Comparison of Supervised-Learning Models and Auditory Discrimination of Infant Cries for the Early Detection of Developmental Disorders

Vergleich von Supervised-Learning Klassifikationsmodellen und menschlicher auditiver Diskriminationsfähigkeit zur Unterscheidung von Säuglingsschreien mit kongenitalen Entwicklungsstörungen

Tanja Fuhr^{1*}, Henning Reetz², Carla Wegener¹

¹Fresenius University of applied Science,
School of Therapy & Social Work, 65510 Idstein, Germany
* tanja.fuhr@hs-fresenius.de

²Johann Wolfgang Goethe-University Frankfurt,
Institut für Phonetik, 60054 Frankfurt, Germany

Received 4 August 2018, accepted 26 October 2018

Abstract

Infant cry classification can be performed in two ways: computational classification of cries or auditory discrimination by human listeners. This article compares both approaches.

An auditory listening experiment was performed to examine if various listener groups (naive listeners, parents, nurses/midwives and therapists) were able to distinguish auditorily between healthy and pathological cries as well as to differentiate various pathologies from each other.

Listeners were trained in hearing cries of healthy infants and cries of infants suffering from cleft-lip-and-palate, hearing impairment, laryngomalacia, asphyxia and brain damage. After training, a listening experiment was performed by allocating 18 infant cries to the cry groups.

Multiple supervised-learning classifications models were calculated on the base of the cries' acoustic properties. The accuracy of the models was compared to the accuracy of the human listeners.

With a Kappa value of 0.491, listeners allocated the cries to the healthy and the five pathological groups with moderate performance. With a sensitivity of 0.64 and a specificity of 0.89, listeners were able to identify that a cry is a pathological one with higher confidence than separating between the single pathologies. Generalized linear mixed models found no significant differences between the classification accuracy of the listener groups. Significant differences between the pathological cry types were found.

Supervised-learning classification models performed significantly better than the human listeners in classifying infant cries. The models reached an overall Kappa value of up to 0.837.

Abstract

Die Diskrimination von gesunden und pathologischen Säuglingsschreien kann auf unterschiedliche Weise erfolgen: über computerbasierte Klassifikation von akustischen Parametern des Schreis oder über auditive Diskriminierung durch das menschliche Gehör. In diesem Artikel wird die Leistungsfähigkeit der beiden Verfahren miteinander verglichen.

Ein Hörexperiment mit unterschiedlichen Hörergruppen (Laien, Eltern, Hebammen und Therapeuten) wurde durchgeführt, um zu evaluieren, ob Hörer in der Lage sind, zwischen gesunden und pathologischen Schreien, sowie zwischen verschiedenen Pathologien zu differenzieren.

In einem Hörtraining wurden den Hörern Schreie von gesunden Säuglingen, von Säuglingen mit Lippen-Kiefer-Gaumenspalte, von Säuglingen mit Hörstörungen, von Säuglingen mit Laryngomalazie, sowie von Säuglingen mit Asphyxie und perinataler intracerebraler Blutung auditiv präsentiert. Im Anschluss fand das Hörexperiment durch Zuordnung von 18 Schreien zu den einzelnen Schreigruppen statt.

Verschiedene computerbasierte Klassifikationsalgorithmen (supervised-learning models) wurden ebenfalls auf den Schreien anhand ihrer akustischen Parameter trainiert.

Die Klassifikationsgenauigkeit der Hörergruppen erzielte einen Kappa-Wert von 0,491 (durchschnittliche Leistungsfähigkeit). Mit einer Sensitivität von 0,64 und einer Spezifität von 0,89 waren die Hörer in der Lage zwischen gesunden und pathologischen Schreien zu differenzieren. Verallgemeinerte lineare gemischte Modelle wurden berechnet, um einen möglichen Einfluss der Gruppenzugehörigkeit der Hörer zu ermitteln. Es konnte kein signifikanter Unterschied in der Klassifikationsgenauigkeit zwischen den Hörergruppen ermittelt werden.

Die computerbasierte Klassifikationsgenauigkeit zeigte signifikant bessere Klassifikationsergebnisse mit einem Kappa-Wert von 0,837.

Keywords

Infant cry discrimination – cry classification – human listening – hearing experiment – auditory discrimination skills

Keywords

Säuglingsschrei-Klassifikation – auditive Diskriminationsfähigkeit für Säuglingsschreie – Hörexperiment

INTRODUCTION

Crying is the earliest way of expressing and communicating needs like hunger, pain, discomfort or tiredness. Additionally, a cry is an acoustic signal containing information that provides insights into the medical status of an infant (Fort and Manfredi, 1998; Orlandi et al., 2015). The research field of infant cry analysis is of high interest for practitioners in the field of pediatric nursing, midwifery, pediatrics and therapists working with infants showing sucking or swallowing difficulties caused by developmental disorders. It is of value for the early detection of severe health issues in infants which might have severe negative effects on the development of children if undetected.

Much research about infant cry analysis focuses on exploring the suitability of infant cry for early diagnostic purposes (Reyes-Garcia et al., 2010; Hadad, 2015). Studies examining the acoustic features of healthy infants and those with developmental disorders showed that infants with medical conditions have different cry characteristics than healthy infants (LaGasse et al., 2005). Diseases like brain damage (Sirviö and Michelsson, 1976; Jobbágy, 2012), asphyxia (VerduzcoMendoza et al., 2012; Michelsson et al., 1977), laryngomalacia (Goberman and Robb, 2005), hearing impairment (Verduzco-Mendoza et al., 2012; Möller and Schönweiler, 1999), or cleft-lip-and palate (CLP) (Etz et al., 2012) were found to influence the cry production process, and therefore, cries of infants with diseases show different acoustic features compared to healthy infants. Studies found an increased fundamental frequency (f_0), more dysphoned or hyperphoned parts as well as a deviation in the f_0 variability, compared to a healthy infant cry. In addition, classification models like neural networks, decision trees and others were used to predict an infant's health status (Etz et al., 2012; Reyes-Galaviz et al., 2005; Reyes-Garcia et al., 2010; Galaviz and García, 2005). In these studies, mainly two approaches for determining the characteristics of infant cries have been applied; the acoustic analysis of cry signals and the auditory discrimination of cries by listeners. This article aims at combining the aspects of both approaches and compares the auditory discrimination skills of listeners with the ability of statistical models to discriminate the types of infant cries. In future research, the findings of this study may be valuable for screening the health state of infants by acoustic parameters of their crying.

Automatic classification of infant cries – that is, using approaches based on statistical algorithms for automatically classifying cries – due to their medical condition is a powerful opportunity to rate the health status of an infant by means of their acoustic features. In a systematic supervised-learning model review, Fuhr et al. (2015) showed the classification models which have

been used in the past to classify infant cries and evaluated the applicability of these classification models. Hence, computer-based algorithms were suited to classify infant cries. But considering the performance of machines in the field of language detection, the human abilities were much better than the machine-based abilities on language and speech identification (Norvig, 2012; Luxton, 2016). Studies investigating the association of specific cry characteristics to the cry perception of listeners showed that multiple acoustic parameters are perceived as negative or abnormal to listeners. An increased or more variable fundamental frequency, as well as more dysphoned or hyperphoned parts of a cry are often perceived as distressing, sick, arousing, aversive and urgent (LaGasse et al., 2005; Möller and Schönweiler, 1999; Schuetze et al., 2003). Listeners were also found to be able to differentiate between the cries of infants with perinatal complications and the cries of healthy infants (Zeskind and Lester, 1978). Hearing the cries of infants with asphyxia, down syndrome, cri-du-chat syndrome or autism, the listeners showed differences in their behavior and their reaction on hearing those cries (Frodi and Senchak, 1990; Venuti et al., 2012; Esposito et al., 2012). These studies indicate that acoustic differences between cries of healthy and non-healthy infants can be perceived by human listeners, and therefore, might allow listeners to 'hear the health status' of infants.

Based on this assumption, Möller and Schönweiler (1999) compared the ability of nurses, parents and otolaryngologists to distinguish the cries of healthy infants from the cries of infants with hearing impairment. Here, the nurses reached significantly better results indicating that experience with hearing infant cries might influence the ability to discriminate between infant cries auditorily. Morsbach and Murphy (1979) also described that the nurses reached better results in classifying healthy infants and infants with hearing impairment than naïve listeners or parents because of their daily contact to various healthy and non-healthy infants. A listening experiment comparing the hearing capacity of mothers as well as the hearing capacity of naïve listeners was able to show that the naïve listeners can reach better results in discriminating healthy neonate cries than mothers (Nolten, 1984). In these studies, the participants were not trained in discriminating infant cries before conducting the listening experiment.

Gladding (1979) tested if listeners can be trained to discriminate cries correctly. Subjects with training showed significantly better results in distinguishing various types of crying than subjects without listening training.

Summarizing, a significant amount of research has been conducted to explore the acoustic properties of infant cries and the potential to identify differences in those properties between healthy and non-healthy cries by computational

models and algorithms as well as by human listeners. However, previous researches did not examine sufficiently if human listeners are able to differentiate not only between healthy and non-healthy cries but also between different types of pathologies. In addition, a comparison of the classification skills of computational models in contrast to the skills of human listeners is still missing.

The present article presents a study aimed at analyzing and comparing the ability of human listeners and automatic classification models to rate the health state of infants by their crying. For the listening experiment, naïve listeners (students and parents) and expert listeners (nurses/midwives and therapists) were trained to auditorily discriminate the cries of healthy infants as well as infants with various pathologies (like hearing impairment (HI), cleft-lip-and palate (CLP), asphyxia (AS), laryngomalacia (LA), brain damage (BD), etc.). After training, the listeners rated cries of infants with different health states and their rating skills were compared to the classification skills of computation models. To achieve a deeper understanding of the ability of human listeners and computation models to classify infant cries, the following research questions were elaborated for this study. In addition to analyzing the classification ability directly (fixed factors), analyzing the influence of random factors like the age of the human listeners (as the human hearing performance might change with increasing age) was added to the research questions:

- RQ 1:** Are human listeners able to discriminate auditorily between healthy infant cries and non-healthy infant cries and are they able to differentiate between different pathologies?
- RQ 2:** Are there differences in the discrimination skills between the listener groups?
- RQ 3:** Are there differences in the listeners' rating performance between the types of crying (e.g., healthy, hearing impaired, ...)
- RQ 4:** Do listeners rate infant cries that were used during training more accurately than unknown cries?
- RQ 5:** Do sociodemographic parameters like age influence the rating skills of human listeners?
- RQ 6:** Do human listeners perform more or less accurately in discriminating between infant cries than computational models?

METHOD

Subjects

Participants of the Listening Experiment

A total of 120 participants were included in the listening experiment and divided into the 4 groups: naïve listeners

(group 1), parents (group 2), nurses/midwives (group 3) and therapists (group 4). Based on the following inclusion and exclusion criteria, these groups were chosen to capture listeners with varying experience in hearing infant cries:

- a) Naïve listeners: no experience in hearing infant crying
- b) Parents: frequent long-term contact to a limited, familiar group of healthy infants
- c) Nurses, midwives: frequent short-term contact to many healthy and rare contact to non-healthy infants
- d) Therapists: frequent long-term contact to many non-healthy infants

General inclusion criteria for all groups were following: all participants were female and German and were without hearing impairments. Because almost all participants in the groups nurses/midwives and therapists were female, the small number of male participants was excluded from the study to avoid any statistical error that might occur in an unbalanced study design. The impact of this decision on generalizing the results is discussed in the Discussion Section.

In addition to the general inclusion criteria, the following criteria were defined per group: group 1 contained 30 female naïve listeners without children and without being in close contact to infants. Group 2 consisted of 30 mothers caring for infants being younger than 2 years old. Participants of the first two groups had jobs not related to the health system. Group 3 included 30 midwives and female pediatric nurses. Group 4 contained 30 female therapists. In this group, physical therapists (11 persons), occupational therapists (5 persons) and speech and language pathologists (14 persons) were included. For all midwives, nurses and all therapists, a professional experience of at least 4 years and a frequent contact to infants and young children with developmental diseases were defined as the inclusion criteria.

Table 1 provides sociodemographic parameters for each listener group. These parameters were selected to capture any random effect on the results of the listening experiment. Especially the age parameter could not be controlled by the study design, as persons with no children (naïve listeners) are likely to be significantly younger than persons with children (parents); a balanced study design with a similar distribution of age across the listener groups was therefore not achievable. The parameters 'number of children' and 'professional experience' were included to test, if the results would be influenced by the personal experience of the participants. A non-parametric Kruskal-Wallis test revealed significant differences in the distribution of age between naïve listeners and the remaining groups, as well as the distribution of the number of children between therapists

Table 1: Sociodemographic parameters of the listener groups

Listener group:	Naive listeners		Parents		Nurses/midwives		Therapists	
(N = 120)	(N=30)		(N=30)		(N=30)		(N=30)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age [years]	23,3	2,6	31,4	5,4	32,4	6,4	35,5	7,1
No. Children			1,6	0,7	1,2	1,2	0,8	1,0
Prof. experience [years]					9,5	5,8	10,3	5,8

and parents. No significant difference was found in the professional experience across groups.¹ Balancing the groups for the parameters age and number of children was not possible with the given pool of participants. Because of the definition of the groups, the participants of the naïve group had to be significantly younger than in the other groups as higher age did highly correlate with a higher number of children. Therefore, most participants with children were older, whereas most younger participants had no children. To cope with this variation of the sociodemographic parameters across groups, a correlation analysis was used to analyze the possible effects of the parameters on the test results as described in the Analysis Section.

Procedure

For exploring the human listener's ability to classify infant cries and for comparing their performance to the rating performance of computational models, both, humans and computational models were applied to the same process of training and prediction. Figure 1 visualizes the training phase and the rating phase for human listeners and computational models.

The ability of human listeners to hear the difference between healthy and pathological cries and between different pathologies was trained in a listening training using 18 training cries. After training, the human listeners predicted the health state of infants on 18 unknown cries. The training and prediction for human listeners is described later in this section in more detail.

To train the computational models, various supervised-learning algorithms were trained on the same training cries as the human listeners. Like humans, the supervised-learning algorithms learn patterns by analyzing the training data for which the result is known (here, acoustic parameters of infant cries, for which the health state of the infant is known, are analyzed). After training,

¹ In the case of the number of children, the naive listeners were not included in the Kruskal-Wallis test and for the professional experience, the naive listeners and parents were excluded, as the differences to these groups result from the characteristics of the groups.

the algorithms create supervised-learning models, that represent the knowledge learned during the training phase. These models are then applied to the same 18 unknown cries that were rated by the human listeners and the models predict the health state of infants based on the acoustic parameters of the cries.

In this setting (Setting A), the supervised-learning algorithms use the same training set of cries as the human listeners. This provides the same setting for both, human listeners and computational algorithms, and allows comparing both. However, supervised-learning models were originally designed to be trained on large datasets to avoid fitting the models too exactly to the training data, losing the ability to predict unknown data correctly (overfitting).

For this reason, a second setting was included in the study. In Setting B, the supervised-learning algorithms were trained on larger infant cry dataset to get more general models and to avoid overfitting to the training data. The resulting supervised-learning models were then applied to the same test cries as before. The training and rating for both settings of supervised-learning models is described later in this section in more detail.

Listening Experiment

The listening experiment was divided into a training phase and a rating phase. Participants were first trained in hearing cries of healthy infants and infants with various pathologies. In the rating phase, listeners had to allocate unknown cries to the different groups of health states.

Training Phase

In the training phase of the listening experiment, the participants had to listen to acoustic cry samples of healthy infants and infants with 5 different pathologies. According to Tsukamoto and Tohkura (1990), 2 to 5 cries build a perceptual unit for infant cry categorization. Therefore, three cries of each cry group were randomly selected from the infant cry dataset described in the Material section, summing up to a total of 18 cry samples for the training phase. All participants were trained on the same set of cry samples.

Figure 1: Overview of the training phase and rating phase for the human listeners and for the computational models. From the infant cry database in Setting B, the 18 cries used in the rating phase were excluded.

The training was held with all participants in a quiet room. The participants were told which cry type they would hear next and then the 3 cries of the cry group were played via speakers. The same 3 cries were then repeated, and the participants took notes of what they thought would be characteristic for the cry group. This procedure was repeated for each cry group. After this session, all cry groups were played again, but without repetitions. Figure 2 visualizes the schema of the training.

Altogether, the listeners heard the 3 cries from each of the 6 cry groups 3 times. This approach was used to ensure that the listeners were able to memorize the cry impressions. For the training phase, listeners were asked to take personal notes about their hearing perception of each cry group to support the training effect.

The listening training was mandatory, as the present listening experiment examined not only the listeners' abilities to discriminate healthy and pathological cries, but also their abilities to discriminate various pathologies. Here, it could not be assumed, that the listeners would know how cries of infants with various pathologies sound. In addition, Gladding (1979) showed, that listeners with training reached significantly better results in distinguishing various types of crying than listeners without training.

Rating Phase

In the rating phase, the participants had to listen to 18 cry samples and had to allocate each sample to one of

the six cry groups. From each cry group, 3 cry samples were presented, but the listeners were not told how many samples from each group were in the set.

One of the three cries was a cry sample that had already been used in the listening training. This approach was chosen to determine if cries known from the training phase can be allocated better to the six groups than unknown cries. The remaining two samples were randomly selected from the infant cry dataset described in the Material section. These cries had not been used in the training phase. All participants listened to the same set of cries.

Computational Classification

Infant cry classification aims at finding a computational model that is able to automatically classify infant cries according to their acoustic properties into given categories of cries. Computational models work similar to human listeners rating infant cries: first, the acoustic properties of a cry must be extracted in an acoustic analysis. Second, a computational model must be trained on a training dataset for which the cry categories are known ('supervised-learning') in order to learn how to categorize the cries. Finally, the computational model can be applied to unknown cries to categorize them.

Following the previous comparisons of models for infant cry classification (Fuhr et al., 2015), the following

Figure 2: Schema of the listening experiment

supervised-learning algorithms suited for infant cry classification were selected for the study.

Artificial neural networks encompass different machine learning approaches following functions of animal brains by simulating information flow through systems of interconnected ‘neurons’. In this study, multilayer perceptrons and radial basis function networks were used. **Bayes classifiers** are probabilistic models based on Bayes’ theorem describing classes by statistical processes.

Linear discriminant analyses identify linear functions to separate groups in data.

Support Vector Machines work similar to linear discriminant analysis, except they can be extended for non-linear discrimination between data sets.

Logistic regressions measure the linear relationship between a categorical target variable and multiple predictors using a logistic probability function.

Decision trees cover different algorithms for computing hierarchical decision rules to decide, to which group data items belong. In this study, C 5.0, CHAID, CRT and QUEST decision tree algorithms were used.

Acoustic Analysis

To extract acoustic properties of the infant cries, Praat software (Boersma and Weenink, 2013) and an automated Praat script were used to compute acoustic parameters. The following parameters have proven to be useful for infant cry classification in previous studies of the authors (Etz et al., 2014): the median as well as lower and upper bounds — represented through the 10th and 90th

percentile — of the fundamental frequency and intensity, the first six formants, jitter and shimmer values as well as the relation of phonated and non-phonated parts, number and degree of voice breaks and the cry duration were measured. These acoustic parameters were automatically extracted for each infant cry and were used as input for the training and application of the computational models.

Training Phase

For training the supervised learning models, two different sets of training data were chosen: In *Setting A*, all supervised-learning models were trained on the same set of 18 infant cries that was used for training the human listeners. In *Setting B*, 526 cry samples from the cry database were used for training; the infant cries used in the rating phase were excluded.

The supervised-learning algorithms described above were applied to the training datasets. Each algorithm follows its own strategy for identifying rules to categorize the cries. All algorithms implement techniques to avoid overfitting of the models to the training data and thus, to be able to categorize unknown cries as correctly as possible.

For training the models, IBM’s SPSS Modeler 18.0 was used. During the training phase, the software automatically varies different parameters of the algorithms to find the best settings of the algorithms.

After training, each algorithm creates a supervised-learning model that represents the classification rules for categorizing infant cries.

Rating Phase

After training, each model was applied to the test set of infant cries that was also presented to the human listeners. Here, the models of both training Settings A as well as B were applied to the same set of test cries.

Based on the rules learned during the training phase, all models categorized the cry samples to predict the health state of the infants.

Material

All infant cry samples used in this study were taken from a dataset of infant cries, built up during a research project of the authors on infant cry classification. The dataset is described in the following section.

Subjects

Cry samples of 69 infants between 1 and 7 months of age were included in the dataset. In total, 6 different infant groups were recorded: 31 infants were healthy, without any developmental disorders, 10 infants had an unilateral cleft-lip-and palate (CLP), 19 infants were hearing impaired (HI, threshold of -60dB hearing loss), 4 infants were suffering from laryngomalacia, 3 were asphyxiated infants and 2 infants had brain damage.

For the healthy infants, the following inclusion criteria were defined: All infants had no complications during birth. Their age, birth weight and gestational age were without pathological findings. APGAR scores ('Appearance, Pulse, Grimace, Activity, Respiration', Apgar, 1953) were documented after 1, 5 and 10 minutes. For all infants, the APGAR scores were 9/10/10. The infants were found to be healthy by pediatricians at postpartum examination. No indication of neurological diseases or further anomalies or any diagnosis that might influence normal development could be found. The hearing function of all infants was assessed for both ears by otoacoustic emissions. No limitation of the hearing function was found. Pediatricians confirmed that there wasn't any indication of an infant's existing cold at the time of recording.

For the infants suffering from developmental disorders, no further anomaly or diseases could be found by pediatricians, except the diagnosed developmental disorder.

All parents of the infants were native speakers of German and gave written informed consent to participate in this study. The study was approved by the Ethic Review Committee of the Fresenius University of Applied Sciences.

Cry recording

The cries of the infants were recorded with a sampling rate of 48 kHz and 24-bit digital resolution on a Zoom

H2n recorder. The Zoom H2n recorder features a built-in microphone. The microphone was held about 30 cm away from the infants' mouths. The infants lay in a supine position during the recording. Recordings were made in similar environments.

One full episode of crying was recorded for each infant. Recordings started with the first cry of the infant (using the H2n's pre-recording function). Recordings were stopped when there was a 15 second pause with no crying. One recording lasted about 10 to 30 seconds.

For acoustic analysis, single cries were extracted from the episodes of crying. Altogether 544 single cry utterances could be extracted from the cry recordings. To guarantee a sufficient quality of the recordings, only recordings with more than 30 dB intensity between the minimum intensity within an episode of crying (corresponding to the noise level) and the maximum intensity within the episode were included in the study. No recordings had to be excluded from this study.

For training and testing the human listeners and the computational models, subsets of cry signals were extracted from the dataset as described in the following subsections.

Analysis

To answer the research questions described in the Introduction section, various statistical methods were used. They are described in the following.

Covariate Analysis

First, a possible influence on the rating performance of the sociodemographic parameters age, number of children and professional experience was analyzed using a correlation analysis between these parameters and the rating correctness (RQ 5). Non-parametric Spearman's rho correlation was computed as the sociodemographic parameters were not normally distributed. As no significant correlations between the parameters and the rating performance were found (c.f. Section 3.3.1), the sociodemographic parameters were not included in any further statistical analyses.

Descriptive Statistics

To analyze the rating performance of the human listeners in the listening experiment (RQ 1 and RQ 3), a confusion matrix was computed to compare the listeners' ratings and the actual cry types.

The following quality coefficients were computed on the confusion matrix to quantify the listeners' performances in discriminating between healthy and pathological cries as well as between the various pathologies:

Table 2: Correlation analysis to analyze the influence of the sociodemographic covariates on the rating correctness

		Age	No. children	Prof. experience	Correctness
Age	Correlation Coefficient ^a	1,000	,633**	,599**	-0,024
	Sig. (2-tailed)		0,000	0,000	0,256
No. children	Correlation Coefficient ^a	,633**	1,000	,205**	0,010
	Sig. (2-tailed)	0,000		0,000	0,632
Prof. experience	Correlation Coefficient ^a	,599**	,205**	1,000	0,019
	Sig. (2-tailed)	0,000	0,000		0,386
Correctness	Correlation Coefficient ^a	-0,024	0,010	0,019	1,000
	Sig. (2-tailed)	0,256	0,632	0,386	

** Correlation is significant at the 0.01 level (2-tailed).

^a Spearman's rho

- *Cohen's kappa coefficient* (κ) was computed to quantify the overall agreement of listener ratings with the actual cry types. In contrast to simple percentage agreement, κ takes into account any agreement occurring by chance.
- *Sensitivity* of the healthy group was computed to rate the listener's ability to identify healthy infant cries correctly.
- *Specificity* of the healthy group was computed to rate the listeners' ability to identify cries with one of the pathologies as not healthy (excluding the ability to differentiate between the various pathologies).

Analysis of Variances for Human Listeners

To analyze the influence of various factors on the classification performance of human listeners and to identify effects between these factors, a Generalized Linear Mixed Model (GLMM) was computed. GLMMs allow to analyze the influence of multiple fixed and random effects as well as effects of their interaction on one target variable that may have any scale or distribution. In this analysis, the correctness of the cry ratings was chosen as a binomial-scaled (0 = wrong rating, 1 = correct rating) target variable. The GLMM was parameterized to use a binomial probability distribution and a Logit link function.

The following nominal variables were included as fixed factors:

The *listener group* was included to test if listeners of any group perform better in infant cry classification than the listeners of other groups (RQ 2).

The *cry type* was included to test if any type of crying (e.g., cries of healthy infants or cries of infants with hearing impairment) was identified more precisely than the other types (RQ 3).

The *knowledge about cries* was included to test if cries that were presented during the training phase are rated more precisely than unknown cries (RQ 4).

To cope with possible differences in the rating performance between single listeners or between single cry samples, these two variables were added as random factors.

For exploring significant differences in more detail, pairwise comparisons with Bonferroni correction were conducted.

Analysis of Variances between Human Listeners and Computer Models

For comparing the rating performance of human listeners and computer models (RQ 6), all ratings of human listeners and computer models were included in an additional analysis of variances. The groups 'Human listeners', 'Models, Setting A' and 'Models, Setting B' were analyzed with the same statistics that were used for exploring variances between the human listener groups.

RESULTS

The statistical methods described in Section 2.4 were computed using IBM's SPSS Statistics 23.0 (IBM, 2016). The results are described in the following subsections.

Covariate Analysis

Table 2 provides the results of the correlation analysis that explores the impact of the covariates *age*, *number of children* and *professional experience* on the rating performance of the human listeners. None of these covariates correlate significantly with the rating correctness.

Table 3: Confusion matrix of the ratings of the participants in the listening experiment

		Listener rating						
		HE	CLP	HI	LA	AS	BD	Total
Real cry type	HE	230	48	55	2	10	15	360
		63,9%	13,3%	15,3%	0,6%	2,8%	4,2%	
	CLP	116	108	74	10	23	29	360
		32,2%	30,0%	20,6%	2,8%	6,4%	8,1%	
	HI	49	73	172	12	9	45	360
		13,6%	20,3%	47,8%	3,3%	2,5%	12,5%	
	LA	4	12	22	308	3	11	360
		1,1%	3,3%	6,1%	85,6%	0,8%	3,1%	
	AS	4	29	14	6	259	48	360
		1,1%	8,1%	3,9%	1,7%	71,9%	13,3%	
	BD	31	58	30	13	61	167	360
		8,6%	16,1%	8,3%	3,6%	16,9%	46,4%	
Total		434	328	367	351	365	315	2160

Table 4: Kappa statistics for the listener groups and for all listeners

Listener group	Kappa Value	Asymptotic Standardized Error ^a	Approximate T ^b	Approximate Significance
Nurses	0,520	0,025	27,092	0,000
Naive listeners	0,498	0,025	25,884	0,000
Parents	0,471	0,026	24,514	0,000
Therapists	0,476	0,026	24,728	0,000
Total	0,491	0,013	51,095	0,000

^a Not assuming the null hypothesis.^b Using the asymptotic standard error assuming the null hypothesis.

Table 5: Sensitivity and specificity values of the human listeners for identifying healthy infants

Listener group	Sensitivity	Specificity
Nurses	0,70	0,88
Naive listeners	0,59	0,89
Parents	0,69	0,89
Therapists	0,58	0,88
Total	0,64	0,89

The ability to differentiate between healthy and non-healthy cries (RQ 1) was quantified by computing the sensitivity and specificity for rating healthy infant cries. Table 5 shows the sensitivity and specificity values for the listener groups. The sensitivity value represents the listeners' ability to identify healthy infants correctly as healthy. The specificity value represents their ability to identify infants with one of the pathologies as non-healthy.

Descriptive Statistics for the Classification Models

The classification performance of the models on the test dataset is shown in Table 6. The models in Setting A were trained on the same 18 cries that were used for training the human listeners. The models in Setting B were trained on the complete dataset described in Section 2.1.

Table 7 presents the Kappa values for the computer models of Setting A and Setting B, representing their overall ability to classify infant cries correctly.

Descriptive Statistics for the Human Listeners

To describe the rating performance of the listeners, the confusion matrix presented in Table 3 was computed. The overall ability of the listeners to correctly rate cries of healthy infants and infants with various pathologies (RQ 1) was computed on the confusion matrix and is represented by the Kappa values shown in Table 4.

Table 6: Confusion matrix presenting the classifications of the supervised-learning models for the training Settings A and B compared to the actual cry types

			Rating						
			HE	CLP	HI	LA	AS	BD	Total
Models, Setting A	RealCryType	HE	23	0	2	2	0	0	27
			85,2%	0,0%	7,4%	7,4%	0,0%	0,0%	100,0%
		CLP	0	26	0	0	0	1	27
			0,0%	96,3%	0,0%	0,0%	0,0%	3,7%	100,0%
		HI	2	1	23	0	0	1	27
			7,4%	3,7%	85,2%	0,0%	0,0%	3,7%	100,0%
		LA	1	0	0	26	0	0	27
			3,7%	0,0%	0,0%	96,3%	0,0%	0,0%	100,0%
		AS	3	1	0	1	19	3	27
			11,1%	3,7%	0,0%	3,7%	70,4%	11,1%	100,0%
		BD	2	1	1	0	0	23	27
			7,4%	3,7%	3,7%	0,0%	0,0%	85,2%	100,0%
		Total	31	29	26	29	19	28	162
Models, Setting B	RealCryType	HE	18	5	4	0	0	0	27
			66,7%	18,5%	14,8%	0,0%	0,0%	0,0%	100,0%
		CLP	4	21	2	0	0	0	27
			14,8%	77,8%	7,4%	0,0%	0,0%	0,0%	100,0%
		HI	3	4	20	0	0	0	27
			11,1%	14,8%	74,1%	0,0%	0,0%	0,0%	100,0%
		LA	7	0	6	14	0	0	27
			25,9%	0,0%	22,2%	51,9%	0,0%	0,0%	100,0%
		AS	1	1	1	1	22	1	27
			3,7%	3,7%	3,7%	3,7%	81,5%	3,7%	100,0%
		BD	0	1	0	0	0	26	27
			0,0%	3,7%	0,0%	0,0%	0,0%	96,3%	100,0%
		Total	33	32	33	15	22	27	162

Table 8 shows the sensitivity and specificity values representing the models' ability to identify healthy infants as healthy and infants with any of the pathologies as non-healthy.

Analysis of Variances for Human Listeners

Computing the Generalized Linear Mixed Model (GLMM) in SPSS resulted in a model with an accuracy value of 69.5%, that is, almost 70% of the variance in the data can be explained by this model.

Table 9 shows the impact of the fixed factors on the rating correctness. The rating correctness does not vary significantly at the $p = 0.05$ level across the listener groups. However, the real cry type (e.g., healthy or CLP cries) and the knowledge about the cry (i.e., if the cry was known because it was already used in the training phase,

or if it was not known) had a significant impact on the rating performance.

To explore the significant fixed factors *RealCryType* and *TestOrUnknownCry* in more detail, pairwise comparisons were computed.

Table 10 shows the pairwise comparisons for the *RealCryType* factor.²

Table 11 shows the contrast between the known cries and the unknown cries. Known cries were rated significantly better than unknown cries, but the effect size is with -0.063 not very large, that is, the rating performance was not that much better for known cries.

Random effect covariances were evaluated to estimate the influence of between-listeners variance and between-cry-samples variance.

² Symmetric contrasts were removed from the table

Table 7: Kappa statistics for the models of Settings A and B

Model Group	Kappa Value	Asymptotic Standard Error ^a	Approximate T ^b	Approximate Significance
Models, Setting A	0,837	0,032	23,872	0,000
Models, Setting B	0,696	0,041	19,939	0,000
Total	0,527	0,012	58,796	0,000

^a Not assuming the null hypothesis.^b Using the asymptotic standard error assuming the null hypothesis.

Table 8: Sensitivity and specificity values of the classification models for identifying healthy infants

Rater Group	Sensitivity	Specificity
Models, Setting A	0,85	0,94
Models, Setting B	0,67	0,90
Total	0,76	0,92

Table 9: Fixed effects impact on the rating correctness

Source	F	df1	df2	Sig.
Corrected Model	30,056	9	2150	0,000
ListenerGroup	0,637	3	2150	0,591
RealCryType	53,099	5	2150	0,000
TestCries	7,078	1	2150	0,008

Probability distribution: Binomial

Link function: Logit

Table 12 shows the random effect covariances. The between-listeners variance as well as the between-cry-samples variance are not significant. Therefore, both have no significant impact on the variance in the data.

Analysis of Variances between Human Listeners and Computer Models

The GLMM model for analyzing the impact of the group factors on the classification correctness reached an overall accuracy of 71.0 %, i.e., 71 percent of the variability in the data can be explained by the model.

The effects of the fixed factors *RaterGroup*, *RealCryType* and *TestOrUnknownCry* are presented in Table 13. All three effects are significant at the $p=0.05$ level.

Table 14 shows the pairwise contrasts of the *RaterGroup* factor. All pairwise contrasts are significant at the $p=0.05$ level. Human listeners are 29 % less precise in rating infant cries than models trained in Setting A, and they are 18 % less precise than models trained in Setting B. Comparing the models trained in the Settings A and

B, models from Setting A are 11 % more precise than models from Setting B.

Table 15 shows the pairwise contrasts of the *RealCryType* factor across all rater groups³.

Table 16 shows the simple contrasts between the rating of unknown cries (UKN cries) and known cries (KN cries). Known cries are rated slightly but significantly better than unknown cries.

Table 17 shows the random effect on the rating performance. The between-rater variance is significant at the $p=0.05$ level.

DISCUSSION

The discussion section is split into two parts: the interpretation of the results is presented first, applying the statistical results to answer the research questions and to interpret the findings and their implications. Hereafter, the approach is compared to the results of previous studies and discussed.

Interpretation of the Results

Are human listeners able to discriminate auditorily between healthy infant cries and non-healthy infant cries (RQ 1) and are they able to differentiate between the different pathologies?

The confusion matrix for the human listeners' ratings (Table 3) provides an overview of the overall rating performance of human listeners. Laryngomalacia cries are identified quite reliably (85.6%). Asphyxia cries and healthy cries also show a good rating accuracy with 71.9% and 63.9%. Although the remaining cry types are rated with lower accuracy, all cry types are identified more accurately than by chance (accuracy by chance is 16.67%, assuming equal chance across all cry types). Hence, training human listeners to hear the health state of an infant seems to be possible. In addition, the performance of identifying healthy infants and distinguishing between various pathologies is better than by chance.

³ Symmetric contrasts were removed from the table

Table 10: Pairwise contrasts of the real cry type groups

Pairwise Contrasts	Contrast Estimate	Std. Error	t	df	Adj. Sig.	95% Confidence Interval	
						Lower	Upper
HE - CLP	0,344	0,035	9,774	2150	0,000	0,241	0,447
HE - HI	0,162	0,037	4,442	2150	0,000	0,075	0,250
HE - BD	0,176	0,037	4,829	2150	0,000	0,080	0,273
HI - CLP	0,182	0,036	5,010	2150	0,000	0,084	0,279
HI - BD	0,014	0,038	0,376	2150	0,707	-0,060	0,088
LA - HE	0,212	0,031	6,860	2150	0,000	0,126	0,298
LA - CLP	0,556	0,031	18,231	2150	0,000	0,467	0,645
LA - HI	0,375	0,032	11,621	2150	0,000	0,282	0,467
LA - AS	0,132	0,029	4,516	2150	0,000	0,059	0,206
LA - BD	0,389	0,032	12,063	2150	0,000	0,297	0,480
AS - HE	0,080	0,034	2,337	2150	0,039	0,003	0,157
AS - CLP	0,424	0,034	12,516	2150	0,000	0,326	0,522
AS - HI	0,242	0,035	6,859	2150	0,000	0,144	0,340
AS - BD	0,256	0,035	7,261	2150	0,000	0,157	0,355
BD - CLP	0,168	0,036	4,622	2150	0,000	0,074	0,261

The sequential Bonferroni adjusted significance level is .05.
Confidence interval bounds are approximate.

Table 11: Simple contrast of the known and unknown cries

Simple Contrasts	Contrast Estimate	Std. Error	t	df	Adj. Sig.	95% Confidence Interval	
						Lower	Upper
UKN cries - KN cries	-0,066	0,024	-2,692	2150	0,007	-0,113	-0,018

The sequential Bonferroni adjusted significance level is .05.
Confidence interval bounds are approximate.

Table 12: Random effect covariances

Random Effect Covariance	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower	Upper
ListenerGroup * ProbandID	0,101	0,064	1,566	0,117	0,029	0,352
RealCryType * TestCries * ListenerGroup * ProbandID * CryNumber	0,030	0,076	0,390	0,696	0,000	4,487

Covariance Structure: Variance components

Table 13: Fixed effects impact on the rating correctness of computer models and human listeners

Source	F	df1	df2	Sig.
Corrected Model	34,340	8	2475	0,000
RaterGroup	26,660	2	2475	0,000
RealCryType	45,894	5	2475	0,000
TestOrUnknownCry	11,497	1	2475	0,001

Probability distribution: Binomial
Link function: Logit

Table 14: Pairwise contrasts of the *RaterGroup* factor

Pairwise Contrasts	Contrast Estimate	Std. Error	t	df	Adj. Sig.	95% Confidence Interval	
						Lower	Upper
Humans - Models, Setting A	-0,290	0,028	-10,179	2475	0,000	-0,358	-0,222
Humans - Models, Setting B	-0,181	0,039	-4,613	2475	8,331E-06	-0,269	-0,093
Models, Setting A - Models, Setting B	0,109	0,045	2,434	2475	0,015	0,021	0,196

The sequential Bonferroni adjusted significance level is .05.
Confidence interval bounds are approximate.

Table 15: Pairwise contrasts of the *RealCryType* factor

Pairwise Contrasts	Contrast Estimate	Std. Error	t	df	Adj. Sig.	95% Confidence Interval	
						Lower	Upper
HE - CLP	0,253	0,032	7,993	2475	0,000	0,163	0,343
HE - HI	0,108	0,027	3,947	2475	0,000	0,038	0,179
HE - BD	0,106	0,027	3,880	2475	0,000	0,038	0,174
HI - CLP	0,145	0,034	4,314	2475	0,000	0,056	0,234
HI - BD	-0,002	0,030	-0,071	2475	0,943	-0,061	0,057
LA - HE	0,116	0,021	5,510	2475	0,000	0,059	0,174
LA - CLP	0,370	0,032	11,420	2475	0,000	0,275	0,465
LA - HI	0,225	0,027	8,202	2475	0,000	0,145	0,304
LA - AS	0,069	0,018	3,832	2475	0,000	0,025	0,112
LA - BD	0,223	0,027	8,155	2475	0,000	0,144	0,301
AS - HE	0,048	0,022	2,185	2475	0,058	-0,001	0,097
AS - CLP	0,301	0,031	9,571	2475	0,000	0,209	0,393
AS - HI	0,156	0,027	5,810	2475	0,000	0,081	0,231
AS - BD	0,154	0,027	5,750	2475	0,000	0,080	0,228
BD - CLP	0,147	0,034	4,384	2475	0,000	0,057	0,237

The sequential Bonferroni adjusted significance level is .05.
Confidence interval bounds are approximate.

Table 16: Simple contrast for the *TestOrUnknownCry* factor across all groups (KN=known cries, UKN=unknown cries).

Simple Contrasts	Contrast Estimate	Std. Error	t	df	Adj. Sig.	95% Confidence Interval	
						Lower	Upper
UKN cries - KN cries	-0,055	0,016	-3,433	2475	0,001	-0,087	-0,024

The sequential Bonferroni adjusted significance level is .05.
Confidence interval bounds are approximate.

Table 17: Random effect covariances

Random Effect Covariance	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower	Upper
RaterGroup * RaterID	0,098	0,045	2,197	0,028	0,040	0,240
RaterGroup * RaterID * RealCryType * TestCries	1,619E-19 ^a					
Covariance Structure: Unknown						
a. This parameter is redundant.						

Cohen's Kappa values as an overall value (Table 4), as the rating performance of human listeners are similar in all listener groups with an overall average value of 0.491. Following Landis and Koch (1977), this Kappa value can be interpreted as medium accuracy, backing the interpretation of the confusion matrix that human listeners have an average performance in identifying healthy infants and infants with various pathologies.

The sensitivity and specificity (Table 5) for identifying healthy infants as indicators for the listeners' performance to distinguish between healthy and non-healthy is similar between the listeners' groups too. The sensitivity value of 0.64 indicates a medium performance in identifying healthy infants as healthy. The specificity value of 0.89 shows that non-healthy infants are identified with high confidence. This observation backs the studies of Bisping (1986), who suspected that humans have the genetic ability to identify pathological states of health.

Summarizing, humans are well able to identify non-healthy infants by their cry. When distinguishing between various pathologies, the performance of humans is only average, but higher than by chance. As clinical implication, our findings suggest that auditory discrimination of human listeners is not reliable enough to be used in clinical applications such as screening approaches. Human infant cry classification can only give first hints on any non-normal infant development and must be examined by more reliable methods later on.

Are there differences in the discrimination skills between the listener groups (RQ 2)?

Analyzing the variances between the listener groups using GLMM showed no significant variances between listener groups. Therefore, the amount of contact of humans to infants with pathologies does not seem to influence the listeners' rating performance. These results contrast the study of Möller and Schönweiler (1999), who found a significant difference in the rating performance of parents and nurses when rating healthy infants and infants with hearing impairment. Although identifying significant differences, this study had only a small effect size in the variances.

Summarizing, previous studies as well as this study did not find differences with large effects between human listeners experienced in infant crying and unexperienced ones. Therefore, our suggestion to not rely on auditory discrimination of infant cries by human listeners applies to all groups of listeners, whether they work with healthy and non-healthy infants on a regular basis or not.

Are there differences in the rating performance between the types of crying (RQ 3), for example, healthy, hearing impaired, ...)?

There are significant differences in the classification correctness across the different cry types. Evaluating the significant contrasts in Table 10, the following statements about the cry types can be made:

1. Cleft lip and palate cries are rated less accurately than the other cry groups.
2. Healthy cries are rated more accurately than CLP, HI and BD cries.
3. Hearing impaired cries are rated more accurately than CLP cries.
4. Laryngomalacia cries are rated more accurately than HE, CLP, HI, AS and BD cries.
5. Asphyxiated cries are rated more accurately than HE, CLP, HI and BD cries.
6. Brain damage cries are rated more accurately than CLP cries.

Cleft lip and palate disorders seem to have fewer auditory cry characteristics that are recognizable by humans than the other cry groups. Deformations in the orofacial tract do not seem to affect the cry signal very much, so an auditory identification is complicated.

Cries of infants suffering from laryngomalacia are rated most accurately. Cries of infants suffering from laryngomalacia are mostly high pitched, showing a lot of variation in the fundamental frequency and showing high intensity. These characteristics and the direct pathological impact of laryngomalacia on the vocal folds and the larynx, seem to result in auditory characteristics of the cries, that are well recognizable by humans.

Summarizing, some pathologies, like laryngomalacia, show high aberrations in the acoustic parameters from physiology of infant crying. For these pathologies, human listeners are able to identify that an infant is not healthy with a higher sensitivity than for other pathologies. For these special pathologies, human auditory discrimination may give first hints on a pathological development in screening processes.

Do listeners rate infant cries that were used during training more accurately than unknown cries (RQ 4)?

Cry samples that were known to human listeners from the training phase were rated significantly better during the rating phase (Table 11). Although, the effect size is with 0.066 (i.e., known cries are rated by 6.7% more accurately than unknown cries) not very high, so human listeners seem to mainly learn the characteristics of the cry groups during training, instead only recognizing certain cry samples they have already heard. For clinical applications, listening trainings therefore seem to be an adequate way of teaching human listeners about other acoustic characteristics in infant cries not related to classifying healthy and non-healthy infants of.

Do sociodemographic parameters like age influence the rating skills of human listeners (RQ 5)?

The correlation analysis (Table 2) did not show any significant correlations between the rating correctness and the sociodemographic parameters' age, number of children and professional experience. Therefore, these parameters do not seem to influence the rating skills of the listeners.

However, the age of the listeners strongly correlates with the number of children and the professional experience, which is somehow expected, as with higher age, it is likely to have one or more children and it is more likely to have a higher professional experience.

Do human listeners perform more or less accurately in discriminating between infant cries than computational models (RQ 6)?

The computational models trained in Setting A as well as those trained in Setting B perform significantly better than the human listeners at the $p = 0.05$ level (Table 14). The confusion matrix in Table 6 for the computational models presents correctness values for the various cry types between 70 and 100% for models trained in Setting A, and between 51 and 100 % for models trained in Setting B.

Kappa values of 0.696 for models trained in Setting B and 0.837 for models trained in Setting A stand for a

substantial agreement between the classification and the actual health state of the infants.

The sensitivity and specificity values are above those of the human listeners, too. However, the specificity values of the classification models are only 0.03 points higher than those of the human listeners. Therefore, human listeners can identify pathological infant cries with a confidence similar to the models.

As for the human listeners, there are significant differences in the classification performance for the different cry types (Table 15). The interpretation of these contrasts is similar to the interpretation for the human listeners. Hence, characteristic acoustic properties, that are relevant for the human listeners when classifying infant cries, seem to be relevant for the computational models, too.

Summarizing, computational models rate healthy infant cries and cries with various pathologies significantly better than human listeners. However, the rating performance in identifying pathological cries in general, is very similar between humans and computer models. For clinical applications, we therefore suggest using computational models instead of human auditory discrimination for reliably rating the health states of infants by acoustic parameters of their crying.

Comparison of the Approach to Previous Studies

Previous studies described that persons with frequent contact to healthy and ill infants perform better in identifying infant cries than persons without daily contact to infants (naïve listeners) or persons having only close contact to one or two infants (parents) (Möller and Schönweiler, 1999; Morsbach and Murphy, 1979). In contrast, this study showed no differences between the listener groups. Here, the listening training seems to be an effective approach to train listeners in classifying infant cries. After a hearing training, experience in listening to infant cries has no impact on the rating accuracy.

In contrast to other studies (Möller and Schönweiler, 1999; Morsbach and Murphy, 1979), which examined how listeners perform in distinguishing between cries of healthy infants and infants with one pathology, this study examined if it is possible to distinguish between various pathologies. Here, a listening training is essential to ensure that listeners can learn to recognize acoustic properties specific for the various pathologies and thus, enable the listeners to distinguish pathologies auditorily. The study could show that computational classification of infant cries reached better results and is more suitable for identifying pathologies by the cries than auditory discrimination by human listeners. Although listeners perform well in identifying cries as pathological, distinguishing between various pathologies seems to be very difficult and leads to bad classification results.

CONCLUSIONS

The study showed that listeners were not able to identify various pathologies with a high accuracy by hearing infants' cry. Especially, distinguishing between different pathologies by hearing was not a reliable method in this study. However, human listeners performed better when deciding if cries were healthy or not healthy (without regards to the specific type of pathology). The highest accuracy in rating infant cries was achieved by computational supervised-learning models. These were able to rate healthy and non-healthy cries and were able to differentiate various pathologies with higher accuracy.

References

- Andersen, N. (1974). On the Calculation of Filter Coefficients for Maximum Entropy Spectral Analysis. *Geophysics*, 39(1):69–72.
- Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Current researches in anesthesia & analgesia*, 32(4):260–267.
- Barr, R. G., Hopkins, B., and Green, J. A., editors (2000). *Crying as a sign, a symptom, & a signal: Clinical, emotional, and developmental aspects of infant and toddler crying*, volume 152 of *Clinics in developmental medicine*. Mac Keith Press and Cambridge University Press, London, 1 edition.
- Bisping, R. (1986). *Der Schrei des Neugeborenen: Struktur und Wirkung*, volume 22 of *Lehr- und Forschungstexte Psychologie*. Springer-Verlag, Berlin, Heidelberg.
- Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences, Amsterdam*, volume 17, pages 97–110.
- Boersma, P. (2009). Should Jitter Be Measured by Peak Picking or by Waveform Matching? *Folia Phoniatica et Logopaedica*, 61(5):305–308.
- Boersma, P. and Weenink, D. (2013). *Praat: doing phonetics by computer: Manual*. Childers, D. G., editor (1978). *Modern spectrum analysis*. IEEE Press, New York.
- Crowe, H. P. and Zeskind, P. S. (1992). Psychophysiological and perceptual responses to infant cries varying in pitch: comparison of adults with low and high scores on the Child Abuse Potential Inventory. *Child abuse & neglect*, 16(1):19–29.
- Esposito, G., Nakazawa, J., Venuti, P., and Bornstein, M. H. (2012). Perceptions of distress in young children with autism compared to typically developing children: a cultural comparison between Japan and Italy. *Research in developmental disabilities*, 33(4):1059–1067.
- Etz, T., Reetz, H., and Wegener, C. (2012). A classification model for infant cries with hearing impairment and unilateral cleft lip and palate. *Folia Phoniatica et Logopaedica*, 64(5):254–261.
- For using the infant cry as screening instrument, human hearing can only give first hints to an existing pathology. For developing a reliable screening-instrument, supervised-learning algorithms are the selection of choice.
- ## ETHICAL APPROVAL, REGISTRATION
- The study was part of a scientific project which was approved en bloc by the Ethic Committee of the Fresenius University of Applied Sciences.
- ## CONFLICT OF INTERESTS
- None.
- Etz, T., Reetz, H., Wegener, C., and Bahlmann, F. (2014). Infant cry reliability: Acoustic homogeneity of spontaneous cries and pain-induced cries. *Speech Communication*, 58:91–100.
- Fort, A. and Manfredi, C. (1998). Acoustic analysis of newborn infant cry signals. *Medical engineering & physics*, 20(6):432–442.
- Frodi, A. and Senchak, M. (1990). Verbal and behavioral responsiveness to the cries of atypical infants. *Child development*, 61(1):76–84.
- Fuhr, T., Reetz, H., and Wegener, C. (2015). Comparison of Supervised-learning Models for Infant Cry Classification. *International Journal of Health Professions*, 2(1):4–15.
- Galaviz, O. F. R. and García, C. A. R. (2005). Infant Cry Classification to Identify Hypo Acoustics and Asphyxia Comparing an Evolutionary-Neural System with a Neural Network System. In Gelbukh, A., de Albornoz, Á., and TerashimaMarín, H., editors, *MICAI 2005: Advances in Artificial Intelligence*, volume 3789 of *Lecture Notes in Computer Science*, pages 949–958. Springer Berlin Heidelberg.
- Gladding, S. T. (1979). Effects of training versus non-training in identification of infant cry-signals: a longitudinal study. *Perceptual and motor skills*, 48(3 Pt 1):752–754.
- Goberman, A. M. and Robb, M. P. (2005). Acoustic characteristics of crying in infantile laryngomalacia. *Logopedics, phoniatrics, vocology*, 30(2):79–84.
- Hadad, A. (2015). VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014, volume 49 of *IFMBE Proceedings*. Springer-Verlag, s.l.
- IBM (2016). *SPSS Statistics 23.0*.
- Jobbágy, Á. (2012). 5th European Conference of the International Federation for Medical and Biological Engineering: 14-18 September 2011, Budapest, Hungary, volume 37 of *IFMBE Proceedings*. Springer-Verlag GmbH Berlin Heidelberg, Berlin, Heidelberg.
- LaGasse, L. L., Neal, A. R., and Lester, B. M. (2005). Assessment of infant cry: acoustic cry analysis and parental perception.

- Mental retardation and developmental disabilities research reviews, 11(1):83–93.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Luxton, D. D. (2016). *Artificial Intelligence in Behavioral and Mental Health Care*. Elsevier Reference Monographs, s.l., 1. edition.
- Michelsson, K., Eklund, K., Leppänen, P., and Lyytinen, H. (2002). Cry Characteristics of 172 Healthy 1- to 7-Day-Old Infants. *Folia Phoniatica et Logopaedica*, 54(4):190–200.
- Michelsson, K., Sirviö, P., and Wasz-Höckert, O. (1977). Pain cry in full-term asphyxiated newborn infants correlated with late findings. *Acta paediatrica Scandinavica*, 66(5):611–616.
- Möller, S. and Schönweiler, R. (1999). Analysis of infant cries for the early detection of hearing impairment. *Speech Communication*, 28(3):175–193.
- Morsbach, G. and Murphy, M. C. (1979). Recognition of individual neonates' cries by experienced and inexperienced adults. *Journal of Child Language*, 6(01):175–179.
- Nolten, G. (1984). Discrimination of neonate cries by mothers, non-mothers and computer analysis.
- Norvig, P. (2012). *Artificial intelligence: Everyday AI*. *New Scientist*, 216(2889):iv–v.
- Orlandi, S., Reyes Garcia, C. A., Bandini, A., Donzelli, G., and Manfredi, C. (2015). Application of Pattern Recognition Techniques to the Classification of Full-Term and Preterm Infant Cry. *Journal of voice : official journal of the Voice Foundation*.
- Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (2002). *Numerical recipes in C: The art of scientific computing*. Cambridge University Press, Cambridge, 2 edition.
- Reyes-Galaviz, O. F., Verduzco, A., Arch-Tirado, E., and Reyes-García, C. A. (2005). Analysis of an infant cry recognizer for the early identification of pathologies. In Chollet, G., Esposito, A., Faundez-Zanuy, M., and Marinaro, M., editors, *Nonlinear Speech Modeling and Applications*, pages 404–409. Springer-Verlag.
- Reyes-García, C. A., Reyes-Galaviz, O. F., Cano-Ortiz, S. D., Escobedo-Becerro, D., Zatarain, R., and Barrón-Estrada, L. (2010). Soft Computing Approaches to the Problem of Infant Cry Classification with Diagnostic Purposes. In Melin, P., Kacprzyk, J., and Pedrycz, W., editors, *Soft Computing for Recognition Based on Biometrics*, volume 312 of *Studies in Computational Intelligence*, pages 3–18. Springer Berlin Heidelberg.
- Schuetze, P., Zeskind, P. S., and Eiden, R. D. (2003). The Perceptions of Infant Distress Signals Varying in Pitch by Cocaine-Using Mothers. *Infancy*, 4(1):65–83.
- Sirviö, P. and Michelsson, K. (1976). Sound-spectrographic cry analysis of normal and abnormal newborn infants: A review and a recommendation for standardization of the cry characteristics. *Folia phoniatica*, 28(3):161–173.
- Tsukamoto, T. and Tohkura, Y. (1990). Perceptual units of the infant cry. *Early Child Development and Care*, 65(1):167–178.
- Venuti, P., Caria, A., Esposito, G., de Pisapia, N., Bornstein, M. H., and de Falco, S. (2012). Differential brain responses to cries of infants with autistic disorder and typical development: an fMRI study. *Research in developmental disabilities*, 33(6):2255–2264.
- Verduzco-Mendoza, A., Arch-Tirado, E., Reyes-García, C. A., Leybon-Ibarra, J., and Licona-Bonilla, J. (2012). Spectrographic cry analysis in newborns with profound hearing loss and perinatal high-risk newborns. *Cirugia y cirujanos*, 80(1):3–10.
- Zeskind, P. S. and Lester, B. M. (1978). Acoustic features and auditory perceptions of the cries of newborns with prenatal and perinatal complications. *Child development*, 49(3):580–589.