richards, U. Yapanel, and S. Gray, "Child vocalization composition as discriminant information for automatic autism detection," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2009*. Minneapolis, MN, USA: IEEE, September 2009, pp. 2518–2522.

[138] S. Orlandi, C. Manfredi, L. Bocchi, and M. Scattoni, "Automatic newborn cry analysis: a non-invasive tool to help autism early diagnosis," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2012*. San Diego, CA, USA: IEEE, August/September 2012, pp. 2953–2956.

[139] B. Schuller, *Intelligent Audio Analysis*. Berlin Heidelberg, Germany: Springer, 2013.

[140] D. B. Bailey, D. Skinner, D. Hatton, and J. Roberts, "Family experiences and factors associated with the diagnosis of fragile X syndrome," *Journal of Developmental and Behavioral Pediatrics*, vol. 21, no. 5, pp. 315–321, 2000.

[141] D. B. Bailey, D. Skinner, and K. L. Sparkman, "Discovering fragile X syndrome: Family experiences and perceptions," *Pediatrics*, vol. 111, no. 2, pp. 407–416, 2003.

[142] R. Hinton, D. B. Budimirovic, P. B. Marschik, V. B. Talisa, C. Einspieler, T. Gipson, and M. V. Johnston, "Parental reports on early language and motor milestones in fragile X syndrome with and without autism spectrum disorders," *Developmental Neurorehabilitation*, vol. 16, no. 1, pp. 58–66, 2013.

[143] D. Zhang, W. E. Kaufmann, J. Sigafoos, K. D. Bartl-Pokorny, M. Krieber, P. B. Marschik, and C. Einspieler, "Parents' initial concerns about the development of their children later diagnosed with fragile X syndrome," *Journal of Intellectual & Developmental Disability*, vol. 42, no. 2, pp. 114–122, 2017.

[144] J. L. Adrien, P. Lenoir, J. Martineau, A. Perrot, L. Hameury, C. Larmande, and D. Sauvage, "Blind ratings of early symptoms of autism based upon family home movies," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 32, no. 3, pp. 617–626, 1993.

[145] E. R. Crais, L. R. Watson, G. T. Baranek, and J. S. Reznick, "Early identification of autism: How early can we go?" *Seminars in Speech and Language*, vol. 27, no. 3, pp. 143–160, 2006.

[146] R. Palomo, M. Belinchón, and S. Ozonoff, "Autism and family home movies: A comprehensive review," *Journal of Developmental & Behavioral Pediatrics*, vol. 27, no. 2, pp. 59–68, 2006.

[147] C. Saint-Georges, R. S. Cassel, D. Cohen, M. Chetouani, M.-C. Laznik, S. Maestro, and F. Muratori, "What studies of family home movies can teach us about autistic infants: A literature review," *Research in Autism Spectrum Disorders*, vol. 4, no. 3, pp. 355–366, 2010.

[148] P. B. Marschik and C. Einspieler, "Methodological note: Video analysis of the early development of Rett syndrome – one method for many disciplines," *Developmental Neurorehabilitation*, vol. 14, no. 6, pp. 355–357, 2011.

[149] M. P. Lynch, D. K. Oller, M. L. Steffens, and E. H. Buder, "Phrasing in prelinguistic vocalizations," *Developmental Psychobiology*, vol. 28, no. 1, pp. 3–25, 1995.

[150] S. Nathani and D. K. Oller, "Beyond ba-ba and gu-gu: Challenges and strategies in coding infant vocalizations," *Behavior Research Methods, Instruments, & Computers*, vol. 33, no. 3, pp. 321–330, 2001.

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

[152] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational linguistics, and Speech Recognition*, 2nd ed. Upper Saddle River, NJ, USA: Pearson Education, 2009.

[153] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. West Sussex, UK: Wiley, 2014.

[154] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association, Interspeech 2013*. Lyon, France: ISCA, August 2013, pp. 148–152.

[155] B. Schuller, "Voice and speech analysis in search of states and traits," in *Computer Analysis of Human Behavior*. London, UK: Springer, 2011, pp. 227–253.

[156] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[157] F. Eyben, F. Weninger, and B. Schuller, "Affect recognition in real-life acoustic conditions – A new perspective on feature selection," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association, Interspeech 2013*. Lyon, France: ISCA, August 2013, pp. 2044–2048.

[158] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013*. Vancouver, Canada: IEEE, May 2013, pp. 483–487.

[159] B. W. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. C. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association, Interspeech 2016*. San Francisco, CA, USA: ISCA, September 2016, pp. 2001–2005.

[160] M. Schedl, E. Gómez, and J. Urbano, "Music information retrieval: Recent developments and applications," in *Foundations and Trends® in Information*

Retrieval. Hanover, MA, USA: Now Publishers, Inc., 2014, vol. 8, no. 2–3, pp. 127–261.

[161] J. R. Deller, J. H. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*. New York, NY, USA: Macmillan Pub. Co., 1993.

[162] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[163] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proceedings of the 17th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1992*, vol. 1. San Francisco, CA, USA: IEEE, March 1992, pp. 121–124.

[164] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[165] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans. 1st International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006. Lecture Notes in Computer Science*, G. J. Stiefelhagen R., Ed. Berlin Heidelberg, Germany: Springer, 2007, vol. 4122, pp. 311–322.

[166] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 evaluation," in *Multimodal Technologies for Perception of Humans. RT 2007, CLEAR 2007. Lecture Notes in Computer Science*, F. J. Stiefelhagen R., Bowers R., Ed. Berlin Heidelberg, Germany: Springer, 2008, vol. 4625, pp. 3–34.

[167] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, "Acoustic event detection and classification," in *Computers in the Human Interaction Loop. HumanComputer Interaction Series.*, S. R. Waibel A., Ed. London, UK: Springer, 2009, ch. 7, pp. 61–73.

[168] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia, MM 2010*. Florence, Italy: ACM, October 2010, pp. 1459–1462.

[169] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia, MM 2013*. Barcelona, Spain: ACM, October 2013, pp. 835–838.

[170] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association, Interspeech 2014*. Singapore, Singapore: ISCA, September 2014, pp. 427–431.

[171] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association, Interspeech 2015*. Dresden, Germany: ISCA, September 2015, pp. 478–482.

[172] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association, Interspeech 2017*. Stockholm, Sweden: ISCA, August 2017, pp. 3442–3446.

[173] F. Eyben, K. R.Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[174] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[175] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.

[176] M. Bland, *An introduction to medical statistics*, 4th ed. Oxford, UK: Oxford University Press, 2015.

[177] R. Rosenthal, *Meta-analytic procedures for social research*. Newbury Park, CA, USA: SAGE Publications, 1991, vol. 6.

[178] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, pp. 249–268, 2007.

[179] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A System for Large-Scale Machine Learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*. Savannah, GA, USA: USENIX, November 2016, pp. 265–283.

[180] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[181] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[182] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support vector Learning*, B. Schölkopf, C. J. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: The MIT Press, 1999, ch. 12, pp. 185–208.

[183] J. R. Quinlan, "Bagging, boosting, and C4.5," in *Proceedings of the 13th National Conference on Artificial Intelligence, AAAI 1996*, vol. 1. Portland, OR, USA: AAAI, August 1996, pp. 725–730.

[184] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[185] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[186] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[187] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010.

[188] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2008.

[189] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments," in *Proceedings of the Workshop on Machine Listening in Multisource Environments, CHiME 2011*. Florence, Italy: ISCA, September 2011, pp. 24–29.

[190] R. Brueckner and B. Schuller, "Hierarchical neural networks and enhanced class posteriors for social signal classification," in *Proceedings of the 2013 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2013*. Olomouc, Czech Republic: IEEE, December 2013, pp. 361–364.

[191] C. Ding, L. Xie, J. Yan, W. Zhang, and Y. Liu, "Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features," in *Proceedings of the 2015 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2015*. Scottsdale, AZ, USA: IEEE, December 2015, pp. 98–102.

[192] Q. Yu, P. Liu, Z. Wu, S. K. Ang, H. Meng, and L. Cai, "Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016*. Shanghai, China: IEEE, March 2016, pp. 5545–5549.

[193] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter, and H. Ney, "A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition," in *Proceedings of the 42st IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017*. New Orleans, LA, USA: IEEE, March 2017, pp. 2462–2466.

[194] H. Niemann, *Klassifikation von Mustern*, 2nd ed. published online: https://www5.cs.fau.de/fileadmin/Persons/NiemannHeinrich/klassifikation-von-mustern/m00-www.pdf (as of 20 June 2018), 2003.

[195] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*, J. F. Kolen and S. C. Kremer, Eds. New York, NY, USA: IEEE Press, 2001, ch. 14, pp. 237–244.

[196] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[197] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*. San Diego, CA, USA: Computational and Biological Learning Society, May 2015.

[198] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[199] G. D. J. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[200] A. J. Viterbi, "A personal history of the Viterbi algorithm," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 120–142, 2006.

[201] H. Phan, M. Maaß, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 20–31, 2015.

[202] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association, Interspeech 2009*. Brighton, UK: ISCA, September 2009, pp. 312–315.

[203] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, Interspeech 2010*. Makuhari, Japan: ISCA, September 2010, pp. 2794–2797.

[204] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech 2011*. Florence, Italy: ISCA, August 2011, pp. 3201–3204.

[205] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. v. Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association, Interspeech 2012*. Portland, OR, USA: ISCA, September 2012, pp. 254–257.

[206] F. B. Pokorny, R. Peharz, W. Roth, M. Zöhrer, F. Pernkopf, P. B. Marschik, and B. W. Schuller, "Manual versus automated: The challenging routine of infant vocalisation segmentation in home videos to study neuro (mal) development," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association, Interspeech 2016*. San Francisco, CA, USA: ISCA, September 2016, pp. 2997–3001.

[207] S. Yamamoto, Y. Yoshitomi, M. Tabuse, K. Kushida, and T. Asada, "Detection of baby voice and its application using speech recognition system and fundamental frequency analysis," in *Proceedings of the 10th WSEAS International Conference on Applied Computer Science, ACS 2010*. Iwate, Japan: WSEAS, October 2010, pp. 341–345.

[208] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[209] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, 2003.

[210] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, 2004.

[211] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, 2005.

[212] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006.

[213] R. Tahmasbi and S. Rezaei, "A soft voice activity detection using GARCH filter and variance gamma distribution," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1129–1134, 2007.

[214] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *Proceedings of the 6th IEEE International Conference on Signal Processing, ICSP 2002*, vol. 2. Beijing, China: IEEE, August 2002, pp. 1124–1127.

[215] Q.-H. Jo, J.-H. Chang, J. Shin, and N. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Processing*, vol. 3, no. 3, pp. 205–210, 2009.

[216] D. Cournapeau, S. Watanabe, A. Nakamura, and T. Kawahara, "Online unsupervised classification with model comparison in the variational Bayes framework for voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1071–1083, 2010.

[217] J. W. Shin, J.-H. Chang, and N. S. Kim, "Voice activity detection based on statistical models and machine learning approaches," *Computer Speech & Language*, vol. 24, no. 3, pp. 515–530, 2010.

[218] J. Wu and X.-L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 466–469, 2011.

[219] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2624–2633, 2011.

[220] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[221] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time voice activity detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018.

[222] F. Eyben, F. Weninger, M. Wöllmer, and B. Schuller, "open-Source Media Interpretation by Large feature-space Extraction. Documentation Version 2.1," audEERING UG, Tech. Rep., 2015.

[223] F. Pernkopf and M. Wohlmayr, "Large margin learning of Bayesian classifiers based on Gaussian mixture models," in *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. Lecture Notes in Computer Science*, J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, Eds. Berlin Heidelberg, Germany: Springer, 2010, vol. 6323, pp. 50–66.

[224] F. B. Pokorny, B. W. Schuller, P. B. Marschik, R. Brueckner, P. Nyström, N. Cummins, S. Bölte, C. Einspieler, and T. Falck-Ytter, "Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association, Interspeech 2017*. Stockholm, Sweden: ISCA, August 2017, pp. 309–313.

[225] F. B. Pokorny, M. Schmitt, M. Egger, M. Feiner, C. Leitinger, C. Einspieler, M. Gugatschka, B. W. Schuller, and P. B. Marschik, "Intelligent infant vocalisation analysis in genetic disorders: Technical underpinnings for an earlier detection of fragile X and Rett syndrome," *Speech Communication*, in preparation.

[226] F. B. Pokorny, P. B. Marschik, C. Einspieler, and B. W. Schuller, "Does she speak RTT? Towards an earlier identification of Rett syndrome through intelligent pre-linguistic vocalisation analysis," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association, Interspeech 2016*. San Francisco, CA, USA: ISCA, September 2016, pp. 1953–1957.

[227] K. Chenausky, C. Nelson, and H. Tager-Flusberg, "Vocalization rate and consonant production in toddlers at high and low risk for autism," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 4, pp. 865–876, 2017.

[228] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[229] F. B. Pokorny, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, R. Vollmann, S. Bölte, M. Gugatschka, B. W. Schuller, and P. B. Marschik, "Typical vs. atypical: Combining auditory Gestalt perception and acoustic analysis of early vocalisations in Rett syndrome," *Research in Developmental Disabilities*, vol. 82, pp. 109–119, 2018.

[230] L. G. Portney and M. P. Watkins, *Foundations of Clinical Research: Applications to Practice*. Upper Saddle River, NJ, USA: Prentice Hall, 2000.

[231] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[232] M. Schmitt and B. Schuller, "openXBOW: Introducing the Passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, no. 96, pp. 1–5, 2017.

[233] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 2002.

[234] F. B. Pokorny, F. Graf, F. Pernkopf, and B. W. Schuller, "Detection of negative emotions in speech signals using bags-of-audio-words," in *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*. Xi'an, China: IEEE, September 2015, pp. 879–884.

[235] C. Einspieler, A. M. Kerr, and H. F. Prechtl, "Abnormal general movements in girls with Rett disorder: The first four months of life," *Brain and Development*, vol. 27, pp. S8–S13, 2005.

[236] ——, "Is the early development of girls with Rett disorder really normal?" *Pediatric Research*, vol. 57, no. 5, Part 1, pp. 696–700, 2005.

[237] C. Einspieler, M. Freilinger, and P. B. Marschik, "Behavioural biomarkers of typical Rett syndrome: Moving towards early identification," *Wiener Medizinische Wochenschrift*, vol. 166, no. 11–12, pp. 333–337, 2016.

[238] D. Zhang, L. Poustka, P. B. Marschik, and C. Einspieler, "The onset of hand stereotypies in fragile X syndrome," *Developmental Medicine & Child Neurology*, vol. 60, no. 10, pp. 1060–1061, 2018.

[239] H. F. Prechtl, Ed., *Continuity of neural functions from prenatal to postnatal life*, ser. Clinics in Developmental Medicine, no. 94. Oxford, UK: Blackwell Scientific Publications, 1984.

[240] H. F. Prechtl, "New perspectives in early human development," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 21, no. 5–6, pp. 347–355, 1986.