Visualisierung kontinuierlicher, multimodaler Schmerz Scors am Beispiel akustischer Signale

Masterarbeit

Franz Anders
HTWK Leipzig

Januar 2017

Abstract

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

2 Grundlagen der medizinischen Schrei-Forschung

2.1 Schmerz Scores

Bei erwachsenen Menschen wird der Schmerzgrad typischerweise durch eine Selbsteinschätzung des Patienten unter der Leitung gezielter Fragen des Arztes vorgenommen. Bei Kindern unter 3 Jahren ist diese Selbsteinschätzung nicht möglich. Schmerz drückt sich in Veränderungen des psychologischen, körperlichen und biochemischen Verhaltens des Säuglings aus. Die für den Arzt am leichtesten feststellbaren Verhaltensänderungen sind von außen wahrnehmbaren Merkmale, wie zum Beispiel ein Verkrampfen des Gesichtsausdrucks, erhöhte Körperbewegungen oder lang anhaltendes Weinen. Um eine weitesgehend objektive Schmerzfeststellung zu ermöglichen, wurden sogenannte *Pain-Scores* entwickelt, die durch ein Punktesystem den insgesamten Schmerzgrad des Babies quantifizieren.[22] Es existieren *eindimensionale* Pain-Scores, die den Schmerz nur Aufgrund der Beobachtung eines Merkmals beurteilen, so wie beispielsweise die reine Beurteilung des Gesichtsausdruckes. *Mehrdimensionale* (auch *multimodale*) Pain-Scores beziehen mehrere Faktoren in das Scoring mit ein.[1]. Tabelle 2.1 zeigt das Scoring-System „Neonatal Infant Pain Scale“(NIPS) als Beispiel für eine multimodale Pain-Score. Der Säugling wird anhand der aufgeführten Kategorien bewertet und alle vergebenen Punkte aufsummiert. Ein insgesamter Wert von > 3 zeigt Schmerz an, ein Wert von > 4 großen Schmerz.[10]

Tabelle 2.1: NIPS-Scoring

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	-

In den meisten mehrdimensionalen Scoring-Systeme werden die Schreigeräusche mit einbezogen. Tabelle 2.2 zeigt eine Übersicht über eine ausgewählte Menge an multimodalen Pain-Scores. Alle Pain-Scores sind für Kleinkinder bis 3 Jahren gedacht. In der Übersicht wird nicht wiedergegeben, welche weiteren Merkmale jeweils in das Scoring mit einbezogen werden, oder welche Insgesamtpunktzahlen auf welche Schmerzintensität hinweisen. Es soll an dieser Stelle nur verdeutlicht werden, welche unterschiedlichen Ansätze zur Bewertung des Schreiens aus medizinischer Sicht im Zusammenhang mit Pain-Scores existieren. Folgende Beobachtungen lassen sich aus der Übersicht ziehen:

1. Die zu beobachtenden Eigenschaften des Weinens werden mit subjektiv behafteten Werten charakterisiert. Beispielsweise wird im N-PASS-System ist ein Schmerz-Schrei

als „High-pitched or silent-continuous crying“ beschrieben. Es wird nicht fest definiert, was als „crying“ gilt oder welche Tonhöhe als „high-pitched“ ist. Auch die Erstquellen geben keine festen Definitionen.

2. Es gibt verschiedene Ansätze zur Bewertung des Weinens. Bei CRIE ist die Tonhöhe, bei BIIP die Länge und bei COMFORT die Art des Weinens entscheidend.

3. Die Beschreibungen sind kurz und prägnant gehalten, der Arzt hat in keinem der Modelle auf mehr als drei Parameter des Schreiens zu achten. Die Begründung liegt darin, dass bei allen Modellen a.) das Schreien nur eines von mehreren Faktoren ist, und b.) Die Schmerzbestimmung in einem vorgegebenen Zeitrahmen durchführbare sein muss.

System	**P.**	**Description**
FLACC[28]	0	No cry (awake or asleep)
	1	Moans or whimpers; occasional complaint
	2	Crying steadily, screams or sobs, frequent complaints
N-PASS[23]	-2	No cry with painful stimul
	-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
BIIP[8]	0	No Crying
	1	Crying <2 minutes
	2	Crying >2 minutes
	3	Shrill Crying >2 minutes
CRIES[4]	0	If no cry or cry which is not high pitched
	1	If cry high pitched but baby is easily consoled
	2	If cry is high pitched and baby is inconsolable
COVERS[14]	0	No Cry
	1	High-Pitched or visibly crying
	2	Inconsolable or difficult to soothe
PAT[11]	0	No Cry
	1	Cry
DAN[6]	0	Moans Briefly
	1	Intermittent Crying
	2	Long-Lasting Crying, Continuous howl
COMFORT[18]	0	No crying
	1	Sobbing or gasping
	2	Moaning

	3	Crying
	4	Screaming
MBPS[21]	0	Laughing or giggling
	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-Scores

2.2 Schmerz-Schrei aus medizinischer Sicht

Die Frage ist: Woher kommen diese unterschiedlichen Bewertungen des Weinens in Tabelle 2.2? Gibt es eine Pain-Score, die aus wissenchaftlicher Sicht „recht hat“? Dieser Fragestellung unterliegen unterliegen zwei grundlegendere Fragen: 1.) Ist es überhaupt möglich, anhand der akustischen Eigenschaften den Grund für den Schrei abzuleiten, also beispielsweise Hunger, Einsamkeit oder Schmerz? Anders formuliert: Gibt es überhaupt so etwas wie einen Schmerz-Schrei? 2.) Ist es möglich, anhand der akustischen Eigenschaften den Schweregrad des Unwohlseins abzuleiten (also beispielsweise den Grad des Schrei-Versursachenden Schmerzes)?

Die Annahme, dass es möglich ist, aus dem Schreien den Grund abzuleiten, wird als „Cry-Types Hypothesis“ bezeichnet. Die berühmtesten Befürworter dieser Hypothese ist eine skandinavische Forschungsgruppe, auch bezeichnet als „Scandinavian Cry-Group“, die diese Idee in dem Buch „Infant Crying: Theoretical and Research Perspectives“[2] publik machte. Die Annahme ist, dass die verschiedenen Ursachen *Hunger, Freude, Schmerz, Geburt und Anderes* klare Unterschiede hinsichtlich ihrer akustischen Merkmale aufweisen, welche an einem Spektogramm ablesbar seien. Entsprechende Beispiele werden in dem Buch gegeben. Nur einige Jahre Später zeigte Müller et al [7] in einem Paper, dass bei leichter Veränderung der Bedingungen der Experimente die Unterscheidung nicht möglich ist. 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 nur starke Hinweise dafür, dass die Plötzlichkeit des Eintretens des Schreigrundes hörbar ist. Ein plötzliches Ereignis, wie ein Nadelstich oder ein lautes Geräuch, führen auch zu einem plötzlich beginnenden Schreien. Ein langsam einretendes Ereignis, wie ein langsam immer stärker werdender physischer Schmerz oder langsam eintretender Hunger führen auch zu einem langsam eintretenden Weinen. Da keine Einigung herrscht, wird empfohlen, den Grund aus dem Kontext abzuleiten.[26]

Die Zweite Frage nach der Ableitung der Stärke des Unwohlseins aus den akustischen Eigenschaften des Geschreis wird in der Fachsprache 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 Unwohlsein des Babies, und nicht der tatsächliche Grad, da dieser ohne die Möglichkeit der direkten Befragung des Kindes nie mit absoluter Sicherheit bestimmt werden kann. Dieser vermutete LoD wird entweder durch das subjektive Empfinden der Beobachter oder durch Pain-Scores festgestellt. Ein hohes Level of Distress hat vor allem

eine schnelle Reaktion der Aufsichtspersonen zur Beruhigung des Babies zur Folge, womit dem Geschrei eine Art Alarm-Funktion zukommt. Es gibt starke Hinweise darauf, dass das Level of Distress anhand objektiv messbarer Eigenschaften des Audiosignals bestimmt werden kann. So herrscht beispielsweise weitesgehend Einigung darüber, dass ein „lang“ anhaltendes Geschrei auf einen hohen Level of Distress hinweist. Insofern aus dem Kontext des Schreiens Schmerz als wahrscheinlichste Ursache eingegrenzt werden kann, kann aus einem hohen Level of Distress ein hoher Schmerz abgeleitet werden. [26] und [25]

Es herrscht wiederum keine Einigung darüber, welche akustischen Eigenschaften im Detail ein hohes Level of Distress anzeigen. Carlo V Bellieni et al [6] haben festgestellt, dass bei sehr hohem Schmerz in Bezug auf die DAN-Scala (siehe Tabelle 2.2) die Tonhöhe des Geschreis steigt. Qiaobing Xie et al [25] haben festgestellt, dass häufiges und „verzerrtes“ Schreien (ohne feststellbares Grundfrequenz, da der Ton stimmlos erzeugt wird) auf einen hohen Level of Distress hinweist.[26] Diese Uneinigkeit hat wahrscheinlich zu den verschiedenen Bewertungen in den Pain-Scores geführt. 2.2.

2.3 Phyisio-Akustische Modellierung des Weinens

Das Ziel dieses Kapitels ist die Schaffung eines einheitlichen Vokabulares, auf den sich bezogen wird, um das Schreien eines Babys zu beschreiben. Die hier vorgestellten Begriffe stammen sowohl aus dem Buch „A Physioacoustic Model of the Infant Cry “ H Golub und M Corwin [12] als auch aus dem Paper „Rythmic organization of the Sound of Infant Cry “ von Zeskind et al.[24]

Die Lautäußerung eines Neugeborenen, umgangssprachlich auch als „Weinen“ oder „Schreien“ bezeichnet, lässt sich im allgemeinen beschreiben als das „rythmische Wiederholen eines beim ausatmen erzeugen Geräusches, einer kurzen Pause, einem Einatmungs-Geräusch, einer zweiten Pause, und dem erneuten Beginnen des Ausatmungs-Geräusches.“[32].

Das Vokabular, welches insbesondere von H Golub und M Corwin geschaffen wurde, ist sehr umfassend. An dieser Stelle wird eine Auswahl grundlegender Begrifflichkeiten vorgestellt, die in dieser Arbeit gebraucht werden. Sie werden in Abbildung 2.1 veranschaulicht.

Expiration beschreibt den Klang, der bei einem einzelnen, ununterbrochenem Ausatmen mit Aktivierung der Stimmbänder durch das Baby erzeugt wird. [24]. Der von Golub et al [12] verwendete Begriff **Cry-Unit** wird in dieser Arbeit synonym verwendet. Umgangssprachlich ist handelt es sich um einen einzelnen, ununterbrochenen *Schrei*.

Inspiration beschreibt den Klang, der beim Einatmen durch das Baby erzeugt wird.

Burst beshreibt die Einheit von einer Expiration und der darauf folgenden Inspiration. Das heisst, dass die zeitliche Dauer eines Bursts sowohl das Expiration-Geräusch, das Inspiration-Geräusch als auch die beiden Pausen zwischen diesen Geräuschen umfasst. Praktisch ergibt sich das Problem, dass vor allem bei stärkerem Hintergrundrauschen die Inspiration-Geräusche häufig weder hörbar noch auf dem Spektrogramm erkennbar sind. Daher wird die Zeitdauer eines Bursts oder Cry-Unit vom Beginn einer Expiration bis zum Beginn der darauf folgenden Expiration definiert und somit allein von den Expirations auf die Bursts geschlossen. Implizit wird somit eine Inspiration zwischen zwei Expirations angenommen.

Cry die insgesamte klangliche Antwort zu einem spezifischen Stimulus. Eine Gruppe meh-

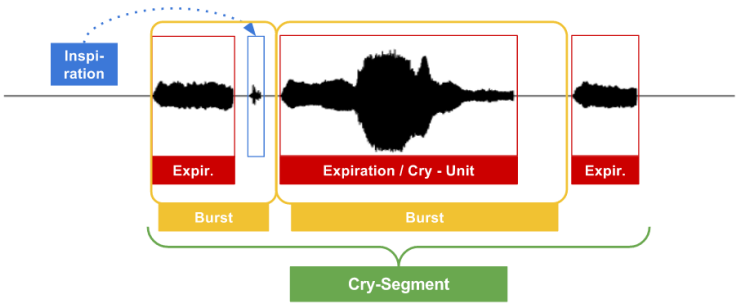

Abbildung 2.1: Veranschaulichung des Grundvokabulars

rerer Cry-Units.[12] In dieser Arbeit wird ein *Cry* als **Cry-Segment** bezeichnet, um Verwechslungen zu vermeiden.

Weiterhin wurden von H Golub und M Corwin [12] Cry-Units in eine der folgenden drei Kategorien eingeführt:

Phonation beschreibt eine Cry-Unit mit einer „vollen Vibration der Stimmbänder“ mit 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-artigem Vibration der Stimmbänder“ mit einer Grundfrequenz zwischen 1000 und 2000 Hz. Entspricht umgangssprachlich einem Weinen mit einem „sehr hohen, aber klaren, hörbaren Ton“.

Dysphonation beschreibt eine Cry-Unit ohne klar feststellbare Tonhöhe, produziert durch Turbulenzen an den Stimmbändern. Entspricht umgangsprachlichen dem „Brüllen oder Krächzen“.

Eine Cry-Unit gehört dabei mindestens einer dieser Kategorien an, kann aber auch in seinem zeitlichen Verlauf die Kategorie wechseln. H Golub und M Corwin [12] stellen weiterhin eine Reihe an charakteristischen Eigenschaften vor, die in Bezug auf ein Cry-Segment berechnet werden.

Latency-Period beschreibt die Dauer zwischen dem zufügen eines Schmerz-Stimulus und dem beginn des ersten Cry-Bursts des Segmentes

Duration beschreibt die insgesamte Zeitdauer des Cry-Segmentes. Es wird keine genaue Definition gegeben, wodurch Beginn und Ende definiert werden. Das Segment endet dort, wo es „scheint, aufzuhören“.

Maximum-Pitch beschreibt die höchste festgetellte Grunfrequenz des Segmentes.

... und viele weitere, die in [12] nachgelesen werden können, aus Platzgründen an dieser Stelle jedoch nicht vollständig genannt werden.

3 Grundlagen der Signalverarbeitung

Ein *Signal* ist eine Funktion eines Parameters mit numerischen Wertebereich. Die Abbildung zwischen Defintions- und Wertebereich kann, aber muss nicht durch eine Formel definiert sein. So fällt $f(x) = \sin(x)$ genauso unter die Definition eines Signals wie eine Folge numerischer Werte, die durch die Aufnahme eines Messgerätes enstanden sind. Weiterhin kommt dem Wertebereich eine gewissen Bedeutung zu, wie *Zeit* oder *Ort*. Ein typisches Beispiel für ein Signal ist die Spannung, die abhängig von der Zeit von einem Mikrofon erzeugt wird. Da in dieser Arbeit nur Signale von Bedeutung sind, deren Wertebereich sich auf die Zeit bezieht, konzetrieren sich alle folgenden Bereich auf diesen Bereich- Im Zusammenhang mit Signalen wird der Definitionsbereich auch als *unabhängiger Parameter* und der Wertebereich auch als *abhängiger Parameter* bezeichnet. [30, S. 11-12] [20, S. 22-23]

Bei einem zeit-kontinuierlichen Signal $x(\,)$ ist der Wertebereich kontinuierlich, wie in Formel 3.1 definiert. Bei einem zeit-diskreten Signal $x[\,]$ ist der Wertebreich diskret, wie in Formel 3.2 definiert. Abbildung 3.1 zeigt Beispiele für ein zeit-kontinuierliches und ein zeit-diskretes Signal. So beschreibt beispielsweise $x[17] = s$ den Wert zur Zeit $n = 17$. „Zeit“ hat in diesem Kontext keine Einheit. Ein Wert wird auch als *Sample* oder *Amplitude* bezeichnet. [20, S. 22 - 23]

$$x(\,) := \quad \forall t \in \mathbb{R} : \; x(t) = s \tag{3.1}$$

$$x[\,] := \quad \forall n \in \mathbb{Z} : \; x[n] = s \tag{3.2}$$

Zeit-diskrete Signale werden häufig dadurch gewonnen, dass ein zeit-kontinuierliches Signal in regelmäßigen Intervallen abgetastet wird. Dieser Prozess wird als *Sampling* bezeichnet und durch Formel 3.3 definiert. Der Parameter T_s wird als *Sampling-Interval* bezeichnet. Das Reziproke des Sampling-Intervals heißt *Sampling-Rate* und wird in der Einheit $\frac{1}{\mathrm{s}} = \mathrm{Hz}$, siehe Formel 3.4. Eine Sampling-Rate von $f_s = 44\,100\,\mathrm{Hz}$ bedeutete beispielsweise, dass ein Signal 44100 mal pro Sekunde abgetastet wurde.[20, S. 24]

$$x[n] = s(n \cdot T_s)\ , -\infty < n < \infty \tag{3.3}$$

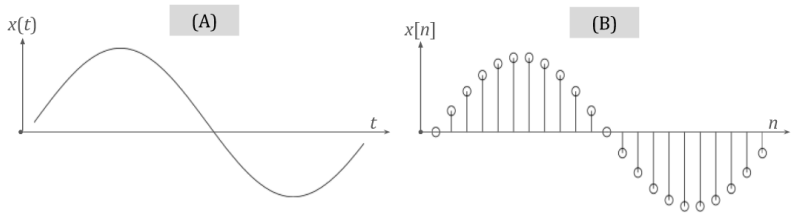

Abbildung 3.1: Ein zeit-kontinuierliches Signal (A) und ein zeit-diskretes Signal (B)

$$\text{Samplingrate:} \quad f_s = \frac{1}{T_s}[\text{Hz}] \tag{3.4}$$

Das so genannte *Nyquist-Shannon-Abtasttheorem* nach Formel 3.5 besagt, dass die Samplingrate mindestens doppelt so hoch sein muss wie die höchste im abgetasteten Signal enthaltene Frequenz. Das bedeutet im Umkehrschluss, dass die höchste im abgetasteten Signal enthaltene Frequenz die Hälfte der Abtastfrequenz entspricht.

$$f_s > 2 \cdot f_{max} \tag{3.5}$$

Da in dieser Arbeit nur zeit-diskrete Signale von Interesse sind, werden ab diesem Punkt die Definitonen für zeit-kontinuierliche Signale ausgelassen. In den beiden Hauptquellen dieser Arbeit, [30] und [20] ist eine Uneinigkeit und Inkonsistent über die symbolische Bezeichung von Signal und Sample festzustellen. In [20] wird die Konvention eingeführt, mit $x[n]$ das gesamte Signal, als auch ein Sample des des Signals zu bezeichnen, was an einigen kritischen Stellen zu unklaren Definitionen führt. In [30] wird eingeführt, dass das gesamte Signal als $x[\,]$, und ein Sample als $x[n]$ bezeichnet wird. Diese Definition wird im Buch jedoch inkonsistent verwendet und an einigen Stellen $x[n]$ als Bezeichung für das gesamte Signal verwendet. In dieser Arbeit wird die Konvention eingeführt, mit $x[\,]$ das gesamte Signal, und mit $x[n]$ ein Sample dieses Signals zu bezeichen. Dies führt zwangsweise zur Abwandlung einiger Formeln der beiden Hauptquellen dieses Grundlagenteils, um die Konsistenz beizubehalten.

Der *Support* ist das kleinst mögliche Zeitintervall, der alle Samples enthält, die nicht den Wert 0 haben, wie Formel 3.6 definiert.

$$\begin{aligned} \text{Sup}(x[\,]) = [sup_s, sup_e] \quad &, sup_s, sup_e \in \mathbb{Z} \\ , x[sup_s] \neq 0 \ \wedge \ x[sup_e] \neq 0 \ \wedge \ \forall n \ \notin [sup_s, sup_e] &: x[n] = 0 \end{aligned} \tag{3.6}$$

Die *Dauer* eines Signales ist die Länge des Supportes nach Formel 3.7. In dieser Arbeit herrscht die Konvention, dass die Länge des Signals kurz mit der Variable N abgekürzt wird. Das Signal $x[n] = \cos(n)\,, 0 \leq n \leq 3$ hat beispielsweise den Support $[0, 3] = \{0, 1, 2, 3\}$ und die Dauer 4. Ein *unendliches Signal* hat einen unendlichen langen Support, das heißt es gilt $\text{Length}(x[\,]) = \infty$. Ein *endliches Signal* hat einen endlichen Support, das heißt $\text{Length}(x[\,]) \neq \infty$. Wird in dieser Arbeit der Support nicht explizit angegeben, gilt bei endlichen Signalen als Konvention $Sup(x[\,]) = [0, N-1]$. Unabhängig von der Endlichkeit oder Unendlichkeit des Supportes wird davon ausgegangen, dass sich alle Signale von negativer bis positiver Unendlichkeit erstrecken. Werden also Berechnungen auf Samples eines Signales durchgeführt, die außerhalb seines Supportes liegen, werden diese Samples mit dem Wert 0 angenommen. [20, S. 24]

$$\text{Length}(x[\,]) = sup_e - sup_s + 1 = N \tag{3.7}$$

Ein Signal gilt als *periodisch*, wenn Formel 3.8 erfüllt ist. Der Parameter N wird als Periode von $x[\,]$ bezeichnet. Wenn ein Signal mit N periodisch ist, dann ist es auch mit $2N, 3N, \ldots$ periodisch. Die Grundfrequenz N_0 ist das kleinste N, für das Formel 3.8 erfüllt

ist. Abbildung 3.2 zeigt ein Beispiel für ein nicht-periodisches und ein periodisches Signal. [20, S. 24]

$$\exists N : \forall n \in Sup : x[n+N] = x[n] \rightarrow \text{Periodisch}(x[n]) = true \tag{3.8}$$

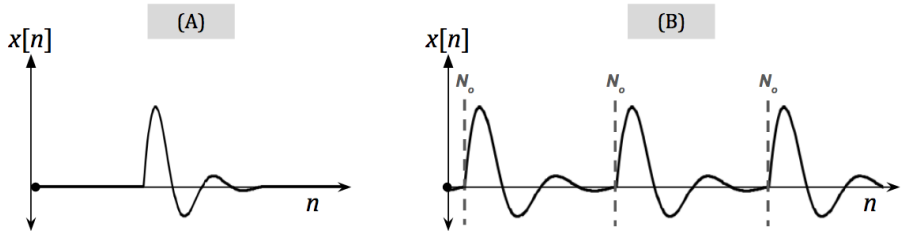

Abbildung 3.2: Ein nicht-periodisches Signal (A) und ein periodisches Signal (B)

3.1 Statistische Merkmale

Im folgenden wird ein überblick über die häufig verwendete Signaleigenschaften gegeben. Abbildung 3.3 visualisiert die Erläuterungen.

1. Der **Maximalwert** / **Minimalwert** beschreibt den höchsten / niedrigsten in $x[\,]$ enthaltenen Wert nach den Formel 3.9.

$$\begin{aligned} \max(x[\,]) &= \max_{n=0\ldots N-1}(x[n]) \\ \min(x[\,]) &= \min_{n=0\ldots N-1} n(x[n]) \end{aligned} \tag{3.9}$$

2. Der **Durchschnittswert** / **Average Value** beschreibt den durchschnittlichen Wert aller Samples von $x[\,]$ nach Formel 3.10. Dieser Durchschnittswert wird über dem Intervall $[n_1, n_2]$ berechnet.

$$\text{AVG}(x[\,]) = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} x[n] \tag{3.10}$$

3. Der **Mean Squared Value** (*MSV*) beschreibt den quadrierten Durchschnittswert über eine bestimmtes Interval nach Formel 3.11. Er wird auch als *durchschnittliche Energie* oder *average Power* bezeichnet.

$$\text{MSV}(x[\,]) = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} x[n]^2 \tag{3.11}$$

4. Das **Root Mean Square** (*RMS*) ist die Wurzel des Mean Squared Value nach Formel3.12. Der RMS findet häufiger Anwendung als der MSV, da er 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.

$$\mathrm{RMS}(x[\,]) = \sqrt{\frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} x[n]^2} \tag{3.12}$$

5. Die **Energie** / **Energy** bezeichnet die „Stärke “ eines Signals über einen bestimmten Intervall nach Formel 3.13. Sie entspricht dem MSV-Wert multipliziert der Länge des Intervalls. [20, S. 27-28]

$$\mathrm{E}(x[\,]) = \sum_{n=n_1}^{n_2} x[n]^2 \tag{3.13}$$

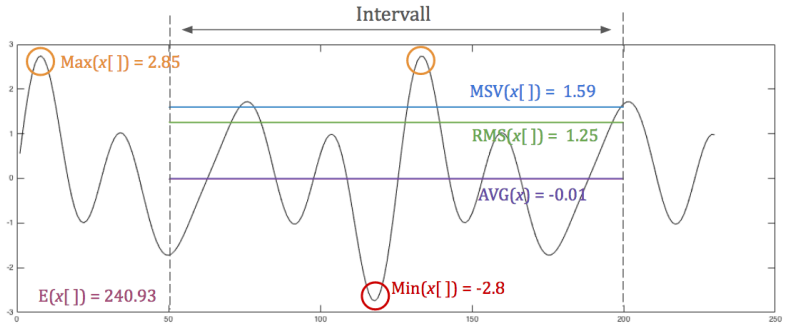

Abbildung 3.3: Statistische Werte eines Signals über das Intervall [50,200]

Die Addition und Multiplikation wird bei Signalen komponentenweise durchgeführt, wie Formel definiert. Abbildung 3.4 visualisiert diese Operationen.

$$\begin{aligned} x_1[\,] + x_2[\,] = y[\,] &:= \forall_{n=n_1}^{n_2} : x_1[n] + x_2[n] = y[n] \\ x_1[\,] \cdot x_2[\,] = y[\,] &:= \forall_{n=n_1}^{n_2} : x_1[n] \cdot x_2[n] = y[n] \end{aligned} \tag{3.14}$$

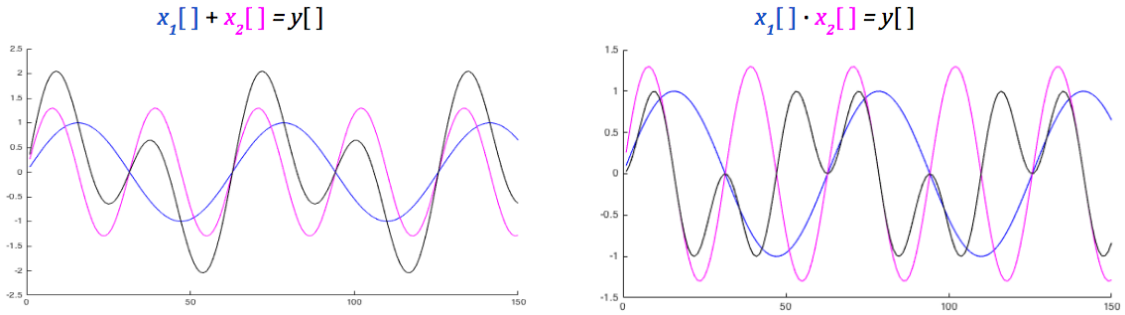

Abbildung 3.4: Komponentenweise Addition und Mulitplikation zweier Signale

3.2 Fehlersignale

Die Addition wird unter anderem für die Modellierung des Einflusses von Störungen benötigt. Angenommen, ein Signal $x[\,]$ wird übertragen, auf dem Übertragungsweg jedoch durch ein anderes Störsignal wie z.B. Rauschen $e[\,]$ überlagert. Dieses Störsignal wird in diesem Zusammenhang auch als „Fehler-Signal bezeichnet. Das resultierende Signal $x'[\,]$ wird nach Formel 3.15 berechnet. Kennt man sowohl das Eingangssignal $x[\,]$ als auch das Ausgangssignal $x'[\,]$, kann das Störsignal $e[\,]$ nach Formel 3.16 berechnet werden.

$$x'[\,] := \quad \overset{n_2}{\underset{n=n_1}{\forall}} \; : \; x'[n] = x[n] + e[n] \tag{3.15}$$

$$e[\,] := \quad \overset{n_2}{\underset{n=n_1}{\forall}} \; : \; e[n] = x'[n] - x[n] \tag{3.16}$$

Errechnet man nun den den MSV- oder RMS-Wert des Störsignales $e[\,]$, gibt das Ergebnis einen Eindrück über die „Stärke “ des Fehler-Signals. Der MSE-Wert des Fehlers wird in diesem Zusammenhang auch als *Mean Squared Error* (*MSE*) und der RMS-Wert als *Root Mean Squared Error* (*RMSE*) oder einfach als *Fehler* oder *Error* bezeichnet. Formel3.17 und 3.18 definierten die Berechnungen des MSE und RMSE. Der RMSE hat im Gegensatz zum MSE den Vorteil, dass er besser ins Verhältnis zu den Werten des Fehlersignals gestetzt werden kann. Ein RMSE $= 0$ heisst, dass $x[\,] = x'[\,]$ und somit kein Störsignal vorliegt. Ein RMSE $=$ RMS(x) heisst, dass Eingangs- und Störsignal den selben Effektivwert und somit die selbe „stärke“ besitzen. Abbildung 3.5 visualisiert die Berechnung des MSE und RMSE. [20, S: 28 - 29]

$$\mathrm{MSE}(x[\,], x'[\,]) = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} (x[n] - x'[n])^2 \tag{3.17}$$

$$\mathrm{RMSE}(x[\,], x'[\,]) = \sqrt{\frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} (x[n] - x'[n])^2} \tag{3.18}$$

Eine weitere Betrachtungsweise bezüglich der Stärke des Rauschens auf das Signal ist, das Eingangssignal ins Verhältnis zum Rauschsignal zu setzen. Formel 3.19 gibt die Definition. Ein $\mathrm{SNR}_{\mathrm{rel}}(x[\,], e[\,]) = 1$ heißt, dass das Eingangssignal den selben MSV wie das Fehlersignal hat. Meistens ist der MSV des Eingangssignals in der Praxis sehr viel höher als der des Fehler-Signals. Um den Zahlenraum zu begrenzen, wird die Pseudo-Einheit dB verwendet. Formel 3.20 den so berechneten *Signal-Rausch-Abstand* (*SNR*, englisch Signal-to-Noise-Ratio). Entgegen des MSE weisst ein *niedriger* SNR-Wert auf ein *starkes* Rauschen hin, und ein *hoher* SNR auf ein *schwaches* Rauschen! Abbildung 3.5 visualisiert die Berechnung des SNR.

$$\mathrm{SNR}_{rel}(x[\,], e[\,]) = \frac{MSV(x[\,])}{MSV(e[\,])} \tag{3.19}$$

$$\mathrm{SNR}(x[\,], e[\,]) = 10 \cdot \lg\left(\frac{MSV(x[\,])}{MSV(e[\,])}\right) \text{ dB} \tag{3.20}$$

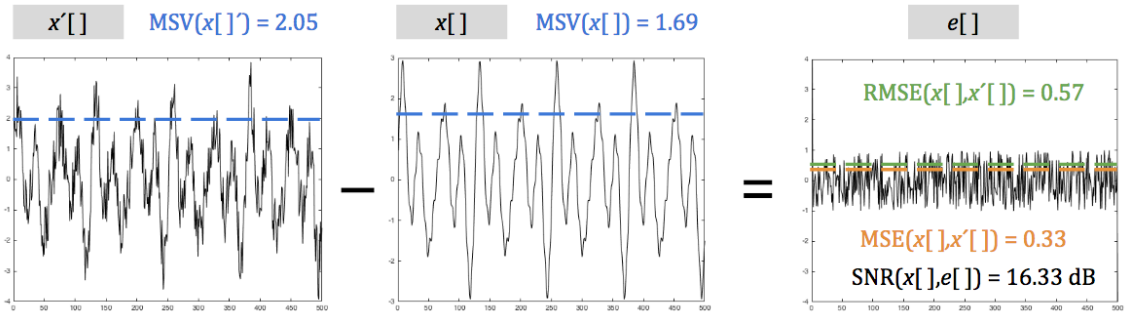

Abbildung 3.5: Berechnung des MSE, RMSE und SNR eines von Rauschen gestörten Signals

3.3 Korrelation

Die *Korrelation* (engl *Correlation*) zweier Signale $x_1[\,]$ und $x_2[\,]$ wird nach Formel 3.21 als die Summe aller Samples des Produktes der beiden Signale über einen bestimmtes Intervall $[n_1, n_2]$ definiert. Das Ergebnis ist eine Wert $\in \mathbb{R}$ welches die „Ähnlichkeit der beiden Signale“ kennzeichnet. Ein Positiver Wert weisst auf eine *positive Korrelation* hin, ein negativer Wert auf eine *negative Korrelation*, und ein Wert von $\mathrm{Corr}(x_1[\,], x_2[\,]) = 0$ auf *keine Korrelation*. Aus der größe des Wertes kann die Stärke der Korrelation jedoch nicht direkt interpretiert werden. Bei der *normalisierten Korrelation* $\mathrm{Corr}_N(x[\,], y[\,])$ wird daher die Korrelationswert ins Verhältnis zu den Energien der beiden Signale gesetzt, wie in Formel 3.22 definiert. Der Wertebereich der normalisierten Autokorrelation ist $-1 \leq \mathrm{Corr}_N(x[\,], y[\,]) \leq +1$. Daraus ergeben sich die in Formel 3.23 definierten Zusammenhänge. Ein Wert von $\mathrm{Corr}_N(x[\,], y[\,]) = 1$ wird auch als *perfekte Korrelation* bezeichnet, ein Wert von $\mathrm{Corr}_N(x[\,], y[\,]) = -1$ als *anti-perfekte Korrelation* [20, S. 46 - 47] Abbildung 3.6 visualisiert die normalisierte Korrelation eines Signales $x[\,]$ mit den Signalen $y[\,]$.

$$\mathrm{Corr}(x[\,], y[\,]) = \sum_{n=n_1}^{n_2} x[n] \cdot y[n] \tag{3.21}$$

$$\mathrm{Corr}_N(x[\,], y[\,]) = \frac{\mathrm{Corr}(x[\,], y[\,])}{\sqrt{\mathrm{E}(x[\,]) \cdot \mathrm{E}(y[\,])}} \tag{3.22}$$

$$\mathrm{Corr}_N(x[\,], y[\,]) = \begin{cases} 1 & \rightarrow x[\,] = y[\,] \\ -1 & \rightarrow x[\,] = -y[\,] \end{cases} \tag{3.23}$$

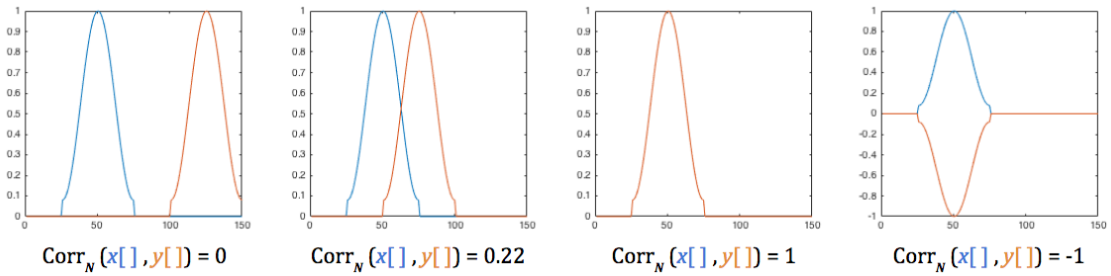

Abbildung 3.6: Correlation der Signale $x[n]$ und $y[n]$

Die Korrelation und die normalisierte Korrelation werden aufgrund ihrer Eigenschaften verwendet, um ein Signal $x[\,]$ in einem Signal $y[\,]$ zu detektieren. Häufig ist das Ziel, ein von einem Rauschen $e[\,]$ überlagerten Signal $x[\,] + e[\,] = y[\,]$ auf das Vorhandensein des erwarteten Signales $x[\,]$ hin zu überprüfen. Wie in Abbildung 3.6 zu sehen ist, ist der Korrelationswert jedoch von der Verzögerung des Signals abhängig. Daher wird in der $Cross - Correlation$ das Signal $y[\,]$ mit einer verzögerten Varianten des Signals $x[\,]$ korreliert, wie in Formel 3.24 definiert. Der parameter k wird als *Lag* bezeichnet und gibt die Verzögerung an.

$$\text{X-Corr}_k(x[\,], y[\,]) = \sum_{n=-\infty}^{\infty} x[n-k] \cdot y[n] \tag{3.24}$$

Im Prozess der so genannten *Running Correlation* nutzt man die Cross-Correlation mit den Lags $k = 0 \cdots k_{max}$ zur Erstellung des *Korrelationssignals* $r[\,]$, wie in Gleichung 3.25 definiert. Das Signal $r[\,]$ gibt Auskunft, zu welchen Verzögerungswerten k die größten Ähnlichkeiten zwischen x und y gefunden wurden.

$$r[\,] := \bigvee_{k=0}^{k_{max}} : \; r[k] = \text{X-Corr}_k(x[\,], y[\,]) \tag{3.25}$$

Abbildung 3.7 zeigt ein Beispiel für die Erzeugung von $r[\,]$ mit der Sliding Correlation. (A) zeigt das zu detektierende Signal $x[\,]$ und (B) das Signal $y[\,]$. (C) zeigt das Korrelationssignal $r[\,]$ mit den Lags $k = 0, \ldots, 1150$. [20, S. 47 - 48]

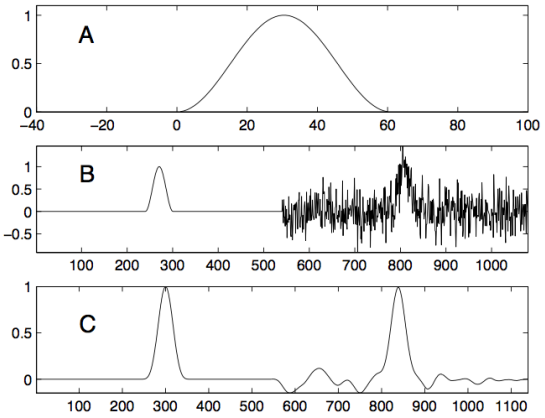

Abbildung 3.7: Beispiel einer Running Correlation

3.4 Diskrete Fourier-Transformation

Die *Fourier-Transformation* ist eine Familie von Transformationen, mit deren Hilfe Signale aus dem Zeit-Bereich in den Frequenz-Bereich transformiert werden. Das heißt, dass der unabhängige Parameter nach der Transformation nicht mehr die Zeit, sondern die Frequenz beschreibt.

Die konkrete Berechnung der Transformation ist abhängig von den Eigenschaften des Signales. Die Variante, die die meiste Anwendung in der digitalen Signalverarbeitung

findet, ist die *Diskrete Fourier-Transformation* (kurz **DFT**). Sie transformiert *zeit-diskrete, periodische, unendliche* Signale (siehe Formel 3.2 und 3.8) in den Frequenz-Bereich. Es exisitert sowohl eine reelle als auch eine complexe Variante der DFT. Die reelle Variante wird mit Hilfe reeller Zahlen, und die komplexe mit Hilfe komplexer Zahlen berechnet. An dieser Stelle werden beide Variante vorgestellt: Die komplexe, da der effizienteste Algorithmus zur Berechnung der DFT, die *Fast-Fourier-Transformation* (**FFT**) auf ihr beruht, und die reelle, da sie das Verständnis der komplexen vereinfacht.[30, S. 142 - 146]

3.4.1 Reelle DFT

Jedes zeitdiskretes, periodisches Signal $x[\,]$ kann erzeugt werden, indem eine endliche Anzahl von Sinus- und Cosinus-Signalen geeigneter Frequenz und Amplitude aufaddiert werden. Der Umkehrschluss ist, dass sich jedes Signal in eine Menge von Sinus- und Cosinus-Signalen zerlegen lässt, ohne das Information für das Signal $x[\,]$ verloren geht. Diese Zerlegung des Signals $x[\,]$ wird als *Dekomposition* bezeichnet, die Kombination der Sinus- und Cosinus-Siganel zu $x[\,]$ als *Synthese*. Genauer gesagt werden für ein Signal $x[\,]$ mit $\text{Length}(x[\,]) = N$ höchstens $\frac{N}{2}+1$ Sinus- und $\frac{N}{2}+1$ Cosinus-Wellen benötigt, also insgesamt $N+2$ Signale. Gleichung 3.26 fasst diese Aussage zusammen. [30, S. 144 - 147]

$$\begin{aligned} x[\,] := \quad \underset{n=0}{\overset{N-1}{\forall}} : \; x[n] = A[0]\cos_{f_0}[n] + \ldots + A[N/2]\cos_{f_{N/2}}[n] \\ + B[0]\sin_{f_{N/2}}[n] + \ldots + B[N/2]\sin_{f_{N/2}}[n] \end{aligned} \tag{3.26}$$

Die Cosinus- und Sinus-Schwingungen, die in Gleichung 3.26 verwendet werden, werden in Gleichung 3.27 definiert. Die Faktoren $A[\,], B[\,]$ geben die Amplitude der entsprechenden Cosinus/Sinus-Schwingung an, der Fatkor f die Frequenz der Schwingung (Perioden pro Sekunde), und f_s die Sampling-Rate (Siehe Gleichung 3.4). [20, S. 62] [30, S. 150] Abbildung 3.8 zeigt ein Beispiel für die Cosinus-Schwingung $[A_f = 2] \cdot cos_{f=4\,\text{Hz}}[\,]$.

$$\begin{aligned} \cos_f[n] &= \cos(2\pi f \frac{n}{f_s}) \\ \sin_f[n] &= \sin(2\pi f \frac{n}{f_s}) \end{aligned} \tag{3.27}$$

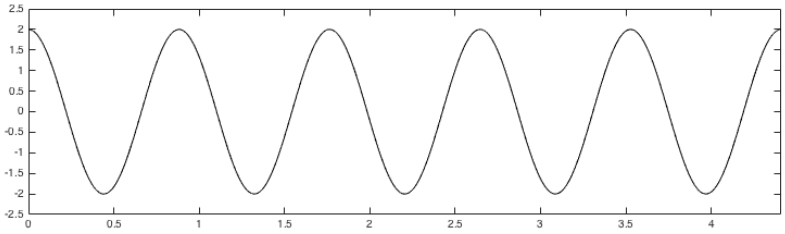

Abbildung 3.8: $1\,\text{s} = 44100$ Samples der Cosinus-Schwingung $[A = 2] \cdot cos_{f=4\,\text{Hz}}[\,] = 2 \cdot \cos(2\pi 4 \frac{n}{f_s})$ bei einer Sampling-Rate von $f_s = 44\,100\,\text{Hz}$

Abbildung 3.9 zeigt ein Beispiel für die Synthese eines Signal $x[\,]$ mit $N = 200$ Samples mit einer Samlingrate von $f_s = 100\,\text{Hz}$. Es werden theoretisch $\frac{N}{2}+1 = 101$ Cosinus und 101 Sinus-Signale für die Synthese benötigt, da aber nur 4 Signale eine Amplitude > 0

haben, werden auch nur diese Signale gezeigt. Die Frage ist: Angenommen, man kennt nur das Signal $x[\,]$, wie errechnet man daraus die die Amplituden $A[\,], B[\,]$ und Frequenzen f der Cosinus- und Sinus-Signale? Anders gesagt: Wie berechnet man die Dekomposition?

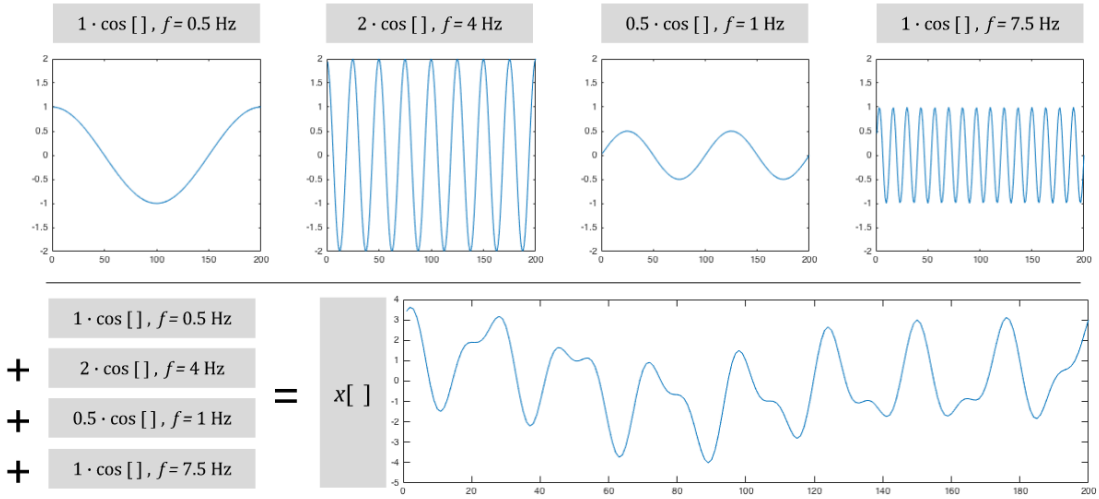

Abbildung 3.9: Synthetisierung eines Signals $x[\,]$ aus vier Cosinus-Funktionen, mit $\text{Length}(x[\,]) = 200$ und $f_s = 100\,\text{Hz}$

Das Problem lässt sich auf die Berechnung der Amplituden $A[\,], B[\,]$ beschränken, in dem man den Fakt, dass höchstens $\frac{2}{N} + 1$ Cosinus- und $\frac{2}{N} + 1$ Sinus-Signale für die Synthese benötigt werden, in Verbindung mit dem Nyquist-Shannon-Abtasttheorem (Gleichung 3.5) bringt. Die niedrigst mögliche Frequenz, die im Signal $x[\,]$ enthalten sein kann, ist $f_0 = 0\,\text{Hz}$, und die höchst mögliche Frequenz $f_{max} = \frac{f_s}{2}$, womit die Frequenzen der Signale $\cos_{f_0=0}[\,], \cos_{f_{max}=f_s/2}[\,], \sin_{f_0=0}[\,], \sin_{f_{max}=f_s/2}[\,]$ bereits feststehen. Die restlichen $\frac{2}{N} - 1$ Cosinus-Signale teilen sich gleichmäßig auf diesen Frequenzraum auf, entsprechendes gilt für die Sinus-Signale. Die Frequenz der insgesamt $N + 2$ Cosinus/Sinus-Signale ergibt sich somit direkt als Funktion der Samplingrate und der Länge des Signals $x[\,]$. Gleichung 3.28 fasst diesen Zusammenhang zusammen. Die dort beschriebenen Funktionen $\cos_k[\,]$ und $\sin_k[\,]$werden als *Basisfunktionen* bezeichnet. Aus dem Index $k = 0, \ldots, \frac{N}{2}$ lässt sich die Frequenz der jeweiligen Basisfunktion nach Gleichung 3.29 berechnen.[30, 140 - 151, S.] In Bezug auf das Beispiel aus Abbildung 3.9 ergeben sich die Indexe $[\cos_{f=0.5\,\text{Hz}}[\,] = \cos_1[\,]]$, $[\cos_{f=4\,\text{Hz}}[\,] = \cos_8[\,]]$, $[\sin_{f=1\,\text{Hz}}[\,] = \sin_2[\,]]$ und $[\sin_{f=7.5\,\text{Hz}}[\,] = \sin_{15}[\,]]$

$$\begin{aligned} \cos_k[\,] := & \quad \forall_{n=0}^{N-1} : \; \cos_k[n] = \cos(2\pi k \frac{n}{N}) \\ \sin_k[\,] := & \quad \forall_{n=0}^{N-1} : \; \sin_k[n] = \sin(2\pi k \frac{n}{N}) \\ & \qquad\qquad N = \text{Length}(x[\,]) \end{aligned} \tag{3.28}$$

$$f = k \frac{f_s}{N} \tag{3.29}$$

Das Problem der Dekomposition wird so auf die Suche der Amplituden-Koeffizienten $A[k], B[k]$ beschränkt, die zu den jeweiligen Basisfunktionen $\cos_k[\,]$ und $\sin_k[\,]$ gehören. Die

Frage ist, vereinfacht formuliert, wie „stark“ jede der Basisfunktionen $\cos_0[\,], \ldots, \cos_{N/2}[\,]$ und $\sin_0[\,], \ldots, \sin_{N/2}[\,]$ in $x[\,]$ enthalten ist. Die Antwort darauf ist die in Kapitel 3.3 vorgestellte Korrelation. Der Korrelationswert einer Cosinus-Basisfunktion mit Eingangssignal $\mathrm{Corr}(x[\,], \cos_k[\,]) = \sum_{n=0}^{N-1} \cdot x[n] cos_k[n]$ gibt somit eine Aussage darüber, wie stark die entsprechende Cosinus-Schwingungen zur Syntehse von $x[\,]$ beiträgt. Ein Wert von 0 spricht für keinen Beitrag, ein hoher oder niedriger Wert für eine positiven oder negativen Beitrag. [30, S. 157 - 158]

Dieses Vorgehen lässt sich sogenannten *forward DFT* nach Formel 3.30 verallgemeinern, kurz als **DFT** bezeichnet. Das Ergebnis sind die Koeffizienten $\bar{A}[\,] = \bar{A}[0] \ldots \bar{A}[N/2]$ und $\bar{B}[\,] = \bar{B}[0] \ldots \bar{B}[N/2]$. Die Koeffizienten werden gemeinsam als der *Frequenz-Bereich* in *kartesischer Notation* $X[\,]$ bezeichnet. Das $X[\,]$ hat im Zusammenhang mit der reellen DFT einen rein symbolisch bezeichnenden Charakter und wird erst im Zusammenhang mit der *komplexen DFT* in Kapitel 3.4.2 zu einem Signal. [30, S. 158]

$$\mathrm{DFT}\{x[\,]\} = X[\,] := \begin{cases} \bar{A}[\,] := \underset{k=0}{\overset{N/2}{\forall}} : \; \bar{A}[k] = \sum_{n=0}^{N-1} x[n] \cos(2\pi k \frac{n}{N}) \\ \bar{B}[\,] := \underset{k=0}{\overset{N/2}{\forall}} : \; \bar{B}[k] = \sum_{n=0}^{N-1} x[n] \sin(2\pi k \frac{n}{N}) \end{cases} \tag{3.30}$$

In Bezug auf das Beispiel aus Abbildung 3.9 ergeben sich bei Anwendung von Formel 3.30 auf das Signal $x[n]$ die Koeffizienten $[\bar{A}_{f=0.5\,\mathrm{Hz}} = \bar{A}[1] = 100]$, $[\bar{A}_{f=4\,\mathrm{Hz}} = \bar{A}[8] = 200]$, $[\bar{B}_{f=1\,\mathrm{Hz}} = \bar{B}[2] = 50]$ und $[\bar{B}_{f=7.5\,\mathrm{Hz}} = \bar{B}[15] = 100]$. Um die eigentlichen Koeffizienten $A[1] = 1, A[8] = 2, B[2] = 0.5$ und $B[[15] = 1$ zu erhalten, muss die Umrechnungsvorschrift nach den Formel 3.31 und 3.32 angewandt werden. Dieser Transformationsschritt $\bar{A}[\,] \longmapsto A[\,]$ und $\bar{B}[\,] \longmapsto B[\,]$ ist ein *Extra-Schritt*, den jeder reelle DFT nach sich zieht. [30, S. 152 - 153]

$$A[\,] := \underset{k=0}{\overset{N/2}{\forall}} : \; A[k] = \begin{cases} \frac{\bar{A}[k]}{N} & , \text{falls } k = 0 \vee k = N/2 \\ \frac{\bar{A}[k]}{N/2} & , \text{sonst} \end{cases} \tag{3.31}$$

$$B[\,] := \underset{k=0}{\overset{N/2}{\forall}} : \; B[k] = \frac{B_k}{N/2} \tag{3.32}$$

Gleichung 3.33 definiert die Synthese des Signals $x[\,]$ aus den Basis-Funktionen mit Hilfe der Koeffizienten $A[\,]$ und $B[\,]$. Die Formel wird auch als *inverse DFT* (**iDFT**) bezeichnet. [30, S. 152 - 153]

$$\mathrm{DFT}\{X[\,]\} = x[\,] := \underset{n=0}{\overset{N-1}{\forall}} : \; x[n] = \sum_{k=0}^{N/2} A[k] \cos(2\pi k \frac{n}{N}) + \sum_{k=0}^{N/2} B[k] \sin(2\pi k \frac{n}{N}) \tag{3.33}$$

Abbildung 3.10 gibt einen Überblick über den Zusammenhang $x[\,]$ und $X[\,]$. Da mit steigender Länge von $x[\,]$ die Anzahl an Basis-Funktionen im Frequenzbereich $X[\,]$ steigt, wird die Auflösung des Frequenz-Bereiches umso höher, je länger $x[\,]$ gewählt wird. Im

Gegenzug sinkt die Auflösung in Bezug auf den Zeit-Bereich: Der Frequenz-Bereich trifft keine Aussage darüber *wann* etwas passiert, sondern nur *welche Frequenzen* daran beteiligt sind.[30, S. 170]

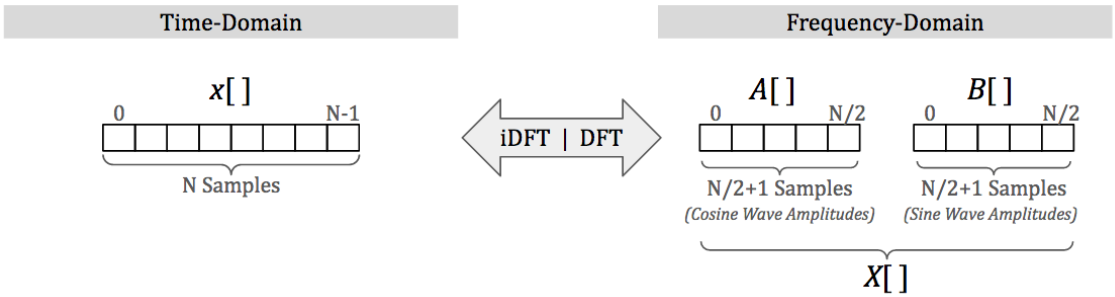

Abbildung 3.10: Überblick über die DFT und die inverse DFT

Aus Formel 3.33 geht hervor, dass bei der Synthese jeweils ein Cosinus-Signal und ein Sinus-Signal der selben Frequenz addiert wird. Diese Addition erzeugt einen sogenannten *Sinusoiden*, eine Cosinuswelle mit der Amplitude M der Phasenverschiebung ϕ nach Formel 3.34.[30, S. 162]

$$\begin{aligned} A\cos(x) + B\sin(x) &= M\cos(x+\phi) \\ , M = \sqrt{A^2+B^2} \quad &, \phi = arctan(B/A) \end{aligned} \tag{3.34}$$

Auf Basis von Formel 3.34 lässt sich der gesamte Frequenz-Bereich in kartesischer Notation als Menge von Sinusoiden-Schwingungen mit den Magnituden $M[\,] = M[0] \ldots M[N/2]$ und den Phasenverschiebungen $\phi[\,] = \phi[0] \ldots \phi[N/2]$ ausdrücken. Formel 3.35 definierte diese Transformation des Frequenz-Bereichs von kartesischer Notation in die *polare Notation*. Formel definiert die dazu inverse Transformation. [30, S. 162] Abbildung 3.11 zeigt den Frequenzbereich in polarer Notation für das Beispiel aus Abbildung 3.9.

$$(A[\,], B[\,]) \longmapsto (M[\,], \phi[\,]) = \begin{cases} M[\,] := \underset{k=0}{\overset{N/2}{\forall}} : \; M[k] = \sqrt{(A[k]^2 + B[k]^2)} \\ \phi[\,] := \underset{k=0}{\overset{N/2}{\forall}} : \; \phi[k] = \arctan\left(B[k]/A[k]\right) \end{cases} \tag{3.35}$$

$$(M[\,], \phi[\,]) \longmapsto (A[\,], B[\,]) = \begin{cases} A[\,] := \underset{k=0}{\overset{N/2}{\forall}} : \; A_k = M_k \cdot \cos(\phi_k) \\ B[\,] := \underset{k=0}{\overset{N/2}{\forall}} : \; B_k = M_k \cdot \sin(\phi_k) \end{cases} \tag{3.36}$$

Die kartesische Notation wird verwendet, um die DFT und die inverse DFT zu berechnen. Die polare Notation hat den Vorteil, dass vor allem die Magnituden $M[\,]$ für den Menschen leichter zu interpretieren sind. Die Magnituden $M[\,]k$ sind Audioingenieuren als *Spectrum* bekannt und werden in dieser Arbeit auch als solches bezeichnet. Die Phasen-Informationen $\phi[\,]$ hingegen haben für das menschliche Gehör kaum Einfluss und wird daher, zumindest bei der Betrachtung durch den Menschen, kaum Einfluss. [9, Signals and transforms, S. 10]

Die Transformation in die polare Notation wird deshalb vor allem dann Angewandt, wenn der Mensch den Frequenz-Bereich interpretieren soll. [30, S. 164]

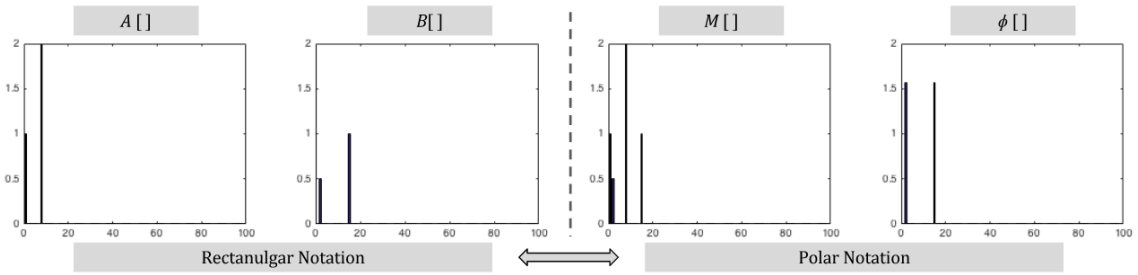

Abbildung 3.11: Frequenz-Bereich des Beispiels aus Abbildung 3.9

Ein Schlussbemerkung: In der Einleitung dieses Kapitels wurde erwähnt, dass die DFT *zeit-diskrete, periodische, unendliche* Signale transformiert, während in der hier vorgestellten Erläuterung das Signal sowohl als endlich angenommen, als auch keine Aussage über die Periodizität gemacht wurde. Bei der DFT wird davon ausgegangen, dass das Signal *außerhalb* des Supportes von x unendlich oft wiederholt wird, um die Vorraussetzungen zu erfüllen. Dabei handelt es sich jedoch um einen „matehmatischen Trick " der nur in Ausnahmefällen Einfluss auf den Frequenz-Bereich hat. Diese Ausnahmefälle tangieren diese Arbeit jedoch nicht, weshalb sie an dieser Stelle nicht weiter erläutert werden.[30, S. 145]

3.4.2 Komplexe DFT

Gleichwohl die in Kapitel 3.4.1 vorgestellte *reelle DFT* hilft beim Verständnis des Frequenz-Bereiches hilft, ist die Berrechnung der DFT nach Formel 3.30 Rechnerisch zu ineffizient, um in Echtzeit durchgeführt zu werden. Der am weitesten verbreitete Algorithmus zur Berechnung des Frequenz-Bereiches, die *Fast Fouerier-Transformation* erlaubt hingegen die Berechnung der DFT in Echtzeit. Da die FFT auf der *complexen* Variante der DFT basiert, wird sie an dieser Stelle vorgestellt. [30, S. 225]

Die Basis der komplexen DFT ist die *Eulerformel*, definiert in Formel 3.37. Sie erlaubt die Darstellung des Funktions-Werte einer Cosinus-Welle und einer Sinus-Welle der selben Frequenz und Amplitude als den Real/Imaginärteil eines komplexen Exponenten der Eulerschen Zahl e. Gleichung 3.38 zeigt, dass die Isolierung des Real/Imaginärteil von e^{ix} Zugriff auf Funktionswert der entsprechenden Cosinus/Sinuswelle erlaubt. Außerdem werden die Funktionen $|e^{ix}|$ und $\phi(e^{ix})$ definiert. Abbildung 3.12 visualisiert diese Zusammenhänge. Auf einen Beweis der Eulergleichung wird an dieser Stelle aus Platzgründen verzichtet. [30, S. 569]

$$e^{ix} = \cos(x) + i\sin(x) \tag{3.37}$$

$$\begin{aligned} \Re(e^{ix}) &= \cos(x) \\ \Im(e^{ix}) &= \sin(x) \\ |e^{ix}| &= \sqrt{\cos(x)^2 + \sin(x)^2} \\ \phi(e^{ix}|) &= \arctan(\frac{\sin(x)}{\cos(x)}) \end{aligned} \tag{3.38}$$

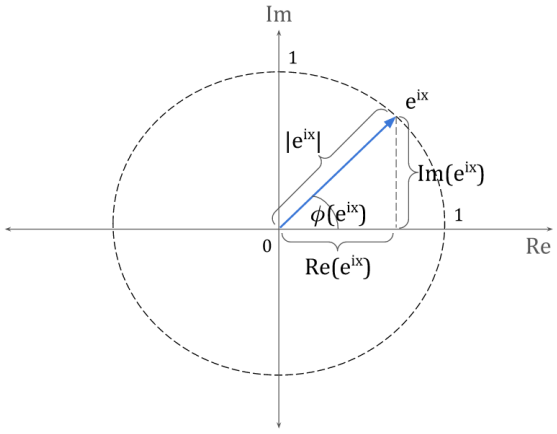

Abbildung 3.12: Visualisierung der Eulergleichung

Dementsprechend lassen sich die Basisfunktionen aus Gleichung 3.28 ebenfalls durch den Real/Imaginärteil der von e^{ix} nach Gleichung 3.39 definieren. Sie werden als die *komplexe Basisfunktionen* bezeichnet. Die Frequenz der jeweiligen Basisfunktion wird, wie in Formel 3.29 definiert, durch $f = k\frac{f_s}{N}$ errechnet.

$$\begin{aligned} \cos_k[\,] := & \quad \forall_{n=0}^{N-1} : \ \Re(e^{i\cdot 2\pi k\frac{n}{N}}) = \cos_k[n] = \cos(2\pi k\frac{n}{N}) \\ \sin_k[\,] := & \quad \forall_{n=0}^{N-1} : \ \Im(e^{i\cdot 2\pi k\frac{n}{N}}) = \sin_k[n] = \sin(2\pi k\frac{n}{N}) \end{aligned} \tag{3.39}$$

Auf Basis dieses Zusammenhanges wird die *komplexe DFT* in *kartesischer Notation* nach Formel 3.40 definiert, das heißt die Transformation vom Zeit-Bereich in den Frequenzbereich mit Hilfe komplexer Zahlen. Formel 3.41 definiert die komplexe DFT in der kompakteren, *polaren Notation*. [30, S. 570]

$$\mathrm{DFT}\{x[\,]\} = X[\,] := \quad \forall_{k=0}^{N-1} : \ X[k] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]\cdot(\cos(2\pi k\frac{n}{N}) - i\sin(2\pi k\frac{n}{N})) \tag{3.40}$$

$$\mathrm{DFT}\{x[\,]\} = X[\,] := \quad \forall_{k=0}^{N-1} : \ X[k] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]\cdot e^{-i2\pi k\frac{n}{N}} \tag{3.41}$$

Die wichtigen Unterschiede zwischen den komplexen DFT nach Formel 3.40 und der reellen DFT nach Formel 3.30 Formelsind (a) die Verwendung des Sinus als komplexe Zahl $i\cdot\sin$, sowie das invertieren seines Vorzeichens und (b) die Summierung über $k = 0\ldots N-1$ anstatt $k = 0\ldots N/2$. Der Frequenz-Bereich wird nun durch das complexe Signal $X[\,]$ ausgedrückt, welcher die Koeffizienten $A[\,], B[\,]$ in seinem Real/Imaginärteil beinhaltet. Der Frequenz-Bereich $X[\,]$ hat die selbe Länge wie der Zeit-Bereich $x[\,]$, das heißt $\mathrm{Length}(x[\,]) = \mathrm{Length}(X[\,])$. [30, S. 571]

Es gelten die die folgenden Zusammenhänge zwischen der reellen und der komplexen DFT 3.42. Es lässt sich ableiten, dass für die Indexe $k = 0\ldots N/2$ die Real/Imaginärteil der komplexen DFT $X[\,]$ den Koeffizienten $A[\,], B[\,]$ der reellen DFT entspricht. Da

$|X[0]|, \ldots, |X[N/2]| = M[\,]$, wird in dieser Arbeit $|X[0]|, \ldots, |X[N/2]|$ ebenfalls als *Spetrum* bezeichnet. [30, S. 225 - 226, 555]

$$\begin{aligned} \Re(X[0]), \ldots, \Re(X[N/2]) &= A[0], \ldots, A[N/2] = A[\,] \\ \Im(X[0]), \ldots, \Im(X[N/2]) &= B[0], \ldots, B[N/2] = B[\,] \\ |(X[0])|, \ldots, |(X[N/2])| &= M[0], \ldots, M[N/2] = M[\,] \\ \phi(X[0]), \ldots, \phi(X[N/2]) &= \phi[0], \ldots, \phi[N/2] = \phi[\,] \end{aligned} \tag{3.42}$$

Gleichung 3.43 definiert die dazu inverse Operation, die *komplexe inverse DFT* (iDFT) in *polrarer Notation*. [30, S. 572]

$$\text{iDFT}\{X[\,]\} = x[\,] := \quad \underset{n=0}{\overset{N-1}{\forall}} : \; x[n] = \sum_{k=0}^{N-1} X[k] \cdot e^{i2\pi k \frac{n}{N}} \tag{3.43}$$

Die Frage ist: Wenn das Signal des Zeit-Bereiches $x[\,]$ aus reellen Zahlen besteht, und das Signal des Frequenz-Bereiches $X[\,]$ aus komplexen Zahlen, wie „verschwinden “ diese komplexen Zahlen wieder bei der Berechnung der inversen DFT?

Genau wie das Signal des Zeitbereiches $x[\,]$ als unendlich und periodische außerhalb des transformierten Bereiches $x[0] \ldots x[N-1]$ angenommen wird, ist auch der Frequenz-Bereich unendlich und periodisch außerhalb des Bereiches $X[0] \ldots X[N-1]$. Gleichung 3.44 definiert diese Aussage. [30, S. 572]

$$\forall k \in \mathbb{Z} : X[n + kN] = X[n] \tag{3.44}$$

Daraus lässt sich Schlussfolgern, dass $X[-N/2], \ldots, X[-1] = X[N/2], \ldots, X[N-1]$. Die Werte $X[-N/2] ... X[-1]$ werden als die *negativen Frequenzen* bezeichnet. Dazu kommt der Fakt, dass der Frequenz-Bereich in Bezug auf das Interval $X[-N/2], \ldots, X[N/2]$ eine Symmetrie aufweist: Der Realteil ist Achsensymmetrisch an der Stelle $X[0]$, und der Imaginäre Teil Punktsymmetrisch. Formel 3.45 definiert diese Symmetrie, und Abbildung 3.13 visualisiert sie Anhand eines Beispiels. Diese Symmetrie tritt nur auf, falls das Signal im Zeitbereich nur aus Reellen Zahlen besteht, was bei der Arbeit mit herkömmlichen akustischen Signalen immer erfüllt ist. Auf die Herleitung dieser Symmetrie wird an dieser Stelle aus Platzgründen verzichtet. [30, S. 574]

$$\underset{n=0}{\overset{N-1}{\forall}} : \; Re(X[n]) = \Re(X[-n]) \; \wedge \; \Im(X[n]) = -\Im(X[-n]) \tag{3.45}$$

Diese Symmetrie ist der Grund dafür, warum bei der inversen DFT nach Formel 3.43 nach der Synthese alle Imaginärteile verschwinden. Zum besseren Verständnis wird die

polare Notation der komplexen inversen DFT zur *kartesischen Notation* nach Formel 3.46 erweitert .[30, S. 573]

$$\begin{aligned}\text{iDFT}\{X[\,]\} = x[\,] := \quad \underset{n=0}{\overset{N-1}{\forall}} :\ x[n] = &\sum_{k=0}^{N-1} \Re(X[k]) \cdot (\cos(2\pi k\frac{n}{N}) + i\sin(2\pi k\frac{n}{N})) \\ &- \sum_{k=0}^{N-1} \Im(X[k]) \cdot (\sin(2\pi k\frac{n}{N}) - i\cos(2\pi k\frac{n}{N}))\end{aligned} \tag{3.46}$$

Das weitere vorgehen wird anhand des Beispiels aus Abbildung 3.13 erklärt. Aus Formel 3.46 geht hervor, dass jeder Realteil von $X[\,]$ zu einer reellen Cosinuswelle und einer imaginären Sinuswelle beiträgt. Jeder Imaginärteil trägt zu einer reellen Sinuswelle und einer imaginären Cosinuswelle bei. Angenommen, im Zeit-Bereich soll eine Cinuswelle mit der Frequenz $f/f_s = 0.23$ und der Amplitude 1 synthetisiert werden, also $x[n] = 1 \cdot \cos(2\pi 0.23n)$. Für die Synthese wird sowohl eine positive als auch eine quenz benötigt. Marker 1 in Abbildung 3.13 erzeugt somit eine relle Cosinuswelle und eine imaginäre Sinuswelle im Zeitbereich mit der Frequenz 0.23 und der Ampltitude 0.5, also $0.5 \cdot \cos(2\pi 0.23n) + i\sin(2\pi 0.23n)$. Genau so trägt die Negative Frequenz von -0.23 bei Marker 2 zu einer reellen Cosinus-Welle und einer imaginären Sinuswelle im Frequenzbereich bei, also $0.5 \cdot \cos(2\pi(-0.23)n) + i\sin(2\pi(-0.23)n)$. Nach der Beziehung $\cos(x) = \cos(-x)$ und $\sin(x) = -\sin(-x)$ lässt sich der Beitrag der negativen Frequenzen umformen zu $0.5 \cdot \cos(2\pi 0.23n) - i\sin(2\pi 0.23n)$. Die Addition dieser beiden Beiträge des reelen und des imaginären Teils, welche bei der Synthese vorgenommen wird, wird in Formel 3.47 zusammengefasst. Wie zu sehen ist, addieren sich die reellen Cosinuswellen, während sich die imaginären Sinuswellen auslöschen. [30, S. 573- 574]

$$\begin{array}{rl} & 0.5 \cdot \cos(2\pi 0.23n) + i\,0.5 \cdot \sin(2\pi 0.23n) \\ + & 0.5 \cdot \cos(2\pi 0.23n) - i\,0.5 \cdot \sin(2\pi 0.23n) \\ \hline & 1 \cdot \cos(2\pi 0.23n) \end{array} \tag{3.47}$$

In der selben weise kann eine reele Sinuswelle im Zeitbereich mit Hilfe der imaginären Werte von $X[\,]$ synthetisiert werden. Marker 3 zeigt die positive Frequenz 0.23 und 4 die Negative Frequenz -0.23 im Imaginärteil von X. Folgt man dem selben Prinzip der oben vorgestellten Rechnung zur Synthese einer Cosinuswelle, ergibt sich die in Gleichung 3.48 Addition. Dieses mal löschen sich die Cosinuswellen aus, während die Sinuswellen addiert werden. Außerdem ist zu sehen, dass das Vorzeichen der Sinuswelle invertiert wird. Diese Invertierung muss entweder bei der Synthese oder der Dekomposition mit einbezogen werden und ist der Grund für die invertierung des Vorzeichens vor dem Sinus in Gleichung 3.40. [30, S. 574]

$$\begin{array}{rl} & -\,0.5 \cdot \sin(2\pi 0.23n) - i\,0.5 \cdot \cos(2\pi 0.23n) \\ & -\,0.5 \cdot \sin(2\pi 0.23n) + i\,0.5 \cdot \cos(2\pi 0.23n) \\ \hline - & \quad 1 \cdot \sin(2\pi 0.23n) \end{array} \tag{3.48}$$

Die hier vorgestellte Beispielrechnung verdeutlicht, wie die Symmetrie des Frequenzbereiches genutzt wird, um bei der Summation einer positiven und der entsprechenden

negativen Frequenz zu einem rein reellwerteigen Signal im Frequenz-Bereich zu führen. Wird der Frequenzbereich jedoch durch den Menschen betrachtet, wird , wie in Kapitel 3.4.1 beschrieben, dass als *Spectrum* bezeichnete $M[\,]$-Signal verwendet, definiert in Formel 3.42 als der Absolutwert des komplexen Frequenz-Bereiches. Dazu wird die Betrachtung auf den Indexbereich $n = 0, \ldots, N/2$ eingeschränkt, was dem Index-Bereich der reellen DFT entspricht, da die negativen Frequenzen aufgrund ihrer Symmetrie bei der Betrachtung redundante Informationen darstellen. Daraus ergibt sich, dass es für die Erzeugung des Spectrums rein mathematisch betrachtet unerheblich ist, ob die reelle oder die komplexe DFT verwendet wird. Abbildung 3.14 gibt einen zusammenfassenden Überblick über die Indexierung der des Zeit- und Frequenzbereiches bei verwendung der komplexen DFT. [30, S. 225 - 226]

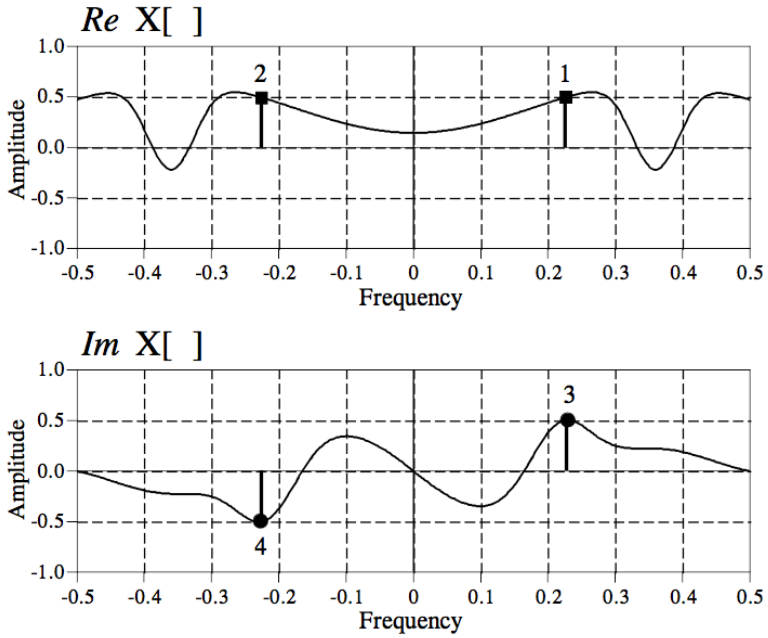

Abbildung 3.13: Symmetrie des Frequenzbereiches in den Indexen $n = (-N/2 = -0.5 * N), \ldots, (N/2 = 0.5 * N)$. Die Marker 1 und 2 haben eine Amplitude von $+0.5$, der Marker 3 hat eine Amplitude von -0.5 und Marker 4 eine Amplitude von $+0.5$

Wie zur Einleitung dieses Kapitels erwähnt wurde, wird die Berechnung der komplexen DFT und inversen DFT mit Hilfe eines Algorithmus implementiert, der als *Fast Fourier Transformation* (**FFT** bzw. **iFFT**). Die Funktionsweise dieses Algorithmus wird an dieser Stelle aus Platzgründen nicht näher erläutert. Schlussendlich ist das Ergebnis der FFT bzw. iFFT der Frequenz-/Zeit-Bereich nach den in diesen Kapitel vorgestellten definitionen. Die FFT fordert als einzige voraussetzung, dass die Länge des Signals als Potenz von zwei gewählt wird, das heißt $\text{Length}(x[\,]) = N = 2^k$. [30, S. 225 - 226]

3.4.3 Short Time Fourier Transform

Wie aus Kapitel 3.4.2 hervorgeht, enthält die Signal des Zeit-Bereiches die Informationen über den zeitlichen Verlauf des Signals, während der Frequenz-Bereich Aufschluss über Frequenzanteile des Signales gibt. Abbildung 3.15 visualisiert den Zusammenhang: Oben

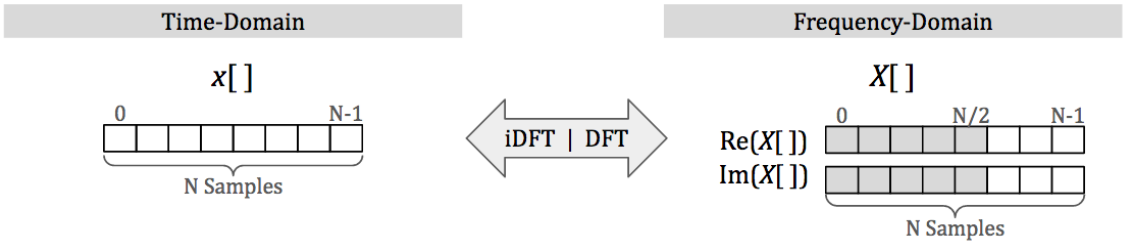

Abbildung 3.14: Überblick über die komplexe DFT und inverse DFT

ist der Zeit-Bereich eines 1.8 Sekunden langen Signals zu sehen. Es können klar drei nacheinander gespielte Töne erkannt werden. Der Zeit-Bereich macht nicht erkennbar, welche Frequenz-Anteile in den Tönen enthalten sind. Unten ist das Frequenz-Spectrum (Magnituden-Signal im Bereich $n = 0, \ldots, N/2$) abgebildet. Die x-Achse bezeichnet die Frequenz von 0 bis 22 050 Hz, die x und die y-Achse werden logarithmisiert dargestellt. Wie zu erkennen ist, gibt das Frequenz-Spectrum einen Eindruck der Frequenz-Anteile des Signals, aber die zeitliche Information über die drei Töne geht verloren.

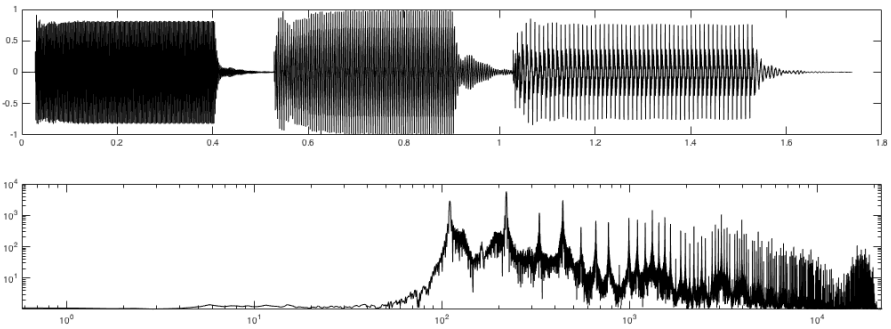

Abbildung 3.15: Ein 1.8-Sekunden langes Signal. Oben: Der Zeitbereich mit drei klar erkennbaren Events. Unten: Das Frequenz-Spectrum des gesamten Signals mit logarithmisierten Achsen.

Es ist wünschenswert, einen Kompromiss aus den Vorteilen beider Bereiche zu finden, in dem man das Frequenz-Spectrum kürzerer Zeitabschnitte des Signals bildet. Dazu wird der Zeit-Bereich $x[\,]$ in Fenster der Länge M zerlegt. Die zeitliche Differenz zwischen zwei Fensterns wird als *Hoptime* R bezeichnet. Gleichung definiert die Bildung des Signal-Fensters $x_m[\,]$.[29]

$$x_m[\,] := \quad \forall_{n=0}^{M-1} : \; x_m[n] = x[n + m \cdot R] \tag{3.49}$$

Abbildung 3.16 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 $M = 22050/f_s = 0.5\,\mathrm{s}$ und die Hoptime $R = M/2 = 0.25\,\mathrm{s}$.

Als Vorbereitungsschritt für die Transformation der Signal-Fenster in den Frequenz-Bereich wird nun jedes Fenster mit einer sogenannten *Fensterfunktion* (engl *window*) $w[\,]$ multipliziert. Gleichung 3.50 definiert eine der am weitesten verbreiteten Fenster-Funktionen,

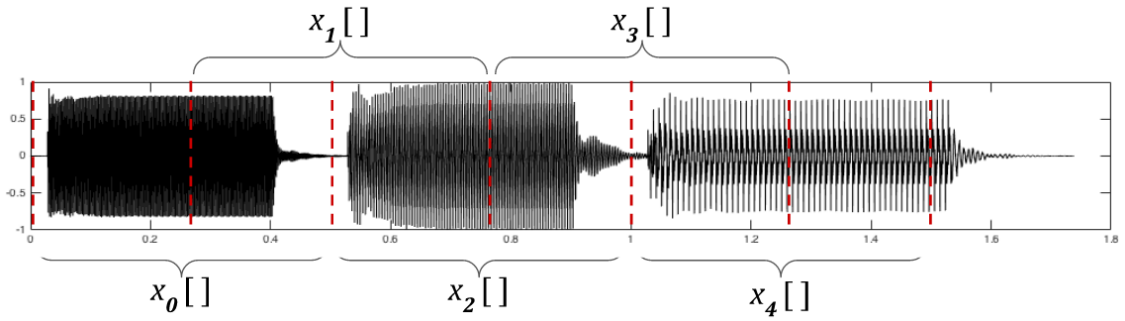

Abbildung 3.16: Zerlegung eines Signals in Singal-Fenster

das *Hamming-Window*. Der Paramter M gibt die länge des Fensters an. Abbildung 3.17 visualisiert das Hamming-Window. [30, S. 286]

$$w[\,] := \quad \underset{n=0}{\overset{M-1}{\forall}} : \; w[n] = 0.54 - 0.46\cos(\frac{2\pi n}{M}) \tag{3.50}$$

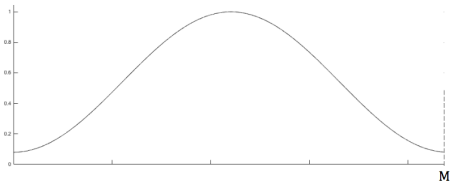

Abbildung 3.17: Das Hamming-Window

Die Gleichung 3.51 definiert die *Kurzzeit-Fourier-Transformation* (engl *Short Time Fourier Transformation*, *STFT*), implementiert mit Hilfe der DFT. Dabei wird das Signalfenster $x_m[\,]$ mit der Fensterfunktion $w[\,]$ multipliziert und in das *Frequenz-Fenster* $X_m[\,]$ transformiert.[5] Abbildung 3.18 visualisiert die STFT des Beispiels aus Abbildung 3.16.

$$\text{STFT}\{x_m[\,]\} = X_m[\,] := \quad \underset{k=0}{\overset{M-1}{\forall}} : \; X_m[k] = \sum_{n=0}^{M-1} x_m[n] \cdot w[n] \cdot e^{-i2\pi k \frac{n}{N}} \tag{3.51}$$

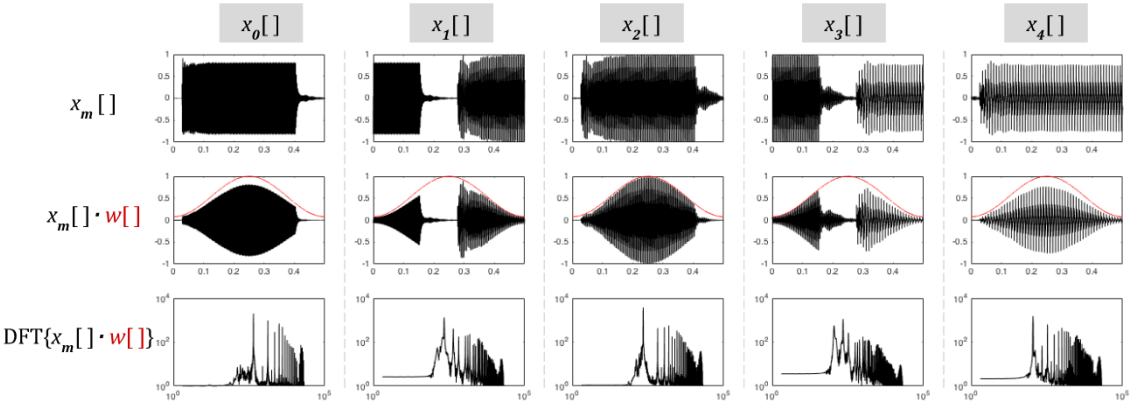

Abbildung 3.18: STFT des Beispiel-Signals aus Abbildung 3.16

3.5 Filter

Ein *Filter* ist eine Funktion, welches ein Eingangs-Signal $x[\,]$ auf ein Sample eines Ausgangssignals $y[n]$ abbildet. Er wird mit Hilfe des sogenannten *Signal Operators* T_n notiert, wie Gleichung 3.52 definiert. [13, „Definition of a Filter“]

$$T_n\{x[\,]\} = y[n] \tag{3.52}$$

Auf Basis des 3.52 wird die *Transformation* $T\{\ \}$ als Operation definiert, welche das Signal $x[\,]$ auf ein Signal $y[\,]$ abbildet, in dem T_n für meherere n angewandt wird, wie Gleichung 3.53 definiert. Diese Notation ist bereits bekannt aus den Gleichungen der DFT und iDFT bekannt (Gleichungen 3.41, 3.40 und 3.43). [13, „Definition of a Filter“]

$$T\{x[\,]\} = y[\,] := \quad \overset{n_2}{\underset{n=n_1}{\forall}} \ : \ T_n\{x[\,]\} = y[n] \tag{3.53}$$

Besonders interesant im Zusammenahng mit Audioverarbeitung ist die Klasse der *Lineare, Zeit-Invariante Filter* (engl. *Linear Time-Invariant Filters*, **LTI**-Filter), notiert mit durch $L_n\{\}$. Dabei handelt es sich um Filter, die die folgenden drei Eigenschaften erfüllen:

Scaling , das heißt, dass eine Skalierung des Input-Signals $x[\,]$ um den Faktor α zur Skalierung des Output-Samples $y[n]$ um den selben Faktor führt, wie Gleichung 3.54 definiert.

$$L_n\{\alpha \cdot x[\,]\} = \alpha \cdot y[n] \tag{3.54}$$

Super-Position , da heißt, dass das Anwenden des Filters auf die Summmer zweier Signale zum selben Ergebnis führt wie die die isolierte Anwendung des Filters auf die beiden Signale mit nachgelagerter Summation, wie Gleichung 3.55 definiert. [13, „Linear Filters“]

$$L_n\{x_1[\,] + x_2[\,]\} = L_n\{x_1[\,]\} + L_n\{x_2[\,]\} \tag{3.55}$$

Zeit-Invariant , das heißt, dass die Verzögerung des Eingangssingals um einen Faktor n_0 zur selben Verzögerung der Ausgangs-Samples führt, wie Gleichung 3.56 definiert.[19, Filters III, S. 2 - 5]

$$L_{n-n_0}\{x[\,]\} = y[n - n_0] \tag{3.56}$$

An dieser Stelle werden verschiedene Betrachtungsweisen von Filtern vorgestellt. Jede dieser *Repräsentationen* definiert den Filtern vollständig, bietet jedoch unterschiedliche Betrachtungsweisen des Einflusses des Filters auf das Eingangssignal. [19, Filters III, S. 1]

3.5.1 Differenzengleichung

Eine Variante der Beschreibung eines Filters ist mit Hilfe einer *Differenzengleichung* nach Gleichung 3.58. In dieser Arbeit wird ein Linearer Filter, welcher durch eine Differenzen-Gleichung implementiert wird, mit dem Operator $ab_n\{\ \}$ gekennzeichnet. Ein solcher Filter

wird vollständig durch die Koeffizienten $a_1, \ldots, a_N$ und $b_0, \ldots, b_M$ definiert. Ein Filter hat somit N a-Glieder und $M + 1$ b-Glieder. [13, „Difference Equation“]

$$\begin{aligned} ab_n\{x[\,]\} = y[n] = & \overbrace{b_0 x[n] + b_1 x[n-1] + \ldots + b_M x[n-M]}^{\text{Feed-Forward}} \\ & \underbrace{- a_1 y[n-1] - \ldots - a_N y[n-N]}_{\text{Feed-Back}} \\ & = \sum_{i=0}^{M} b_i x[n-i] - \sum_{j=1}^{N} a_j y[n-j] \end{aligned} \tag{3.57}$$

Die Glieder $b_0, \ldots, b_M$ sind die Koeffizienten des aktuellen Samples und vergangener Samples des Einganssignals, und werden auch als *Feed-Forward*-Koeffizienten bezeichnet. Die Glieder $a_1, \ldots, a_N$ sind die Koeffizienten vergangener Samples der Antwort $y[\,]$. Sie werden daher auch als *Feed-Back*-Koeffizienten oder *rekursive Glieder* bezeichnet. Verwendet ein Filter nur b-Glieder und kein a-Glied, wird er als *Finite Impulse Response*-Filter (**FIR**-Filter) bezeichnet. Verwendet ein Filter zumindest ein a-Glied, wird er als *Infinite Impulse Response*-Filter (**IIR**-Filter) bezeichnet.[13, „Difference Equation“]

$$\begin{aligned} \text{FIR} := &\quad \forall j = [1, N] : \; a_j = 0 \\ \text{IIR} := &\quad \exists j \in [1, N] : a_j \neq 0 \end{aligned} \tag{3.58}$$

Die *Ordnung* eines Filters wird definiert als die $\max(M, N)$. [13, „Filter Order“].

$>$ Hier folgt ein Beispiel für Filter $<$.

3.5.2 Faltung

Die *Faltung* (engl. *Convolution*) ist eine der Zentralen Operationen zwischen zwei Signalen, so wie die Addition oder die Mulitplikation. Sie wird mit dem Symbol $*$ notiert. Sie wird notiert mit $x[\,] * h[\,] = y[\,]$.

Die Faltung basiert auf der sogenannten *Faltungs-Summe*, definiert in Gleichung 3.59. In diesem Zusammenhang wird $x[\,]$ Eingangs- und $y[\,]$ als Ausgangs-Signal bezeichnet. Je nach Anwendungsfall bekommt $h[\,]$ den Namen *Faltungs-Kernel*, *Filter-Kernel* oder einfach *Kernel*. In dieser Arbeit wird die Faltungs-Summe mit dem Operator $h_n\{\,\}$ definiert. [30, S. 107-108]

$$h_n\{x[\,]\} = y[n] = x[n] * h[n] = \sum_{i=1}^{M} h[i] \cdot x[n-i] \tag{3.59}$$

Wird die Faltung auf alle Samples des Signal $x[\,]$, wird das Ausgangs-Signal $y[\,]$ im Vergleich zum Eingangs-Signal um die Länge des Faltungskern verlängert, wie Gleichung 3.60 definiert. $x[\,]$ ist ein Signal mit $\text{Support}(x[\,]) = [0, N-1]$ und $\text{Length}(x[[\,]) = N$. $h[\,]$ ist ein Signal mit $\text{Support}(h[\,]) = [0, M-1]$ und $\text{Length}(h[\,]) = M$ und $y[\,]$ ist ein Signal mit $\text{Support}(y[\,]) = [0, N+M-2]$ und $\text{Length}(y[\,]) = N + M - 1$. um die Länge

des Faltungskerns verlängert wird. Abbildung 3.20 zeigt ein Beispiel für die Faltung.[30, S. 115-120]

$$h\{x[\,]\} = x[\,] * h[\,] = y[\,] := \quad \overset{N+M-2}{\underset{n=0}{\forall}} : \; x[n] * h[n] = y[n] \tag{3.60}$$

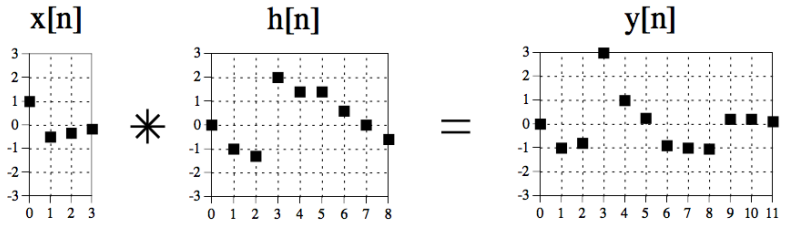

Abbildung 3.19: Beispiel für die Faltung

Das neutrale Element der Faltung ist der *Delta-Funktion* $\delta[\,]$, definiert in Gleichung 3.61 . Das heißt, dass $x[\,] * \delta[\,] = x[\,]$. Die Faltung ist kommutativ, das heißt $x[\,] * h[\,] = h[\,] * x[\,] = y[\,]$. [30, S. 107, 113] Weitehrin ist die Faltung assoziativ, das heißt $(x[\,] * y[\,]) * z[\,] = x[\,] * (y[\,] * z[\,])$. [30, S. 133]

$$\delta[\,] := \quad \overset{\infty}{\underset{n=-\infty}{\forall}} : \; \delta[n] = \begin{cases} 1 & , n = 0 \\ 0 & , n \neq 0 \end{cases} \tag{3.61}$$

Die Faltung ist eine der Varianten der Umsetzung von linearen, zeit-invarianten Filtern, neben den in Kapitel 3.5.1 vorgestellten Differenzengleichungen. Während ein Filter $L_n\{\,\}]$ in der Repräsentation als Differenzen-Gleichung vollständig durch die Koeffizienten a und b beschrieben wird, wird bei der Faltungs-Repräsentation ein Filter vollständig durch die Impulsantwort $h[\,]$ beschrieben. Um die Impulsantwort für einen Filter zu erhalten, der zunächst durch einen anderen linearen Filter repräsentiert wird, filtert man den Delta-Impuls mit diesem Filter, wie Gleichung 3.62 definiert. Wird also beispielsweise ein Filter als Differenzengleichung definiert, aber soll mit Hilfe der Faltung umgesetzt werden, so erhält man die Impulsantwort durch $h[n] = ab_n\{\delta[n]\}$.[13, Impulse-Response Representation]

$$h[\,] := \quad \overset{\infty}{\underset{n=0}{\forall}} : \; h[n] = L_n\{\delta[n]\} \tag{3.62}$$

Daraus ergibt sich, dass bei FIR-Filtern die Impuls-Antwort endlich ist, das heißt $\text{Length}(h[\,]) = M \neq infty$, womit bei einem endlichen Eingangssignal $x[\,]$ auch das Ausgangssignal $y[\,]$ endlich wird. Bei IIR Filtern ist die Impulsantwort hingegen unendlich, da heißt $\text{Length}(h[\,]) = M = infty$, womit auch das Ausngangssignal $y[\,]$ zwangsweise eine unendliche Länge erhält. [13, „The Finite in FIR“]

3.5.3 Multiplikation im Frequenz-Bereich

Die Faltung im Zeit-Bereich nach den in Kapitel 3.5.3 vorgestellten Prinzipien entspricht einer Multiplikation im Ortsbereich, wie Formel 3.63 definiert. Das Prinzip wird auch als *FFT-Convolution* bezeichnet. Dazu werden zunächst das Eingangssignal und die Impulsantwort mit Hilfe der DFT in den Frequenz-Bereich transformiert, also $\text{DFT}\{x[\,]\} = X[\,]$ und

$\text{DFT}\{h[\,]\} = H[\,]$. Diese beiden Frequenz-Bereiche werden nun miteinander Multipliziert, um den Frequenz-Bereich des Ausganssignals zu erzeugen, das heißt $\forall n = [0, N-1] : X[n] \cdot H[n] = Y[n]$. Mit Hilfe der inversen DFT wird der Frequenz-Bereich in den Zeit-Bereich zurücktransformiert, also $\text{iDFT}\{Y[\,]\} = y[\,]$. Das so erzeugte Ausgangs-Signal entspricht dem Signal, welches durch die Faltung $x[\,] * h[\,] = y[\,]$ entstanden wäre.[30, S. 182]

$$x[\,] * h[\,] = y[\,] = \text{iDFT}\Big\{\ \text{DFT}\{x[\,]\} \cdot \text{DFT}\{h[\,]\}\ \Big\} \tag{3.63}$$

In Kapitel wurde erwähnt, dass sich ein Signal $x[\,]$ durch die Faltung um die Länge der Impulantwort ausdehnt. Daher muss sichergestellt werden, dass das Signal $x[\,]$ vor der Transformation in den Frequenz-Bereich mindestens $\text{Length}(h[\,]) = M$ 0-Samples an seinem Ende hat, welche gegebenenalls zuerst angehangen werden müssen. Dadurch wird der Frequenz-Bereich von $x[\,]$ nur dahingehend beeinflusst, dass seine Auflösung erhöht wird. Damit die Frequenz-Bereiche $X[\,]$ und $H[\,]$ multipliziert werden können, müssen sie die selbe Länge haben. Da die Impulsantwort meistens kürzer als das Eingangssignal ist, müssen ihr vor der DFT meist 0-Samples angehangen werden, um sie auf die selbe länge wie das Eingangsssignal „zu strecken“.[30, S. 183 -184]

Ein weiterer Vorteil der FFT-Convolution ist, dass Einblick über den Einfluss des Filters auf den Frequenz-Bereich des Eingangs-Signals zur Erzeugung des Ausgangs-Signals gibt. Abbildung 3.20 verdeutlicht dieses Prinzip am Beispiel des „windowed Sync Filters“. Links ist der Zeit-Bereich des Filter-Kernels zu sehen, rechts der Frequenzbereich $\text{DFT}\{h[\,]\} = H[\,]$. Der Frequenz-Bereich macht deutlich dass es sich bei der Impuls-Antwort um einen Tiefpass-Filter handelt, also ein Filter, welcher hohe Frequenzen des Eingangs-signals $x[\,]$ bei seiner Anwendung blockiert, aber tiefe Frequenzen jedoch passieren lässt.[30, S. 180]

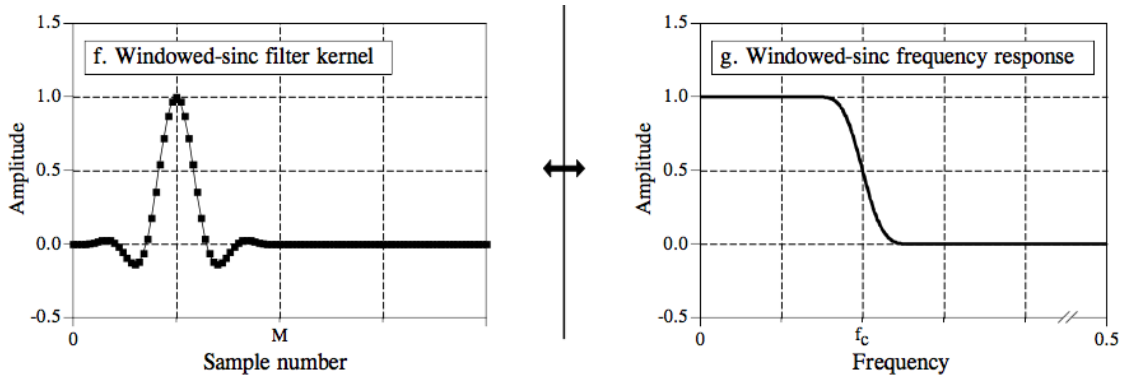

Abbildung 3.20: Links: Zeit-Bereich einer Impulsantwort $h[\,]$ des „windowed Sync Filters“. Rechts: Frequenz-Bereich dieser Impulsantwort $\text{DFT}\{h[\,]\} = H[\,]$ [30, S. 287]

Da ein FIR-Filter eine endlich lange Impulsantwort hat, folgt daraus auch ein endlichlanger Frequenz-Bereich. Ein IIR-Filter hingegen hat einen unenlich lange Impulsantwort und somit auch einen unendlich langen Frequenz-Bereich.

3.6 akustische Modellierung der menschlichen Stimme

An dieser Stelle wird die ein akustisches Modell der menschlichen Stimme vorgestellt. Folgende Organe sind an der Produktion der menschlichen Stimme beteiligt. Abbildung 3.21 visualisiert diese Organe schematisch.

1. **Lunge / Lung**, welcher einen Luftstrom als Grundlage der Lutäußerung erzeugt
2. **Stimmbäner / Vocal Chords**, platziert im Kehlkopf (engl. Larynx). Sie erzeugen auf Basis des Luftstroms einen Ton, bezeichnet als *Glottal Source*. Werden die Stimmbänder leicht gespannt, vibrieren sie und erzeugen einen Ton, dass heißt ein im zeit-bereich (quasi-) periodisches Signal (engl. *periodic Source*). Sind die Stimmbänder stark gespannt und nur leicht geöffnet, erzeugen Sie ein Rauschen (engl. *turbulance Source*). Ein durch vibration erzeugter Ton wird als *stimmhaft* bezeichnet, ein durch Rauschen erzeugter Ton als *stimmlos*. Der Ton wür über den Schlund weitergegeben (engl Pharynx) weitergegeben.
3. **Vokaltrakt / Vocal-Tract**, bestehend aus der Mundhöhle und den Namenraum. Das Velum bestimmt, ob der Ton der Stimmbänder in die Mundhöhle oder den Nasenraum weitergeleitet wird. Je nach Stellung der Zunge, Kiefer, Lippen etc. wird der von den Stimmbändern erzeugte Ton verschieden moduliert. Resonanzen, die durch den Vokaltrakt entstehen, werden als *Formanten* bezeichnet. [12, S. 62] [16]

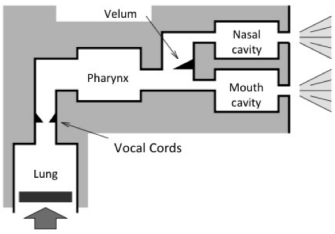

Abbildung 3.21: Schematische Übersicht über die Organe der Spracherzeugung [16]

Die menschliche Laut-Prudktion wird nach dem so genannten *Source-Filter-Modell* modelliert. Der periodische Ton, der durch die Stimmbänder erzeugt wird, wir angenähert durch einen Impuls-Zug (produziet durch die Stimmbänder), welcher durch den Schlund als linearen Filter leicht 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 Vocaltrakt weitergeben, welcher als linearer, zeitinvarianter Filter mit der Impulsantwort $v[\,]$ angenommen wird. Diese Impulsantwort ist abhängig von der Konfiguraiton der Organe des Vokaltraktes. Die Lippen werden als zweiter linearer, zeit-invarianter Filter mit der Impulsantort $r[\,]$. Das Ausgangssignal $y[\,]$ entsteht also aus der glottal Source $u[\,]$ und den zwei linearen, zeit-invarianten Filtern nach Gleichung 3.64. dargestellt als Faltung im Zeit-Bereich oder Mulitplikation im Frequenz-Bereich. Abbildung 3.22 visualsiert dieses Vorgehen Schematisch. [12, S. 62 - 63] [16]

$$\begin{aligned} u[\,] * v[\,] * r[\,] &= y[\,] \\ U[\,] \cdot V[\,] \cdot R[\,] &= Y[\,] \end{aligned} \tag{3.64}$$

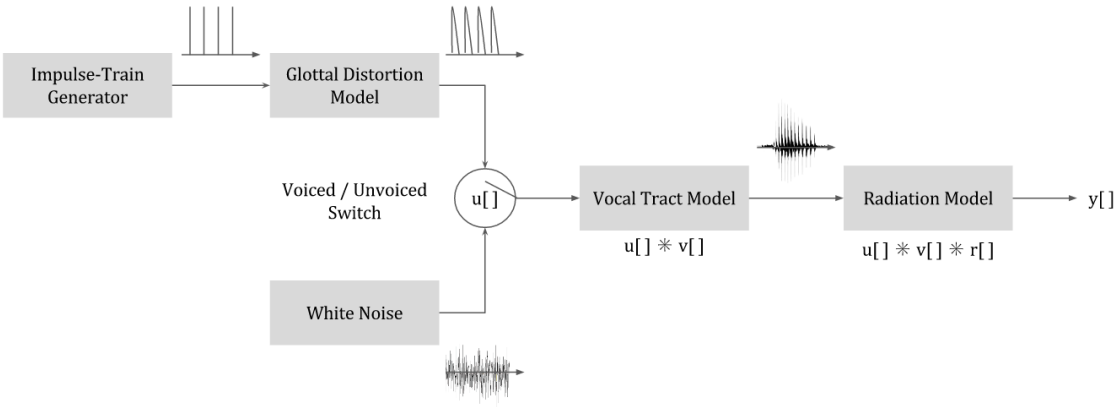

Abbildung 3.22: Schematische über das Source-Filter-Model [9, S. „Source estimation“, S. 17]

Abbidlung **??** zeigt die Zeitbereiche der periodic und der turbulance Source im Detail. Wie zu sehen ist, bestimmt der zeitliche Abstand zwischen den Impulsen die Grund-Frequenz der Stimme. Dieses Signal $p[\,]$ wird durch die Stimmbändern gefiltert und um den Zeit-Bereich der periodic Source ensteht $G\{p[\,]\} = u_p[\,]$. Darunter ist der Zeit-Bereich des weißen Rauschen zu sehen. [17, Source]

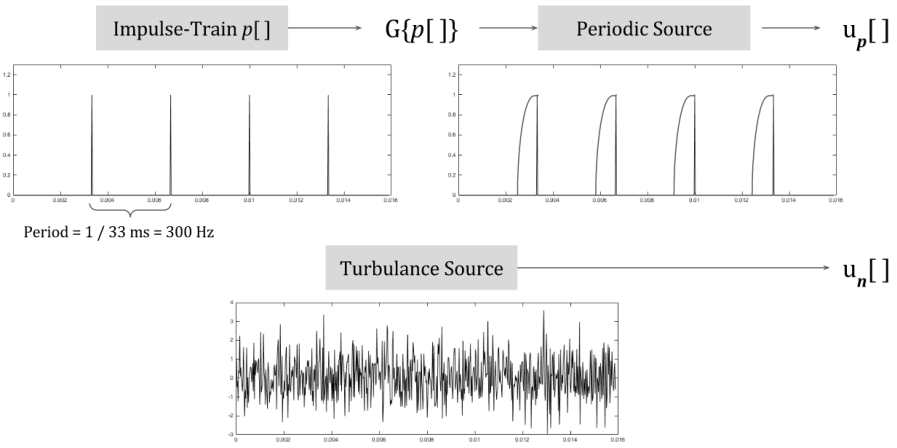

Abbildung 3.23: Zeit-Bereiche der periodic und der turbulance Source [17, Source]

Abbildung 3.24 zeigt die Betrachtung der Frequenz-Bereiche des Source-Filter-Modells. Die periodic Source ($U[\,]$ links) zeichnet sich im Frequenz-Bereich durch gleichmäßig verteilte Spitzen aus, die mit steigender Frequenz an Amplitude verlieren. Jeweils zwei Spitzen haben einen Abstand von f zueinander, also in diesem Fall 300 Hz. Die Frequenz-Antwort des Vocal Tract Model $V[\,]$ zeichnet sich durch Resonanz-Frequenzen aus, in diesem Beispiel sind 4 erkennbar. Das Radiation Model $R[\,]$ wird als Hochpass-Filter angenähert, also ein Filter, welcher hohe Frequenzen passieren lässt und tiefe blockiert. Das Ausgangssignal $U[\,] \cdot V[\,] \cdot R[\,] = Y[\,]$ zeigt den Einfluss der Filter auf das Eingangssignal.[9, Source estimation], [17, Vocal Tract Resonance]

Abbildung 3.25 zeigt anhand des Frequenz-Spectrums eines stimmhaften Sprachsignals, wie die die Grundfrequenz und die harmonischen Obertonwellen erkannt werden: Die Grundfrequenz N_0 ist Bezüglich des Zeit-Bereiches nach Formel 3.8 definiert. Im Frequenz-Bereich

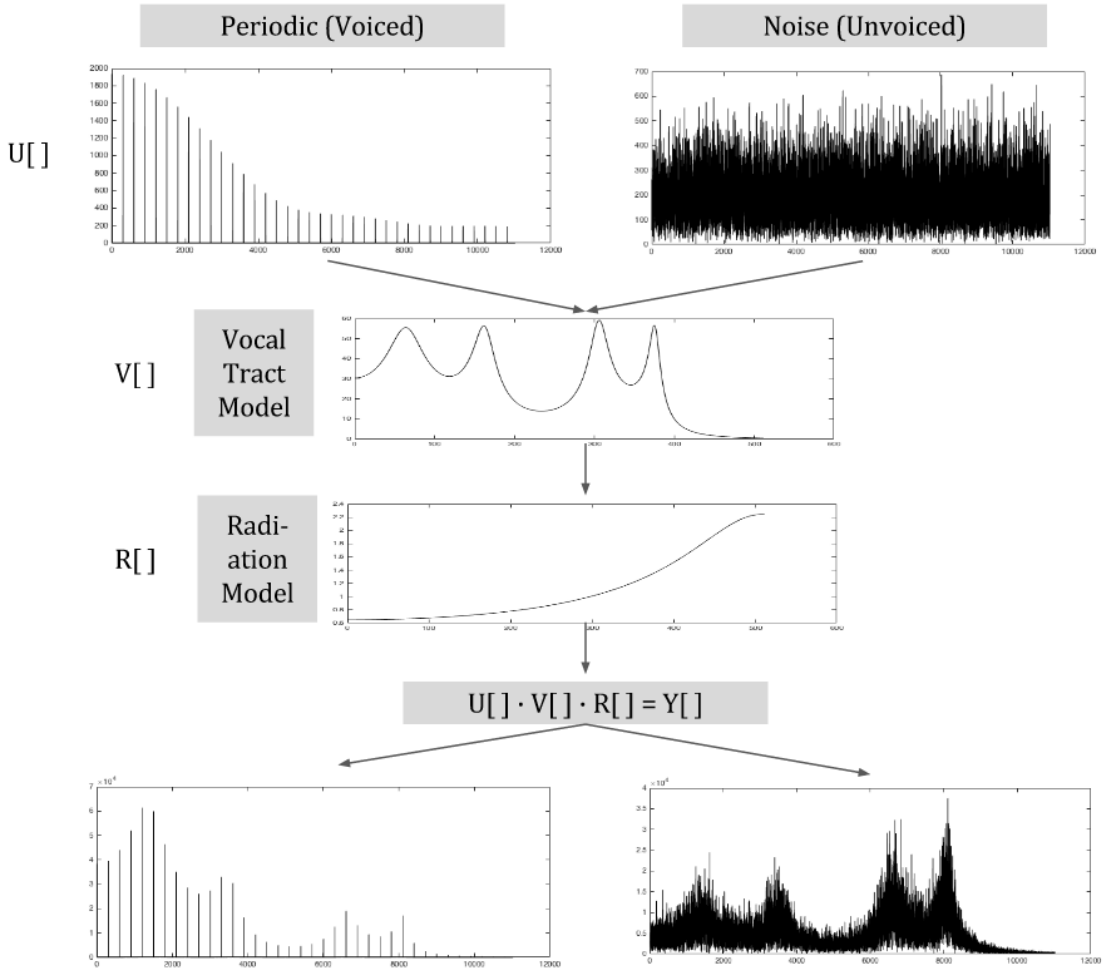

Abbildung 3.24: Betrachtung der Frequenz-Bereiche des Source-Filer-Modell

sind die Grund-Frequenz und die harmonischen Oberwellen als die „vielen, kurzen Spitzen“ erkennbar. Die Frequenz der ersten dieser Spitzen entspricht der Grund-Frequenz N_0, in diesem Beispiel 250.7 Hz. Die harmonsichen Oberwellen entsprechen dem doppelten, dreifachen, ... dieser Grundfrequenz und werden bezeichnet hals $H_1, H_2, \ldots$. Die Grundfrequenz ist *nicht zwingend* die Spitze der höchsten Amplitude! Durch den Einfluss des Vokaltrackts als Filter können harmonische Oberwellen eine höhere Amplitude als die Grundfrequenz haben. Vielmehr lässt sich durch den kleinsten gemeinsamen Teiler der Frequenzspitzen auf die Grundfrequenz schließen.[3, S. 52 - 53]

Abbildung 3.27 [1] verdeutlicht, wie das als linearer, Zeit-Invarianter Filter modellierte Vokaltrakt mit Hilfe von Formanten beschrieben wird. Die Formanten spielen vor allem bei der Beschreibung von Vokalen eine Rolle. Formanten sind lokale Maxima im Spektrum, die dadurch erzeugt werden, dass der Vokaltrakt Resonanzfrequenzen 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. Mit steigender Frequenz nimmt die Amplitude der Formanten ab, der dominanteste Formant ist also immer der erste. Daher werden

[1] Bildquelle: `http://hyperphysics.phy-astr.gsu.edu/hbase/index.html`

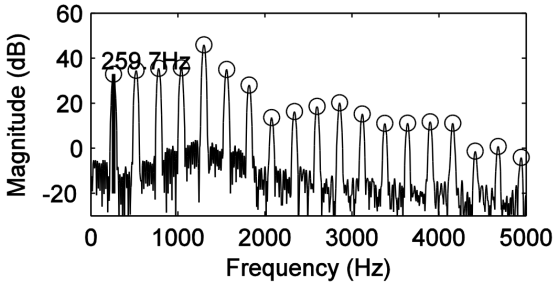

Abbildung 3.25: Grundfrequenz und Harmonische Oberwellen eines Sprachsignals.

meist nur die ersten 2 oder 3 Formanten zur Beschreibung eines Vokals verzeichnet. Für verschiedene Sprachen sind allerlei Tabellen zu finden, welche die Formantenfrequenzen der Vokale listen.[3, S. 19]

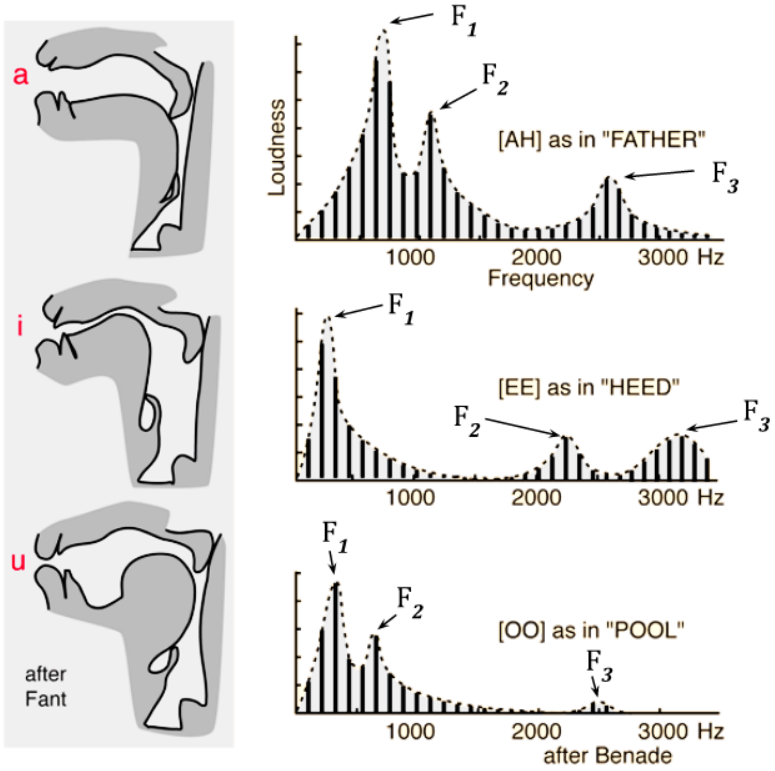

Abbildung 3.26: Formanten im Sprach-Signal

Da Sprache etwas zeitlich dynamisches ist, befinden sich sowohl die Glottal Source als auch der Filter des Vokaltraktes und der Lippen in ständiger Veränderung. Da die Informationen der Sprache vor allem im Frequenz-Bereich codiert sind, wird die in Kapitel 3.4.3 vorgestellte Short Time Fourier Transformation für die Visualisierung von Sprache eingesetzt. Dabei wird auf der x-Achse die Zeitpunkte der Fenster, auf der y-Achse die Frequenz dargestellt. Die Frequenz-Fenster werden sozusagen „auf die Seite gelegt“, damit ihr zeitlicher Verlauf übersichtlich betrachtet werden kann. Die Amplitude der entsprechenden Frequenz wird farblich oder durch Helligkeiten codiert, abhängig von der konkreten Implementierung des Spectograms. Je länger das Zeitfenster der STFT, desto besser die Auflösung bezüglich des Frequenz-Bereiches, desto schlechter ist jedoch die zeitliche Auflösung. Je kürzer die

Zeitfenster der STFT, desto besser wird der zeitliche Verlauf bei sinkender Frequenz-Auflösung erkennbar.[3, S. 48 - 50] [17, Acoustic Representations of Speech]. Abbildung 3.27 zeigt ein Beispiel für zwei Spectogramme mit unterschiedlichen Fenstergrößen der STFT, angewandt auf einem 9 sekunden langen Signal mit Baby-Weinen. Es ist zu erkennen, wie bei sinkender Fensterlänge der zeitliche Verlauf besser erkennbar, jedoch die einzelnen harmonsichen Obertöne weniger gut voneinander unterscheidbar sind. Inbesondere bei der Analyse von Audiosignalen des Weinens der Neugeborenen von Golub und Corwin waren [12] waren Auswertungen von Spectogrammen von Zentraler Bedeutung.

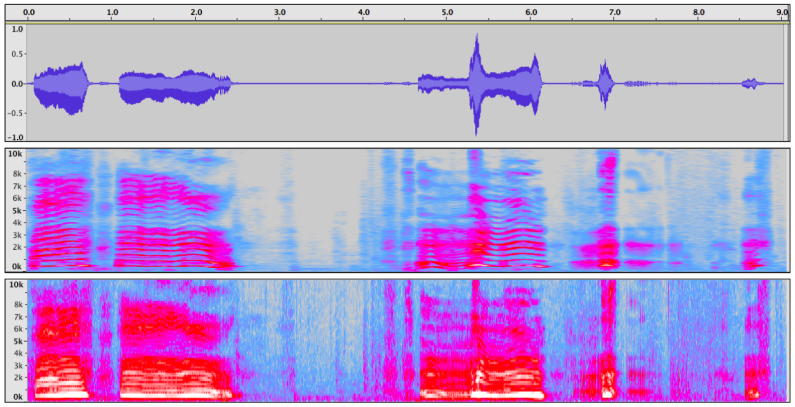

Abbildung 3.27: Spectogram von Baby-Weinen. Rot = Hohe Amplituden, Blau = niedrige Amplituden. Oben: Zeit-Bereich. Mitte: Spectogram mit einer Fensterlänge von 185 ms(8192-Sample DFT). Unten: Spectogram mit einer Fensterlänge von 5 ms
(265-Sample DFT).

3.7 Feststellung von Periodizität in Signalen

Die Feststellung von Periodizität in Zeit-Bereich in einem Signal hat eine besondere Rolle in der Sprachverarbeitung. Ein Ensatzgebiet, welches in dieser Arbeit weiter von Bedeutung sein wird, ist die Voice-Activity-Detection, die Feststellung des Vorhandenseins von Stimme in einem Signal. [31] Ein weiteres Einsatzgebiet ist die Tonhöhenerkennung. [15] [**?** , S. 1 - 2]

Ein Sprachsignal ist 1.) nur über kurze Zeitfenster periodisch, da sich die Tonhöhe jederzeit ändern kann, und 2.) selbst bei „perfekt gewählter Fensterlänge“ nicht perfekt periodisch, sondern nur annhäernd periodisch (*quasi periodisch*).[**?** , S. 1 - 2] Formel 3.65 definiert diese Aussage als Formel für den Zeit-Bereich.

$$x[n+N] \approx x[n] \tag{3.65}$$

Über die Jahre wurde eine Vielzahl an Methoden zur Feststellung der Periodizität entwickelt. An dieser Stelle wird eine Auswahl vorgestellt, die für die Voice-Activity-Deteciton in Kapitel xxx von Bedeutung sind.

3.7.1 Autokorrelation

Die Korrelation wurde in Kapitel 3.3 vorgestellt. Bei der Autokorrelation wird ein Signal mit einer vezögerten Variante von sich selber korrelliert. Gleichung 3.66 definiert die Autokorrelation des Signals eines N-Samples langen Signals $x[\,]$ mit einer um das Lag k verzögerten Variante von sich selber.

$$\text{A-Corr}_k(x[\,]) = \sum_{n=k}^{N} x[n-k] \cdot x[n] \tag{3.66}$$

Wie bei der in Kapitel 3.3 vorgestellten Cross-Correlation gibt es verschiedene Möglichkeiten der Normalisierung der Korrelationswertes in Bezug auf die Signalenergien. Gleichung 3.67 definiert die normalisierte Autokorrelation.[31]

$$\text{NA-Corr}_k(x[\,]) = \frac{\sum_{n=k}^{N} x[n-k] \cdot x[n]}{\sqrt{\sum_{n=1}^{N-k} x[n]^2} \cdot \sqrt{\sum_{n=k}^{N} x[n]^2}} \tag{3.67}$$

Nun wird das Autokorrelations-Signal $a[\,]$ erstellt, in dem Gleichung 3.66 oder 3.67 für verschiedene $k = k_min \ldots k_max$ angewandt wird, wie Gleichung 3.68 definiert.

$$a[\,] := \overset{k_{max}}{\underset{k=k_{min}}{\forall}} : \; a[k] = \text{NA-Corr}_k(x[\,]) \tag{3.68}$$

Ein hoher Wert des Signals $a[\,]$ an der Position k spricht für eine Periodizität des Signals mit der Frequenz $f = f_s/k$. Es ist üblich, den Bereich $[k_{min}, k_{max}]$ so einzuschränken, dass die Autokorrelation nur für den Frequenz-Raum durchegeführt wird, in dem man Periodizität erwartet.

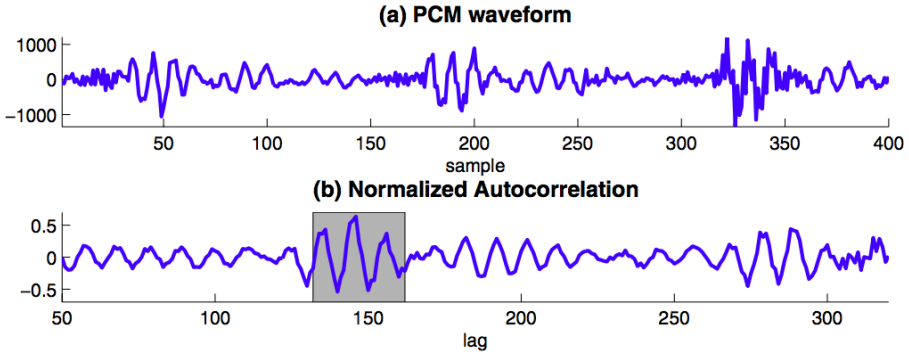

Abbildung 3.28: Autokorrelation eines Signals

Abbildung 3.28 verdeutlicht das Vorgehen an einem Beispiel. Gezeigt wird in (a) der Zeit-Bereich eines quasi-periodischen Signals $x[\,]$, welches mit einer Sampling-Frequenz von $f_s = 16.000\,\text{Hz}$ aufgenommen wurde. Es wird eine Periodizität im Bereich $50\,\text{Hz} - 400\,\text{Hz}$ vermutet, daher wird die normalisierte Autokorrelation nach Formel 3.68 durchgeführt mit den Lags $k_{min} = 16.000\,\text{Hz}/400\,\text{Hz} = 40$ bis $k_{max} = 16.000\,\text{Hz}/50\,\text{Hz} = 320$. (b) zeigt das so entstandene Signal $a[\,]$. Das Maximum an der Stelle $k = 146$ weisst auf eine Grundfrequenz von $N_0 \approx 109\,\text{Hz}$ hin. Dabei kann es sich jedoch auch um die Frequenz einer Harmonische Schwingung handeln, welche durch den Einfluss des Filters des Vokaltraktes verstärkt wurde.

Es kann sich auch um die Hälfte der Grundfrequenz handeln, da wie in Kapitel 3 erläutert, ein Signal mit der Periode N ebenfalls eine Periodizität bezüglich $2N, 3N, \ldots$ aufweist. [31] [20, S. 24]

3.7.2 Cepstrum

Das Cepstrum wird nach Gleichung 3.69 als die inverse DFT des Logarithmus des Magnitudensignals des Frequenz-Bereiches definiert.

$$ceps[\,] = \text{iDFT}\Big\{ \log\Big(\big|\ \text{DFT}\{x[\,]\}\big|\ \Big)\Big\} \tag{3.69}$$

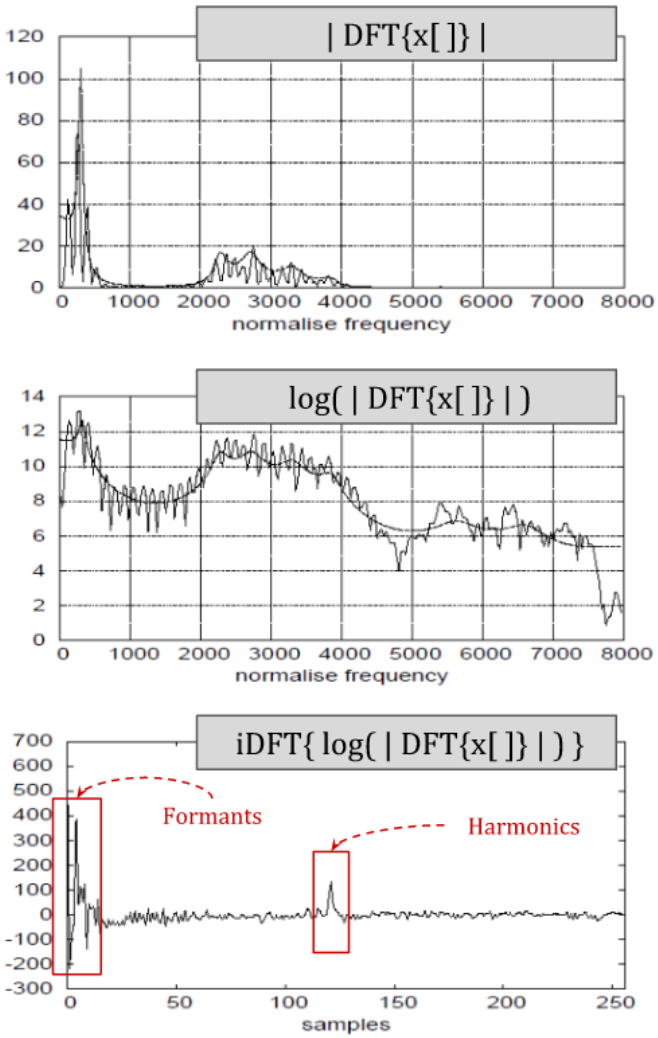

Abbildung 3.29: Berechnung des Cepstrums

Das Vorgehen wird mit Hilfe des Beispiels aus Abbildung 3.29 erläutert. $|\ \text{DFT}\{x[\,]\}|$ zeigt das Spectrum (Magnituden-Signal) eines „typischen stimmhaften“ Signals $x[\,]$. Es sind die in Kapitel 3.6 erläuterten harmonischen Obertöne zu sehen, welche mit steigender Frequenz an Amplitude verlieren. Durch das logarithmieren des Spectrums $\log\Big(\big|\ \text{DFT}\{x[\,]\}\big|\ \Big)$

wird die Dynamic des Frequenz-Bereiches verringert und somit der Amplituden-Verlust der höheren Obertöne verringert. Nun stellt man sich vor, dieses Spectrum wäre ein Signal des Zeit-Bereiches. Dieses Signal würde man als ein Quasi-Periodisches Signal mit einer Amplituden-Modulation interpretieren, das heißt ein Signal mit hoher Frequenz, addiert mit einem Signal mit nierdiger Frequenz. Um diese beiden Komponenten voneinander zu trennen, würde man wieder die DFT anwenden, um das Spectrum zu bilden. Diese DFT kommt in dem Fall einer inversen DFT gleichkommt, da das Magnituden-Signal verworfen wird. Man erwartet in diesem „Spectrum vom Spectrum“ einen Peak im „oberen Frequenz-Bereich“, bedingt durch die harmonischen Oberwellen, sowie einen Peak im „unteren Frequenz-Bereich“, bedingt durch die Formanten.[9, Cepstral analysis]

Der Bereich dieser „Fouriertransformation der Foueriertransformation“ wird als *Cepstrum* bezeichnet. Cepstrum ist ein ein Wortspiel, welches durch die Umkehrung der ersten vier Buchstaben des Wortes SSpectrumënsteht. Die Unabhängige Variable des Cepstrum folgt dem Wortspiel und wird als *Quefrency* bezeichnet. Damit wird verdeutlicht, dass die unabhängige Variable des Cepstrum zwar mathematisch betrachtet die Zeit darstellt, jedoch als Frequenz interpretiert wird.[9, S. 7]

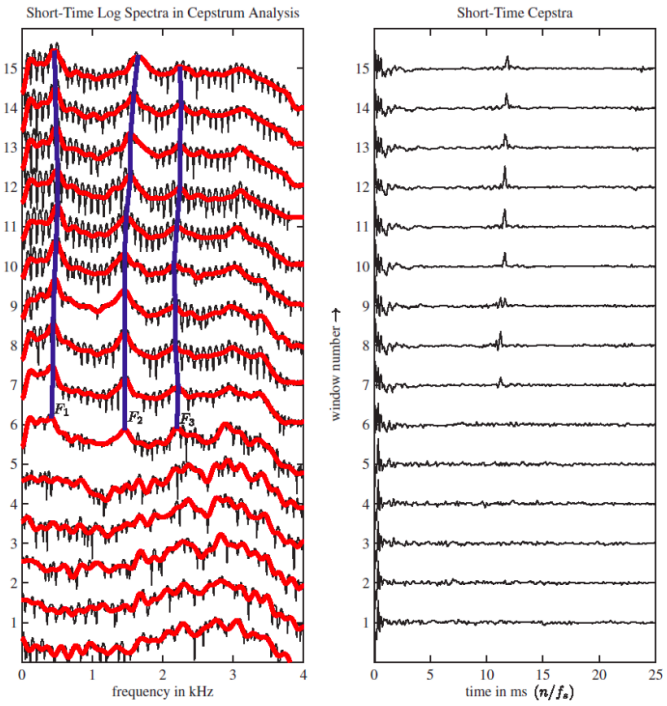

Abbildung 3.30: Aufkommen eines Peaks im oberen Quefrency-Bereich bei stimmhaften Signalfenstern [9, S. 17]

Insbesondere ein Auftauchen eines Peakes im oberen Quefrency-Bereich $> 3\,\text{ms}$ spricht für das vorhandensein von harmonischen Obertönen und somit für periodizität im Signal, wie sie durch Stimme entsteht. Abbildung 3.30 verdeutlicht das Prinzip an einem Beispiel. Zu sehen ist die STFT eines Signals mit einer Fensterlänge von 50 ms und einer Hopsize von 12.5 ms. Links wird das Logarithmisierte Spektrum abgebildet, und rechts das Cepstrum. Die Frames 1-5 sind stimmlos, die Frames 8-15 sind stimmhaft, und die zwischen-Frames eine Mischung. Man sieht das Aufkommen eines Peaks bei einer Quefrency $q = 12\,\text{ms}$.[9, S. 16]

Abbildung 3.31 verdeutlicht, wie eine Grundfrequenz f_0 im Zeit-Bereich einen Peak im Cepstrum erzeugt. So weist ein Peak an bei der Quefrency q auf eine Grundfrequenz von $q = f_s/f_0$ hin. Dabei kann es sich, wie bei der Autokorrelation, um einen Oktaven-Fehler handeln, das heißt, dass der höchste Peak das Doppelte oder die Hälfte der eigentlichen Grundrequenz beträgt. [27]

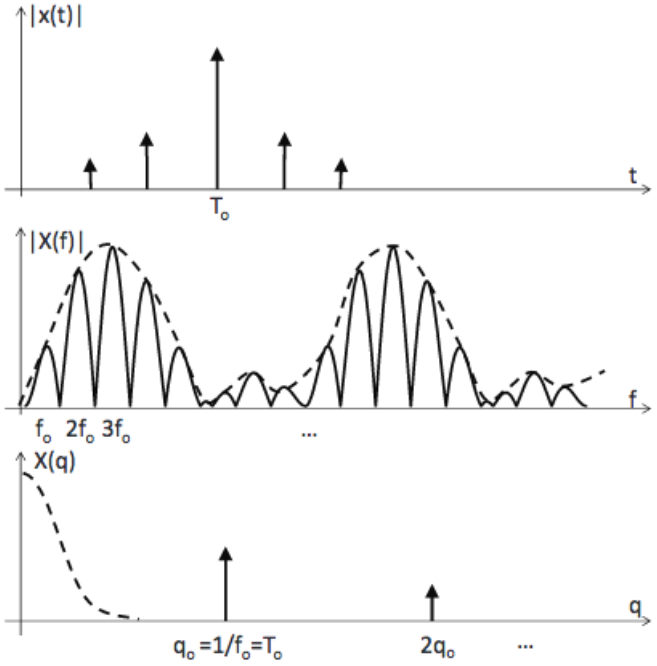

Abbildung 3.31: Feststellung der Grundfrequenz aus dem Cepstrum[27]

4 Zusammenfassung

Literaturverzeichnis

[1] K J S Anand. *Pain in Neonates and Infants*. Elsevier, 2007.

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

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

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

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

[6] 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`.

[7] 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`.

[8] Jan Hamers & Peter Gessler Eva Cignac, Romano Mueller. Pain assessment in the neonate using the Bernese Pain Scale for Neonates. *Early Human Development*, 78(2):125–131, 2004. Online unter `http://www.sciencedirect.com/science/article/pii/S0378378204000337`.

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

[10] Health Facts For You. *Using Pediatric Pain Scales Neonatal Infant Pain Scale (NIPS)*, 2014. Online unter `https://www.uwhealth.org/healthfacts/parenting/7711.pdf`.

[11] Hodgkinson. Neonatal Pain Assessment Tool , 2012. Online unter `http://www.rch.org.au/uploadedFiles/Main/Content/rchcpg/hospital_clinical_guideline_index/PAT%20score%20update.pdf`.

[12] 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.

[13] Julian O Smith III. Introduction to digital filters with audio applications, 2007. Online erhältlich unter: `https://ccrma.stanford.edu/~jos/filters/filters.html`.

[14] 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/.

[15] Aaron Rosenberg Lawrence Rabiner, Michael Cheng and Carol McGonegal. A Comparative Performance Study of Several Pitch Detection Algorithms. *IEEE Transactions on acoustics, speech, and signal processing*, 24(5):399–417, 1976. Online unter http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1162846.

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

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

[18] 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.

[19] D L Neuhoff. *Filters in the Time-Domain I, II and III.* Online erhältlich unter: http://www.eecs.umich.edu/courses/eecs206/archive/spring02/notes.dir/filters1.pdf, http://www.eecs.umich.edu/courses/eecs206/archive/spring02/notes.dir/filters2.pdf.

[20] 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.

[21] Taddio Nulman. A revised measure of acute pain in infants. *J Pain Symptom Manage*, 10:456–463, 1995. Online unter http://geriatricphysio.yolasite.com/resources/Modified%20Behavioral%20Pain%20Scale%20(MBPS)%20in%20infants.pdf.

[22] 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.

[23] Steven Creech & Marc Weiss. Patricia Hummel, Mary Puchalski. 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.

[24] 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.

[25] 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/.

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

[27] Juan Ignacio Godino-Llorente Ruben Fraile. Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control*, 14(1):42–54, 2014. Online unter https://www.researchgate.net/publication/264084923_Cepstral_peak_prominence_A_comprehensive_analysis.

[28] 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.

[29] 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.

[30] 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.

[31] Sabine Deligne & Peder Olsen Trausti Kristjansson. Voicing Features for Robust Speech Detection. Interspeech Lisboa, September 2005. Online unter http://papers.traustikristjansson.info/wp-content/uploads/2011/07/KristjanssonRobustVoicingEurospeech2005.pdf.

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

Appendices

Tabelle .1: Accuracy-Werte der Grenzwertfindung mit REPTree

$S_{Training}$	3 dB				50 dB				50+3 dB			
A_{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%

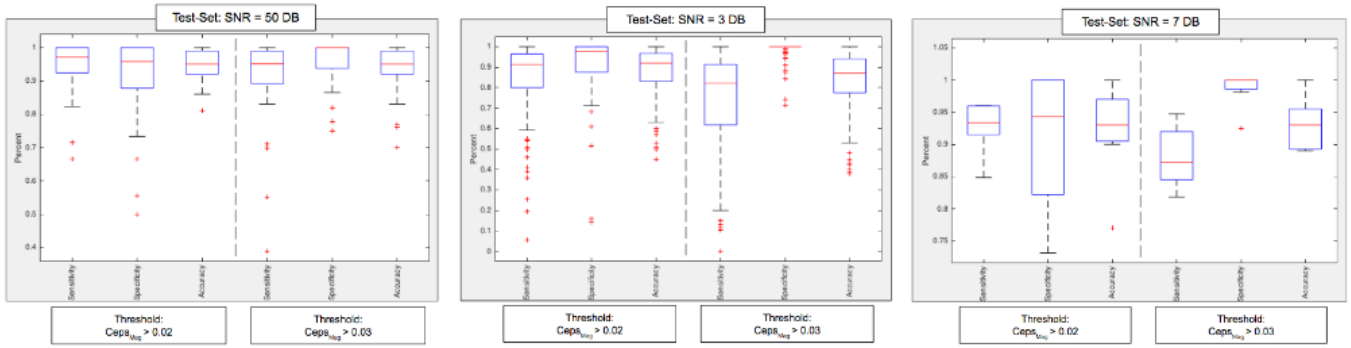

Abbildung .1: Boxplot-Auswertung über Sensitivity, Specificity und Accuracy der beiden VAD-Modelle