Visualisierung kontinuierlicher, multimodaler Schmerz Scores am Beispiel akustischer Signale

Masterarbeit

Franz Anders
HTWK Leipzig

Januar 2017

Abstract

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

2 Grundlagen der Schmerzbewertung mit Hilfe akustischer Signale

Das Ziel dieses Kapitels ist es, wichtig Grundlagen zum Verständnis der Schmerzbewertung bei Neugeberonenen auf Basis akustischer Signale zu legen. Dazu wird in Kapitel 2.1 zunächst Erläutert, wie die Schmerzbewertung aus Sicht medizinischer Fachkräfte im klinischen Alltag durchgeführt wird. Der Fokus liegt dabei insbesondere auf die Schlüsse, die man aus dem Weinen eines Babys auf dessen Schmerz machen kann. Um die menschliche Stimme automatisiert analysieren zu können, werden Methoden der Signalverarbeitung verwendet. Daher werden in Kapitel 2.2 technische Grundlagen erläutert, die zu diesem Zweck unerlässlich sind. In Kapitel 2.3 wird eine Einführung in die „klassische Schreiforschung". Dabei handelt es sich um Wissenschaftsgebiet, bei dem versucht wird, mit Hilfe von Methoden der Signalverarbeitung ein tieferes Verständnis über die Bedeutung des Weinens von Babys zu erhalten. Da sich in dieser Arbeit erstellte Konzept zur automatisierten Analyse des Weinenes als Erweiterung der klassischen Methoden versteht, ist ein Verständnis des Wissenschaftsgebietes unerlässlich. In Kapitel 2.4 werden Grundlagen des Überwachten maschienellen Lernens erläutert, da diese zur automatisierten Interpretation von Audiosignalen von Bedeutung sind.

2.1 Schmerz und Weinen bei Neugeborenen aus medizinischer Sicht

Schmerz wird definiert als eine „ein unangenehmes Sinnes- oder Gefühlserlebnis, das mit tatsächlicher oder potenzieller Gewebeschädigung einhergeht".[32, S. 438] Abseits von dieser theoretischen Definition hat der Mensch ein intuitives Verständnis für Schmerz, da jeder ihn in seine Leben erfahren musste. In der ersten Hälfte des 20sten Jahrhunderts war die vorherrschende Meinung, dass Neugeborene keinen Schmerz empfinden können. Beispielsweise bekam sie nach Operationen keine Schmerzmittel verabreicht. Die aktuell vorherrschende Meinung ist, dass Neugeborene im selben Maße wie Erwachsene Schmerz empfinden können. Die freien Nervenenden, die in der Lage sind, physische Schäden am Körper festzustellen, sind bei Neugeborenen ebenso wie bei Erwachsenen über den Körper verteilt. Die hormonelle Reaktion ist ebenfalls vergleichbar. [18, S. 402] [32, S. 438]

2.1.1 Pain Scales

Es gibt diverse Gründe, warum Neugeborene Schmerz empfinden können. Sie reichen über physische Schäden, aufgrund von komplikationen bei der Geburt oder Gewalteinwirkungen, über Erkrankungen, wie Kopfschmerzen oder Infektionen, bis hin zu therapeutischen Prozeduren, wie Injektionen oder Desinfektionen von Wunden. Das Vorhandensein von Schmerz

ist anhand diverser physiologischer, biochemischer, verhaltensbezogener und psychologischer Veränderungen messbar.[32, S. 441]

Schlussendlich ist Schmerz jedoch ein subjektives Empfinden. Daher wird der Schmerzgrad bei Erwachsenen typischerweise durch eine Selbsteinschätzung des Patienten unter der Leitung gezielter Fragen des Arztes festgestellt. Bei Kindern unter 3 Jahren ist diese Selbsteinschätzung nicht möglich. Diese Einschätzung muss daher von anderen Personen vorgenommen werden. Im klinischen Kontext sind dies medizinische Fachkräfte, wie beispielsweise Ärzte, Krankenpfleger oder Geburtshelfer. Die von außen am leichtesten feststellbaren Indikatoren von Schmerz sind die verhaltensbasierten Merkmale, wie zum Beispiel ein Verkrampfen des Gesichtsausdrucks, erhöhte Körperbewegungen oder lang anhaltendes Weinen.[32, S. 438] Die Schmerzdiagnostik durch eine andere Person ist etwas inherent subjektives und abhängig von Faktoren wie dem Alter, Geschlecht, kulturellen Hintergrund, persönlichen Erfahrungen mit Schmerz etc.[13, S. 3] Um die Schmerzdiagnostik objektiver zu gestalten, wurden daher sogenannte *Pain Scales* entwickelt, mit Hilfe eines Punktesystems den Schmerzgrad des Babys quantifizieren.[32, S. 438 - 439] Es existieren *monomodale* oder *unidimensionale* Pain Scales, bei denen der Schmerzgrad aus der Beobachtung *eines* Merkmals geschlossen wird, wie beispielsweise der Gesichtsausdruck. Ein Merkmal wird in diesem Zusammenhang als *Schmerzindikator* bezeichnet. *Multimodale* oder auch *Multidimensionale* Pain Scales beziehen mehrere Schmerzindikatoren in das Scoring mit ein.[20, S. 69 - 71].

Tabelle 2.1 zeigt das Scoring-System „Neonatal Infant Pain Scale“(NIPS) als Beispiel für eine multimodale Pain Scale. Diese Pain Scale ist für Babys von 0 bis 1 Jahr geeignet. Sie wurde auf der Basis der Erfahrungen von Krankenschwestern erarbeitet. Das Baby soll bei der Anwendung dieser Pain Scale für eine Minute beobachtet werden. Dies Schmerzfeststellung kann im häufigsten Fall alle 30 Minuten durchgeführt werden. Für jede der aufgeführte Kategorie werden ein, zwei oder drei Punkte vergeben und anschließend aufsummiert. Ein insgesamter Wert von > 3 zeigt moderaten Schmerz an, ein Wert von > 4 großen Schmerz.[15]

Tabelle 2.1: Neonatal Infant Pain Scale [15]

NIPS	**0 points**	**1 point**	**2 points**
Facial Expr.	Relaxed	Contracted	-
Cry	Absent	Mumbling	Vigorous
Breathing	Relaxed	Different than basal	-
Arms	Relaxed	flexed/stretched	-
Legs	Relaxed	flexed/stretched	-
Alertness	Sleeping	uncomfortable	-

Nach dem Muster der NIPS existieren viele weitere Pain Scales. Sie unterscheiden sich hinsichtlich der Schmerzindikatoren, die betrachtet werden, dem Punktesystem oder der Art des Schmerzes, die festzustellen ist. Einige Pain Scales sind beispielsweise auf die Schmerzdiagnostik während eines Eingriffes spezialisiert, andere auf den darauf folgenden Heilungsprozess. In den meisten multimodalen Pain Scales wird das Weinen oder Schreien der Babys als Schmerzindikator mit einbezogen. In der englischen Fachliteratur ist von „Cry“ die Rede.[30, S. 97 - 98] In dieser Arbeit wird „Cry“ mit „Weinen“ oder mit dem neutraleren Begriff „kindliche Lautäußerungen“ übersetzt. Tabelle 2.2 zeigt eine Übersicht über einige multimodale Pain Scales. Die Übersicht zeigt vor allem, nach welchen Kritieren das Weinen in den jeweiligen Scales bewertet wird. Außerdem wird für jede Pain Scale

angegeben, für welches Alter sie geeignet ist, welcher Schmerz-Typ diagnostiziert wird, sowie der zur Diagnose vorgesehene Beobachtungs-Zeitraum und Intervall. Angaben, die mit einem *?* verzeichnet wurden, konnten nicht in Erfahrung gebracht werden. Es handelt sich hier nur eine Übersicht über die Wichtigsten Fakten. Die Anleitungen der jeweiligen Pain Scales geben weitere Anweisungen zur Benutzung.

System	**P.**	**Description**	**other Ind.**	**Comments**
FLACC	0	No cry (awake or asleep)	Face, Legs, Activity, Consolability	Age: 2 months - 7 years Observe for: 1 - 5 minutes Observe every: ? Pain-Type: Ongoing
	1	Moans or whimpers; occasional complaint		
	2	Crying steadily, screams or sobs, frequent complaints		
N-PASS	-2	No cry with painful stimul	Behaviour, Facial Expr., Extremities, Vital Signs	Age: 0 - 100 days Observe for: ? Observe every: 2 - 4 hours Pain-Type: Ongoing
	-1	Moans or cries minimally with painful stimuli		
	0	Appropiate Crying		
	1	Irritable or Crying at Intervals. Consolable		
	2	High-pitched or silent-continuous crying. Not consolable		
BPSN	0	No Crying	Alertness, Skin Color, Eyebrows, ...	Age: ? Observe for: ? Observe every: Pain Type: ?
	1	Crying less than 2 minutes		
	2	Crying more than 2 minutes		
	3	Shrill Crying more than 2 minutes		
CRIES	0	If no cry or cry which is not high pitched	O2, Vital Signs, Expression, Sleeplessness	Age: 0 - 6 Months Observe for: ? Observe every: 1 hour Pain-Type: Post Operative
	1	If cry high pitched but baby. is easily consoled		
	2	If cry is high pitched and baby is inconsolable		
COVERS	0	No Cry	O2, Vital Signs, Expression, ...	Age: ? Observe for: ? Observe every: ? Pain Type: Procedural
	1	High-Pitched or visibly crying		
	2	Inconsolable or difficult to soothe		
PAT	0	No	Posture, Sleep Pattern, Expression, ...	Age: 0 - 3 months Observe for: 15 - 30 sec Observe every: 30 min Pain Type: Post Operative
	1	When disturbed, doesn't settle after handling, loud, whimper, whining		
DAN	0	Moans Briefly	Facial Exp., Limb Mov.	Age: 0 - 2 years Observe for: ? Observe every: ? Pain Type: Procedural
	1	Intermittent Crying		
	2	Long-Lasting Crying, Continuous howl		
COMFORT	0	No crying	Alertness, Calmness, Respiration, ...	Age: 0 - 3 years Observe for: ? Observe every: ? Pain: Post Operative
	1	Sobbing or gasping		
	2	Moaning		
	3	Crying		
	4	Screaming		
MBPS	0	Laughing or giggling	Facial Exp., Movement	Age: ? Observe for: ? Observe every: ? Pain Type: Procedural
	1	Not Crying		
	2	Moaning quiet vocalizing gentle or whimpering cry		

3	Full lunged cry or sobbing
4	Full lunged cry more than baseline cry

Tabelle 2.2: Übersicht über Pain-Scales. [30, S. 98] [38] [33] [9] [3] [19] [16] [5] [29] [9]

Da die Begriffe *Pain Scale* und *Pain Score* in einigen Veröffentlichungen inkonsistent verwendet werden, wird in dieser Arbeit die Konvention getroffen, dass mit *Pain Scale* das System zur Schmerzdiangostik gemeint ist und mit *Pain Score* die auf Basis der Pain Scale vergebene Punktzahl. *NIPS* ist also beispielsweise eine Pain Scale, und 3 eine Pain Score.

Folgende Anmerkungen werden bezüglich der Pain Scales aus Tabelle 2.2 gemacht:

1. Die Kriterien zur Bewertung des Weinens werden zum größten Teil mit *subjektiv behafteten Begriffen* beschrieben. Beispielsweise wird bei dem *N-PASS*-System ein Score von drei für „High-pitched or silent-continuous crying“ vergeben. Die Begriffe „high-pitched“ und „silent-continuous“ werden nicht näher definiert. Auch in die Anwendungsvorschriften der Pain Scales werden keine festen Definitionen gegeben. Dies erleichtert den praktischen Einsatz der Pain Scales, führt jedoch zu einem Interpretationsspielraum und somit zu einem von der diagnostizierenden Person abhängigen Scoring. Die *BPSN*-Scale nutzt als einzige der vorgestellten Scales objektiv messbare Eigenschaften.
2. Die Pain Scales fokussieren unterschiedliche Eigenschaften. Bei *CRIES* ist die Tonhöhe, bei *BPSN* die Länge und bei *COMFORT* die Art des Weinens ausschlagebend für das Scoring.
3. Die Beschreibungen sind kurz und prägnant gehalten, die diagnostizierende Person hat bei keiner Pain Scale auf mehr als drei Eigenschaften des Weinens zu achten.

2.1.2 Weinen bei Neugeborenen

An dieser Stelle stellt sich der Leser eventuell die Frage, woher die unterschiedlichen Bewertungskriterien für das Weinen in den Pain Scales stammen. Gibt es eine „beste“ Pain Scale? Dieser Frage unterliegen zwei grundlegendere Fragen:

1. Ist es möglich, aus den akustischen Eigenschaften den motivierenden Grund für die Lautäußerung abzuleiten? Klingt ein durch Hunger bedingtes Weinen anders als ein durch Schmerz bedingtes?
2. Ist es möglich, anhand der akustischen Eigenschaften den Schweregrad dieses motivierenden Grundes abzuleiten?

Die Annahme, dass es möglich sei, aus den Eigenschaften des Weinens den Grund ablesen zu können, wird als „Cry-Types Hypothesis“ bezeichnet. Die berühmtesten Befürworter dieser Hypothese ist eine skandinavische Forschungsgruppe, auch bezeichnet als „Scandinavian Cry-Group“, die die Idee in dem Buch „Infant Crying: Theoretical and Research Perspectives“[22] publik machte. Die Hypothese besagt, dass die Empfindungen *Hunger, Freude, Schmerz, Geburt* sowie Sonstiges klare Unterschiede hinsichtlich der akustischen Merkmale des Weinens aufweisen. Diese Unterschiede seien im Spektogramm sichtbar. Wenige Jahre Später zeigten Müller et al. [8], dass bei leichter Veränderung des Experimentes die Unterscheidung nicht mehr möglich sei. Die Gegenhypothese ist, dass Weinen „nichts als undifferenziertes

Rauschen“ sei. 50 Jahre später liegt kein anerkannter Beweis für die eine oder andere Hypothese vor. Es gibt lediglich starke Hinweise dafür, dass die Plötzlichkeit des Eintretens des Grundes sich in den akustischen Eigenschaften bemerkbar macht. Ein plötzliches Ereignis, wie ein Nadelstich oder ein lautes Geräusch, führen auch zu einem plötzlich beginnenden Weinen. Ein langsam eintretendes Ereignis, wie ein langsam zunehmender Schmerz oder Hunger führen auch zu einem langsam eintretenden Weinen. Da nach Kenntnis des Autors bis heute keine wissenschaftlich belastbarer Beweis vorgelegt wurde, wird empfohlen, den Grund aus dem Kontext abzuleiten.[37, S. 9 - 13, 17 - 19]

Die Zweite Frage nach der Ableitung der Stärke des Unwohlseins aus den akustischen Eigenschaften des Weinens wird in der Fachliteratur unter dem Begriff *Cry as a graded Signal* subsumiert. Je „stärker“ das Weinen, desto höher das Unwohlsein (*Level of Distress (LoD)*) des Säuglings. Tatsächlich bemessen wird dabei der von dem Beobachter vermutete Grad des Unwohlseins des Babys, und nicht der tatsächliche Grad, da dieser ohne die Möglichkeit der direkten Befragung des Kindes nie mit absoluter Sicherheit bestimmt werden kann. Ein hohes Unwohlsein hat vor allem eine schnelle Reaktion der Aufsichtspersonen zur Beruhigung des Babys zur Folge, womit dem Weinen eine Art Alarmfunktion zukommt. Es gibt starke Hinweise darauf, dass das Level of Distress anhand objektiv messbarer Eigenschaften des Audiosignals bestimmt werden kann. So herrscht beispielsweise weitestgehend Einigung darüber, dass ein „lang“ anhaltendes Wein auf einen hohen Level of Distress hinweist. Insofern aus dem Kontext des Weinens Schmerz als die wahrscheinlichste Ursache eingegrenzt werden kann, kann aus einem hohen Level of Distress ein hoher Schmerz abgeleitet werden. [37, S. 13 - 17] [36] Es herrscht wiederum keine Einigung darüber, welche akustischen Eigenschaften im Detail ein hohes Level of Distress anzeigen. Carlo V Bellieni et al. [5] haben festgestellt, dass bei sehr hohem Schmerz in Bezug auf die DAN-Scale (siehe Tabelle 2.2) die Tonhöhe steigt. Qiaobing Xie et al. [36] haben festgestellt, dass häufiges und „verzerrtes“ Schreien (ohne feststellbares Grundfrequenz, da der Ton stimmlos erzeugt wird) auf einen hohen Level of Distress hinweist.

2.2 Signalverarbeitung

In Kapitel 2.1 wurde erläutert, wie Weinen von Neugeborenen mit Hilfe subjektiv behafteter Begriffe beschrieben werden kann. Möchte man das Weinen objektiv beschreiben und messbar machten, so verwendet man die Methoden der digitalen Signalverarbeitung. An dieser Stelle wird eine Einführung in die wichtigsten Themen dieses Wissenschaftsgebietes gegeben, die im Zusammenhang mit der Audiosignalverarbeitung größere Bedeutung haben. Es wird ein grundlegendes Verständnis der Signalverarbeitung vorausgesetzt, da die Erläuterungen in diesem Kapitel eher der Definition der verwendeten Begriffe dient, aus Platzgründen jedoch keine für Neulinge geeignete Einführung in das Themengebiet gewährleisten kann. Falls dieses Wissen nicht vorhanden ist, wird zur Einarbeitung das Buch „The Scientist and Engineer's Guide to Digital Signal Processing“ von Steven W. Smith empfohlen.[40], welches kostenlos als E-Book bereitgestellt wird.

2.2.1 Grundlegende Definitionen

In dieser Arbeit sind nur *digitale Signale* von Bedeutung. Ein digitales Signal $x[\,]$ ist nach Formel 2.1 ein beliebige Zahlenfolge mit diskreten Definitionsbereich. Dem Definitionsbereich

kommt die Bedeutung *Zeit* zu.[40, S. 11-12] In dieser Arbeit gilt die Konvention, dass mit $x[\,]$ das gesamte Signal gemeint ist und mit $x[n]$ *ein* Wert des Signals (in diesem Zusammenhang auch als *Sample* bezeichnet) an dem Index $\hat{=}$ Zeitpunkt n. Die Samplingfrequenz des digitalen Signales wird mit f_s bezeichnet.

$$x[\,] := \quad \forall n \in \mathbb{Z} : \; x[n] = s \tag{2.1}$$

Der Definitionsbereich eines Signals erstreckt sich implizit immer von negativer bis positiver Unendlichkeit. Das heißt nicht, dass alle Samples des Signals auch Informationen enthalten müssen. Der *Support* ist das kleinst mögliche Zeitintervall, der alle Samples enthält, die nicht den Wert 0 haben, wie Formel 2.2 definiert. Wird also auf ein Sample zugegriffen, das außerhalb des Supportes liegt, hat dieses Sample den Wert 0 (ein „0-Sample“)[31, S. 24]

$$\begin{aligned} \mathrm{Sup}(x[\,]) = [sup_s, sup_e] \quad , sup_s, sup_e \in \mathbb{Z} \\ , x[sup_s] \neq 0 \ \wedge \ x[sup_e] \neq 0 \ \wedge \ \forall n \ \notin [sup_s, sup_e] : x[n] = 0 \end{aligned} \tag{2.2}$$

Die *Dauer* eines Signales ist die Länge des Supportes nach Formel 2.3. In dieser Arbeit herrscht die Konvention, dass die Länge des Signals kurz mit der Variable N abgekürzt wird. Wenn nicht anders definiert, erstreckt sich der Support eines Signals von $0, \ldots, N-1$.[31, S. 24]

$$\mathrm{Length}(x[\,]) = sup_e - sup_s + 1 = N \tag{2.3}$$

2.2.2 Statistische Merkmale

Im folgenden wird ein Überblick über die häufig verwendete Signaleigenschaften gegeben. Abbildung 2.1 visualisiert die Erläuterungen.

1. Der **Maximalwert** / **Minimalwert** beschreibt den höchsten / niedrigsten in $x[\,]$ enthaltenen Wert nach den Formel 2.4.

$$\begin{aligned} \max(x[\,]) &= \max_{n \in \mathrm{Sup}(x[\,])} \{\ x[n]\ \} \\ \min(x[\,]) &= \min_{n \in \mathrm{Sup}(x[\,])} \{\ x[n]\ \} \end{aligned} \tag{2.4}$$

2. Der **Durchschnittswert** / **Average Value** beschreibt den durchschnittlichen Wert aller Samples von $x[\,]$ nach Formel 2.5. Dieser Durchschnittswert wird über ein beliebiges Intervall $[n_1, n_2]$ berechnet.

$$\mathrm{AVG}(x[\,]) = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} x[n] \tag{2.5}$$

3. Der **Mean Squared Value** (MSV) beschreibt den quadrierten Durchschnittswert über

eine bestimmtes Intervall nach Formel 2.6. Er wird auch als *durchschnittliche Energie* oder *average Power* bezeichnet.

$$\mathrm{MSV}(x[\,]) = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} x[n]^2 \tag{2.6}$$

4. Das **Root Mean Square** (*RMS*) wird definiert als die Wurzel des Mean Squared Value nach Formel2.7. Der RMS kann im Vergleich zum MSV besser ins Verhältnis zu den Werten des Signals gesetzt werden kann. Er wird im Deutschen auch als **Effektivwert** oder **Durchschnittsleistung** bezeichnet. Da die deutschen Begriffe in einigen Quellen jedoch auch für den MSV verwendet werden, wird an dieser Stelle nur mit den englischen Begriffen gearbeitet.

$$\mathrm{RMS}(x[\,]) = \sqrt{\frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} x[n]^2} \tag{2.7}$$

5. Die **Energie / Energy** eines Signals wird nach Formel 2.8 definiert. Sie entspricht dem MSV-Wert multipliziert mit der Länge des Intervalls. [31, S. 27-28]

$$\mathrm{E}(x[\,]) = \sum_{n=n_1}^{n_2} x[n]^2 \tag{2.8}$$

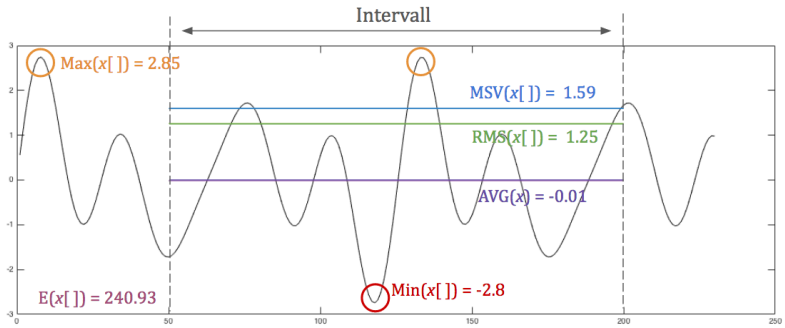

Abbildung 2.1: Statistische Werte eines Signals über das Intervall [50,200]

2.2.3 Fehlersignale

Angenommen, ein Signal $x[\,]$ wird übertragen, auf dem Übertragungsweg jedoch durch ein anderes Störsignal wie z.B. Rauschen $e[\,]$ überlagert, auch bezeichnet als Fehlersignal. Das resultierende Signal $x'[\,]$ wird nach Formel 2.9 berechnet.

$$x'[\,] := \quad \underset{n=n_1}{\overset{n_2}{\forall}} \; : \; x'[n] = x[n] + e[n] \tag{2.9}$$

Kennt man sowohl das Eingangssignal $x[\,]$ als auch das Ausgangssignal $x'[\,]$, kann das Störsignal $e[\,]$ nach Formel 2.10 berechnet werden.

$$e[\,] := \quad \underset{n=n_1}{\overset{n_2}{\forall}} : \; e[n] = x'[n] - x[n] \tag{2.10}$$

Eine Möglichkeit der Quantifizierung der Stärke des Rauschens auf das Signal ist, das Eingangssignal ins Verhältnis zum Rauschsignal zu setzen. Formel 2.11 gibt die Definition.

$$\mathrm{SNR}_{rel}(x[\,], e[\,]) = \frac{MSV(x[\,])}{MSV(e[\,])} \tag{2.11}$$

In der Praxis ist der MSV des Eingangssignals meist sehr viel höher als der des Fehlersignals. Um den Zahlenraum zu begrenzen, wird die Pseudoeinheit dB verwendet. Formel 2.12 definiert den *Signal/Rausch-Abstand* (*SNR*, englisch Signal-to-Noise-Ratio). Ein *niedriger* SNR-Wert auf ein *starkes* Rauschen hin, und ein *hoher* SNR auf ein *schwaches* Rauschen. Im Zusammenhang mit der Spracherkennung ist der Signal/Rausch-Abstand von Bedeutung, da ein höheres Rauschen die Verarbeitung des Nutzssignals, der Sprache, erschwert.

$$\mathrm{SNR}(x[\,], e[\,]) = 10 \cdot \lg\left(\frac{MSV(x[\,])}{MSV(e[\,])}\right) \text{ dB} \tag{2.12}$$

2.2.4 Kurzzeit-Fourier-Transformation

Das Signal $x[\,]$ beschreibt den Zeitbereich des Signals, da die unabhängige Variable die Zeit definiert. Gleichung 2.13 definiert die *komplexe diskrete Fouriertransformation*, kurz *DFT*, die das diskrete Signal $x[\,]$ aus dem Zeitbereich in den Frequenzbereich $X[\,]$ transformiert. Das Signal des Frequenzbereiches ist, ebenso wie das Signal des Zeitbereiches, N punkte Lang und hat den Support $0, \ldots, N-1$. Jedes Sample des Frequenzbereiches ist eine komplexe Zahl, deren Realteil $\Re(x[k])$ die Amplitude der entsprechenden Sinuswelle mit der Frequenz $f = k\frac{f_s}{N}$ bezeichnet und deren Imaginärteil $\Im(x[k])$ die Amplitude der entsprechenden Kosinuswelle bezeichnet.[40, S. 149, S. 567 - 571] [1, S. 60]

$$\mathrm{DFT}\{x[\,]\} = X[\,] := \quad \underset{k=0}{\overset{N-1}{\forall}} : \; X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j2\pi k \frac{n}{N}} \tag{2.13}$$

Das *Spektrum* wird nach Gleichung 2.14 definiert als der Absolutwert des Frequenzbereiches im Bereich $0, \ldots, N/2$.

$$\text{Spektrum} := \quad |X[0]|, \; \ldots \; , |X[N/2]| \tag{2.14}$$

Abbildung 2.2 visualisiert die Transformation in den Frequenzbereich: In der Abbildung ist oben der Zeitbereich eines 1.8 Sekunden langen Signals zu sehen. Es können klar drei nacheinander gespielte Töne erkannt werden. Der Zeitbereich lässt klar erkennen, zu welchen Zeitpunkten die Töne beginnen und Enden, aber nicht, welche Frequenzen in den Tönen enthalten sind. Unten ist der Spektrum abgebildet. Die x-Achse bezeichnet die Frequenz von 0 bis 22 050 Hz und die x-Achse die Amplitude der entsprechenden Frequenz. Beide Achsen werden logarithmisiert dargestellt. Das Frequenzspektrum zeigt, welche Frequenzen

im dem Signal enthalten sind. So sind beispielsweise keine Frequenz unterhalb von 1000 Hz enthalten. Das Spektrum acht jedoch nicht erkennbar, welche Frequenzen enthalten sind.

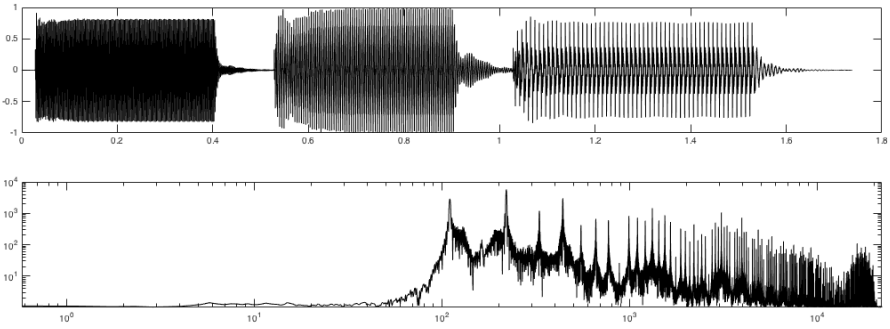

Abbildung 2.2: Ein 1.8-Sekunden langes Signal. Oben: Der Zeitbereich mit drei klar erkennbaren Events. Unten: Das Frequenz-Spectrum des gesamten Signals mit logarithmisierten Achsen.

Es ist wünschenswert, einen Kompromiss aus den Vorteilen beider Bereiche zu finden, in dem man das Spektrum kürzerer Zeitabschnitte des Signals bildet. Dazu wird der Zeitbereich $x[\,]$ in Fenster der Länge M zerlegt. Die zeitliche Differenz zwischen zwei Fenstern wird als *Hoptime* R bezeichnet. Gleichung definiert die Bildung des Signalfensters $x_i[\,]$. Dieser Prozess wird als *Windowing* bezeichnet.[39]

$$x_i[\,] := \quad \forall_{n=0}^{M-1} : \; x_m[n] = x[n + i \cdot R] \tag{2.15}$$

Abbildung 2.3 gibt ein Beispiel für die Zerlegung von x in Signalfenster $x_0[\,], \ldots, x_4[\,]$. Die Samplingrate des Signals ist $f_s = 44100$, die Fensterlänge beträgt $M = 22050/f_s = 0.5\,\text{s}$ und die Hoptime $R = M/2 = 0.25\,\text{s}$.

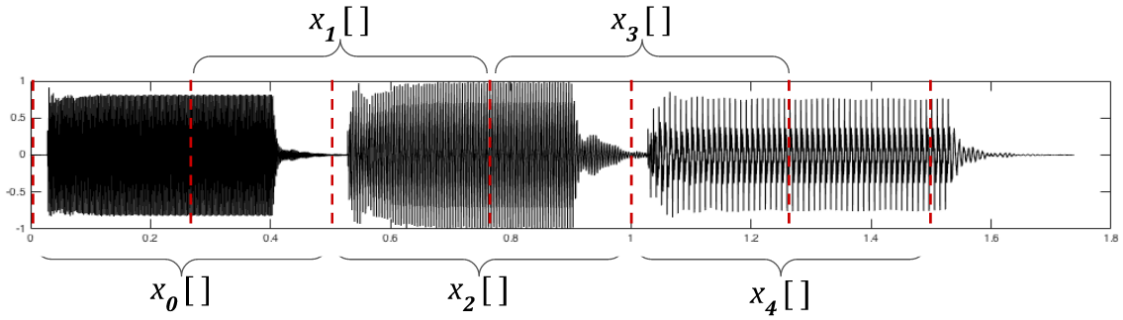

Abbildung 2.3: Windowing: Die Zerlegung eines Signals in kürzere Fenster.

Als Vorbereitungsschritt für die Transformation der Signalfenster in den Frequenzbereich wird nun jedes Fenster mit einer sogenannten *Fensterfunktion* (engl *window*) $w[\,]$ multipliziert.[1, S. 69] Gleichung 2.16 definiert eine der am weitesten verbreiteten Fenster-Funktionen, das *Hamming-Window*. Der Paramter M gibt die länge des Fensters an. Abbildung 2.4 visualisiert das Hamming-Window. [40, S. 286]

$$w[\,] := \quad \forall_{n=0}^{M-1} : \; w[n] = 0.54 - 0.46\cos(\frac{2\pi n}{M}) \tag{2.16}$$

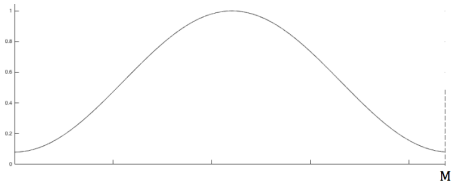

Abbildung 2.4: Das Hamming-Window

Die Gleichung 2.17 definiert die *Kurzzeit-Fourier-Transformation* (engl *Short Time Fourier Transformation*, *STFT*), implementiert mit Hilfe der DFT. Dabei wird das Signalfenster $x_i[\,] = x[n + i \cdot R]$ mit der Fensterfunktion $w[\,]$ multipliziert und in das *Frequenz-Fenster* $X_i[\,]$ transformiert.[1, S. 69] [4] Abbildung 2.5 visualisiert die STFT des Beispiels aus Abbildung 2.3.

$$\mathrm{STFT}_i\{x[\,]\} = X_i[\,] := \quad \underset{k=0}{\overset{M-1}{\forall}} : \; X_i[k] = \sum_{n=0}^{M-1} x[n + i \cdot R] \cdot w[n] \cdot e^{-j2\pi k \frac{n}{N}} \tag{2.17}$$

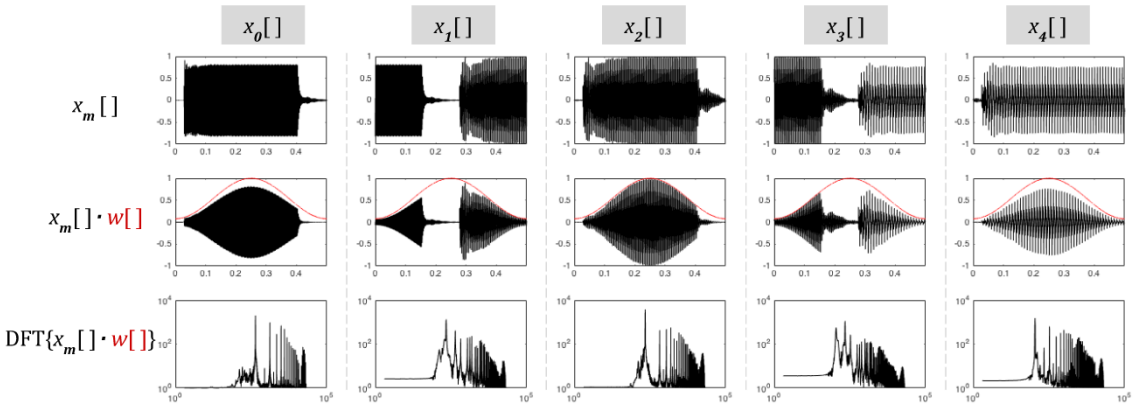

Abbildung 2.5: STFT des Beispiel-Signals aus Abbildung 2.3

2.2.5 Akustische Modellierung der menschlichen Stimme

Der menschliche Sprechapparat wird in die folgenden Komponenten Unterteilt:

Schallproduktion: Die Lunge stößt Luft aus, welche die Stimmbänder passieren. Sind die Stimmbänder leicht gespannt, so wird der Luftstrom periodisch unterbrochen. Die Schwingfrequenz beträgt bei Männern etwa 120 Hz und bei Frauen 220 Hz. Die Frequenz kann während des Sprechens um bis zu einer Oktave variieren. Es wird so ein periodisches, akustisches Signal produziert, bezeichnet als „periodische Quelle“ (engl. „periodic Source“). Sind die Stimmbänder stark gespannt, so entstehen Turbulenzen, die sich akustisch als ein zischendes Geräusch ohne identifizierbare Tonhöhe äußert. Dieses stimmlose Signal wird bezeichnet als „Turbulenzquelle“ (engl. „turbulance Source“)

Klangformung: Das Signal der Stimmlippen passiert den Rachen, Mund- und Nasenraum, welche gemeinsam als „Vokaltrakt“ beschrieben werden. Das Halszäpfchen bestimmt,

ob der Luftstrom in den Mund- oder Nasenraum geleitet wird. Die Stellung der Artikulatoren, bestehend aus Kiefer, der Zunge usw. bestimmen die Beeinflussung des Klanges, der durch die Stimmbänder erzeugt wurde. Diese Beeinflussung wird als Filter angenähert. [17, S. 62] [1, S. 13] Abbildung 2.6 visualisiert diese Komponenten.

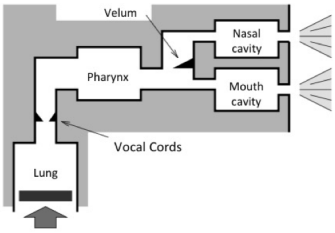

Abbildung 2.6: Schematische Übersicht über die Organe der Spracherzeugung. Lung = Lunge, Vocal Chords = Stimmbänder, Pharynx = Rachen, Velum = Halszäpfchen, Mouth Cavity = Mundraum, Nasal Cavity = Nasenraum [25]

Aus Sicht der Signalverarbeitung wird die menschliche Lautproduktion durch das sogenannte *Source-Filter-Modell* modelliert. Der durch die Stimmbänder erzeugte periodische Ton wird angenähert durch einen Impuls-Zug, welcher durch den Schlund als linearen Filter moduliert wird. Der stimmlose, nicht-periodische Ton wird durch weißes Rauschen angenähert. Der so erzeugte periodische oder nicht-periodische Ton wird als das Eingangs-Signal $u[\,]$ bezeichnet. Dieses Signal wird daraufhin an den Vokaltrakt weitergeben, welcher als lineares, zeitinvariantes Filter mit der Impulsantwort $v[\,]$ modelliert wird. Diese Impulsantwort ist abhängig von der Konfiguration der Organe des Vokaltraktes. Die Lippen werden als zweites lineares, zeitinvariantes Filter mit der Impulsantwort $r[\,]$ modelliert. $r[\,]$ wird auch als „radiant Model“ bezeichnet. Das tatsächliche Sprachsignal $y[\,]$ entsteht somit als die Faltung des Signals $u[\,]$ und den beiden linearen, zeitinvarianten Filtern nach Gleichung 2.18. Gleichung 2.19 definiert den Frequenzbereich des Ausgangssignals $Y[\,]$ durch die Multiplikation der Frequenzbereiche dieser drei Komponenten. Abbildung 2.7 visualisiert diesen Prozess schematisch. [17, S. 62 - 63] [25]

$$u[\,] * v[\,] * r[\,] = y[\,] \tag{2.18}$$

$$U[\,] \cdot V[\,] \cdot R[\,] = Y[\,] \tag{2.19}$$

Abbildung 2.8 zeigt die Zeitbereiche des stimmhaften und turbulenten Signals im Vergleich. Wie zu sehen ist, bestimmt der zeitliche Abstand zwischen den Impulsen die Grundfrequenz der Stimme. Dieses Signal $p[\,]$ wird durch den Schlund als Filter $G\{\,\}$gefiltert, wodurch der Zeitbereich der periodischen Quelle entsteht $G\{p[\,]\} = u_p[\,]$. Darunter ist der Zeitbereich des weißen Rauschen zu sehen. [26, Source]

Abbildung 2.9 zeigt die Frequenzbereiche der Komponenten des Source-Filter-Modells. Die periodische Quelle ($U[\,]$ links) zeichnet sich im Frequenzbereich durch gleichmäßig verteilte Spitzen aus, die mit steigender Frequenz an Amplitude verlieren. Rechts daneben ist der Frequenzbereich des weißen Rauschen zu sehen. Die Frequenzantwort des Vokaltraktes $V[\,]$ zeichnet sich durch Resonanzfrequenzen aus, von denen in diesem Beispiel vier erkennbar sind. Die Übertragungsfunktion der Lippen $R[\,]$ wird als Hochpassfilter angenähert. Das

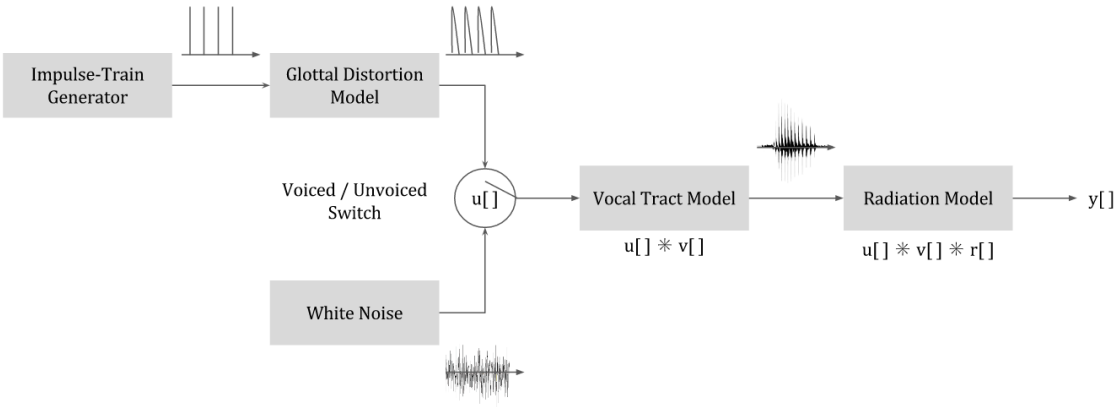

Abbildung 2.7: Schematische über das Source-Filter-Model [14, nach Source estimation, S. 17]

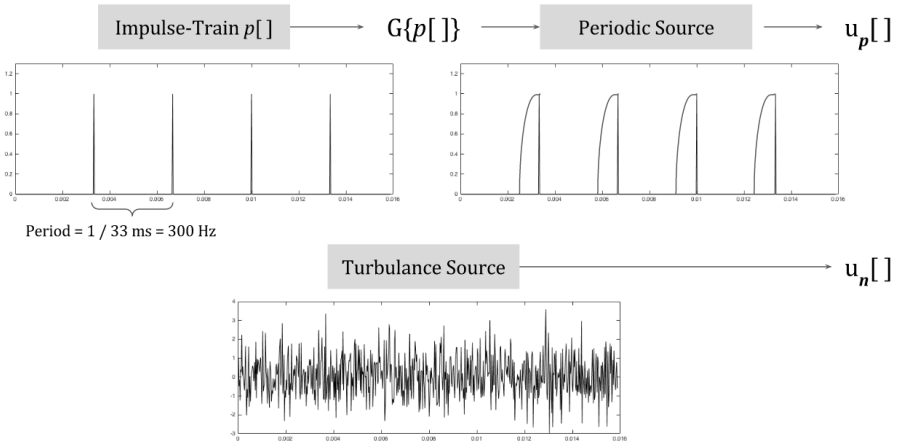

Abbildung 2.8: Zeit-Bereiche der periodic und der turbulance Source [26, Source]

Ausgangssignal $Y[\ ] = U[\ ] \cdot V[\ ] \cdot R[\ ]$ zeigt den Einfluss der Filter auf das jeweilige Eingangssignal.[14, Source estimation], [26, Vocal Tract Resonance]

Abbildung 2.10 zeigt schematisch das Spektrum eines stimmhaften Sprachsignals. Sowohl die Grundfrequenz als auch die harmonischen Obertonwellen sind rein visuell als „vielen, kurzen Signalspitzen“ im Spektrum erkennbar. Der kleinste gemeinsame Teiler der Frequenzen dieser Signalspitzen entspricht der Grundfrequenz f_0 dieses Stimmsignals, in diesem Beispiel 250.7 Hz. Die Grundfrequenz ist ebenfalls an der Signalspitze mit der tiefsten Frequenz ablesbar. Die harmonsichen Obertöne entsprechen der doppelten, dreifachen, ... Frequenz dieser Grundfrequenz, das heißt $2 \cdot f_0, 3 \cdot f_0, \ldots$ und werden bezeichnet mit $H_1, H_2, \ldots$. Die Grundfrequenz ist *nicht zwingend* die Spitze der höchsten Amplitude! Durch den Einfluss des Vokaltraktes als Filter können harmonische Oberwellen eine höhere Amplitude als die Grundfrequenz erhalten. Auf Basis des Spektrums lässt sich somit rein visuell ein stimmhaften Signal von einem nicht stimmhaften (Rausch-)Signal unterscheiden, in dem das Spektrum nach dem Vorhandensein dieser regelmäßigen Signalspitzen geprüft wird (vergleiche mit Abbildung 2.9).[1, S. 52 - 53]

Abbildung 2.11 verdeutlicht, wie der als lineares, zeitinvariantes Filter modellierte Vokaltrakt durch Formanten beschrieben wird. Diese Formanten spielen vor allem bei der

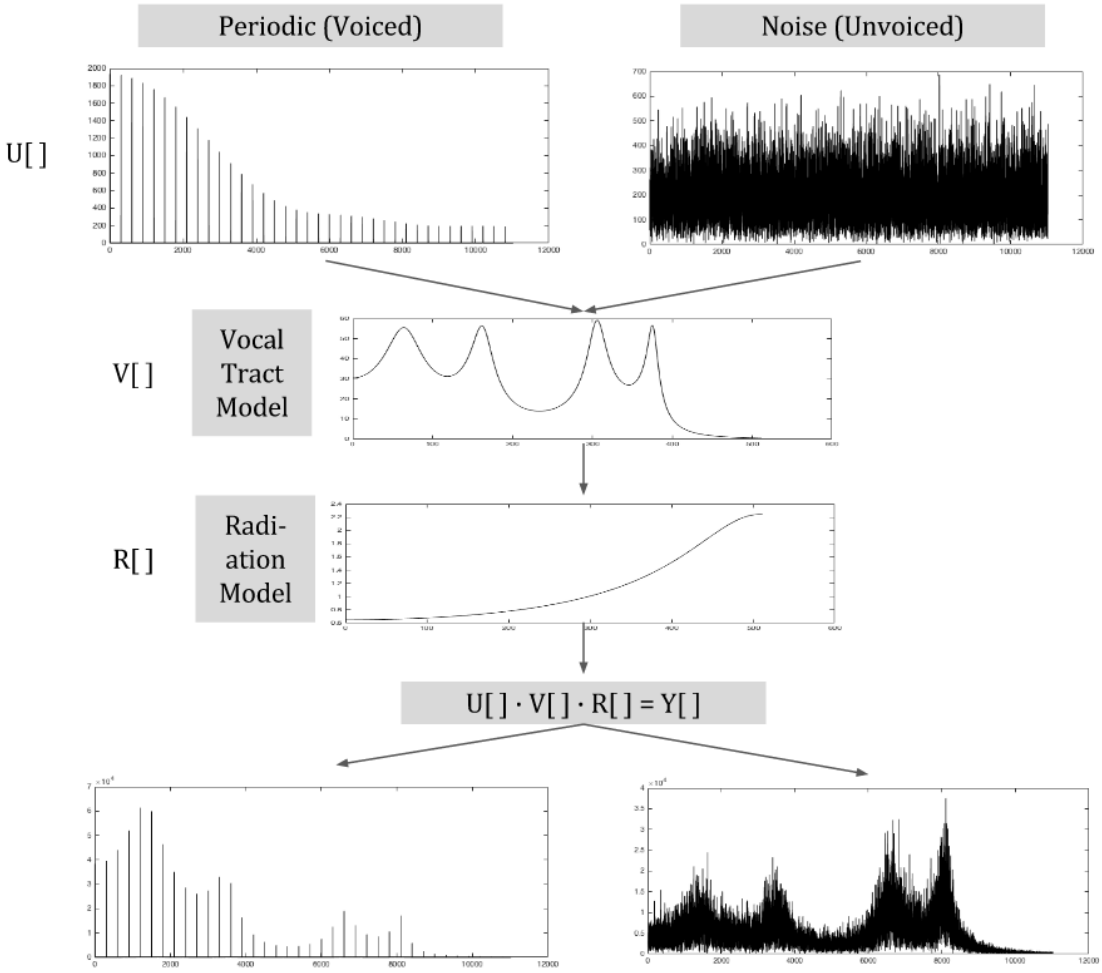

Abbildung 2.9: Betrachtung der Frequenz-Bereiche des Source-Filer-Modell (nach: [14, Source Estimation, S. 3])

Beschreibung von Vokalen eine Rolle. Formanten sind lokale Maxima im Spektrum der Transferfunktion, die dadurch erzeugt werden, dass der Vokaltrakt Resonanzen erzeugt. Die Formanten werden von links nach rechts durchnummeriert, von $F_1, \ldots F_n$. Jeder Formant wird durch seine Mittenfrequenz, seine Bandbreite und seine Amplitude beschrieben. Das wichtigste Merkmal ist jedoch die Mittenfrequenz, da sie vom menschlichen Gehör am stärkster zur Identifikation und Unterscheidung der Vokale genutzt werden. Mit steigender Frequenz nimmt die Amplitude der Formanten ab, der dominanteste Formant ist somit immer der erste. Daher werden meist nur die ersten 2 oder 3 Formanten zur Beschreibung eines Vokals angegeben, auch, wenn theoretisch weitaus mehr vom Vokaltrakt erzeugt werden. Für verschiedene Sprachen sind allerlei Tabellen zu finden, welche die Formantenfrequenzen der Vokale auflisten.[1, S. 19]

Beim Sprechen befinden sich sowohl das Signal der Stimmbänder als auch das Filter des Vokaltraktes und der Lippen in ständiger Veränderung. Ein stimmhaftes Sprachsignal gilt über kurze Zeitbereiche weniger Millisekunden als periodisch. Schlussendlich ist die Stimme nie perfekt periodisch, sondern nur annähernd periodisch. Da die Informationen der Sprache vor allem im Frequenzbereich codiert sind, wird die in Kapitel 2.2.4 vorgestellte Kurzzeit-Fourier-Transformation zur Analyse von Sprache eingesetzt. Die Visualisierung

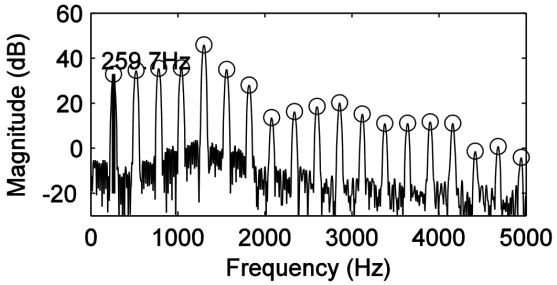

Abbildung 2.10: Grundfrequenz und harmonische Obertöne eines Sprachsignals.

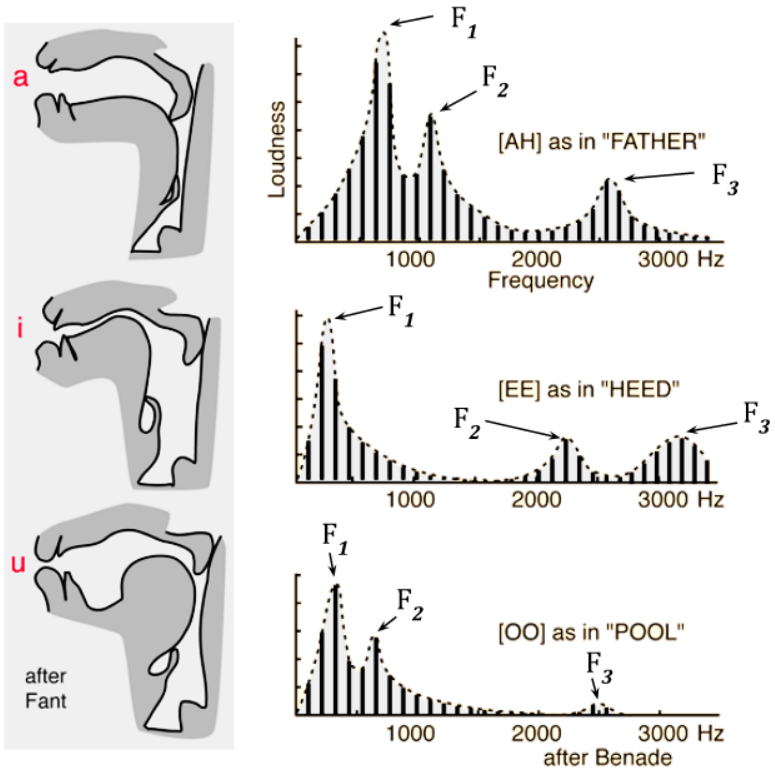

Abbildung 2.11: Formanten im Sprach-Signal (nach: [2])

der STFT wird als *Spektogramm* bezeichnet. Dabei werden auf der x-Achse die Zeitpunkte der Fenster und auf der y-Achse die Frequenz dargestellt. Die Frequenzfenster werden „auf die Seite gelegt“, damit ihr zeitlicher Verlauf übersichtlich betrachtet werden kann. Die Amplitude der entsprechenden Frequenzen wird farblich oder durch Helligkeiten codiert, abhängig von der konkreten Implementierung des Spektograms. Je länger das Zeitfenster der STFT, desto höher ist die Auflösung bezüglich des Frequenzbereiches und desto niedriger die Auflösung bezüglich der Zeitbereiches. Je kürzer die Zeitfenster der STFT, desto höher ist die Auflösung bezüglich des Zeitbereiches, und desto niedriger die Auflösung des Frequenzbereiches.[1, S. 45 - 50] [26, Acoustic Representations of Speech].

Abbildung 2.12 zeigt ein Beispiel für zwei Spektogramme mit unterschiedlichen Fensterlängen der STFT, angewandt auf einer 9 Sekunden langen Aufnahme eines weinenden Babys. Es ist zu erkennen, wie bei der geringeren Fensterlänge der zeitliche Verlauf besser erkennbar, jedoch die einzelnen harmonsichen Obertöne weniger gut voneinander unterscheidbar sind.

Bei der längeren Fensterlänge sind die Formanten leichter zu unterscheiden, der Beginn und das Ende der Lautäußerungen jedoch schwerer zu lokalisieren.

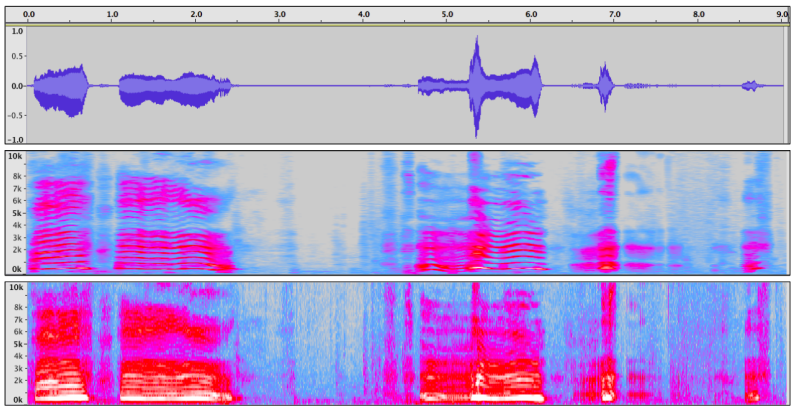

Abbildung 2.12: Spectogram von Baby-Weinen. Rot = Hohe Amplituden, Blau = niedrige Amplituden. Oben: Zeit-Bereich. Mitte: Spectogram mit einer Fensterlänge von 185 ms(8192-Sample DFT). Unten: Spectogram mit einer Fensterlänge von 5 ms
(265-Sample DFT).

2.3 Schreiforschung

Das Wissenschaftsgebiet, welches sich mit der Analyse und Interpretation von Lautäußerungen Neugeborener auseinandersetzt, wird als „Schreiforschung“ bezeichnet. Das bis heute wohl prominenteste Forschungsgruppe dieses Wissenschaftsgebietes ist die im vergangenen Kapitel erwähnte „Scandinavian Cry-Group“[22], welche zwischen 1960 und 1990 die Laute von Babys systematisch erforscht haben. Das wichtigste Werkzeug zur Analyse der Lautäußerungen war das eben vorgestellte Spektogramm, welches damals auf analogen Technologien basierte. Das Ziel der frühen Schreiforschung war es, mit Hilfe des Spektogramms Muster zur Unterscheidung eines abnormalem Weinen von einem normalen Weinen zu finden, um beispielsweise Krankheiten erkennen zu können.[37, S. 142]

Teil der Scandinavian Cry-Group waren H Golub und M Corwin, die in der Veröffentlichung „A Physioacoustic Model of the Infant Cry“[17] ein Vokabular zur Beschreibung typischer, im Spektogramm erkennbarer Muster festgelegt haben. Da das Vokabular bis heute Einsatz findet, wird eine Teilmenge dieses Vokabulars an dieser Stelle vorgestellt. Weiterhin werden Begriffe eingeführt, die von Zeskind et al. in „Rythmic organization of the Sound of Infant Cry “ veröffentlicht wurden.[34]

2.3.1 Phyisio-Akustische Modellierung des Weinens

Das Weinen von Babys lässt sich im allgemeinen als das „rythmische Wiederholen eines beim Ausatmen erzeugen Geräusches, einer kurzen Pause, einem Einatmungs-Geräusch, einer zweiten Pause, und dem erneuten Beginn des Ausatmungs-Geräusches“beschreiben. [43].

Die folgenen Begriffe werden in Abbildung 2.13 veranschaulicht.

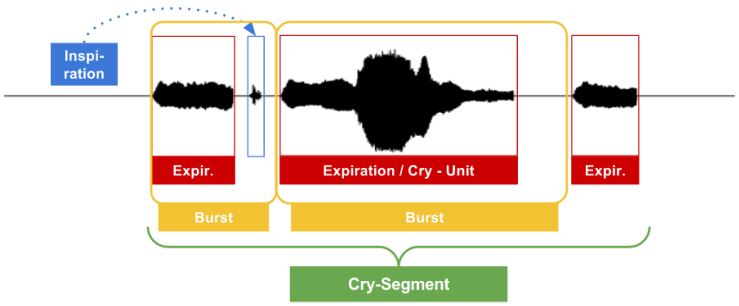

Abbildung 2.13: Veranschaulichung des Grundvokabulars

- **Expiration (Ausatmung):** Der Klang, der bei einem einzelnen, ununterbrochenen Ausatmen mit Aktivierung der Stimmbänder durch das Baby erzeugt wird. [34]. Der von Golub et al. [17, S. 61] verwendete Begriff **Cry-Unit** wird in dieser Arbeit synonym verwendet. Umgangssprachlich ist handelt es sich um einen einzelnen, ununterbrochenen *Schrei*.
- **Inspiration (Einatmung):** Der Klang, der beim Einatmen durch das Baby erzeugt wird.
- **Burst:** Die Einheit einer Ausatmung und der darauf folgenden Einatmung. Das heisst, dass die zeitliche Dauer eines Bursts sowohl die Ausatmung, die Einatmung als auch die beiden Pausen zwischen diesen Geräuschen umfasst. Praktisch ergibt sich das Problem, dass vor allem bei stärkerem Hintergrundrauschen die Einatmung häufig weder hörbar noch auf dem Spektrogramm erkennbar ist. Daher wird die Zeitdauer eines Bursts von Beginn einer Ausatmung bis zum Beginn der darauf folgenden Ausatmung definiert und somit allein von den Ausatmungsgeräuschen auf die Bursts geschlossen. Implizit wird somit eine Einatmung zwischen zwei Ausatmungen angenommen.
- **Cry:** Die gesamte klangliche Antwort zu einem spezifischen Stimulus. Eine Gruppe mehrerer Cry-Units.[17, S. 61] In dieser Arbeit wird ein *Cry* auch als **Cry-Segment** bezeichnet, um Verwechslungen zu vermeiden.

Cry-Units werden von H Golub und M Corwin in eine der drei folgenen Kategorien eingeordnet, bezeichnet als *Cry-Types*: [17, S. 61 - 62]

- **Phonation** beschreibt eine Cry-Unit mit einer „vollen Vibration der Stimmbänder“ und einer Grundfrequenz zwischen 250 und 700 Hz. Entspricht umgangssprachlich einem Weinen mit einem „klaren, hörbaren Ton“.
- **Hyper-Phonation** beschreibt eine Cry-Unit mit einer „falsetto-artigen Vibration der Stimmbänder“ mit einer Grundfrequenz zwischen 1000 und 2000 Hz. Entspricht umgangssprachlich einem Weinen mit einem „sehr hohen, aber klar hörbaren Ton“.
- **Dysphonation** beschreibt eine Cry-Unit ohne klar feststellbare Tonhöhe, produziert durch Turbulenzen an den Stimmbändern. Entspricht umgangsprachlich dem „Brüllen oder Krächzen“.

Die folgenden weiteren Eigenschaften können für einzelne Cry-Units extrahiert werden:

- **Duration:** Die zeitliche Dauer der Cry-Unit.
- **Duration of Inspiration:** Die zeitliche Dauer der Pause zwischen zwei Cry-Units.

- **Grundfrequenz:** Für eine Cry-Unit kann die durchschnittliche, die höchste und die niedrigste Grundfrequenz sowie die Varianz festgestellt werden.
- **Frequenz der Formanten:** Wie bei der Grundfrequenz kann der Durchschnitt, das Maximum, Minimum etc. für eine Cry-Unit berechnet werden.
- **Ratio2:** Verhältnis zwischen den Energien der Frequenzen unterhalb von 2000 Hz zu den Frequenzen oberhalb von 2000 Hz
- **Cry-Mode Changes:** Häufigkeit des Wechsels des Cry-Modes innerhalb einer Cry-Unit.
- **Amplitude:** Die Lautstärke der Cry-Unit, gemessen in Dezibel. [23, S. 85] [10, S. 156]

Golub et al. haben weiterhin eine Reihe von Features vorgestellt, die das zeitliche Verhalten der Grundfrequenz und der harmonischen Obertöne innheralb einer Cry-Unit beschreiben. [17, S. 73]

- **Pitch of Shift:** Grundfrequenz nach einem schnellen Anstieg zu Beginn der Cry-Unit
- **Glide:** Kurzes, starkes ansteigen der Grundfrequenz
- **Glottal Roll:** Dysphonation, die häufig am Ende einer Cry-Unit nach einem Abfall der Grundfrequenz beobachtet wird.
- **Vibrato:** Mehr als vier starke Schwankungen der Grundfrequenz innerhalb einer Cry-Unit.
- **Melody-Type:** einer Cry-Unit. Meist: fallend, steigend/fallend, steigend, fallend/-steigend, flach.
- **Continuity:** Verhältnis zwischen stimmhaften und nicht-stimmhaften Bereichen der Cry-Unit
- **Double Harmonic Break:** Das Aufkommen einer zweiten Serie von harmonischen Obertönen zwischen den eigentlichen harmonischen Obertönen der Cry-Unit.
- **Biphonation:** Das Aufkommen einer zweiten Grundfrequenz mit eigenen harmonischen Obertönen zusätzlich zu der eigentlichen Grundfrequenz.
- **Noise Concentration:** Starke Energiespitzen zwischen 2000 und 2300 Hz.
- **Furcation:** Plötzliches Aufteilen der Grundfrequenz und harmonsichen Obertöne in mehrere, schwächere Obertöne.

Abbildung 2.14 visualsiert diese Grundfrequenz bezogenen Features in einem schematisch dargstellten Spektogramm.

Die folgende Features werden in Bezug auf das gesamte Cry-Segment, oder zumindest auf eine Menge aufeinander folgender Cry-Units berechnet:

- **Cry Latence:** Zeit zwischen Stimulus, wie zum Beispiel einem Nadelstich, und erster Cry-Unit.
- **Utterances:** Anzahl der Cry-Units im Segment.
- **Short Utterances:** Anzahl stimmloser Cry-Units im Segment.
- .... und statistische Auswertungen bezüglich aller oben genannten Features, die sich auf

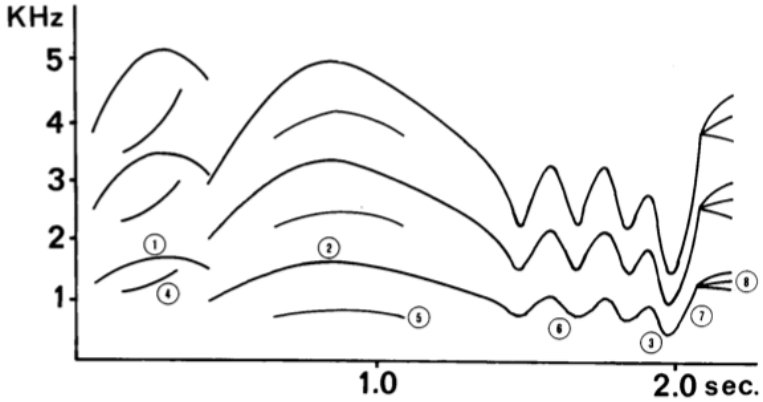

Abbildung 2.14: (1) Pitch of Shift (2) Maximale Grundfrequenz (3) Minimum der Grundfrequenz (4) Biphonation (5) Double Harmonic Break (6) Vibrato (7) Glide (8) Furcation [37, S. 142]

eine Cry-Unit beziehen, wie beispielsweise der Durchschnitt aller durchschnittlichen Tonhöhen, Anzahl des Vorkommens bestimmter Melodiekonturen, Varianz der Länge der Cry-Units etc.[23, S. 85]

Verschiedene Krankheitsbilder wurden in Zusammenhang mit dem vorkommen bestimmter Features des Cry-Segmentes gebracht. So wurde eine Korrelation zwischen dem Anstieg der durchschnittlichen Grundfrequenz, häufiger Biphonation und geringer Duration in Zusammenhang mit Gehirnschäden gebracht. Tendenziell niedrige Grundfrequenzen zeigen eine Korrelation mit Trisomie 13, 18 und 21[23, S. 85]

2.3.2 Diskussion

Bis heute bleibt die Analyse von kindlichen Lautäußerungen weitestgehend unstandartisiert [37, S. 142]:

- Es gibt keine Einigung darüber, welche der zahlreichen vorgestellten Eigenschaften die wichtigsten sind. Beispielsweise konzentrierten sich Golub et al. [17] vermehrt auf die Erkennung von Mustern im Melodieverlauf, Zeskind et al. auf zeitliche Eigenschaften. [34]. Die Eigenschaft, die am häufigsten mit Schmerz, Krankheiten und sonstigen Abnormalitäten in Verbindung gebracht wird, ist eine abnormal hohe oder niedrige Tonhöhe. Bei einigen Features, die vor allem von Golub et al. verwendet wurden [17], ist nicht einmal gesichert, ob es sich nicht doch um technische Artefakte der damals verwendeten Analogtechnik handelt. [23, S. 84 - 85]
- Zusammenhänge, die zwischen bestimmten Eigenschaften des Weinens und bestimmten Krankheitsbildern festgestellt wurden, haben häufig eine hohe Spezifität, aber niedrige Sensitivität. So wurde zum Beispiel festgestellt, dass Kinder, die am plötzlichen Kindstot verstarben, fast immer eine Erhöhung der Frequenz des ersten Formanten in Verbindung mit häufigen Cry-Mode-Changes zeigen. Viele Babys, die nicht am plötzlichen Kindstot versterben, zeigen jedoch die selben Merkmale.[23, S. 85]
- Selbst, wenn in verschiedenen Studien die selbe Eigenschaft verwendet wird, wie zum Beispiel die durchschnittliche Tonhöhe, ist nicht standardisiert, wie dieses exakt zu

berechnen ist. Mit „durchschnittliche Tonhöhe des Segmentes“ kann gemeint sein: (1) die Durchschnittliche Tonhöhe, errechnet aus den durchschnittlichen Tonhöhen der der Cry-Units (2) Die durchschnittliche Tonhöhe aller festgestellten Tonhöhen (3) die durchschnittliche Tonhöhe nur von Ausatmungslauten etc.

- Golub et al. behaupten, bereits in den achziger Jahren ein System zur computergestützten und voll automatisierten Analyse von Cry-Segmenten implementiert zu haben. Das System nimmt (1.) eine Audioaufnahme, gespeichert auf einer Kasette an, (2.) berechnet Formanten, Grundfrequenz und Amplitude gegen die Zeit, (3.) samplt die Grundfrequenz-Kontur (4.) berechnet insgesamt 88 akkumulierte Features für das gesamte Segment und (5.) zieht Schlussfolgerungen aus den 88 Features, wie zum Beispiel die Diagnose einer bestimmten Krankheit.[17, S. 75 - 76] Abseits der kurzen Erwähnung der Existenz dieser “Mutter aller automatisierten Analysesysteme für das Weinen von Babys“ konnte der Autor dieser Arbeit keine Implementierungsdetails oder sonstige genaueren Ausführungen finden, welche für diese Arbeit von höchstem Interesse gewesen wären.

2.4 Klassifizierung und Regression

Klassifzierung und Regression sind Teilgebiete des Wissenschaftsgebietes des *Überwachten Lernens*, einem Teilgebiet des Wissenschaftsgebietes des *maschienellen Lernens*. Das Ziel der Überwachten Lernen ist es, ein *Prädikator* (*Modell*) zu entwerfen, der aus den Eigenschaften einer Instanz dessen Kategorie oder Wert ableiten kann. Im Zusammenhang mit der Schreiforschung könnte eine Instanz eine Baby sein, dessen Eigenschaften (1.) das Gewicht und (2.) die Augenfarbe ist. Der Prädikator hat nun die Aufgabe, aus diesen beiden Eigenschaften eine Klasse abzuleiten, wie zum Beispiel das Geschlecht des Babys, oder einen Wert, wie beispielsweise das Alter. Das Lernen basiert dabei aus dem Generalisieren einer Liste von Beispielen, die der Algorithmus zur Verfügung gestellt bekommt. In diesem Zusammenhang wäre dies eine Liste an Babys, bei der für jede Instanz das Geschlecht oder das Alter bereits bekannt ist. Der Algorithmus versucht nun, diese Beispiele soweit zu Verallgemeinern, dass er für neue, bisher unbekannte Babys die Klasse oder den Wert korrekt voraussagen kann.[27, S. 6 - 7]

Eine Instanz x ist ein Vektor $x = (f_1 \in F_1, \ldots, f_n \in F_n)$. f_i wird in diesem Zusammenhang als *Eigenschaft*, *Feature* oder *Attribut* bezeichnet werden. In Bezug auf das eben genannte Beispiel wäre das erste Feature $F_1 = \textit{Gewicht}$ und das zweite Feature $F_2 = \textit{Augenfarbe}$. Eine Instanz wäre in diesem Fall ein Tupel mit zwei beliebigen Werten dieser Attribute, wie zum Beispiel $x = (3\,\text{kg}, \text{Blau})$). Features, die einen kontinuierliche Wertebereich mit einem quantitativem Charakter haben, wie zum Beispiel das Gewicht, werden als *kontinuierliche* Features bezeichnet. Features, die einen diskreten Wertebereich mit einem qualitativem Charakter haben, wie zum Beispiel die Augenfarbe, werden als *diskrete* Features bezeichnet. Die Menge aller möglichen Kombination der Features $F_1 \times, \ldots, F_n$ wird als *Feature-Raum* bezeichnet. Der Trainings-Datensatz D_{Traing} besteht aus einer Liste an Instanzen, wobei für jede Instanz die Kategorie oder der Wert, gemeinsam Bezeichnet als *Output* oder *Target* $y \in Y$, bekannt ist. Y bezeichnet die Menge aller möglichen Outputs des Problems. Das heißt, $D_{Training} = \big((x_1, t_1), \ldots (x_N, t_N) \big)$. Der Prädiktor P ist nun eine Funktion, die von einer Instanz auf den Output abbildet, also $P : X \mapsto Y$. Die Fehlerfunktion E berechnet, wie häufig sich der Prädiktor bei der Bestimmung der Targets eines bisher bekannte

oder unbekannten Trainings-Datensatzes D_{Test} irrt. Der Test- und der Trainingsdatensatz können die selben Instanzen, teilweise die selben oder gar keine gemeinsamen Instanzen beinhalten.[27, S. 6 - 7, 18 - 19] [7, S. 8 - 9]

Bei der **Klassifizierung** wird eine Target als *Klasse* bezeichnet. Die Menge aller möglichen Klassen eines bestimmten Problems $Y = \{y_1, \ldots, y_n\}$ ist dabei diskret und hat einen *qualitativen* Charakter. Das heißt, dass keine Klasse „besser“ oder „höher“ ist als eine andere. Ein Beispiel für ein Klassifizierungsproblem wäre die also die Ableitung des Geschlechtes für eine Instanz, also $Y = \{m, w\}$. Der Prädiktor wird in diesem Fall als Klassifikator C bezeichnet.[1] [11, S. 28, 127]

Bei der Regression *Regression* ist die Menge der möglichen Targets eines bestimmten Problem *kontinuierlich* und hat einen „quantitativen Charakter. Das heißt, es kann eine interne Ordnung in der Menge der Outputs festgelegt werden. Ein Beispiel für ein Regressionsproblem wäre die also die Ableitung des Alters des Babys, also $Y = \{0, \ldots, 130\}$. De Prädiktor wird in diesem Fall auch als *Regressor* R bezeichnet.[7, S. 24] [27, S. 8] [11, S. 28]

Es gibt eine Vielzahl an Algorithmen zum Finden des Klassfikators oder Prediktors. Welcher Algorithmus der „beste“ ist, das heißt für einen Test-Datensatz eine möglichst hohe *Genaugigkeit* oder einen möglichst geringen *Klassifikationsfehler* erzeugt, ist abhängig von der konkreten Problemstellung. Auf die Bestimmung der Genauigkeit wird weiter in Kapitel 2.4.2 eingegangen. Ein Algorithmus, der in dieser Arbeit zur Klassifizierung eingesetzt wird, ist der *ID3*-Algorithmus, welcher genauer in Kapitel 2.4.1 beschrieben wird.

2.4.1 ID3

Es gibt drei Algorithmen zur Erzeugung von Entscheidungsbäumen, die weitreichende Einsatz finden: *ID3*, *C.45* und *CART*, wobei die letzteren Erweiterungen der grundlegenden Idee des *ID3*-Algorithmus darstellen. Daher wird an dieser Stelle zuerst der *ID3*-Algorithmus vorgestellt.

Es wird zunächst davon ausgegangen, dass alle Features diskret und nicht kontinuierlich sind. Tabelle 2.3 gibt einen Beispieldatensatz, an dessen Beispiel ein Classificator mit Hilfe des ID3 erzeugt wird. Es geht ähnlich dem Beispiel aus Tabelle **??** um die Frage, ob Federball-Spielen abhängig von Temperatur und Tageszeit Spaß macht, nur sind in diesem Fall alle Features diskret.

Tabelle 2.3: Beispieldatensatz D für die Kassfikation mit ID3

x_i	Temperatur	Tageszeit	c_i = Spaß?
x_1	warm	Tag	Ja
x_2	kalt	Tag	Ja
x_3	normal	Nacht	Nein
x_4	kalt	Nacht	Nein
x_5	normal	Tag	Ja
x_6	warm	Nacht	Ja

[1] In vielen Quellen werden die Begriffe *Klassifizierung* und *Klassifikation* inkonsitent verwendet. Die *Klassifizierung* ist ein Prozess, dessen Ergebnis die *Klassifikation* ist. Daher wird von einem *Klassifizierungs-Algorithmus* gesprochen, da sich der Algorithmus auf den Prozess des Klassifizerens konzentriert, aber vom *Klassifikationsfehler*, da der Fehler des Ergebnisses der Klassifzierung bestimmt wird.

Abbildung 2.15 zeigt einen Klassifikator, den der ID-3 Algorithmus für diesen Datensatz baut. Es handelt sich um einen Entscheidungsbaum. In Jedem Knoten steht ein Feature, welches einen Ast für jeden möglichen Wert dieses Features bildet. In den Blättern stehen die Klassen.[27, S. 134]

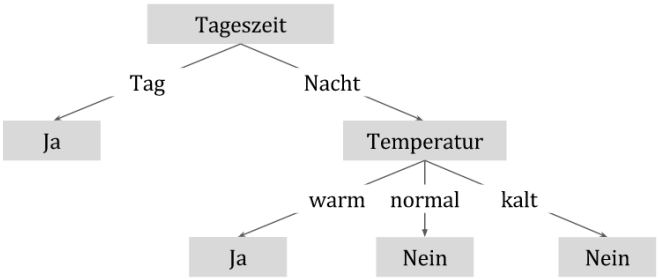

Abbildung 2.15: Entscheidungsbaum, der durch den ID3-Algorithmus für den Datensatz aus Beispiel 2.3 erzeugt wurde.

Der Entscheidungsbaum lässt sich in eine Reihe von `if ... then ...` -Regeln transformieren. Jeder Weg von der Wurzel bis zu einem Blatt ergibt eine Entscheidungsregel, bei der Feauture-Werte der betretenen Kanten konjunktiv Verknüpft werden und die Klasse implizieren. Die Entscheidungsregeln für den Baum aus Abbildung 2.15 sind: [27, S. 134]

- `if` *Tageszeit* = *Tag* `then` *Spaß* = *Ja*
- `if` *Tageszeit* = *Nacht* `and` *Temperatur* = *warm* `then` *Spaß* = *Ja*
- `if` *Tageszeit* = *Nacht* `and` *Temperatur* = *normal* `then` *Spaß* = *Nein*
- `if` *Tageszeit* = *Nacht* `and` *Temperatur* = *kalt* `then` *Spaß* = *Nein*

Der Klassifikator, das heißt der Entscheidungsbaum, wird beim ID3 Algorithmus nach folgenden Muster erstellt: Der Baum wird Top-Down erzeugt, dass heisst beginnend bei der Wurzel bis zu den Blättern. Da in jedem Knoten genau ein Feature aufgespalten wird, wird an der Wurzel die Frage gestellt „*Welches Feature sollte zuerst getestet werden?*“. Um diese Frage zu beantworten, wird jedes Feature einem statistischen Test unterzogen und festzustellen, wie „gut“ es zur Klassifikation der Trainings-Daten beiträgt. Das „beste“ Attribut wird ausgewählt und als Wurzel festgelegt. Nun wird ein Kind für jeden möglichen Wert des Features gebildet. Der Datensatz des Elternknotens wird in disjunkte Teilmengen aufteilt, wobei jedes Kind die Untermenge erhält, die den jeweiligen Feature-Wert bestitzt. Daraufhin beginnt für jedes Kind der Prozess des Auswählen des „besten“ Attributes von vorn. Ein Kind wird dann zu einem Blatt, wenn seine Teilmenge an Daten nur noch aus Instanzen einer Klasse besteht und somit kein weiteres Aufteilen notwendig ist.[28, S. 55]

Das Wort „gut“ wurd ein dieser Beschreibung in Anführungsstrichen geschrieben, da es subjektiv ist und quantifiziert werden muss. Zur Quantifizierung der Information wird die Entropie nach Formel 2.20 als Hilfsmittel definiert. p_i ist die Wahrscheinlichkeit, dass in einem Datensatz D eine Instanz mit der Klasse $i \in C$ angetroffen wird.

$$H(p) = -\sum_{i \in C} p_i \cdot \log_2 p_i \tag{2.20}$$

Die Entropie quantifiziert die *Unreinheit des Datensatzes*. Angenommen, ein Datensatz hat zwei Klassen, $C = \{+, -\}$. Existiert der gesamte Datensatz nur aus einer der beiden

Klasse, ist die Entropie $-p_+ \log_2 p_+ - p_- \log_2 p_- = -1log_2 1 - 0log_2 0 = 0$. Das heißt, dass die *Unreinheit des Datensatzes* 0 beträgt. Ist die *Unreinheit des Datensatzes* hingegen maximal, das heißt es liegen exakt 50% positive und 50% negative Samples vor, ist die Entropie $-p_+ \log_2 p_+ - p_- \log_2 p_- = -0.5log_{0.5} 1 - 0.5log_2 0.5 = 1$. [27, S. 135]

Es ist das Attribut in einem Knoten zu wählen, welches den höchsten *Informatoinsgewinn* gewährleistet, das heißt, zu einer bestmöglichen *Reinheit* bei der alleinigen Unterteilung des Datensatzes auf Basis dieses Attributs führt. Der informationsgewinn eines Features f für den Datensatz D wird nach Formel 2.21 definiert. v sind alle möglichen Werte dieses Features. $|D|$ beschreibt die Anzahl an Instanzen des Datensatzes. D_v ist die Untermenge an Instanzen, die für das Feature f den Wert v besitzen.[27, S. 136 - 137]

$$\text{Gain}(D, f) = H(D) - \sum_{v \in dom(f)} \frac{|D_v|}{|D|} H(D_v) \tag{2.21}$$

Für das Beispiel aus Tabelle 2.15 ergibt sich für den ersten Test folgende Berechnung des Informationsgewinnes der beiden Features *Temperatur* und *Tageszeit*. Da die Tageszeit den höheren Informationsgewinn gewährleistet, wird dieses Features in der Wurzel gewählt.

$$H(D) = -p_+ \log_2 p_+ - p_- \log_2 p_- = -\frac{4}{6}\log_2(\frac{4}{6}) - \frac{2}{6}\log_2(\frac{2}{6}) = 0.91 \tag{2.22}$$

$$Gain(D, Tageszeit) = 0.91 - \Big(\overbrace{\frac{3}{6} \cdot (-\frac{3}{3}\log_2\frac{3}{3} - -\frac{0}{3}\log_2\frac{0}{3})}^{Tag} \underbrace{\frac{3}{6} \cdot (-\frac{1}{3}\log_2\frac{1}{3} - -\frac{2}{3}\log_2\frac{2}{3})}_{Nacht} \Big) = 0.86 \tag{2.23}$$

$$Gain(D, Temperatur) = 0.91 - \Big(\overbrace{\frac{2}{6} \cdot (-\frac{2}{2}\log_2\frac{2}{2} - -\frac{0}{2}\log_2\frac{0}{2})}^{warm} \overbrace{\frac{2}{6} \cdot (-\frac{1}{2}\log_2\frac{1}{2} - -\frac{1}{2}\log_2\frac{1}{2})}^{normal} \overbrace{\frac{2}{6} \cdot (-\frac{1}{2}\log_2\frac{1}{2} - -\frac{1}{2}\log_2\frac{1}{2})}^{kalt} \Big) = 0.66 \tag{2.24}$$

Algorithmus1 zeigt den Ablauf des ID-3 in Pseudocode. D ist die Menge aller Test-Examples, X ist die Menge aller Features, C ist die Menge aller Klassen, f_{parent} das Feature des momentanen Eltern-Knotens und v_{parent} der Wert des zum momentan konstruierten Knotens eingehenden Kante. [27, S. 139] [28, S. 56]

Algorithm 1 ID3-Algorithmus in Pseudocode

```
1: tree = {}
2: function ID3(D, X, C, f_parent, v_parent )
3:                                   ▷ If all Examples have the same label, return a leaf with that Label
4:     if ∀e ∈ D : ∃k ∈ C : e.c = k then
5:         tree = tree ∪ {(f_parent, v_parrent, k)}
6:         return
7:     else
8:                                                ▷ If there are no Features left to test, return a leaf with
9:                                              ▷ the most common Label of the Examples remaining in D
10:        if isEmpty(X) then
11:            tree = tree ∪ {(f_parent, v_parrent, most common Label in D)}
12:            return
13:        else
14:                          ▷ Choose the feature that maximizes the Information-Gain to be the next node
15:            f_best = max_{f∈X} Gain(D, f)
16:                                                                  ▷ Add a Branch to this node
17:            tree = tree ∪ {(f_parent, v_parrent, f_best)}
18:                                                  ▷ Remove the feature from the set of features
19:            X_/f ← X/f_dom
20:            for v ∈ f_best do
21:                ▷ Calculate the new Dataset D_/f by removing all instances with the corresponding value
22:                D_/f ← ∀e ∈ D : e.f_best = v
23:                                                              ▷ Recursivly call the algorithm
24:                ID3(D_/f, X_/f, f_dom, v)
25:            end for
26:        end if
27:    end if
28: end function
```

Der ID3-Algorithmus hat folgende **Vorteile**:

Kurze Entscheidungsbäume Der Klassifizierer versucht, möglichst kurze Entscheidungsbäume zu bauen, indem Features mit hohem Informationsgewinn bevorzugt werden. Dies ist eine Umsetzung von *Ocam's Razor*: „Bevorzuge die kürzeste Hypothese“

Verständlichkeit Der Klassifikator ist für den Menschen verständlich, da er sich in Regeln übersetzen lässt (im Gegensatz zu zum Beispiel Neuronalen Netzen). Es existiert die unbewiesene Hypothese, dass der Mensch bei der Klassifizierung intuitiv ähnlich vorgeht wie der ID3-Algorithmus.[28, S. 63 - 65]

Der ID3-Algorithmus hat folgende **Nachteile**

Nur Diskrete Werte Der Algorithmus akzeptiert keine kontinuierlichen Werte [28, S. 72]

Overfitting Der Algorithmus neigt zu *Overfitting*. Overfitting bedeutet, dass der erzeugte Klassfikator c zwar einen möglichst geringen Fehler in Bezug auf den *Trainings-Datensatz* hat, es jedoch einen anderen Klassifikator c' gibt, welcher in Bezug auf den Trainings-Datensatz einen höheren Fehler erzeugt, jedoch einen geringeren Fehler als c in Bezug auf *alle möglichen Instanzen dieses Typs* erzeugt. Anders formuliert bedeutet Overfitting, dass der Klassifikator den Trainings-Datensatz „auswendig gelernt hat“ und nicht mehr genügend generalisiert, um auf im Training nicht enthaltene Instanzen angewandt werden zu können. Overfitting im Zusammenhang mit dem ID-3 Algorithmus wird durch *Rauschen im Trainings-Datensatz* bedingt. Es gibt keinen festen Beweis für das vorhandensein von Overfitting. Methoden zum Feststellen von Overfitting sind:

- Verwendung eines seperaten Test-Datensatzes, welcher bestätigt, dass der für den Trainings-Datensatz erzeugte Klassifikationsfehler auch bei bisher unbekannten Instanzen erzeugt wird.
- Verwendung von Statistischen Tests, die eine signifikante Reduktion des Klassifikationsfehlers bei Erweiterung des Entscheidungsbaumes beweisen.
- Expertenwissen über applikationstypischen Tiefen von Entscheidungsbäumen.[28, S. 66 - 70]

Lokale Maxima Der Algorithmus bevorzugt greedy Attribute, die zum Zeitpunkt der Berechnung den höchsten Inforamtionsgewinn gewährleisten. Dabei besteht die Gefahr, dass der Algorithmus in ein lokales Maximum läuft.[28, S. 66 - 70]

2.4.2 Gütemaße binärer Klassifikatoren

Ein binärer Klassifikation ist eine, bei dem es nur zwei Klassen gibt, das heißt $|C| = 2$. Applikationsabhängig werden die beiden Klassen als *Positive* und *Negative*, 1 und 0 oder *True* und *False* beschrieben. Eine Klassifikation, bei der ein tatsächliches Positive richtig als Positive vorhergesagt wird, spricht man von einem *True Positive* [TP]. Wird hingegen ein tatsächliches Positive fälschlicherweise als Negative vorhergesagt, spricht man von einem *False-Negative* [FN]. Das System wird entsprechend für die Klassifikation tatsächlicher Negatives angewandt und ergibt. *True-Negatives* [TN] und *False-Positives* [FP]. Die *Confusion Matrix* in Abbildung 2.16 gibt eine Übersicht über die vier möglichen Klassifikations-Ergebnisse. [21, S. 213 - 214]

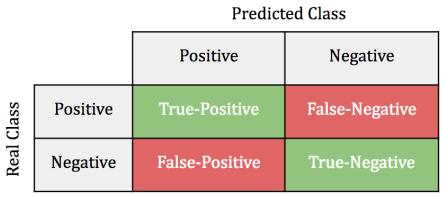

Abbildung 2.16: Confusion-Matrix (nach: [21, S. 214])

Die insgesamte Güte einer Klassifikation wird durch die *Accuracy* nach Formel 2.25 bestimmt. Eine Accuracy von 100% bedeutet, dass *alle* Instanzen richtig klassifiziert werden, eine Accuracy von 50% bedeutet, dass die Hälfte aller Instanzen richtig klassifiziert werden, was der Güte einer rein zufälligen Wahl entspricht. [21, S. 214]

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{2.25}$$

Die Accuracy beziffert die insgesamte Performance des Klassifikators, gibt jedoch keinen Aufschluss darüber, ob der Klassifikator eher eine Tendenz zur falschen Klassifizierung von Positives oder Negatives hat. Bei einer Datenbank mit der selben Anzahl an Positives und Negatives kann eine Accuracy von 50% beispielsweise dadurch entstehen, dass *alle* Instanzen als Positives markiert werden, also sowohl die Positives richtigerweise als Positives, aber die Negatives fälschlicherweise ebenfalls als Positives. Im Umgedrehten Fall ergibt die Klassifizierung aller Instanzen als Negatives ebenfalls eine Accuracy von 50%. In einem dritten Fall irrt sich die Klassifikator gleich oft bei der Einordnung der Negatives

und Positives. Die Maße *Sensitivity* und *Specificity* geben Aufschluss über die Güte der Klassifikation hinsichtlich der Positives und Negatives. Die *Sensitivity*, auch bezeichnet als *True-Positive-Rate*, bemisst den Anteil tatsächlicher Positives, die auch als solche erkannt wurden, nach Formel 2.26. Eine Sensitivity von 100% bedeutet, dass alle Positives durch den Klassifikator erkannt wurden. Die Erkennungsrate der Negatives hat keinen Einfluss auf die Sensitivity. Eine hohe Sensitivity lässt sich somit „einfach“ erzielen, in dem man *alle* Instanzen immer als Positives klassifiziert. Die Specificity nach Formel 2.27 bestimmt analog zur Sensitivity den Anteil der korrekt als Negatives bestimmten Instanzen. Ein Klassifikator, der alle Instanzen als Positives markiert, hat zwar eine Sensitivity von 100%, aber eine Specificity von 0%. Ergeben zwei verschiedene Klassifikationsmodelle sehr ähnliche Accuracies, hilft die Bestimmung der Sensitivity und Specificity bei der Auswahl des für den Anwendugnsfall Adäquteren Klassifikators. So ist beispielsweise bei der Bestimmung von schweren Krankheiten eventuell ein Klassifikator mit höherer Sensitivity wünschbar, um die Wahrscheinlichkeit zu minimieren, dass die entsprechende Krankheit nicht erkannt wird. [24] [21, S. 222]

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2.26}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2.27}$$

3 Konzept zur Visualisierung von Schmerz Scores aus akustischen Signalen

In Kapitel 2.1 wurde vorgestellt, wie die Schmerzdiagnose mit Hilfe von Pain Scales durch medizinisches Fachpersonal durchgeführt wird. Es wurde eine Reihe an Pain Scales vorgestellt und dabei insbesondere der Schmerzindikator „Weinen“ beleuchtet. Unabhängig von der konkret eingesetzten Pain Scale wird in jedem Fall das Baby für eine bestimme Zeit beobachtet, für jeden Indikator, wie das Weinen oder der Gesichtsausdruck, Punkte vergeben, diese Aufsummiert und aus der Summe der insgesamte Schmerz Score bestimmt.

In Kapitel 2.3 wurde vorgestellt, wie in der klassischen Schreiforschung mit Hilfe von Mitteln der Signalverarbeitung das Weinen von Babys tiefergehend Analysiert wurde. Es wurde die Möglichkeit gezeigt, aus den objektiv messbaren Eigenschaften des Weinens eines Babys Rückschlüsse auf dessen Zustand gemacht werden können.

Ziel dieser Arbeit ist der Entwurf eines Systems zur automatisierten Feststellung und Visualisierung von Pain Scores beliebiger Pain Scales mit dem Fokus auf den Schmerzindikator „Weinen“. Das System muss folgenden Anforderungen erfüllen:

1. Das System muss dazu in der Lage sein, aus den akustischen Eigenschaften des Weinens eines Babys den Schmerz Score bezüglich einer Pain Scale abzuleiten.
2. Das System muss dazu in der Lage sein, die abgeleiteten Schmerz Scores zu visualiseren.
3. Das System muss dazu in der Lage sein, beliebige Pain Scales einzubinden.
4. Die System muss dazu in der Lage sein, die Analyse auch bei nicht-optimalen akustischen Bedingungen durchzuführen.
5. Das System muss dazu in der Lage sein, die Analyse kontinuierlich durchzuführen.

Der **Input** des Systems ist folglich ein Audiosignal, welches kontinuierlich in das System gegeben wird. Der **Output** ist eine Visualisierung der abgeleiteten Pain Score, welche kontinuierlich erzeugt wird.

In Kapitel 3.1 wird zunächst ein Überblick über einige Veröffentlichungen gegeben, in denen ebenfalls Konzepte zur Analyse und Auswertung kindlicher Lautäußerungen vorgestellt wurden. In Kapitel 3.2 wird daraufhin das in dieser Arbeit entworfene Konzept in Form einer Verarbeitungs-Pipeline vorgestellt. Diese Verarbeitungs-Pipeline kombiniert das Vorgehen bei der Schmerzdiagnostik auf Basis der Pain Scales, Methoden der klassischen Schreiforschung sowie einige Ideen nun vorgestellten Veröffentlichungen.

3.1 Literaturüberblick

Der größere Teil der Veröffentlichungen, die sich in das Feld der Analyse von Audioaufnahmen Neugeborener einordnen lassen, stellten Algorithmen zur Klassifizierung einzelner Cry Units vor, entweder bezüglich der Weinursache (Hunger, Angst, Schmerz, ...) oder zur Diagnose bestimmter Krankheiten. Diese Methoden waren in den meisten Fällen nicht für die kontinuierliche Analyse geeignet, sondern hatten das Ziel, eine gegebenen Cry-Unit mit einer möglichst hohen Genauigkeit bezüglich des jeweiligen Sachverhaltes zu klassifizieren. Probleme wie Hintergrundrauschen, Berechnungsaufwand oder kontextuelle Informationen haben eine untergeordnete Rolle gespielt. Beispiele für solche Veröffentlichungen sind die von Abdulaziz et al. [44] oder Fuhr et al. [41].

Várallyay stellte in seiner Dissertation „Analysis of the Infant Cry with Objective Methods“ [42] Methoden zur automatisierten Analyse kindlicher Lautäußerungen vor. Das primäre Ziel der Dissertation war die Erforschung der Unterschiede zwischen den Lautäußerungen gesunder und tauber Neugeborener. Die Algorithmen zur automatisierten Analyse der Audiosignale waren ein „Nebenprodukt“ zur schnelleren Datenauswertung. Die Auswertung musste nicht kontinuierlich erfolgen. In der vorgestellten Verarbeitungs-Pipeline wurde das Eingangssignal in Zeitfenster weniger Millisekunden zerlegt und jedes Fenster nach Entscheidungsregeln als *stimmhaft* oder *nicht stimmhaft* markiert. Die stimmhaften Signalfenster wurden zu *Segmenten* zusammengefasst (welche in Kapitel 2.3.1 als Cry-Units bezeichnet werden). Auf Basis der Segmente wurden Auswertungen bezüglich des Zeitbereiches (Durchschnittliche Segmentlänge, Pausenlängen etc.), des Frequenzbereiches (Grund-Frequenz, Formanten-Frequenzen etc.) und des Melodieverlaufes angestellt. Analysiert wurden Audioaufnahmen von Babys mit einer Länge von 10 bis 100 s. Aus den Auswertungsergebnisse stellte Varallyay die wichtigsten Unterscheidungsmerkmale zwischen tauben und gesunden Babys fest. In der Dissertation [42] wird ein Überblick über das Vorgehen und die Ergebnisse gegeben. Die Verarbeitungsschritte wurden detaillierter in einzelnen Veröffentlichungen beschrieben, wobei der Autor dieser Arbeit nur den Zugriff auf einige dieser Veröffentlichungen erhalten konnte.

Cohen et al. haben 2012 in der Veröffentlichung „Infant Cry Analysis and Detection“ [6] ein System zur Analyse der akustischen Signale von Neugeborenen vorgestellt. Dieses System klassifizierte die Audiosignale in eine der drei Klassen *Cry, No Cry* und *No Activity*. Die Klasse *Cry* bezeichnet Lautäußerungen, die eine potentiell Gefahr für das Baby anzeigen, wie z.B. wie Schmerz oder Hunger. Die Klasse *No Cry* bedeutete, dass das Baby zwar Laute von sich gibt, diese aber keine potentielle Gefahr anzeigen. Die Klasse *No Activity* bezeichnete keinerlei Lautäußerung. Die Verarbeitungs-Pipeline wurde detailliert vorgestellt und war für die kontinuierliche Verarbeitung mit einer gewissen Verzögerungszeit spezialisiert. Das Signal wird in überlappende *Segmente* à 10 s zerlegt. Die Stimmaktivität in den Segmenten wird algorithmisch festgestellt. Wenn Aktivität vorliegt, wird das Segment in Sektionen à 1 s zerlegt und die Stimmaktivität für jede Sektion gemessen. Wird genügend Stimmaktivität in einer Sektion festgestellt, wird die Sektion in *Frames* à 32 ms zerlegt und Attribute für jeden Frame errechnet. Mit Hilfe von Entscheidungsregeln werden die Frames in *Cry, No-Cry* oder *No Activity* klassifiziert, wobei kontextuelle Informationen der umliegenden Frames mit einbezogen werden. Aus den Klassen der Frames wird auf die Klasse der Sektion geschlossen, und aus den Klassen der Sektionen auf die Klasse des Segmentes. Das System hat mit den Anforderungen dieser Arbeit gemeinsam, dass ebenfalls die kontinuierliche Verarbeitung im Vordergrund steht. Der Nachteil an dieser Methode ist,

dass die zeitliche längste Einheit, für die die Klassifizierung vorgenommen wird, unflexibel auf 10 s festgelegt ist. Daher müsste diese Verarbeitungs-Pipeline abgewandelt werden, um anstelle der Ableitung der drei genannten Klassen einen Pain Score ableiten zu können, die einen längeren Beobachtungszeitraum als 10 s benötigt.

Pal et al. haben 2006 in der Veröffentlichung „Emotion detection from infant facial experessions and cries“ [35] ein System vorgestellt, welches aus den akustischen Eigenschaften des Weinens die Emotion ableitet. Die zu erkennenden Emotionen sind *Traurigkeit, Wut, Hunger, Angst und Schmerz*. Es wird nicht erwähnt, ob die Analyse kontinuierlich oder nicht kontinuierlich erfolgt. Bei der Verarbeitung der akustischen Signale werden die Attribute *Grundtonhöhe* und die *Frequenz der ersten drei Formanten* extrahiert und mit einem Klassifizierungsalgorithmus klassifiziert. Es wurde nicht beschrieben, inwiefern die Eigenschaften aus kurzen Signalfenstern oder längeren Signalabschnitten errechnet werden, welche Vorverarbeitungsschritte angewandt werden und ob die Klassifizierung auf Ebene der Signalfenster oder über längere Zeitabschnitte hinweg geschieht.

Zamzi et al. haben 2016 in der Veröffentlichung „An Approach for Automated Multimodal Analysis of Infants' Pain“ [12] ein System zur automatisierten und kontinuierlichen multimodalen Analyse von Neugeborenen zur Ableitung des Schmerzes vorgestellt. Das System trägt den Namen *MPAS*. Der Schmerzgrad wird aus den Analyseergebnissen der monomodalen Schmerzindikatoren für *Gesichtsausdruck, Körperbewegung, Vitalfunktionen und Weinen* errechnet. Das System kommt der Aufgabenstellung dieser Masterarbeit am nächsten, da es ebenfalls um die Ableitung von Schmerz in einem multimodalen Verbund geht. Der Schmerz wurde hier „direkt“ abgeleitet, ohne den Weg über Pain Scales zu wählen. Während in der Veröffentlichung die Analyse der ersten drei genannten Schmerzindikatoren angekündigt wurde, wurden daraufhin die Methoden zur Analyse der akustischen Signale *nicht* erläutert. Auch die ersten Validierungsergebnisse beziehen sich nur auf den Gesichtsausdruck, die Körperbewegung und die Vitalfunktionen. Es ist nicht klar, ob die Miteinbeziehung akustischer Signale fallen gelassen wurde. Die Ausführungen konzentrieren sich dazu vermehrt auf die Methoden zur Kombination der Auswertungsergebnisse der monomodalen Schmerzindikatoren.

3.2 Verarbeitungs-Pipeline

In Kapitel 3.1 wurden verschiedene Konzepte vorgestellt, deren Zielstellungen ebenfalls die Analyse und Auswertung von Audioaufnahmen kindlicher Lautäußerungen ähnelt. Keines der präsentierten Konzepte eignet sich, um mit nur leichten Anpassungen übernommen werden zu können: Entweder wurden die Verarbeitungsschritte nicht für die kontinuierliche Verarbeitung konzipiert [44] [41] [42], nicht genügen abstrahiert, um für andere Klassifizierungen als die ursprünglich geplanten abgewandelt werden zu können [6], oder die Verarbeitungs-Pipeline wurde nicht vorgestellt. [35] [12].

In dieser Arbeit wurde die folgende Verarbeitungs-Pipeline entworfen. Sie wird in in Abbildung 3.1 visualisiert.

1. **Input:** Ein Audiosignal, das möglicherweise kindliche Lautäußerungen enthält. Es wird kontinuierlich hinzugegeben.
2. **Vorverarbeitung** (engl. *Pre-Processing*) des Signals.
3. **Voice Activity Detection**. Zunächst muss festgestellt werden, ob und wo in dem

Signal kindliche Lautäußerungen vorhanden sind. Ein Algorithmus zur Feststellung von Stimmaktivität, bezeichnet als Voice Activity Detection, untersucht das Signal und markiert die Cry-Units. Die gefunden Cry-Units bilden die Grundlage aller darauf folgenden Verarbeitungsschritte. Der vorgestellte Algorithmus kombinierte herkömmliche Methodne der Voice Activity Detection mit Ideen, die von Varallyay [42] vorgestellt wurden.

4. **Segmentierung** (engl *Segmenting*). Eine Pain Score wird nicht aus der Beobachtung einer einzigen, sondern einer Reihe von Cry-Units abgeleitet. In Kapitel 2.1.1 wurde gezeigt, dass bestimmte Pain Scores die Beobachtung über mehrere Minuten hinweg erfordern. Zu diesem Zweck werden die Cry-Units vorgruppiert. Da keines der in Kapitel 3.1 vorgestellten Konzepte Methoden zur Segmentierung vorstellte, wurde ein eigener Algorithmus für diese Aufgabe entworfen.
5. **Extrahierung von Eigenschaften und Ableitung der Schmerz Score** (engl. *Feature Extraction* und *Prediction of Pain Score*). Für jedes Segment werden Eigenschaften bezüglich des Weinens berechnet, wie zum Beispiel die durchschnittliche Tonhöhe, durchschnittliche Pausenlänge usw. Dieses Vorgehen implementiert Ideen der in Kapitel 2.3 vorgestellten klassischen Schreiforschung. Auf Basis dieser Eigenschaften wird die Pain Score abgeleitet.
6. **Output: Visualisierung** (engl. *Visualisation*) der abgeleiteten Schmerz Score. Es werden mehrere Varianten vorgeschlagen, welche die Höhe des Schmerz Score in seinem zeitlichen Verlauf auf Ampelfarben abbildet.

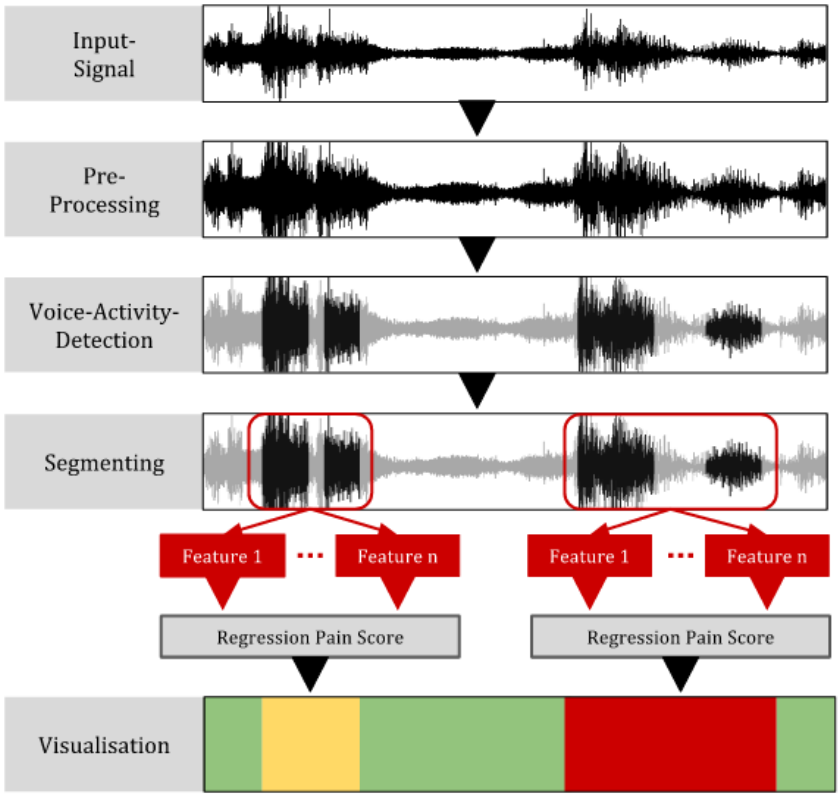

Abbildung 3.1: Überblick über die Verarbeitungs-Pipeline dieser Arbeit

4 Zusammenfassung

Literaturverzeichnis

[1] Tobias Kaufmann Beat Pfister. *Sprachverarbeitung*. Springer, Berlin, 2008.

[2] Arthur H Benade. Fundamentals of Musical Acoustics. 1976.

[3] Judy Bildner. *CRIES Instrument Assessment Tool of Pain in Neonates*. City of Hope Pain, 1997. Online unter `http://prc.coh.org/pdf/CRIES.pdf`.

[4] Richard Brown. The short time fourier transform, 2014. Online erhältlich unter: `http://spinlab.wpi.edu/courses/ece503_2014/12-6stft.pdf`.

[5] R Sisto & Giuseppe Buonocore Carlo Bellieni, Franco Bagnoli. Cry features reflect pain intensity in term newborns: An alarm threshold. *Pediatric Research*, 5:142–146, 1. Online unter `https://www.researchgate.net/publication/297827342_Cry_features_reflect_pain_intensity_in_term_newborns_An_alarm_threshold`.

[6] Rami Cohen and Yizhar Lavner. Infant Cry Analysis and Detection. In *27th Convention of Electrical and Electronics Engineers in Israel*. IEEE, 2012. Online unter `https://www.researchgate.net/publication/261116332_Infant_cry_analysis_and_detection`.

[7] Alin Dobra. Introduction to classification and regression, 2005. Online erhältlich unter: `https://www.cise.ufl.edu/~adobra/datamining/classif-intro.pdf`.

[8] H. Hollien & T Murry E Müller. Perceptual responses to infant crying: identification of cry types. *Journal of Child Language*, 1(1):89–95, 1974. Online unter `https://www.cambridge.org/core/journals/journal-of-child-language/article/perceptual-responses-to-infant-crying-identification-of-cry-types/4F0F8088116FCE381851D8D560697A5F`.

[9] Jan Hamers Eva Cignac, Romano Mueller and Peter Gessler. Pain assessment in the neonate using the Bernese Pain Scale for Neonates. *Early Human Development*, 78(2):125–131, 2004. Online unter `https://www.researchgate.net/publication/8485535_Pain_assessment_in_the_Neonate_using_the_Bernese_Pain_Scale_for_Neonates`.

[10] Barbara Fuller. Acoustic Discrimination of three Cry Types. *Nursing Research*, 40(3), 1991. Online erhältlich unter: `https://www.researchgate.net/publication/21125005_Acoustic_Discrimination_of_Three_Types_of_Infant_Cries`.

[11] Trevor Hastie Gareth James, Daniela Witten and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[12] Dmitry Goldgof Rangachar Kasturi Terri Ashmeade Ghada Zamzmi, Chih-Yun Pai and Yu Sun. An Approach for Automated Multimodal Analysis of Infants' Pain. In *23rd International Conference on Pattern Recognition*, Cancun, Mexico, 2016.

[13] Dmitry Goldgof Rangachar Kasturi Yu Sun Ghada Zamzmi, Chih-Yun Pai and Terri Ashmeade. Machine-based Multimodal Pain Assessment Tool for Infants: A Review,

2016. Online unter https://arxiv.org/ftp/arxiv/papers/1607/1607.00331.pdf.

[14] Ricardo Gutierrez-Osuna. Introduction to Speech Processing. Online unter http://courses.cs.tamu.edu/rgutier/csce689_s11/.

[15] Health Facts For You. *Using Pediatric Pain Scales Neonatal Infant Pain Scale (NIPS)*, 2014. Online unter https://www.uwhealth.org/healthfacts/parenting/7711.pdf und unter https://com-jax-emergency-pami.sites.medinfo.ufl.edu/files/2015/02/Neonatal-Infant-Pain-Scale-NIPS-pain-scale.pdf.

[16] Hodgkinson. Neonatal Pain Assessment Tool, 2012. Online unter http://www.rch.org.au/rchcpg/hospital_clinical_guideline_index/Neonatal_Pain_Assessment/#The%20Pain%20Assessment%20Tool.

[17] Michael J Corwin Howard L Golub. A Physioacoustic Model of the Infant Cry. In *Infant Crying - Theoretical and Research Perspectives*, chapter 3, pages 59 – 82. Plenung, 1985.

[18] Bonnie Stevens Huda Huijer Abu-Saad, Gerrie Bours and Jan Hamers. Assessment of pain in Neonates. *Seminars in Perinatology*, 2(5):402–416, 1998. Online unter https://www.ncbi.nlm.nih.gov/pubmed/9820565.

[19] Donna Geiss Laura Wozniak & Charles Hall Ivan Hand, Lawrence Noble. COVERS Neonatal Pain Scale: Development and Validation. *International Journal of Pediatrics*, 2010, 2010. Online unter https://www.hindawi.com/journals/ijpedi/2010/496719/.

[20] Bonnie J. Stevens K. J. S. Anand and Patrick J. McGrath. *Pain in Neonates and Infants*. Elsevier, 2007.

[21] Miroslav Kubat. *An Introduction to Machine Learning*. Springer, 2015.

[22] Barry Lester and Zachariah Boukydis. *Infant Crying: Theoretical and Research Perspectives*. Springer, 1985.

[23] A. Rebecca Neal Linda L. LaGasse and Barry M. Lester. Assessment of infant cry: Acoustic cry analysis and parental perception. *Mental retardation and developmental disabilities*, 11(1):83–93, 2005. Online unter https://www.ncbi.nlm.nih.gov/pubmed/15856439.

[24] Tze-Wey Loong. Understanding sensitivity and specificity with the right side of the brain. *BMJ*, 327(7417), 2003. Online unter https://www.ncbi.nlm.nih.gov/pmc/articles/PMC200804/.

[25] Michael Lutter. Speech production, 2015. Online erhältlich unter: http://recognize-speech.com/speech/speech-production.

[26] Robert Mannell. Acoustic theory of speech production, 2015. Online erhältlich unter: http://clas.mq.edu.au/speech/acoustics/frequency/acoustic_theory.html.

[27] Stephen Marsland. *Machine Learning - An Algorithmic Perspective*. Chapman & Hall / CRC, 2009.

[28] Tom M Mitchell. *Machine Learning*. WCB McGraw-Hill, 1997.

[29] Hans M Koot Dick Tibboel Jan Passchier & Hugo Duivenvoorden Monique van Dijk, Josien de Boer. The reliability and validity of the COMFORT scale as a postoperative pain instrument in 0 to 3-year-old infants. *Pain*, 84(2):367—-377, 2000. Online unter http://www.sciencedirect.com/science/article/pii/S0304395999002390

und unter.

[30] Sinno Simons Monique van Dijk and Dick Tibboel. Pain assessment in neonates. *Paediatric and Perinatal Drug Therapy*, 6(2):97–103, 2004. Online unter http://www.sciencedirect.com/science/article/pii/S0304395999002390.

[31] D L Neuhoff. *Signal and Systems I - EECS 206 Laboratory*. The University of Michigan, 2002. Online erhältlich unter: http://www.eecs.umich.edu/courses/eecs206/archive/spring02/ abgerufen am 11. Januar 2016.

[32] J L Mathew P J Mathew. Assessment and management of pain in infants. *Postgrad Med J*, 79:438–443, 2003. Online unter http://pmj.bmj.com/content/79/934/438.full.

[33] Steven Creech Patricia Hummel, Mary Puchalski and Marc Weiss. N-PASS: Neonatal Pain, Agitation and Sedation Scale – Reliability and Validity. *Pediatrics/Neonatology*, 2(6), 2004. Online unter http://www.anestesiarianimazione.com/2004/06c.asp.

[34] Susan Parker-Price & Ronald Barr Philip Zeskind. Rythmic organization of the Sound of Infant Cry. *Dev Psychobiol*, 26(6):321–333, 1993. Online unter https://www.ncbi.nlm.nih.gov/pubmed/8119482.

[35] Ananth N. Iyer Pritam Pal and Robert E. Yantorno. Emotion detection from infant facial experessions and cries. In *Acoustics, Speech and Signal Processing*. IEEE, 2006.

[36] R Ward & C Laszlo Qiaobing Xie. Automatic Assessment of Infants Levels-of-Distress from the Cry Signals. *IEEE Transanctions on Speech and Audio Processing*, 4(4):253–265, 1996. Online unter http://ieeexplore.ieee.org/document/506929/.

[37] Brian Hopkins & James Green Ronald Barr. *Crying as a Sign, a Symptom, and a Signal*. Mac Keith Press, 2000.

[38] J R Shayevitz & Shobha Malviya Sandra Merkel, Terri Voepel-Lewis. The FLACC: A Behavioral Scale for Scoring Postoperative Pain in Young Children. *Pediatric Nursing*, 23(3):293–7, 1996. Online unter https://www.researchgate.net/publication/13998379_The_FLACC_A_Behavioral_Scale_for_Scoring_Postoperative_Pain_in_Young_Children.

[39] Julius Smith. *Spectral Audio Signal Processing*. Center for Computer Research in Music and Acoustics (CCRMA), 1993. Online unter https://www.dsprelated.com/freebooks/sasp/Short_Time_Fourier_Transform.html.

[40] Steven W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, 1999. Online erhältlich unter: http://www.dspguide.com/pdfbook.htm.

[41] Henning Reetz & Carla Wegener Tanja Fuhr. Comparison of Supervised-learning Models for Infant Cry Classification. *InternatIonAl Journal of HealtH Professions*, 2015. Online unter https://www.degruyter.com/view/j/ijhp.2015.2.issue-1/ijhp-2015-0005/ijhp-2015-0005.xml.

[42] Gyorgy Ivan Varallyay. *Analysis of the Infant Cry with Objective Methods*. PhD thesis, Budapest University of Technology and Economics, 2009. Online erhältlich unter: https://pdfs.semanticscholar.org/5c38/b368dc71d67cbfabc3077a50536b086d8eec.pdf.

[43] P H Wolff. The role fo biological rhythms in early psychological development. *Bulletin of the Menninger Clinic*, 31(1):197–218, 1967.

[44] Syed Ahmad Yousra Abdulaziz, Sharrifah Mumtazah. Infant Cry Recognition System: A Comparison of System Performance based on Mel Frequency and Linear Prediction Coefficients. In *Information Retrieval & Knowledge Management*, 2010. Online unter `http://ieeexplore.ieee.org/document/5466907/`.

Appendices

Tabelle .1: Accuracy-Werte der Grenzwertfindung mit REPTree

$SNR_{Training}$	3 dB				50 dB				50+3 dB			
SNR_{Test}	3 dB	50 dB	7 dB*	Mean	3 dB	50 dB	7 dB*	Mean	3 dB	50 dB	7 dB*	Mean
Zeit	77.81%	79.02%	86.04%	80,96%	49.33%	94.70%	48.66%	64,23%	77.54%	92.47%	84.38%	84,80%
Freq	82.05%	89.28%	82.71%	84,68%	70.52%	94.37%	55.06%	73,31%	81.75%	91.22%	74.90%	82,62%
Ceps	88.98%	94.72%	92.96%	**92,22%**	86.83%	94.68%	92.83%	**91,45%**	88.98%	94.72%	92.96%	**92,22%**
Corr	80.45%	73.47%	84.89%	79,60%	73.07%	87.14%	77.98%	79,39%	77.90%	84.88%	82.84%	81,87%
Zeit+Freq	82.05%	89.28%	82.71%	84,68%	70.52%	94.37%	55.06%	73,31%	81.75%	91.22%	74.90%	82,62%
Zeit+Ceps	88.98%	94.72%	92.96%	**92,22%**	86.83%	94.68%	92.83%	**91,45%**	88.98%	94.72%	92.96%	**92,22%**
Zeit+Corr	80.45%	73.47%	84.89%	79,60%	49.33%	94.70%	48.66%	64,23%	80.32%	92.35%	88.22%	86,96%
Freq+Ceps	88.98%	94.72%	92.96%	**92,22%**	70.65%	94.75%	55.06%	73,49%	88.98%	94.72%	92.96%	**92,22%**
Freq+Corr	82.05%	89.28%	82.71%	84,68%	70.52%	95.60%	95.60%	87,24%	81.75%	94.42%	74.90%	83,69%

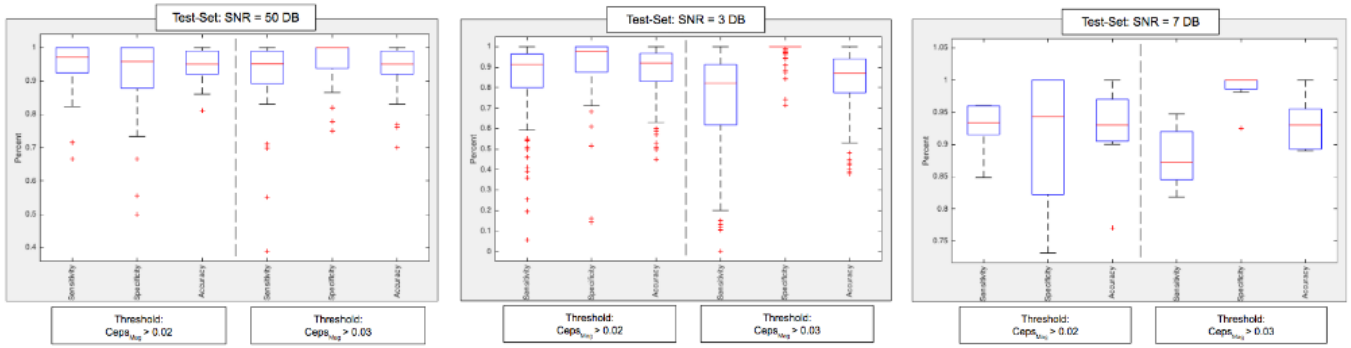

Abbildung .1: Boxplot-Auswertung über Sensitivity, Specificity und Accuracy der beiden VAD-Modelle