

Grundlagen der Sprachsignalverarbeitung

1. Einführung

1.1 Grundbeobachtungen am Sprachsignal

Bild 1.1 zeigt das Beispiel eines Sprachsignals. Ohne viel von seiner Entstehung zu wissen, können wir sofort mehrere Grundeigenschaften des Sprachsignals an diesem Beispiel ablesen.

1) Das Sprachsignal ist *zeitveränderlich*. In der Art und Weise, wie die Zeitveränderung erfolgt, liegt die Information, die das Sprachsignal zum Kommunikationsträger für den Menschen macht. Im Vergleich zu dem Informationsfluss, der in der übermittelten Nachricht enthalten ist, wenn sie in Schriftform vorliegt, ist der Informationsgehalt des akustischen Sprachsignals sehr hoch.

2) Das Sprachsignal ist in einzelnen Abschnitten (quasi-)periodisch. In anderen Abschnitten besitzt es die zeitliche Struktur eines Rauschens. Weiterhin existieren Pausen im Verlauf des Sprachsignals, die teilweise deutlich kürzer sind als die vom Sprecher beabsichtigten Pausen zur Gliederung der Rede, die der Hörer sofort als solche wahrnimmt.

3) Im Vergleich zur Schriftform, wo wir eine *Zeichenkette* vorliegen haben, stellt das Sprachsignal ein Kontinuum, einen *Datenstrom* dar. Von Ausnahmen abgesehen, erfolgen Änderungen der Signalstruktur stetig und fließend, nicht abrupt.

1.2 Bau des menschlichen Sprechtrakts (kurze Einführung)¹

Bild 1.2 zeigt eine schematische Übersicht des menschlichen Sprechtraktes. Im Prinzip durchläuft das Sprachsignal zwei Stufen: 1) die *Anregung*, sowie 2) die *Signalformung*. Diese beiden Stufen sind funktionell in erster Näherung trennbar.

1.2.1 Anregung des Sprachsignals

Die Anregung ist im wesentlichen auf dreierlei Arten möglich.

1) *Stimmhafte Anregung (Phonation)*. Die Luft durchströmt beim Ausatmen den Kehlkopf und passiert dort die Stimmbänder (Glottis). Diese werden bei entsprechender Spannung zu quasiperiodischen Schwingungen angeregt, die in Periode und Amplitude veränderlich sind.

2) *Stimmlose Anregung*. Die Luft durchströmt die geöffnete Glottis und passiert eine Engstelle im Mund- oder Rachenraum. An dieser Stelle entsteht eine turbulente Strömung und somit ein Rauschen, dessen Spektrum von der Lage der Engstelle bestimmt wird.

3) *Transiente Anregung (Plosivanregung)*. Die Luft staut sich hinter einer Verschlussstelle im Mund- oder Rachenraum. Durch plötzliches Öffnen des Verschlusses wird der Druck abgebaut.

Die drei Anregungsformen äußern sich im Zeitverlauf des Sprachsignals auf charakteristische Weise: 1) die stimmhafte Anregung durch die Periodizität, 2) die stimmlose Anregung durch die geräuschartige Signal-

¹ Der Sprechtrakt wird in anderen Lehrveranstaltungen im Detail behandelt (z.B. Grundlagen der Phonetik, Phonetik und akustische Kommunikation).

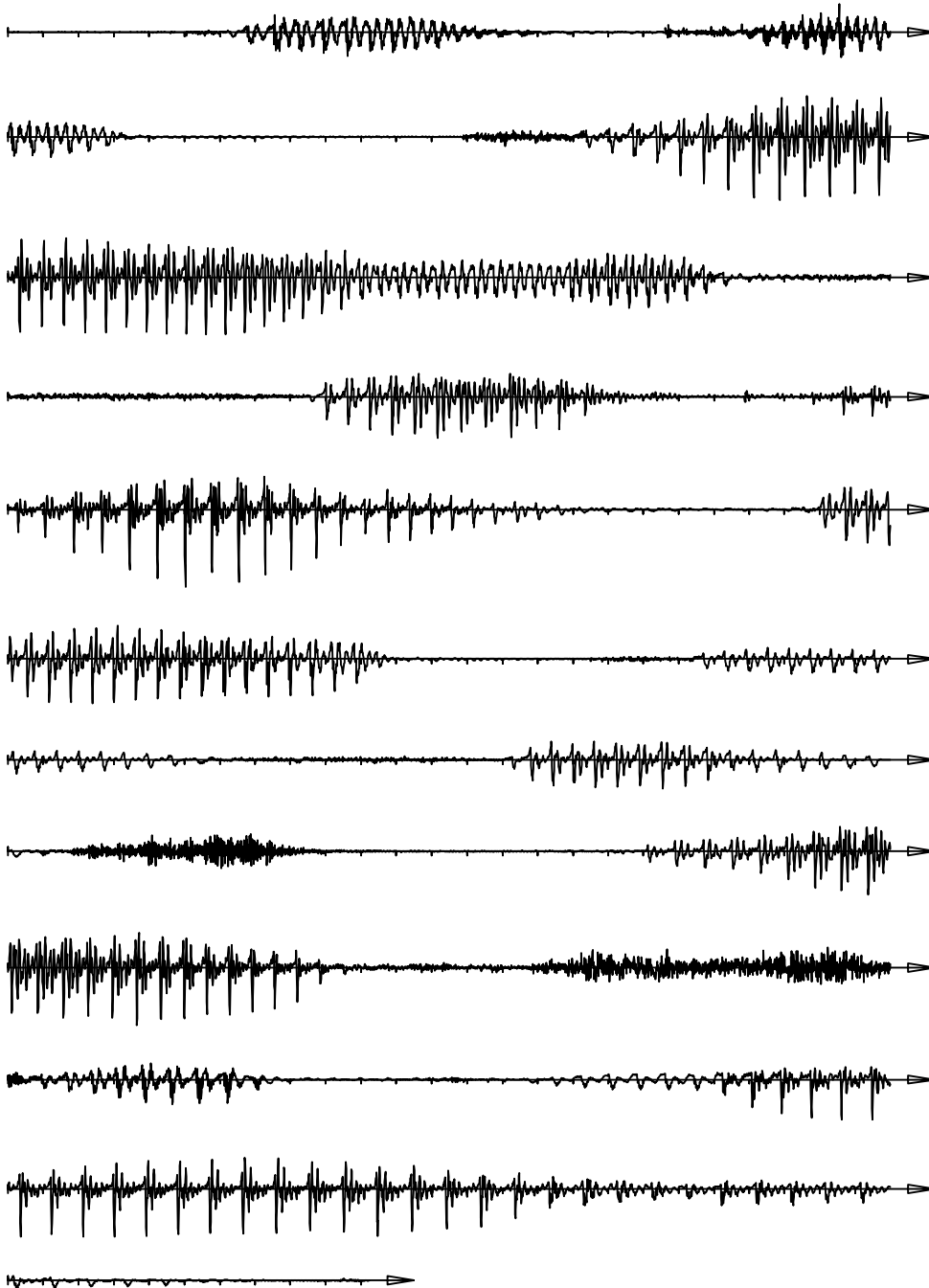
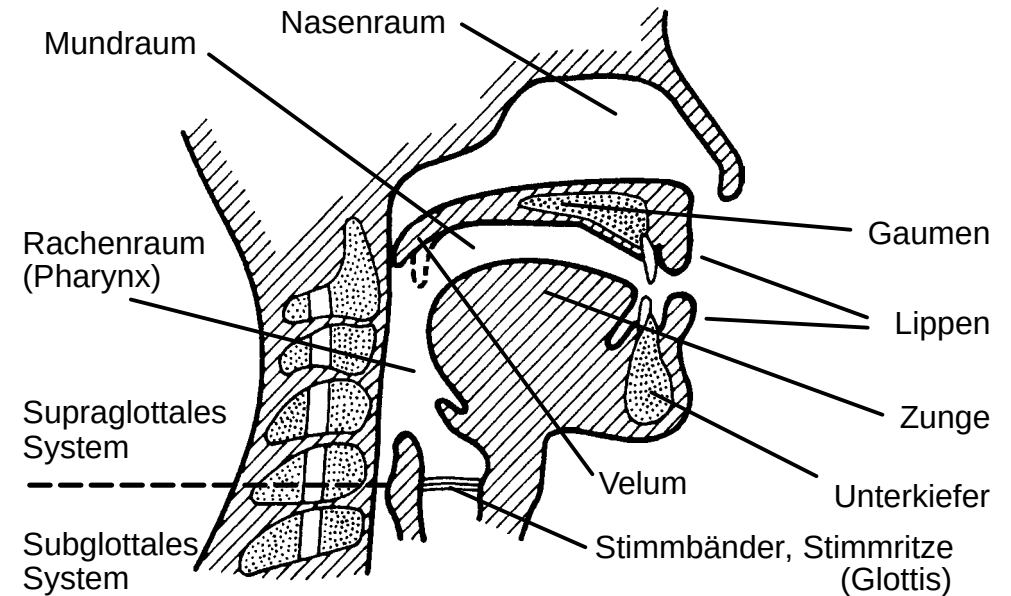


Bild 1.1 (gegenüber). Beispiel eines Sprachsignals. Maßstab: 250 ms je Zeile; Sprecher WGH (männlich)

Bild 1.2. Schematische Übersicht über den Bau des menschlichen Sprechtraktes



form, 3) die transiente Anregung durch eine bestimmte Folge im Zeitverhalten. Zunächst wird hier durch den Verschluss jedes akustische Signal für eine bestimmte Zeit unterbunden; es entsteht eine kurze Pause. Höchstens eine leise Stimmbandschwingung kann während dieser Zeit noch bestehen (engl.: "voice bar"). Das Öffnen des Verschlusses im Mund- oder Rachenraum bewirkt das plötzliche Entweichen der angestauten Luft mit dem zugehörigen Geräusch, dessen spektrale Eigenschaften wiederum von der Stelle des Verschlusses bestimmt werden. Stimmlose und stimmhafte Anregung können gleichzeitig auftreten sowie (jede für sich oder beide gemeinsam) für einige Zeit beibehalten werden. Die transiente Anregung, die außer der erwähnten Stimmbandschwingung keine weitere gleichzeitige Anregungsform duldet, verläuft kurzzeitig; die Pause liegt in der Größenordnung von 20 bis 100 ms; das anschließende Öffnungsgeräusch des Verschlusses dauert 20 bis 50 ms.

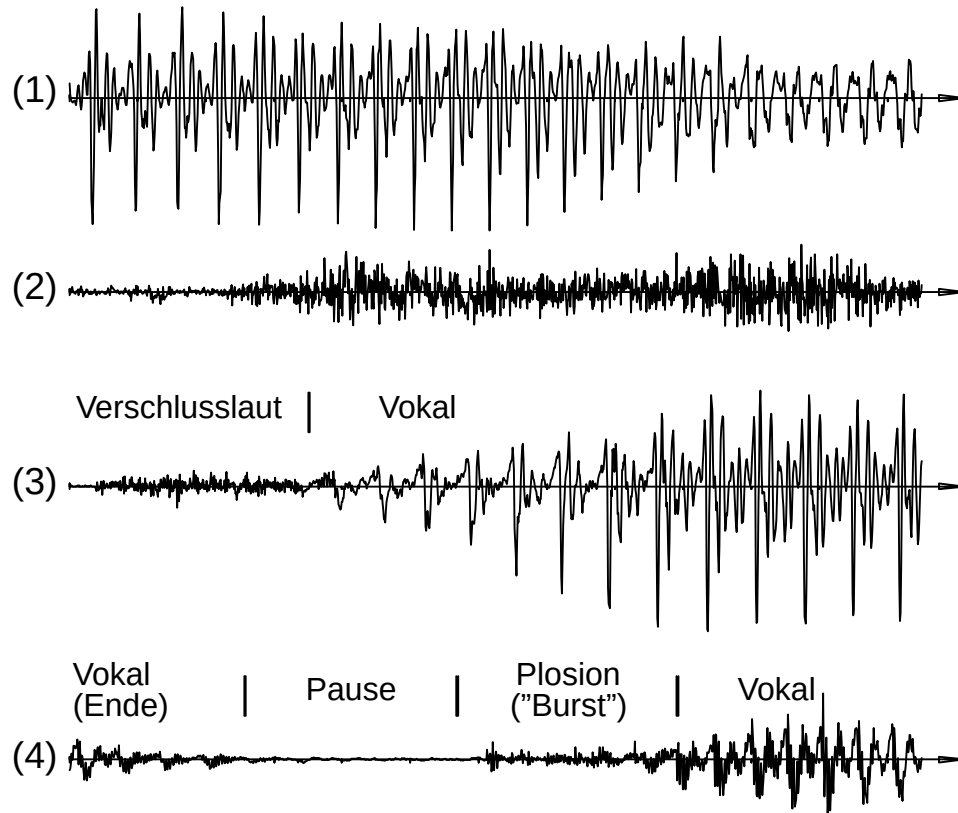


Bild 1.3. Beispiele für die einzelnen Lautklassen. (1) Stimmhaft (Vokal), (2) stimmlos (Frikativ), (3) Übergang Plosivlaut-Vokal, (4) Übergang Vokal-Plosivlaut-Vokal. Maßstab: 100 ms/Zeile

Falls ein anderer Laut folgt, verbindet es sich mit diesem. Bild 1.3 zeigt ein charakteristisches Beispiel für jede der drei Anregungsformen.

Entsprechend der Art der Anregung lassen sich elementare Sprachsignalparameter (oder Sprachmerkmale) wie folgt definieren:

1) *Art der Anregung*: a) stimmhafte Anregung; b) stimmlose Anregung. Wenn diese beiden Eigenschaften als *voneinander unabhängige, binäre Merkmale* definiert werden, so kann damit auch die *Pause* (keines der Merkmale vorhanden) sowie die *transiente Anregung* (aufgrund des Zeitverhaltens) beschrieben werden.

2) Bei stimmhafter Anregung: die *Periodendauer* bzw. *Grundfrequenz*. Dieser Parameter ist innerhalb bestimmter Grenzen ("Stimmumfang") stetig veränderlich.

Höhere Merkmale, wie z.B. die Klangfarbe der Stimme, sollen hier zunächst außer acht gelassen werden. Es ist in der Sprachsignalverarbeitung üblich, die *Amplitude* des Signals als unabhängigen Parameter zu betrachten, da die Amplitude sowohl von der Anregung als auch von der nachfolgenden Signalformung beeinflusst wird.

1.2.2 Signalformung

Das Quellen- oder *Anregungssignal*, wie es nach der Anregung vorliegt, ist als Informationsträger für Sprache noch nicht ausreichend. Es bedarf noch der Signalformung. Diese wird im *Vokaltrakt* durchgeführt. Unter dem Vokaltrakt im engeren Sinn versteht man den Rachenraum über der Glottis (Epiglottis, Pharynx) sowie den Mundraum. Die wesentlich zur Klangfarbe des Sprachsignals beitragenden Komponenten im Vokaltrakt sind: 1) die Stellung der Zunge, 2) die Stellung der Lippen. Beide tragen in Form stetig veränderlicher Merkmale zum Sprachsignal bei. Die Stellung des Gaumensegels (Velum) hingegen, das den Nasenraum an- und abkoppelt, ist ein im wesentlichen binäres Merkmal.

Die Gesamtheit der Bewegungen der Organe des Sprechtrakts und ihre zeitliche Aufeinanderfolge werden als *Artikulation* bezeichnet.

Die Sprachsignalparameter, die die Signalformung betreffen (sie werden üblicherweise als dem Vokaltrakt zugehörige Parameter bezeichnet), sind auf mancherlei Weise zu definieren und zu beschreiben.

1.3 Signaldarstellung im Rechner

1.3.1 Der Begriff des Signals

Ein *Signal* ist "das Erscheinungsbild einer physikalischen Information" (Kunt, 1980). Darstellbar ist es als Funktion einer oder mehrerer Veränderlicher; hierbei wird der Begriff der Funktion jedoch wesentlich weiter gefasst als in der Mathematik. Ein Signal ist beispielsweise beschreibbar durch

- eine mathematische Funktion im strengen Sinn, also einen analytischen Zusammenhang (in geschlossener Form), oder
- ein Verteilungsgesetz (z.B. für ein stochastisches Signal). Hier ist der Augenblickswert des Signals nicht bekannt, da die Verteilungsfunktion nur globale Signaleigenschaften beschreibt, oder
- empirisch durch eine Messreihe. Die meisten Signale in der Praxis gehören zu dieser Kategorie.

Ist ein Signal Funktion nur einer Veränderlichen, so sprechen wir von einem *eindimensionalen Signal*, ansonsten von einem *mehrdimensionalen Signal*. Eindimensionale Signale liegen meist als *Zeitfunktionen* [”Vorgänge” (Meyer-Eppler, 1969)] vor. Sie seien im folgenden als $s(t)$ bezeichnet. Der Wert $s(t)$ stellt die *Augenblicksamplitude*, den *Momentanwert* oder einfach den *Wert* des Signals zu einem bestimmten Zeitpunkt t dar. Beispiele für eindimensionale Signale sind: Sprachsignale, akustische Signale aller Art (Musik, Geräusch), biologische Signale (Elektrokardiogramm, Elektroenzephalogramm usw.), physikalische Signale (z.B. seismische Wellen) oder Radarsignale. Den wichtigsten Fall mehrdimensionaler Signale stellen zweidimensionale Bildsignale dar.

Bild 1.4 zeigt einige Signale, die in der Signalverarbeitung eine besondere Bedeutung haben.

1.3.2 Darstellung digitaler Signale

Die Darstellung eines *digitalen Signals* kann nur in Form einer *Zahlenfolge* erfolgen:

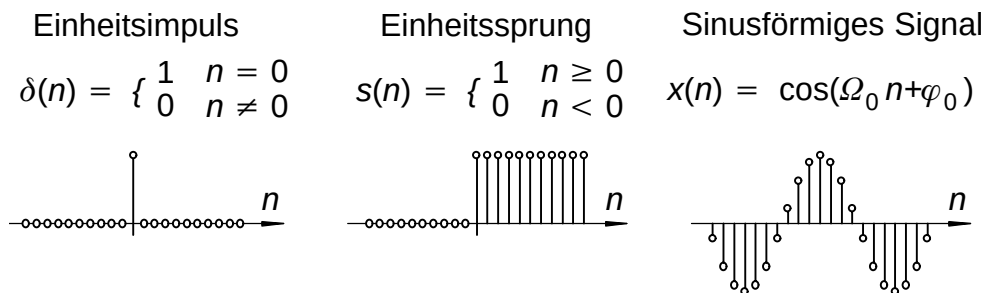


Bild 1.4. Beispiele für digitale Signale (hier als abgetastete Signale gezeichnet)

$$s(t) \rightarrow \{s_i\}. \quad (1.1)$$

Das digitale Signal (Bild 1.5) ist demnach *diskret* im doppelten Sinn: zum einen bezüglich der unabhängigen Variablen t , zum anderen bezüglich der abhängigen Variablen s . Der Übergang von der kontinuierlichen zur diskreten Darstellung wird allgemein als *Quantisierung* bezeichnet.

Quantisierung der Werte der unabhängigen Variablen ergibt *Abtastung* und zerlegt das Signal in eine Folge von *Abtastwerten*:

$$s(t) \rightarrow \{ \dots s(t_1), s(t_2), \dots \}. \quad (1.2)$$

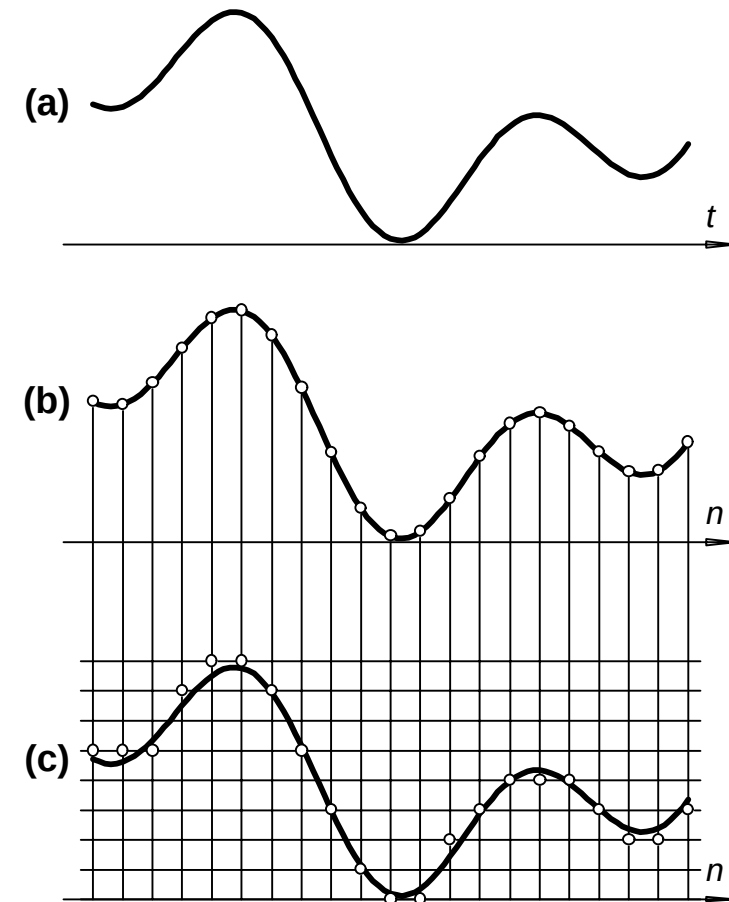


Bild 1.5a-c. Vom kontinuierlichen (”analogen”) Signal zum digitalen Signal. (a) Analoges Signal, (b) Abtastsignal, (c) digitales Signal

Als zusätzliche voraussetzende Vereinbarung gelte, dass die Abtastung *gleichförmig*, d.h., stets in gleichen Intervallen der unabhängigen Variablen erfolgen soll:

$$s(n) := s(nT); \quad n = -\infty \dots -1, 0, 1 \dots +\infty \text{ [ganzzahlig];} \quad (1.3)$$

T : Abtastintervall

Das *Abtastintervall* T kann in der Bezeichnung $s(n)$ für das Signal weggelassen werden. Dementsprechend ist n nur noch indirekt ein Maß für die Zeit; daher sei n als *Index* oder *Adresse* des Abtastwertes (*Messwertadresse*) bezeichnet. Der Kehrwert des Abtastintervalls wird als *Abtastfrequenz* bzw. *Abtastrate* bezeichnet:

$$F := 1/T. \quad (1.4)$$

Welche Werte das Abtastintervall T bzw. die Abtastfrequenz F für ein gegebenes Signal annehmen müssen, sagt uns das *Abtasttheorem*.² Dieses verlangt, dass ein Signal mit einer Abtastfrequenz F abgetastet werden muss, die größer ist als das Doppelte der höchsten im Signal enthaltenen Frequenz. Missachtung dieser Regel führt zu groben Verzerrungen, die sich aber ausschließlich auf die Signalanteile auswirken, die das Abtasttheorem verletzen.

Sprachsignale in guter Qualität besitzen Frequenzanteile bis ungefähr 8 kHz. Zur digitalen Darstellung von Sprachsignalen in guter Qualität ist daher eine Abtastfrequenz von 16 kHz ausreichend, aber auch notwendig. Sprache in Telefonqualität ist auf 4 kHz bandbegrenzt; deshalb wird hier eine Abtastfrequenz von 8 kHz verwendet. Musiksignale umfassen den gesamten Hörbereich, also Frequenzen bis mindestens 16 kHz. Gängige Abtastfrequenzen sind hier 44,1 kHz (CD) oder 48 kHz (sonstige digitale Tonaufzeichnungsverfahren)

Wegen endlicher Wortlänge bei Verarbeitung und Speicherung abgetasteter Signale erfolgt *Quantisierung* auch im Bereich der *abhängigen Variablen*:

² Das Abtasttheorem (Shannon, 1949), wird in der Vorlesung *Signalverarbeitung 1* hergeleitet. Hier sind nur die Ergebnisse von Bedeutung.

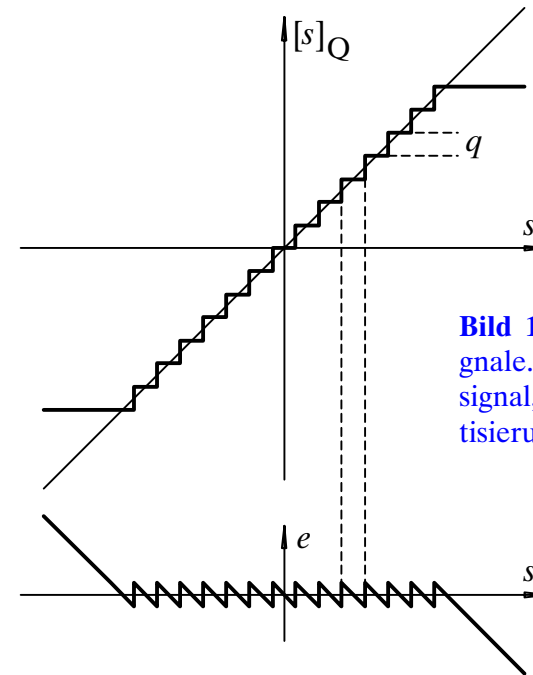


Bild 1.6. Zur Quantisierung digitaler Signale. (s) Eingangssignal, ($[s]_Q$) Ausgangssignal, (q) Quantisierungsstufe, (e) Quantisierungsfehler

$$s(n) \rightarrow [s(n)]_Q = k(n) \cdot q; \quad k \text{ ganzzahlig.} \quad (1.5)$$

Bei *linearer* oder (besser) *gleichförmiger Quantisierung* ist q eine Konstante, die *Quantisierungsstufe*. Die Bezeichnung *Quantisierung* wird bei der abhängigen Variablen im Gegensatz zur unabhängigen Variablen beibehalten. Im allgemeinen ist k auch noch betragsmäßig begrenzt; bei binärer Darstellung beispielsweise ergibt sich die Zahl der unterscheidbaren Amplitudenstufen zu

$$N = 2^W; \quad (1.6)$$

W ist die *Wortlänge* (in bit). Durch die Quantisierung im Bereich der abhängigen Variablen ergibt sich ein *Quantisierungsfehler* als Differenz des tatsächlichen und des quantisierten Wertes (siehe Bild 1.6).

Im Unterschied zu Fehlern bei der Wahl der Abtastfrequenz machen sich Quantisierungsfehler allmählich bemerkbar. Als Kriterien für die Anforderungen gelten a) Hörbarkeit des Fehlers sowie b) Verständlichkeit der Sprache. Bei steigendem Quantisierungsfehler wird in akusti-

schen Signalen die Qualität zunehmend rauher. Musiksingale (CD-Qualität) verlangen eine Wortlänge von 16 bit (65536 Amplitudenstufen); Sprachsignale in guter Qualität erfordern etwa 12 bit (4096 Amplitudenstufen), in Telefonqualität 8 bit (256 Amplitudenstufen). Bei Telefonqualität (insbesondere wenn das Signal dann über einen Lautsprecher wiedergegeben wird) sind die Quantisierungsgeräusche bereits deutlich hörbar. Jedoch selbst bei einer Wortlänge von 1 bit (2 Amplitudenstufen; "totale Begrenzung") ist Sprache noch weithin verständlich.

Erst Abtastung *und* Quantisierung machen aus einer Zeitfunktion ein *digitales Signal*. Bei theoretischen Betrachtungen in der Signalverarbeitung werden die Probleme der Quantisierung zunächst meist vernachlässigt, nie jedoch die Probleme, die sich aus der Abtastung ergeben.³

1.4 Der Informationsgehalt der Sprache

Die Sprache als Kommunikationsmittel zwischen Menschen bedient sich im wesentlichen zweier Formen: 1) der *akustischen Form* als gesprochene Sprache, 2) der *Textform*. Die der Sprache (akustische Form wie auch Textform) zugrundeliegende Information ist enthalten als zeitliche bzw. örtliche *Sequenz diskreter Einheiten*, die sich in verschiedener Komplexität und in verschiedenen Informationsebenen darstellen lassen.

1.4.1 Linguistische Einheiten

Tabelle 1.1 zeigt eine mögliche Zusammenstellung auf. Kennzeichen dieser Einheiten ist die hierarchische Gliederung. Jede größere Einheit enthält eine ganze Zahl (≥ 1) Elemente der nächstkleineren Einheit. In der Textform der Sprache lassen sich sämtliche dieser Einheiten in der Regel gut trennen. Sätze und oftmals auch Wortgruppen werden durch Satzzeichen auseinandergehalten, Wörter werden durch Zwischenräume getrennt. Die Zerlegung in Morpheme und Silben erfolgt gemäß den für die jeweilige Sprache geltenden Regeln der Grammatik. Bei den

Tabelle 1.1. Einige Einheiten der Sprache

Laut (Phonem)
Silbe
Wortteil: Stamm, Präfix, Suffix (Morphem)
Wort
Wortgruppe (Satzglied, Phrase)
Satz
Satzgruppe
....

Phonemen wird eine 1:1-Zuordnung zwischen dem einzelnen Schriftzeichen und dem zugehörigen Phonem angestrebt. (Dies gilt, soweit die jeweilige Sprache eine Schrift auf Lautbasis besitzt. Ist dies nicht der Fall, so trifft diese Aussage entsprechend auf die kleinste Schrifteinheit zu.) Der Übergang zwischen Textform und akustischer Form der Sprache erfolgt durch die *Rechtschreibungs-* bzw. *Ausspracheregeln*. In zweierlei Hinsicht unterscheiden sich die Textform und die akustische Form der Sprache jedoch grundlegend:

1) Die akustische Form der Sprache ist reicher an Ausdrucksmöglichkeiten als die Textform. Die Textform enthält nur die Information, die sich mit Hilfe der Einheiten von Tabelle 1.1 darstellen lässt. Die akustische Form lässt dagegen auch zusätzliche Information *semantischer* bzw. *prosodischer* Art (z.B. Betonung und Hervorhebung einzelner Einheiten) und *sprecherspezifischer* Art zu (Identität des Sprechers, Stimmung usw.). Aus diesem Grund sind Satzbau und Wortwahl für (spontane) gesprochene Sprache und geschriebenen Text grundsätzlich verschieden.

2) Die Textform stellt die einzelnen Elemente, insbesondere die Phoneme, in Form einer örtlichen *diskreten Sequenz* dar. Verbindungen zwischen den Elementen sind unwichtig oder gar störend (Druckschrift ist leichter zu lesen als Schreibschrift!). Die Ausspracheregeln streben eine entsprechende Realisierung auch in der akustischen Form an, indem jedem Phonem eine bestimmte Artikulations- und Anregungsart, d.h., eine bestimmte Zielkonfiguration von Anregung und Vokaltrakt, zugeordnet wird. Gesprochene Sprache als Folge diskreter Artikulationsstellen-

³ Bei graphischer Darstellung digitaler oder abgetasteter Signale wird sogar häufig eine (pseudo-)analoge Darstellung gewählt, wenn es nur darauf ankommt, die Signalform als solche darzustellen, nicht aber die einzelnen Abtastwerte. Hiervon wird auch in diesem Skript Gebrauch gemacht.

gen mit vernachlässigbaren Übergängen ist aus folgenden Gründen nicht möglich:

- die Geschwindigkeit der Träger der Artikulation (insbesondere Lippen und Zunge) ist begrenzt, so dass die Übergänge einen großen Teil der gesamten Sprechdauer ausmachen;
- nachdem die Anregung zeitkontinuierlich erfolgt, werden die Übergänge nicht unterdrückt;
- viele Laute, insbesondere die, die sich der transienten Anregung bedienen, besitzen de facto keine Zielkonfiguration des Vokaltrakts, d.h., eine solche wird entweder nie erreicht, oder aber sie erlaubt keine hörbare stationäre Äußerung, wie z.B. ein Verschluss).

Im Unterschied zur Textform ist damit in der akustischen Form eine diskrete Folge von Elementen nicht unmittelbar gegeben; die Hauptinformation liegt in den *Übergängen*. Diese werden damit für die akustische Informationsübertragung wichtiger als die Zielstellungen des Vokaltraktes selbst. Es ist daher im Prinzip unmöglich – zumindest auf Phonemebene – eine *eindeutige* Zuordnung zwischen der Folge akustischer Ereignisse und einer Folge gesprochener Phoneme vorzunehmen.

1.4.2 Prinzip der Informationsebenen

Eine alternative Gliederung der Sprachinformation, die Text- und akustische Form gleichermaßen umfasst, erhält man, wenn man die Sprache in *Informationsebenen* oder *-quellen* gliedert, die nicht notwendigerweise hierarchisch miteinander verknüpft sein müssen (es ist aber für didaktische Zwecke sinnvoll, dies hier als gegeben anzunehmen). Die verschiedenen Ebenen sind in Tabelle 1.2 aufgelistet. Im Zusammenhang mit der Sprachsignalverarbeitung werden alle Informationsebenen jenseits der Ebene "Akustik-Phonetik" als *höhere Informationsebenen* bezeichnet. Grob gesprochen, sind die akustische, die parametrische und die akustisch/phonetische Ebene der akustischen Form der Sprache zugeordnet; die lexikale, syntaktische und (bedingt auch) die semantische Ebene gehören zur Textform; die phonologische Ebene stellt die Verbindung zwischen den beiden Formen her.

Tabelle 1.2. Die verschiedenen Informationsebenen der Sprache

Informationsebene	Repräsentation
Akustik	Sprachsignal
Parameter	Quellen- und Vokaltraktparameter
(Akustik/)Phonetik	Realisationen von Phonemen ("Phone") oder anderer phonetischer Einheiten
(Phonetik/)Phonologie	Phoneme oder entsprechende elementare Einheiten (z.B. Silben)
Lexikalische Ebene	Morpheme oder Wörter bzw. deren Realisierung
Syntaktische Ebene	Wortgruppe und/oder Satz
Semantische Ebene	Inhalt und Bedeutung einer Aussage
Pragmatik	"Umfeld" oder "Weltmodell", in das die Aussage eingebettet ist.
Prosodie	Betonung, Sprachmelodie (Grundfrequenzverlauf), Dauer (Rhythmus) und Intensität einzelner Laute, Silben etc. (in der Textform der Sprache so gut wie nicht vertreten)
Individuelle Merkmale	Klangfarbe der Stimme, Dialekt, Tonfall, Stimmung, Wortwahl und Sprechstil

Die Darstellung der Sprache mit Hilfe der höheren Informationsebenen soll an einem Beispiel erläutert werden. Gesprochen werde der Satz: "Der schwarze König geht von Feld H4 nach G3." Sollte er von einem Sprachinterpretationssystem automatisch verarbeitet werden, so würde – immer richtige Verarbeitung vorausgesetzt – das System auf der lexikalischen Ebene erkennen, dass der Satz aus 11 Wörtern besteht; weiterhin würde festgestellt, um welche Wörter es sich handelt, und welche Wortarten (Substantiv, Verb usw.) jeweils vorliegen. Auf der syntaktischen Ebene würde festgestellt, dass ein einzelner Aussagesatz vorliegt, der ein Subjekt, ein Prädikat und zwei adverbiale Bestimmungen

aufweist; außerdem würde festgestellt, dass der Satz syntaktisch richtig ist. In der lexikalen und syntaktischen Ebene steuert die Prosodie wertvolle Information bei. Die semantische Ebene, die nach dem Sinn der Aussage fragt, stellt fest, dass mit dem Satz ein Schachzug definiert wird, und dass der Satz folglich einen Sinn besitzt. Die pragmatische Ebene schließlich steuert das Wissen bei, das für das Schachspiel spezifisch ist: die Bezeichnung der einzelnen Felder, die Spielregeln, die Bezeichnung der Figuren, aber auch die momentane Stellung der Figuren in einem speziellen Spiel. Nur mit diesem Wissen lässt sich endlich der Satz vollständig interpretieren, d.h., es lässt sich feststellen, ob der Zug legal (d.h., den Spielregeln entsprechend), ausführbar und sinnvoll ist. (Nicht ausführbar wäre der Zug beispielsweise, wenn auf Feld G3 bereits eine andere schwarze Figur stünde; nicht sinnvoll wäre er, wenn sich die schwarze Partei beispielsweise durch diesen Zug in Schach begäbe.)

1.4.3 Der Informationsfluss auf den verschiedenen Informationsebenen

Eine wesentliche Eigenschaft des akustischen Sprachsignals ist der hohe *Informationsfluss*, der in einem krassen Missverhältnis zur übertragenen höheren Information steht. Tabelle 1.3 gibt die entsprechende Übersicht.

In dieser Tabelle steht die akustische Form ganz obenan. Die Textform (sofern eine Sprechgeschwindigkeit von ca. 10 Phonen je Sekunde vorausgesetzt ist) liegt bei etwa 60 bit/s. Durch weitere Reduktion der Information werden die ca. 10 bis 20 bit/s bei der Verarbeitung im menschlichen Bewusstsein gewonnen, die zum Verstehen und Verarbeiten der Information durch den Menschen genügen. Auch beim stillen Lesen eines Textes führt der Mensch eine Informationsreduktion durch. Dass die Information, die zum Verstehen ausreicht, den menschlichen Bewusstseinskanal nicht auslastet (besonders wenn der Hörende mit dem Inhalt, also der semantischen Komponente, schon vertraut ist), lässt sich experimentell leicht verifizieren (Schnellsprechen; Schnelllesen; Verrichten von Handlungen, während jemand spricht).

Tabelle 1.3. Informationsfluss der Sprache in verschiedener Repräsentation

Informations- ebene	Informations- fluss (bit/s)	Repräsentation
Akustik	240000	HiFi-Sprachsignal (Abtastfrequenz 20 kHz, Wortlänge 12 bit/Abtastwert)
	64000	Sprachsignal in Telefon-PCM-Qualität (Standardwerte: Abtastfrequenz 8 kHz, Wortlänge 8 bit/Abtastwert)
	2400	Untere Grenze für Repräsentation auf akustischer Ebene (mit aufwendigen Codierverfahren)
	4800	Parametrische Darstellung in guter Qualität oder gemischte Darstellung akustisch-parametrisch
Parameter	2400	Parametrische Darstellung in guter Qualität
	1200	Parametrische Darstellung in mäßiger Qualität
	400	Untergrenze für parametrische Darstellung bei sehr grober Quantisierung
Akustik/Phonetik	100-500	Gemischte parametrisch-phonetische Darstellung (z.B. "Phonetischer Vocoder")
Phonetik/Phonologie	60	Darstellung als Phonemfolge. Angenommen: rund 40 Phoneme, d.h., 6 bit/Phonem; Sprechgeschwindigkeit 10 Phone/s; keine einschränkenden linguistischen Bedingungen
Phonologie u. höhere	10-20	Enthaltene Information, die zum Verstehen der Nachricht notwendig ist, unter Einbeziehung aller Einschränkungen linguistischer Art (Syntax, Semantik, Pragmatik).

1.5 Die wesentlichen Aufgaben der Sprachsignalverarbeitung

Zu den Aufgaben der Sprachsignalverarbeitung rechnet man alle Aufgaben, die das Sprachsignal als Ausgangspunkt und/oder als Endpunkt besitzen. Je nach Zahl und Art der zusätzlich berührten Informationsebenen ergeben sich die verschiedenen charakteristischen Aufgaben, wie sie in Bild 1.7 zusammengestellt sind.

Beim Übergang von einer niederen zu einer höheren Informationsebene ist stets eine Datenreduktion durchzuführen; beim Übergang in umgekehrter Richtung erfolgt eine Expansion des Datenflusses. Aufgaben mit Ausgangspunkt Sprachsignal, deren Ziel in einer höheren Informationsebene liegt, werden der *Sprachanalyse* zugerechnet, die sich zur *Sprach-* oder *Sprechererkennung* ausweitet, wenn die entsprechende Informationsebene berührt wird. In umgekehrter Richtung spricht man von *Sprachsynthese*. Ist das Sprachsignal Ausgangspunkt und zugleich auch Endpunkt der Verarbeitung, so hat man ein *Sprachübertragungssystem* vor sich. Je nach durchlaufener Ebene spricht man hierbei von einem einfachen *Sprachcodierverfahren*, einem (parametrischen) *Vocoder*, oder einem *phonetischen Vocoder*. Einzelheiten hierzu folgen in einem späteren Kapitel.

Unter Sprachanalyse und Sprachsynthese im engeren Sinn versteht man Systeme und Algorithmen zum Übergang von der akustischen auf die parametrische Ebene bzw. umgekehrt. Im folgenden sollen die beiden Begriffe nur noch in diesem engeren Sinn interpretiert werden.

Die in dieser Vorlesung behandelten Aufgaben beschränken sich im wesentlichen auf die akustische, parametrische und akustisch/phonetische Informationsebene. Dementsprechend sind hier die folgenden Aufgaben zu behandeln: 1) *Sprachanalyse* zum Übergang von der akustischen auf die parametrische Ebene; 2) *Sprachsynthese* als Umkehrung der Analyse; 3) *Segmentation* zum Übergang von der parametrischen auf die akustisch/phonetische Ebene; sowie 4) *Verkettung* (Konkatenation) als Umkehrung der Segmentation. Jede dieser Aufgaben hat ihre spezifischen Probleme und Lösungsmethoden.

Jeder Übergang von einer Informationsebene zur nächsthöheren ist mit Verlusten an (wesentlicher) Information verbunden. Die Trennung

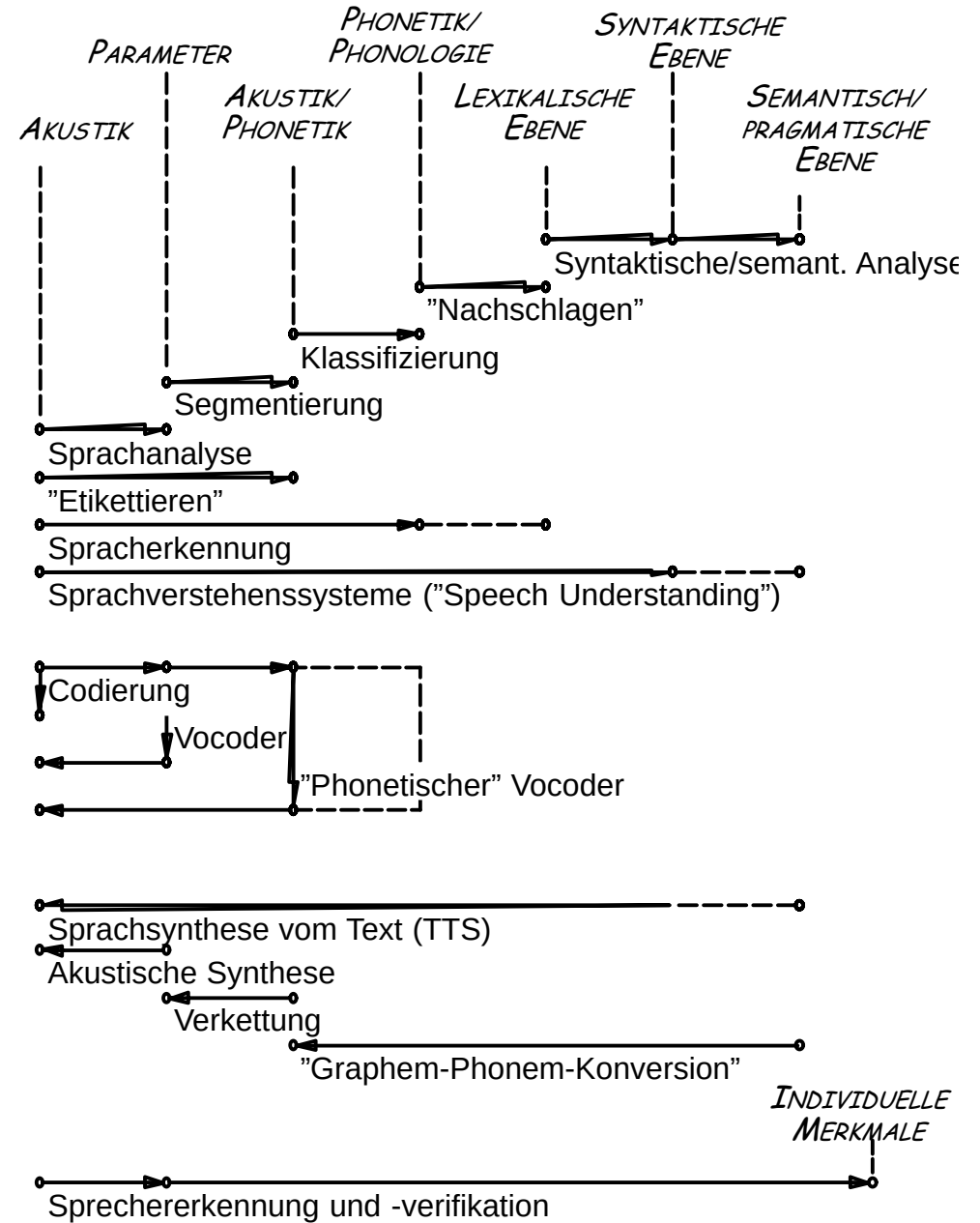


Bild 1.7. Zusammenstellung der wesentlichen Aufgaben der Sprachsignalverarbeitung

von Nutzinformation einerseits und Redundanz/Irrelevanz andererseits gelingt nie ganz! Dementsprechend wird die Qualität eines Sprachübertragungs- bzw. Sprachsynthesystems um so schlechter sein, je höher die Informationsebene ist, von der die Verarbeitung ausgeht bzw. die im Lauf der Verarbeitung berührt wird. Als erstes gehen hierbei die sprecherspezifischen Merkmale verloren; es folgt die Natürlichkeit der Sprache (besonders letzteres weitgehend eine subjektiv definierte Größe), dann letztlich ihre Verständlichkeit; mit der Verständlichkeit allerdings steht und fällt die Brauchbarkeit eines solchen Systems.

1.6 Phonetische Einheiten und die Sprachsignalverarbeitung

Als kleinste (bedeutungsunterscheidende) phonetische Einheit ist in der Linguistik und Phonetik das *Phonem* definiert. Akustisch steht hierbei das Phonem (als Aussprachevorschrift betrachtet) für eine bestimmte *Zielstellung* des Sprechtraktes (Artikulation sowie Anregungsform). Gesprochene Sprache ergibt sich durch *sequentielles Aneinanderfügen* von Phonemen. Auch die Artikulation ist noch zumindest teilweise darauf ausgerichtet; in der akustischen Realisierung verwischen sich die Grenzen dann allerdings weitgehend (siehe hierzu die Lehrveranstaltungen *Grundlagen der Phonetik* sowie *Phonetik und akustische Kommunikation 1*).

Die akustische Realisierung eines Phonems ist das *Phon*. Zwischen einem Phonem und der Menge der zugehörigen Phone besteht über die Ausspracheregeln kein eindeutiger Zusammenhang; es können kontextabhängig (und auch teilweise nach Belieben des Sprechers) verschiedene Realisationen vorgeschrieben bzw. möglich sein. Man spricht in solchen Fällen von *Allophonen* bzw. *freien Varianten*.

Bild 1.8 zeigt das gleiche Sprachsignal wie Bild 1.1, nur mit eingezeichneten Lautgrenzen (allerdings in Klartext, nicht in phonetischen Symbolen). Offensichtlich ändern sich die Eigenschaften des Sprachsignals beim Übergang von einem Laut zum nächsten nicht schlagartig. Dies ist im wesentlichen durch die Trägheit der Artikulationsorgane bestimmt, die ihre Stellung nicht sprunghaft wechseln können. Wesentliche Information liegt daher in den *Lautübergängen*. Die Lautgrenzen dürfen nur als Anhaltspunkte dafür dienen, wo ungefähr ein Laut endet

und der nächste beginnt. Den wechselseitigen Einfluss benachbarter Phone infolge der artikulatorischen Randbedingungen und Einschränkungen bezeichnet man als *Koartikulation*. Die Koartikulation beschränkt sich nicht nur auf den Übergang benachbarter Phone, sondern sie beeinflusst darüber hinaus die tatsächlich realisierte (oder auch nicht realisierte) zugehörige Zielstellung des Sprechtraktes. Sie bildet somit eines der wesentlichen Probleme beim Übergang vom Signal auf die akustisch-phonetische Informationsebene (und umgekehrt) und stellt eine erhebliche Fehlerquelle dar für die Systeme, die diesen Übergang durchführen müssen. Andererseits wird auch der Artikulationsaufwand um so größer, je weiter voneinander entfernt die Zielstellungen aufeinanderfolgender Phone sind.

Die Phoneme sind als Einheiten zur Beschreibung der Sprache gut geeignet, weil ihre Zahl verhältnismäßig klein ist: ca. 40 im Deutschen und in anderen europäischen Sprachen. Phoneme eignen sich darüber hinaus vorzüglich zur Beschreibung der Steuerung der Artikulation, d.h., zur Beschreibung, wie der Vorgang des Sprechens, also der Übergang von der Textform zur akustischen Form der Sprache erfolgen soll. Für den Übergang in umgekehrter Richtung sowie für maschinelle Sprachsynthese sind diese Einheiten weniger günstig, weil sie auf die Koartikulation, die vom Sprecher mehr oder weniger unbeabsichtigt ausgeführt wird, keine Rücksicht nehmen. Besser verhalten sich für diese Aufgaben solche Einheiten, die mit Koartikulationsgrenzen zusammenfallen, d.h., sich gegenseitig nicht so stark beeinflussen wie z.B. benachbarte Phone. Solche Einheiten sind am ehesten auf Silbenebene zu finden.

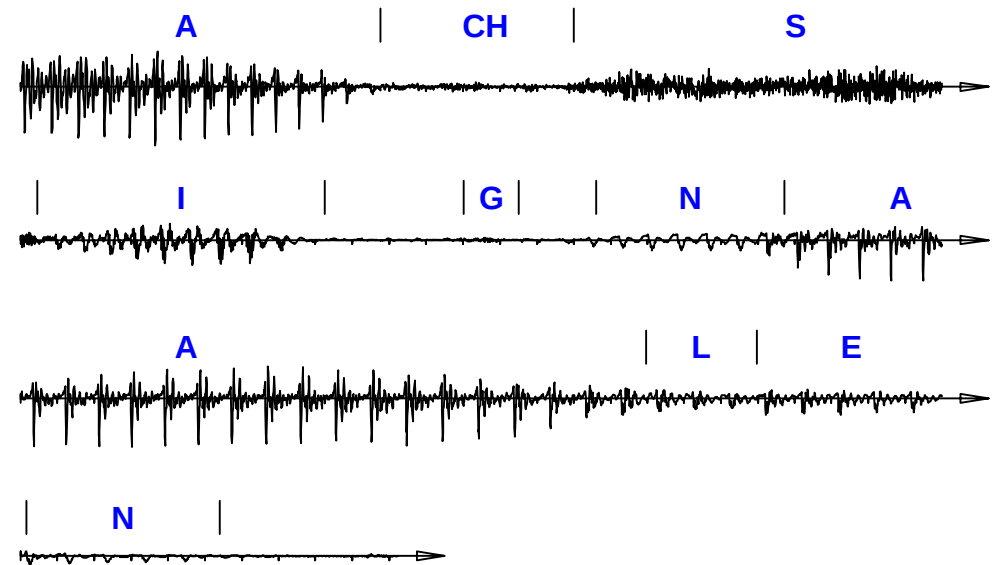
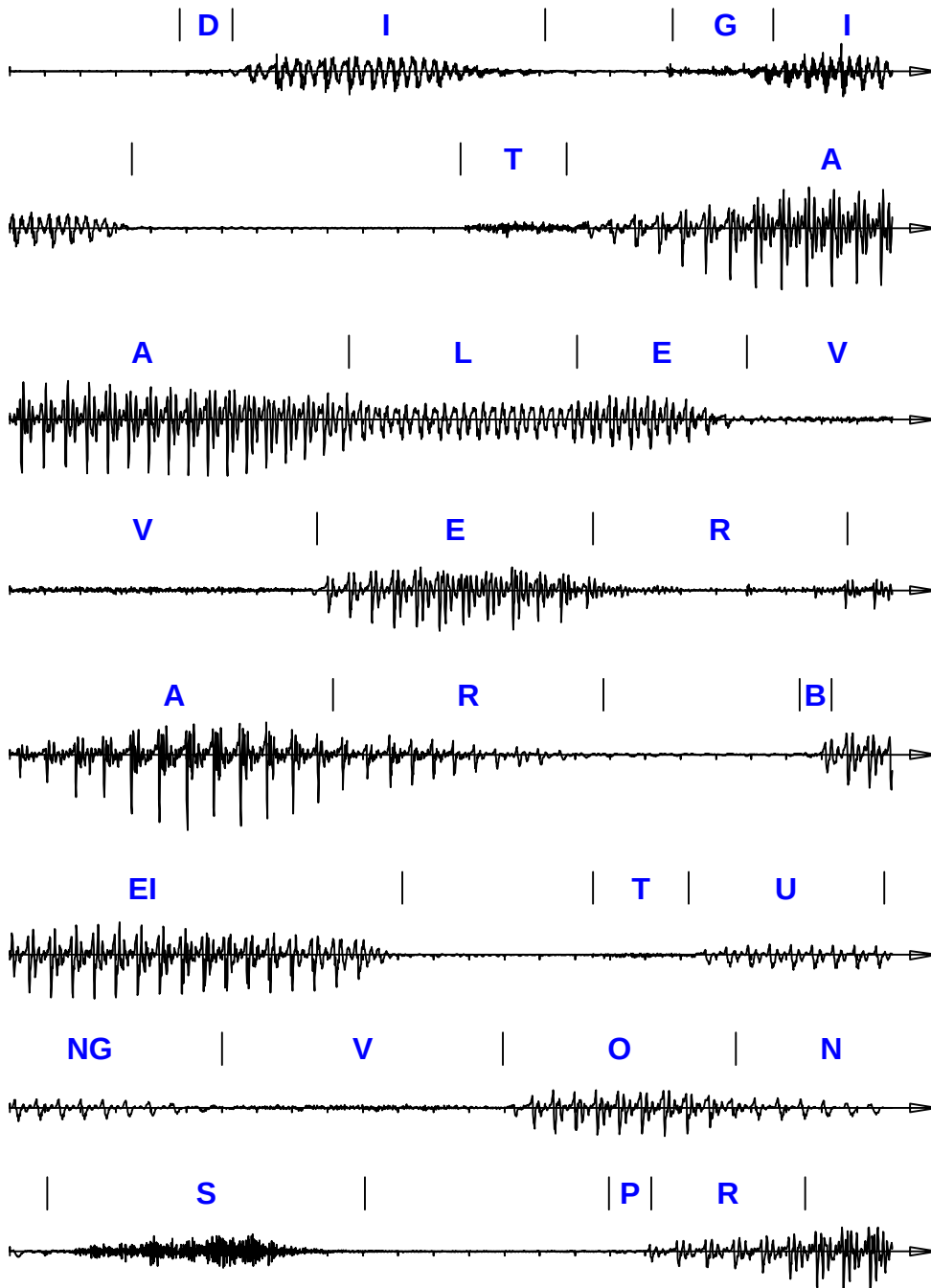


Bild 1.8. Beispiel eines Sprachsignals. Maßstab: 250 ms je Zeile; Sprecher WGH (männlich). Mit orthographischer Transkription und eingezeichneten Lautgrenzen. Das Signal ist mit dem in Bild 1.1 gezeigten identisch

Literatur⁴

- Kunt, Murat (1980): *Traitement numérique des signaux*. Traité d'Electricité, vol.20 (Editions Georgi, St.Saphorin, Suisse)
- Meyer-Eppler, Werner (²1969): *Grundlagen und Anwendungen der Informationstheorie*. (2. Aufl. 1969; mit einer Erweiterung von G. Heike; 1. Aufl. 1959) (Springer, Berlin)
- Rosen, Stuart / Howell, Peter (1991): *Signals and systems for speech and hearing* (Academic Press, London)
- Shannon, Claude E. (1949): "Communication in the presence of noise." Proc. IRE 37, 10-21
- Vary, Peter / Heute, Ulrich / Hess, Wolfgang (1998): *Digitale Sprachsignalverarbeitung* (Teubner, Stuttgart)

⁴ Diese Literatur dient primär der Quellenangabe. Als Vertiefungsliteratur dringend empfohlen sei das Buch von Rosen und Howell. Das Buch von Vary et al. ist als Lektüre für Fortgeschrittene geeignet. Die Bücher von Kunt und Meyer-Eppler sowie der Artikel von Shannon haben mit Sprachsignalverarbeitung ursprünglich nichts zu tun.

Grundlagen der Sprachsignalverarbeitung

2. Kurzzeitanalyse; elementare Parameter des Sprachsignals

2.1 Einige Grundbegriffe der Systemtheorie¹

Unter einem System (Bild 2.1) verstehen wir eine Vorrichtung, die nach einer bestimmten Rechenvorschrift S ein Signal x eindeutig in ein Signal y umwandelt. Hierbei wird x als Eingangssignal, y als Ausgangssignal bezeichnet. Die Signale x und y können Funktionen der gleichen unabhängigen Variablen sein; dies ist jedoch nicht unabdingbar. In verallgemeinerter Form kann ein System auch mehrere Ein- oder Ausgänge haben.²

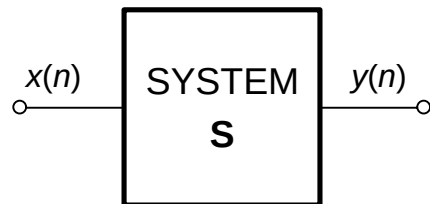


Bild 2.1. Blockdiagramm (Blockschaltbild) eines Systems mit Rechenvorschrift S . $[x(n)]$ Eingangssignal, $[y(n)]$ Ausgangssignal

Die Systemtheorie betrachtet und behandelt die Auswirkung der Rechenvorschrift S auf die Signale unabhängig von der tatsächlichen Realisierung des Systems. S wird auch gelegentlich als *Transformationsvorschrift* oder als *Operator* bezeichnet:

¹ Die Begriffe *System* und *Systemtheorie* werden je nach Wissenschaftszweig verschieden definiert und verwendet. In diesem Abschnitt werden die in der *Nachrichtentechnik* gebräuchlichen Definitionen eingeführt. Diese umreißen den Systembegriff sehr eng.

² Schüller (1984, S.2) definiert das *System* wie folgt: "Ein System ist ein mathematisches Modell eines Gebildes mit Eingängen und Ausgängen, das die Beziehungen zwischen den Eingangs- und Ausgangsgrößen beschreibt."

$$y(n) = S \{ x(n) \} . \quad (2.1)$$

$\{x(n)\}$ bedeutet die endliche bzw. abzählbar unendliche Menge der Abtastwerte des Eingangssignals, wohingegen $y(n)$ den einzelnen Abtastwert des Ausgangssignals zum Zeitpunkt n kennzeichnet.

Bezüglich Ein- und Ausgangssignal gilt das Prinzip von Ursache und Wirkung. Das Ausgangssignal $y(n)$ ergibt sich als Wirkung, deren Ursache das Eingangssignal $x(n)$ ist. Ursache und Wirkung sind über die Rechenvorschrift S miteinander verknüpft. Das Signal $y(n)$ wird auch als *Antwort* des Systems auf das Eingangssignal $x(n)$ bezeichnet (*Systemantwort*). Besondere Bedeutung besitzt die *Impulsantwort* $h(n)$, die sich ergibt, wenn man an den Eingang des Systems den Einheitsimpuls $\delta(n)$ anlegt:

$$\delta(n) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 . \end{cases} \quad (2.2)$$

Die Systemtheorie behandelt im einfachsten Fall Systeme, die *linear*, *zeitinvariant*, *stabil* und *kausal* sind. Die Bedeutung dieser Grunddefinitionen wird im folgenden erklärt.

Linearität. Ein System ist *linear*, wenn es das *Superpositionsprinzip* erfüllt. D.h., sofern

$$y_1(n) = S \{ x_1(n) \} \quad \text{ sowie } \quad y_2(n) = S \{ x_2(n) \} ,$$

dann ist das Superpositionsprinzip erfüllt, wenn gilt

$$k_1 y_1(n) + k_2 y_2(n) = S \{ k_1 x_1(n) + k_2 x_2(n) \} ; \quad (2.3)$$

hierbei sind k_1 und k_2 Konstante. Legt man also an den Eingang eines Systems die (gewichtete) Summe zweier Ursachen an, so addieren sich

am Ausgang die zugehörigen Wirkungen. Systeme, die das Superpositionsprinzip nicht erfüllen, werden als *nichtlinear* bezeichnet.

Die Linearität ist eine der wichtigsten Voraussetzungen für zahlreiche Aufgaben in der Nachrichtentechnik und der Signalverarbeitung, insbesondere auch bei der Übertragung und Verarbeitung von Sprachsignalen. Von ihr hängen ab: der Typ der Rechenvorschrift S und ihrer Anwendbarkeit sowie der Typ der zugehörigen Übertragungsfunktion (siehe Kap. 3). Unter zusätzlichen einschränkenden Voraussetzungen (die aber in der Sprachsignalverarbeitung meist gegeben sind) ist es möglich, einzelne Komponenten eines linearen Systems zu Teilsystemen zusammenzufassen oder ihre Reihenfolge zu ändern.

In der Theorie linearer Systeme wird davon ausgegangen, dass die Bedingung der Linearität streng eingehalten ist. In der Praxis digitaler Systeme treten durch begrenzte Wortlängen bei der Darstellung von Signalen und bei arithmetischen Operationen jedoch Effekte auf, die das realisierte System insgesamt *schwach nichtlinear* machen. Diese Tatsache werden wir allerdings weitgehend vernachlässigen können.

Nichtlineare Systeme zeichnen sich dadurch aus, dass sie ein Signal *nichtlinear verzerren*. Insbesondere entstehen durch nichtlineare Verzerrungen Signale mit neuen, in den Eingangssignalen nicht enthaltenen Frequenzen, aus denen sich die ursprünglichen Frequenzverhältnisse nicht mehr ohne weiteres rekonstruieren lassen.

Beispiel für ein nichtlineares System. Gegeben sei ein Signal $s(t) = \cos at + \cos bt$; es werde über ein System geschickt, das das Eingangssignal quadriert; wir erhalten also

$$S: y(t) = [x(t)]^2,$$

und im Fall unseres Signals $s(t)$

$$\begin{aligned} y(t) &= [\cos at + \cos bt]^2 = \cos^2 at + \cos^2 bt + 2 \cos at \cos bt \\ &= 1 + \frac{1}{2} \cos 2at + \frac{1}{2} \cos 2bt + \cos[(a+b)t] + \cos[(a-b)t]. \end{aligned}$$

Das Superpositionsprinzip gilt hier nicht. Gegenüber dem Eingangssignal umfasst das Ausgangssignal fünf bisher nicht vorhandene Kreisfrequenzen

$$\omega_1 \dots \omega_5 = 2a, 2b, a+b, a-b \text{ sowie Null.}$$

Die ursprünglichen Kreisfrequenzen a und b sind im Ausgangssignal nicht mehr enthalten.

Zeit(in)varianz. Ein System ist *zeitinvariant*, wenn ein gegebenes Eingangssignal $x(n)$ zu allen Zeiten auf das gleiche Ausgangssignal $y(n)$ führt, d.h., wenn

$$y(n) = S\{x(n)\}, \quad \text{dann gilt } y(n-k) = S\{x(n-k)\}, \quad (2.4)$$

und zwar unabhängig von der (als beliebig, aber fest angenommenen) Verzögerung k .

Sprachsignale sind, wie wir bereits gesehen haben, in der Regel *nicht* zeitinvariant.

Stabilität. Ein System ist *stabil*, wenn es auf jedes wertemäßig beschränkte Eingangssignal $x(n)$ mit einem beschränkten Ausgangssignal $y(n)$ reagiert, d.h., wenn

$$|x(n)| < M_x < \infty, \quad \text{dann gilt } |y(n)| < M_y < \infty \quad (2.5)$$

Diese Stabilitätsbedingung ist für wertkontinuierliche lineare Systeme notwendig und hinreichend. Quantisierungseffekte und Nichtlinearitäten können bei realen digitalen Systemen (also z.B. in wertdiskreter Realisierung) jedoch bewirken, dass ein System nicht stabil ist, obwohl die theoretische Stabilität nach (2.5) gewährleistet ist. In solchen Fällen muss die Stabilität anders definiert werden; man gelangt dabei zum Begriff der *asymptotischen Stabilität* (*Ljapunow-Stabilität*). Hier wird verlangt, dass, sofern das Eingangssignal zu Null wird, auch das Ausgangssignal gegen Null strebt:

$$\text{wenn } x(n) \equiv 0 \text{ für } n > n_1, \quad \text{dann gilt } y(n) \rightarrow 0 \text{ für } n \rightarrow \infty. \quad (2.6)$$

Kausalität. Ein System ist *kausal*, wenn jeder Abtastwert des Ausgangssignals $y(n)$ nur von zeitlich zurückliegenden, höchstens aber gleichzeitigen Abtastwerten des Eingangssignals $x(n)$ abhängt:

$$y(n) = y[x(k), k = \dots (1) n]. \quad (2.7)$$

Anders ausgedrückt: Sind alle Abtastwerte eines Eingangssignals $x(n)$ Null für $n < n_0$, so kann auch das Ausgangssignal eines kausalen Systems für $n < n_0$ keine von Null verschiedenen Abtastwerte besitzen.

Es ist üblich geworden, den Kausalitätsbegriff auch auf *Signale* auszuweiten. Ein Signal wird dann als *kausal* bezeichnet, wenn es für $n < 0$ identisch verschwindet.

Alle in der Praxis realisierten Systeme sind kausal. Für theoretische Betrachtungen oder für Messungen an *gespeicherten*, zeitlich begrenzten Signalen ist es in bestimmten Fällen sinnvoll, das Kausalitätsprinzip (vorübergehend) zu verletzen. Dies ist dann möglich, wenn die Kausalität durch Hinzufügen einer endlichen Verzögerung wiederhergestellt werden kann.

2.2 Das Prinzip der Kurzzeitanalyse

Das Sprachsignal gehört zu den Signalen, bei denen die Veränderlichkeit mit der Zeit eine wesentliche Rolle spielt. Die hauptsächlichen Parameter des Signals werden damit auch *zeitabhängig*. Sprachsignalanalyse ist daher vornehmlich die Analyse eines *zeitvarianten* Signals.

Grundannahme: Das Signal ist zeitveränderlich, aber die relevanten Parameter ändern sich *langsam* gegenüber den Abtastwerten des Signals. Damit wird der Zeitverlauf des gemessenen Parameters ein neues Signal, das jedoch mit einer wesentlich niedrigeren Frequenz abgetastet werden kann als das ursprüngliche Signal. Die Abtastfrequenz des Parametersignals wird im folgenden als *Parameterabtastfrequenz* (bzw. -rate, englisch: "frame rate") bezeichnet, im Unterschied zur *Signalabtastfrequenz*, die den einfachen Namen Abtastfrequenz behält.

Ein beliebiger Parameter P wird damit an einzelnen *Stützstellen* $n=q$ in nicht notwendigerweise konstanten, künftig aber stets als konstant vereinbarten Intervallen (*Messintervallen*, englisch: "frame") bestimmt, die durch die *Parameterabtastrate* vorgegeben sind. Bei der Messung ist zu gewährleisten, dass auch wirklich der Augenblickswert $P(q)$ des Parameters zum Zeitpunkt $n=q$ gemessen wird. Dies erreicht man durch die *Gewichtung* des Signals im Zeitbereich mit Hilfe der *Gewichtungsfunktion* $w(k)$, die auch als *Fensterfunktion* oder einfach *Fenster* bezeichnet

wird (engl.: "weighting function", oder "window"). Diese Gewichtung betont die Abtastwerte des Signals in der Umgebung des Zeitpunkts $n=q$ und unterdrückt die Abtastwerte, die weit von diesem Zeitpunkt entfernt sind (Bild 2.2). Das Signal $s(n)$ wird also aufgeteilt in eine Folge von Teilsignalen $s_K(n, q)$ derart, dass

$$s_K(n, q) = s(n) w(q-n). \quad (2.8)$$

wird. Hierbei verändert sich die Stützstelle q des Parameters in Schritten von Q (Parameterabtastintervall, kurz als *Messintervall* bezeichnet).

In der digitalen Darstellung ist die Fensterfunktion $w(k)$ stets *zeitbegrenzt*, besitzt also eine endliche Dauer K . Hiermit nimmt auch das ge-

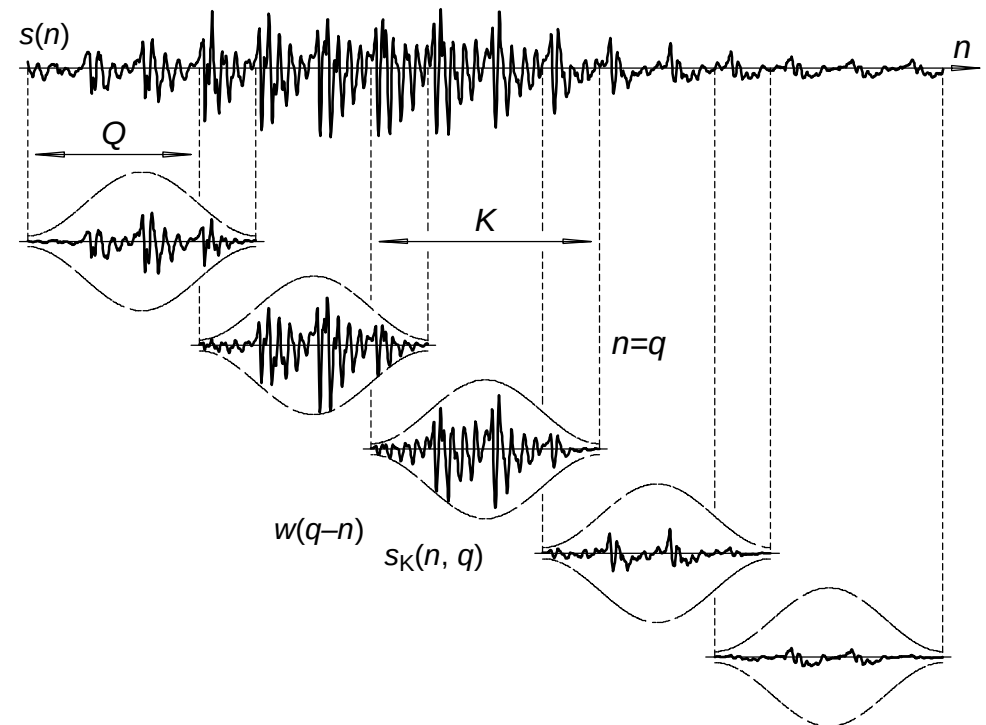


Bild 2.2. Prinzip der Kurzzeitanalyse: Unterteilung des Signals in Teilsignale mit Hilfe der Gewichtungsfunktion (Fenster). (Q) Messintervall (frame); (q) Beispiel für eine Stützstelle; (K) Fensterlänge; (w) Gewichtungsfunktion; (s_K) Kurzzeitsignal

wichtete Signal $s_K(n, q)$ eine endliche Länge an, und wir können dieses Signal z.B. wie folgt definieren:

$$s_K(n, q) = \begin{cases} s(n) w(q-n); & n = q-K+1 \text{ (1) } q \\ \text{sonst (zunächst) undefiniert, d.h. beliebig.} \end{cases} \quad (2.9)$$

Die Fensterlänge K ist so zu wählen, dass der zu bestimmende Parameter 1) zuverlässig genug gemessen werden kann, aber 2) auch noch mit hinreichender Genauigkeit als konstant angenommen werden darf. Da diese beiden Bedingungen sich widersprechen, wird die Messung stets von einer gewissen Unschärfe begleitet sein. In der Sprachsignalverarbeitung gebräuchliche Werte sind:

- für das Messintervall Q : 5 bis 25 ms
- für die Fensterlänge K : 20 bis 50 ms.

Von einem Fenster zum nächsten tritt also bei Sprachsignalen in der Regel eine bestimmte Überlappung ein.

Zur Behandlung der undefinierten Werte außerhalb des Fensters werden je nach Messmethode verschiedene Vereinbarungen getroffen:

1) Das Signal s_K wird als *einmaliger Vorgang* angesehen. In diesem Fall werden die Signalwerte von s_K außerhalb des Fensters zu Null gesetzt. Damit lassen sich alle Methoden anwenden, die zur Analyse eines einmaligen bzw. zeitinvarianten Vorgangs herangezogen werden können ("stationärer" Ansatz).

2) Das Signal s_K wird als *periodischer Vorgang* angesehen. In diesem Fall wird das Signal außerhalb des Fensters periodisch fortgesetzt; die Periode ist in der Regel gleich der Fensterlänge K (ist sie größer, so muss mit Nullwerten aufgefüllt werden). Damit lassen sich alle Methoden anwenden, die zur Analyse eines periodischen zeitinvarianten Vorgangs herangezogen werden ("synchroner" bzw. "periodischer" Ansatz).

3) Das Signal s_K wird als *Ausschnitt* aus einem *zeitvarianten Vorgang* angesehen. In diesem Fall ist eine besondere Definition der verwendeten *Messmethode* für zeitlich begrenzte Signale erforderlich. Das Signal s_K außerhalb des Fensters wird hierbei nicht gebraucht und kann undefi-

nirt bleiben ("nichtstationärer" Ansatz). Für den Fall, dass s_K außerhalb des Fensters doch einmal gebraucht wird, wählt man meist $s_K(n, q) = s(n)$, nimmt also die Originalwerte.

Ein Sonderfall ergibt sich, wenn die Fensterlänge K gleich der Dauer T_0 der Grundperiode wird. In diesem Fall lässt sich die natürliche Periodizität des Signals ausnützen. Die Methode hat den Vorteil, dass Verzerrungen durch das Fenster auf ein Minimum reduziert werden können (*grundperiodensynchrone* oder auch "grundfrequenzsynchrone" Analyse).

2.3 Elementare Sprachsignalparameter

Wichtigste elementare Kenngrößen akustischer Signale sind *Lautstärke und Frequenz*(-anteile). Dieser Abschnitt ist der *groben* Bestimmung dieser und daraus abgeleiteter Parameter gewidmet. Detaillierte Auskunft über das Frequenzverhalten des Sprachsignals liefert die *Spektralanalyse*, die in Kapitel 3 behandelt wird.

Abschnitt 2.3.1 stellt zunächst die analogen Bezugsgrößen aus der Akustik und Elektrizitätslehre zusammen. In den Abschnitten 2.3.2 und 2.3.3 werden dann die elementaren Sprachsignalparameter im digitalen Bereich definiert.

2.3.1 Physikalische Grundgrößen in der Sprachsignalverarbeitung

Energie, Leistung, Effektivwert. In der Physik sind Energie (Arbeit) und Leistung vor allem aus Mechanik und Elektrizitätslehre wohlbekannt:

$$E = F \cdot s; \quad [(\text{mechanische}) \text{ Arbeit gleich Kraft mal Weg}]; \quad (2.10)$$

$$P = U \cdot I; \quad [(\text{elektrische}) \text{ Leistung gleich Spannung mal Stromstärke}]. \quad (2.11)$$

Energie und Leistung sind verbunden durch die ebenfalls wohlbekannte Beziehung

$$P = E/t; \quad [\text{Leistung gleich Energie durch Zeit}]. \quad (2.12a)$$

oder umgekehrt

$$E = P \cdot t; \quad [\text{Energie gleich Leistung mal Zeit}] . \quad (2.12b)$$

Ein Grundgesetz der Elektrizitätslehre ist das *Ohm'sche Gesetz*, das die Größe des elektrischen *Widerstandes* definiert,

$$U/I = R; \quad [\text{Spannung durch Strom gleich Widerstand}]. \quad (2.13)$$

Der Widerstand ist dadurch gekennzeichnet, dass er eine Kenngröße des *Materials* ist, in dem der Strom fließt, d.h., der Widerstand hängt ab von der chemischen Zusammensetzung, dem Querschnitt und der Länge des verwendeten elektrischen Leiters; darüber hinaus hängt er zumeist auch von der Temperatur ab. *Nicht* (zumindest nicht unmittelbar) abhängig ist er jedoch von der anliegenden Spannung oder von der Stärke des Stromes, der in dem Leiter fließt (höchstens mittelbar auf dem Umweg z.B. über Erwärmungseffekte).

Unter Verwendung des Widerstandes lässt sich die elektrische Leistung allein durch *eine* der beiden Größen Spannung und Strom ausdrücken, wobei der Zusammenhang quadratisch wird:

$$P = U \cdot I = I^2 \cdot R = U^2/R . \quad (2.14)$$

Diese Beziehungen sind in dieser Form zunächst für den (einfachsten) Fall definiert, dass alle beteiligten Grundgrößen (Kraft, Spannung, Stromstärke und insbesondere auch die Leistung) *keine Funktionen der Zeit* sind. In der Mechanik bedeutet dies eine konstante Kraft, in der Elektrizitätslehre Gleichspannung bzw. Gleichstrom. Ist aber im Fall der Elektrizitätslehre – und nur dieser soll hier weiter betrachtet werden – die elektrische Spannung eine Funktion der Zeit, so errechnet sich beispielsweise die in einem Verbraucher während eines Zeitintervalls ($t_2 - t_1$) umgesetzte elektrische Energie zu

$$E = \int_{t_1}^{t_2} u(t) i(t) dt . \quad (2.15a)^3$$

³ Sind Spannung und Stromstärke Funktionen der Zeit, so werden sie – als Augenblickswerte – üblicherweise mit *kleinen* Buchstaben bezeichnet, Gleichspannungen, Gleichströme und Effektivwerte dagegen mit großen.

Der *Effektivwert* ist nun definiert z.B. als der Gleichspannungswert U_{eff} , der bei gleichem Widerstand R die gleiche (mittlere) Leistung ergibt wie die zugehörige zeitabhängige Spannung $u(t)$; dies bedeutet insbesondere auch gleiche Energie im betrachteten Zeitintervall:

$$E = \int_{t_1}^{t_2} u(t) i(t) dt = U_{\text{eff}} \cdot I_{\text{eff}} \cdot (t_2 - t_1) . \quad (2.15b)$$

Die haushaltsübliche elektrische Spannung von 220 V stellt einen Effektivwert dar.

In analoger Weise werden auch Effektivwerte für die im folgenden definierten akustischen Größen gebildet; hieran angelehnt auch ein (Pseudo-)Effektivwert⁴ für digitale Signale.

Grundgrößen der Akustik: Schalldruck, Schallschnelle, Schallkennimpedanz. Der *Schalldruck* p gibt an, wie sich durch einen Schall am Ort der Messung der Luftdruck (bzw. allgemein der Druck des Mediums, in dem sich eine Schallwelle bewegt) ändert. Gemessen wird er in *Pascal* (Pa). In der älteren Literatur ist auch noch die Einheit *Mikrobar* (μb) zu finden,

$$1 \text{ Pa} = 1 \text{ N/m}^2 = 10 \mu\text{b} .$$

Zu unterscheiden ist auch hier zwischen dem *Augenblickswert* $p(t)$ und dem *Effektivwert* $\tilde{p}$, der für die Berechnung der Schallenergie und -leistung herangezogen wird.

Die *Schallschnelle* v gibt die Geschwindigkeit an, mit der sich ein Luftteilchen infolge eines Schallereignisses bewegt. Bei beliebiger Welle-

⁴ Wichtig ist, dass in realen physikalischen Systemen zur Bestimmung von Energie- bzw. Leistungsmaßen immer *zwei* miteinander korrespondierende physikalische Größen verschiedener Dimensionen erforderlich sind, also Kraft und Weg in der Mechanik, Spannung und Stromstärke in der Elektrizitätslehre, Schalldruck und Schallschnelle in der Akustik. Dies gilt nicht für digitale Signale und Systeme; bei unserem digitalisierten Sprachsignal $s(n)$ beispielsweise steht nur *eine* Größe bzw. ein Wert zur Verfügung. Deshalb können wir hier – genau gesprochen – nur *Pseudoleistungen*, *Pseudoenergien* und *Pseudoeffektivwerte* definieren; diese Feinheit bleibt im täglichen Sprachgebrauch jedoch in der Regel unbeachtet.

nausbreitung ist $\mathbf{v}$ ein Vektor; bei dem uns in der Sprachsignalverarbeitung ausschließlich interessierenden Fall der *ebenen Welle* (d.h., einer Welle, die sich nur in einer Richtung ausbreitet – im Unterschied etwa zu einer sich kugelförmig nach allen Richtungen ausbreitenden Welle) kann $\mathbf{v}$ als skalare Größe v dargestellt werden.

Nicht zu verwechseln ist die Schallschnelle mit der *Schallausbreitungsgeschwindigkeit* (oder kurz *Schallgeschwindigkeit*) c , mit der sich die Schallwelle als Ganzes ausbreitet. In Luft ist c wesentlich größer als v .

Entsprechend dem Ohm'schen Gesetz in der Elektrizitätslehre kann auch in der Akustik eine Widerstandsgröße gebildet werden; sie ist im allgemeinen Fall definiert als *Schallfeldimpedanz*

$$Z_s := p/v \quad [\text{Einheit } \text{Ns/m}^3 \text{ oder } \text{Pa s/m}] \quad (2.16)$$

und kennzeichnet die akustischen Eigenschaften des Übertragungsmediums. Für eine ebene Welle im Ausbreitungsmedium Luft gilt (Zwicker und Zollner, 1985, S. 3)

$$Z_s = Z_0 = 414 \text{ Ns/m}^3 = 414 \text{ Pa s/m} . \quad (2.16a)$$

Diese Größe wird *Schallkennimpedanz* genannt. Sie lässt sich für jedes Schallübertragungsmedium angeben. Übertragungsmedien mit niedriger Schallkennimpedanz (beispielsweise Luft oder andere Gase) werden als *schallweich*, solche mit hoher Schallkennimpedanz (z.B. Flüssigkeiten oder feste Körper) als *schallhart* bezeichnet.

Pegelmaße, Schalldruckpegel. Da die in der Akustik üblichen Werte für den Schalldruck in weiten Grenzen (bis zu 7 Zehnerpotenzen zwischen Ruheschwelle und Schmerzgrenze) variieren, ist es üblich, den Schalldruck nicht direkt in Pa oder μPa , sondern als logarithmisches Verhältnismaß in der dimensionslosen Größe *Dezibel* (dB) anzugeben. Hierbei lässt sich dieses Maß sowohl für Leistungs- oder Energiegrößen als auch für Amplitudengrößen (elektrische Spannung oder Stromstärke, insbesondere aber auch Schalldruck) angeben. Dieses logarithmische Verhältnismaß ist für Leistungsgrößen definiert als

$$L := 10 \lg (P/P_0) \text{ [dB]} \quad (2.17a)$$

und für Amplitudengrößen, wie z.B. den Schalldruck p , als

$$L := 20 \lg (p/p_0) \text{ [dB]} \quad (2.17b)$$

hierbei bezeichnet $\lg$ den Zehnerlogarithmus. Da die (hier nicht explizit definierte) Schall-Leistung – entsprechend Beziehung (2.14) in der Elektrizitätslehre – quadratisch mit dem Schalldruck wächst, liefern die beiden Definitionen übereinstimmende Werte; 10 dB entsprechen einer Erhöhung der (Schall-)Leistung um den Faktor 10; die Erhöhung des Schalldruckes um den Faktor 10 bedeutet Anhebung des zugehörigen Pegels um 20 dB.

Die Einheit Dezibel bezeichnet keine physikalische Größe (wie z.B. m, kg oder s), sondern ist eine Konvention zur einfacheren Schreibweise. Daher ist streng darauf zu achten, dass bei der Berechnung von Pegelmaßen aller Art das Argument des Logarithmus dimensionslos ist. Dies wird erreicht durch Einführung eines *Bezugswertes* P_0 bzw. p_0 . Für den Schalldruck ist ein Wert $p_0 = 20 \mu\text{Pa}$ als Bezugswert international festgelegt und normiert. Dieser Wert liegt knapp unterhalb der Ruheschwelle bei 1000 Hz.

Spezielle akustische Signale. Ein *reiner Ton* ist ein Ton, bei dem der Schalldruck dem Gesetz

$$p(t) = \hat{p} \cos(\omega t + \varphi) \quad (2.18)$$

folgt. Der reine Ton wird auch als *Sinuston* oder allgemein als *Sinoidal-signal* bezeichnet. Für den Effektivwert gilt hier

$$\tilde{p} = \frac{\hat{p} \cdot \sqrt{2}}{2} . \quad (2.18a)$$

Als *Ton* bezeichnen wir ein Gemisch aus reinen Tönen, deren Frequenzen die *Grundfrequenz* $F_0 = \omega_0/2\pi$ als größten gemeinsamen Teiler besitzen:

$$p(t) = \sum_{\kappa=1}^k \hat{p}_\kappa \cos(\kappa \omega_0 t + \varphi_\kappa) . \quad (2.19)$$

Die Summe aus zwei (oder mehr) gleichzeitig erklingenden verschiedenen Tönen ergibt einen *Klang*.

Als *weißes Rauschen* wird ein Geräusch bezeichnet, das keine Tonhöhe und keinen Rhythmus besitzt und sich von einem Frequenzabschnitt

zum nächsten und von einem Zeitabschnitt zum nächsten nicht unterscheidet. Ein Spitzenwert $\hat{p}$ des Schalldrucks kann hier nicht angegeben werden, wohl aber ein Effektivwert $\tilde{p}$.

2.3.2 Amplitude, Pegel, Energie, Leistung des Sprachsignals in digitaler Darstellung

Eine naheliegende Methode der Grobbeschreibung des Sprachsignals ist die Bestimmung der Einhüllenden oder eines vergleichbaren Maßes für Pegel bzw. Amplitude. Dementsprechend wird der *Kurzzeiteffektivwert* des Sprachsignals wie folgt definiert:

$$u_K(q) := \sqrt{\frac{1}{K} \sum_{n=q-K+1}^q s_K^2(n, q)} \quad (2.20)$$

In der quadratischen Form wird dieser Parameter insbesondere in der angelsächsischen Literatur als *Kurzzeitleistung*

$$P_K(q) := \frac{1}{K} \sum_{n=q-K+1}^q s_K^2(n, q) \quad (2.21)$$

bzw. als *Kurzzeitenergie* bezeichnet:

$$U_K(q) := \sum_{n=q-K+1}^q s_K^2(n, q) \quad (2.22)$$

Daneben wird auch gern der *Betragsmittelwert* des Signals verwendet. In der Kurzzeitdefinition ergibt er sich zu

$$|\bar{s}_K(q)| := \frac{1}{K} \sum_{n=q-K+1}^q |s_K(n, q)| \quad (2.23)$$

Der Betragsmittelwert berechnet sich in der Regel leichter (d.h. schneller) als der Effektivwert, da keine Multiplikation vonnöten ist. Als weiteres Maß für die Einhüllende gilt auch die *Kurzzeitamplitude*, d.h. der

Kurzzeitspitzenwert des Signals, der betragsmäßig, als Wert von Spitze zu Spitze oder auch für jede Polarität einzeln definiert werden kann:

$$A_b(q) := \max [|s_K(n, q)|]; \quad (2.24a-d)$$

$$A_+(q) := \max [s_K(n, q)], \quad A_-(q) := \min [s_K(n, q)];$$

$$A_{ss}(q) := \max [s_K(n, q)] - \min [s_K(n, q)];$$

$$n = q-K+1 \text{ (1) } q.$$

Für das in Kap. 1 vorgestellte Signalbeispiel ist der Betragsmittelwert (in logarithmischer Darstellung) in Bild 2.3 aufgezeichnet. Hinzu kommt zur besseren Lesbarkeit eine vorläufige und automatisch vorgenommene (daher nicht ganz fehlerfreie) Unterteilung in stimmhafte und stimmlose Segmente.

2.3.3 Grobmessung des Frequenzverhaltens

Das *Frequenzverhalten* des Signals lässt sich grob messen, wenn man auch das *differenzierte Signal* analysiert:

$$s_D(n) \approx s(n) - s(n-1). \quad (2.25)$$

Erwähnt sei, dass mit (2.25), welches ein sehr einfaches digitales Filter darstellt, das Signal nicht exakt, sondern nur sehr annähernd differenziert wird. Anhand der Formel ist aber leicht festzustellen, dass tiefe Frequenzen unterdrückt und hohe Frequenzen verstärkt werden.

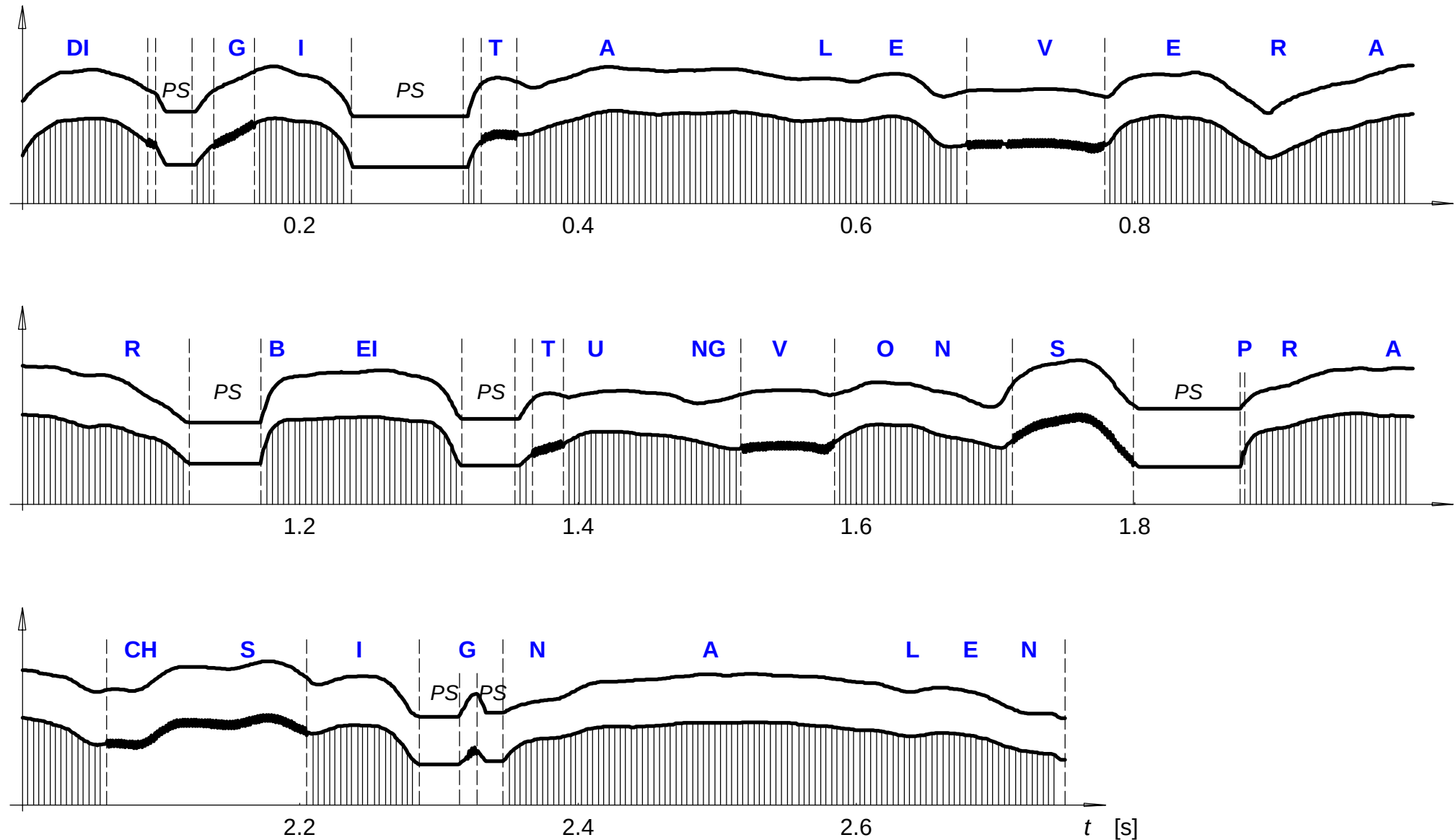
Auch für das differenzierte Signal lassen sich die elementaren Parameter bestimmen. Eine grobe Übersicht über das Frequenzverhalten gewinnt man mit Hilfe des Verhältnisses z.B. der Betragsmittelwerte von differenziertem Signal und Originalsignal:

$$Q_S := \frac{|\bar{s}_D(q)|}{2 |\bar{s}(q)|} \quad (2.26)$$

Dieser Wert ergibt sich für die Frequenz $f=0$ zu Null, für $f=F/2$ (halbe Abtastfrequenz) zu 1.

Bild 2.3. Betragsmittelwert des Signals. Beispiel: Text "Digitale Verarbeitung ...", Sprecher WGH. (Oberer Parameterverlauf) Betragsmittelwert des differenzierten Signals; (unterer Parameterverlauf) Betragsmittelwert des Signals (beide in logarithmischer Darstellung). Das Signal selbst wurde bereits in Kap.1 dargestellt. Beigegeben ist die Zuordnung Parameterverlauf-Text in orthographischer Transkription

Betragsmittelwert (logarithmische Darstellung)



Ein anderer Parameter, der einen groben Überblick über das Frequenzverhalten des Signals erlaubt, ist die Zahl bzw. relative Häufigkeit von *Nulldurchgängen*:

$$H_N(q) := \frac{1}{K} \sum_{n=q-K+1}^q x_S(n) \quad (2.27)$$

mit
$$x_S(n) = \begin{cases} 0, & \text{sgn}[s(n)] = \text{sgn}[s(n-1)] \\ 1, & \text{sgn}[s(n)] \neq \text{sgn}[s(n-1)] \end{cases}$$

Bei allen in diesem Abschnitt definierten Parametern ist es gleichgültig, welcher Ansatz (stationär, nichtstationär, oder periodisch) für die Kurzzeitanalyse gebraucht wird, da keiner der Parameter auf Werte außerhalb des Fensters Bezug nimmt.

Bisher haben wir in den Abschnitten 2.2 und 2.3 die Kurzzeitanalyse und die elementare Parametermessung als kausale Operationen dargestellt. Gerade bei diesen Operationen wird jedoch häufig die Stützstelle $n=q$ an den Beginn oder in die Mitte des Fensters gelegt. Liegt die Stützstelle am Beginn des Fensters, so müssen die Summationen bzw. Operationen in (2.9), (2.20-2.24) sowie (2.27) zwischen den Grenzen $n=q$ (1) $q+K-1$ laufen; liegt $n=q$ in der Mitte des Intervalls, so laufen die Operationen zwischen den Grenzen $n=q-K/2$ und $q+K/2-1$ ab.

2.4 Vereinfachte Modellierung des Sprechtrakts als lineares System

Wie bereits in Abschnitt 1.2 berichtet, lässt sich der menschliche Sprechtrakt funktional grob in zwei Teilkomponenten aufteilen: die *Anregung* und die *Signalformung*. Um nun das Sprachsignal über die Stufe der elementaren Parameter hinaus analysieren zu können, erweist es sich als notwendig, den Prozess der Spracherzeugung im menschlichen Sprechtrakt in einem einfachen Modell nachzubilden bzw. zu simulieren. Dies führt zum *linearen Modell der Spracherzeugung*, das im folgenden kurz (und qualitativ) beschrieben wird. Eine mehr quantitative Herangehensweise werden wir in Kap. 3 kennenlernen.

2.4.1 Einfache Modellierung des Sprechtrakts für stimmhafte Signale

Die beiden Komponenten der Spracherzeugung, Anregung und Signalformung, lassen sich in erster Näherung für stimmhafte Signale als voneinander unabhängige, in Reihe geschaltete lineare Systeme definieren. Damit wird auch das gesamte Modell linear. Es ist zweckmäßig, die (Kurzzeit-)Amplitude als unabhängigen multiplikativen Faktor hinzutreten zu lassen. Eine der wesentlichen Eigenschaften linearer Systeme besteht darin, dass derartige Komponenten an beliebiger Stelle in das Gesamtsystem eingefügt werden können; die Stellung zwischen Anregung und Signalformung ist für die weiteren Überlegungen besonders günstig. Bild 2.4 zeigt den grundsätzlichen Aufbau dieses Modells. Alle drei Komponenten sind zeitveränderlich.

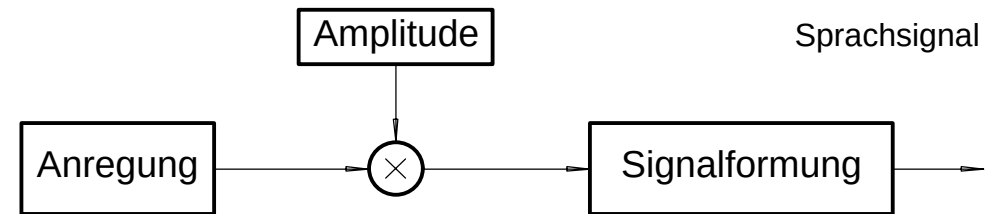


Bild 2.4. Einfachstes lineares Modell der Spracherzeugung für stimmhafte Signale

Anregung und Signalformung sind in dieser Form sehr komplexe Teilsysteme, die von zahlreichen verschiedenen Umgebungsbedingungen abhängen. Es erweist sich daher als günstig, beide noch weiter aufzuteilen. So lässt sich beispielsweise im Teilsystem Anregung die reine Periodizität (Grundfrequenz F_0 bzw. Grundperiode T_0) von einer weiteren Komponente trennen, die im wesentlichen die Klangfarbe der individuellen Stimme modelliert. Diese Komponente, die wir im folgenden als *Korrekturglied Anregung* bezeichnen wollen, ist sprecherspezifisch und kann – wiederum in erster Näherung – als *zeitinvariantes* Teilsystem angesehen werden, so dass mit dieser Komponente die Anregung in einen zeitveränderlichen und einen nicht zeitveränderlichen Anteil aufgespalten werden kann.

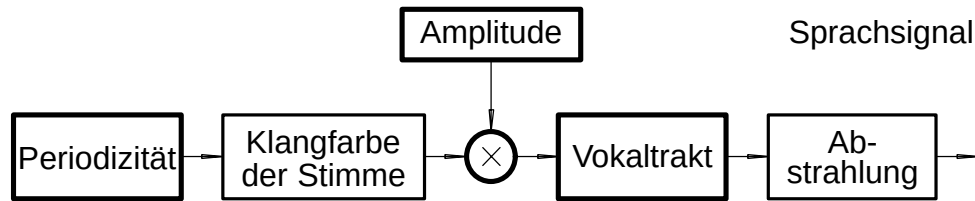


Bild 2.5. Das lineare Modell der Spracherzeugung für stimmhafte Laute

Gleiches lässt sich im Teilsystem Signalformung durchführen. Hier kann ein – ebenfalls in erster Näherung zeitinvariantes – *Korrekturglied* *Abstrahlung* abgespaltet werden, das die akustische Anpassung der von Mund bzw. Nase abgestrahlten Schallwelle an das freie Schallfeld nachbildet. Es verbleibt dann der *Vokaltrakt* bzw. das *Ansatzrohr* als eigenes zeitveränderliches Teilsystem. Dieses modelliert den wesentlichen Beitrag zur Signalformung.

Dem Ansatz der Modellierung der Signalformung (Vokaltrakt und Signalabstrahlung an Mund oder Nase) als lineares System liegt die *akustische Theorie der Vokalartikulation*⁵ (Fant, 1960; Ungeheuer, 1962; vgl. auch Flanagan, 1972) zugrunde. Diese modelliert das Ansatzrohr als akustische Leitung.

Es ist nun wünschenswert, die verschiedenen Komponenten noch weiter aufzuteilen, insbesondere zeitvariante und zeitinvariante Anteile zu trennen. Hierbei wird die Vokaltraktkomponente zerlegt in ein als zeitinvariant angenommenes Korrekturglied, das die *Abstrahlung* berücksichtigt, und die eigentliche (zeitvariante) *Übertragungsfunktion des Vokaltraktes*. Die stimmhafte Quelle wird zerlegt in die *Periodizität*, ein Korrektur- bzw. *Formungsglied* und – als multiplikativen Faktor – die *Amplitude* A , die auch den Einfluss des Vokaltrakts berücksichtigen soll (Bild 2.5). Die Periodizität sei definiert als reiner Impulsgenerator, der

eine Folge von Einheitsimpulsen abgibt. Der Abstand benachbarter Impulse ist hierbei die Periodendauer T_0 des Signals; diese ist zeitveränderlich. Als annähernd zeitinvariant (jedoch sprecherspezifisch) ist dagegen das Formungsglied anzusehen, das aus der reinen Impulsfolge das Anregungssignal entstehen lässt, wie es als Stimmbandschwingung bekannt ist. Hierbei kann das Formungsglied mit dem Anregungsspektrum, also der Klangfarbe der Stimme, gleichgesetzt werden.

2.4.2 Erweiterung des Modells auf stimmlose Signale

Das in Bild 2.5 gezeigte Modell gilt zunächst nur für stimmhafte Sprachsignale. Es lässt sich aber durch Hinzunahme eines stimmlosen Zweiges leicht auch auf stimmlose und transiente Anregung erweitern. Hierbei wird analog zu stimmhaften Lauten die stimmlose Anregung dargestellt durch eine Standardrauschquelle mit weißem Rauschen, das Korrekturglied und die Amplitude (stimmlos) als multiplikativen Faktor. Die beiden Zweige können additiv verknüpft werden, wenn vereinbart ist, dass die Umschaltung der Zweige mit Hilfe der Amplitudenregelung vorgenommen wird. Damit erhalten wir als Komponenten des Modells (Bild 2.6):

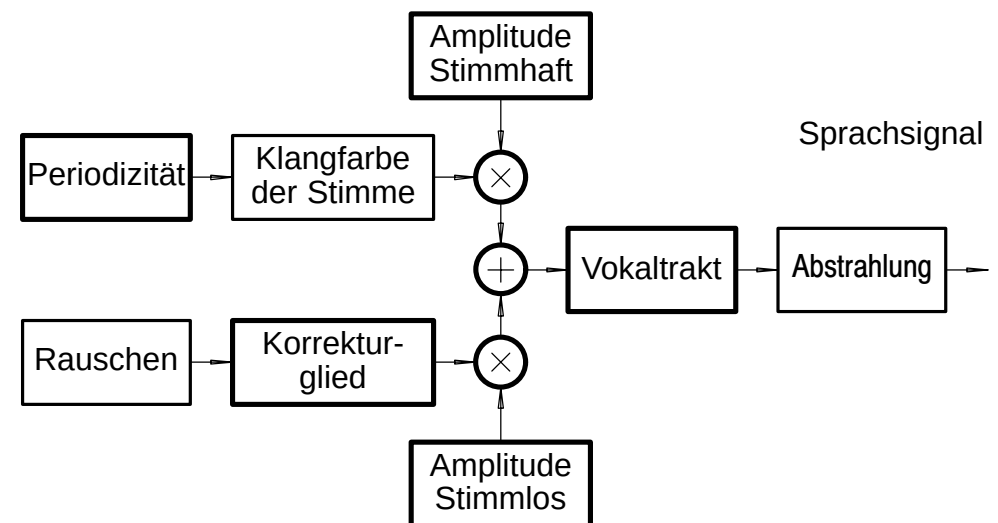


Bild 2.6. Das komplette lineare Modell der Spracherzeugung

⁵ Die akustische Theorie der Vokalartikulation ist anderen Lehrveranstaltungen vorbehalten. Sie kann in dieser Vorlesung aus Zeitgründen sowie wegen fehlender mathematischer Voraussetzungen nur angedeutet werden. Wir erschließen uns die wesentlichen Eigenschaften des Sprachsignals ersatzweise aufgrund von Beobachtungen und Messungen.

- 1) die beiden Anregungen als Impuls- bzw. Rauschgeneratoren mit "weißem", d.h. dem Betrag nach frequenzunabhängig konstantem Spektrum;
- 2) Korrektur- bzw. Formungsglieder, in denen die Anregungsfunktion so hergestellt wird, wie sie tatsächlich auftritt;
- 3) Amplitudensteuerung(en) A, die auch das Ein- und Ausschalten der Quellen übernehmen;
- 4) den Vokaltrakt (Ansatzrohr) als Hauptträger der Sprachinformation; sowie
- 5) ein Korrekturglied zur Berücksichtigung der Abstrahlung.

Literatur⁶

- Fant, Gunnar (²1970, ¹1960): *Acoustic theory of speech production* (Mouton, 's-Gravenhage)
- Flanagan, James L. (²1972, ¹1965): *Speech analysis, synthesis, and perception* (Springer, Berlin)
- Meyer-Eppler, Werner (²1969): *Grundlagen und Anwendungen der Informationstheorie*. (2. Aufl. 1969; mit einer Erweiterung von G. Heike; 1. Aufl. 1959) (Springer, Berlin)
- Rosen, Stuart / Howell, Peter (1991): *Signals and systems for speech and hearing* (Academic Press, London)
- Schüßler, Hans-Wilhelm (³1993, ¹1984): *Netzwerke, Signale und Systeme. Band II – Theorie kontinuierlicher und diskreter Signale und Systeme* (Springer, Berlin)
- Shannon, Claude E. (1949): "Communication in the presence of noise." Proc. IRE 37, 10-21
- Ungeheuer, Gerold (1962): *Elemente einer akustischen Theorie der Vokalartikulation* (Springer, Berlin)

- Vary, Peter / Heute, Ulrich / Hess, Wolfgang (1998): *Digitale Sprachsignalverarbeitung* (Teubner, Stuttgart)
- Zwicker, Eberhard / Zollner, Manfred (1984): *Elektroakustik*. Hochschultext (Springer, Berlin)

⁶ Diese Literatur dient primär der Quellenangabe. Als Vertiefungsliteratur dringend empfohlen sei das Buch von Rosen und Howell; die übrigen hier genannten Werke dürften für den Hörerkreis dieser Vorlesung noch zu anspruchsvoll sein. Das Buch von Schüßler, das Lehrbuch von Zwicker und Zollner sowie der Artikel von Shannon haben mit Sprachsignalverarbeitung ursprünglich nichts zu tun.

Grundlagen der Sprachsignalverarbeitung

3. Kurzzeitspektralanalyse

Die (Kurzzeit-)Spektralanalyse stellt das in der Phonetik mit Abstand wichtigste Messverfahren für Sprachsignale dar. Von den verschiedenen Prinzipien soll hier dasjenige diskutiert werden, das auf dem Rechner Anwendung findet, die *diskrete Fouriertransformation*. Wegen fehlender mathematischer Kenntnisse ist die folgende Einführung eher qualitativ und phänomenologisch gehalten, ohne dass allerdings auf die Mathematik ganz verzichtet werden kann.

3.1 Harmonische Analyse (diskreter) periodischer Signale

3.1.1 Sinoidalschwingung (Sinusschwingung) als Grundbaustein akustischer Signale

Grundlage jedes akustischen Signals ist die *Sinoidalschwingung* (Sinuston, reiner Ton, vgl. Abschnitt 2.2.1), also eine Schwingung, deren zeitlicher Verlauf durch eine Kosinus- oder Sinusfunktion beschrieben ist,

$$x_g(t) = A_g \cos \omega_0 t; \quad x_u(t) = A_u \sin \omega_0 t. \quad (3.1)$$

Hierbei ist $x_g(t)$ eine gerade, $x_u(t)$ eine ungerade Zeitfunktion. Die in (3.1) definierten Sinoidalschwingungen haben die *Kreisfrequenz* ω_0 ; diese ist durch die Beziehung

$$\omega_0 = 2\pi f_0 \quad (3.2)$$

mit der *Frequenz* der Schwingung verbunden. Unter der Frequenz versteht man die Zahl der Schwingungen je Zeiteinheit, in der Akustik

durchweg je Sekunde. Bekanntlich ist die Schwingung nach (3.1) mit der *Periode* oder *Periodendauer*

$$t_0 = 1 / f_0 \quad (3.3)$$

periodisch.¹

In der Akustik wird eine Sinoidalschwingung als *reiner Ton* bezeichnet. Im allgemeinen Fall hat ein reiner Ton einen Kosinus- und einen Sinusanteil; diese lassen sich zusammenfassen:

$$\begin{aligned} x_r(t) = x_{rg}(t) + x_{ru}(t) &= A_g \cos \omega_0 t + A_u \sin \omega_0 t \\ &= A \cos(\omega_0 t + \varphi) \end{aligned} \quad (3.4)$$

mit

$$A = \sqrt{A_g^2 + A_u^2} \quad \text{sowie} \quad \varphi = \arctan \frac{A_u}{A_g}. \quad (3.5a,b)$$

A ist die *Amplitude* der resultierenden Schwingung; φ stellt die *Phase* der Schwingung dar.

Die Sinoidalschwingung resultiert aus der gleichförmigen Kreisbewegung. Bewegt sich ein Punkt auf dem Umfang eines Einheitskreises, so führt seine Projektion auf die Abszisse des zugehörigen Koordinatensystems eine Kosinusschwingung aus (Bild 3.1).

Im Unterschied zum reinen Ton verstehen wir unter einem *Ton* in der Akustik eine i.A. nichtsinusförmige Schwingung $x_p(t)$, die aber wiederum mit der Periode t_0 periodisch sein soll. *Eine solche Schwingung lässt sich*

¹ Man beachte die doppelte Bedeutung des Wortes *Periode*: mit bestimmtem Artikel als Synonym für die Periodendauer, also eine die gesamten Schwingung global charakterisierende Eigenschaft, zum anderen mit unbestimmtem Artikel als Kennzeichnung eines beliebigen Abschnitts des Signals, dessen Dauer mit der Periodendauer identisch ist.

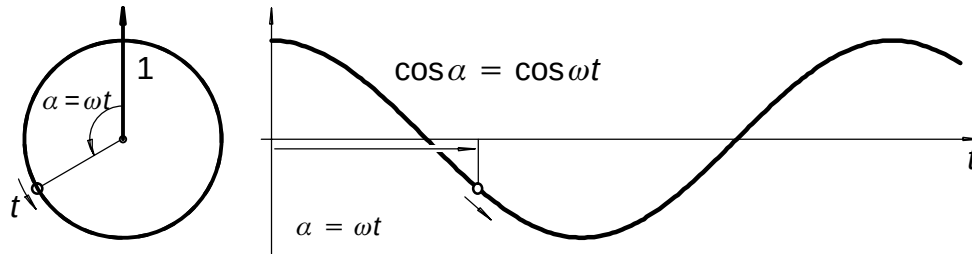


Bild 3.1. Zur Entstehung einer Sinoidalschwingung aus einer gleichförmigen Kreisbewegung. (Links) Einheitskreis mit dem sich gleichförmig (d.h., mit konstanter Winkelgeschwindigkeit) bewegenden Punkt; (rechts) daraus resultierende Kosinusschwingung als Projektion der Lage des Punktes auf die Abszisse. Aus Platzgründen ist das Bild um 90° gedreht gezeichnet (Abszisse des Kreiskoordinatensystems nach oben)

durch Zerlegung in eine Reihe als Summe von (i.A. unendlich vielen) Sinoidalschwingungen exakt darstellen (siehe Beispiel in Bild 3.2). Die zugehörige Reihe heißt (nach einem ihrer Entdecker, dem Mathematiker Joseph B. Fourier, benannt) *Fourierreihe*.

Die einzelnen Sinoidalschwingungen, die die Glieder der Reihe ausmachen, sind in ihren Frequenzen nicht beliebig wählbar. Ist t_0 die Periode der Gesamtschwingung, und gilt nach wie vor $f_0 = 1/t_0$ [(3.3)], so können die Glieder der Reihe nur Frequenzen von

$$0, f_0, 2f_0, 3f_0, \dots, mf_0, \dots, \quad m \text{ ganzzahlig} \quad (3.6)$$

annehmen. Anteile jeder anderen Frequenz würden die Bedingung verletzen, dass $x_p(t)$ mit der Periode t_0 periodisch ist (siehe Beispiel in Bild 3.3). Das Signal $x_p(t)$ stellt sich also als Fourierreihe wie folgt dar:

$$x_p(t) = \frac{a_0}{2} + a_1 \cos(\omega_0 t + \varphi_1) + a_2 \cos(2\omega_0 t + \varphi_2) + \dots + a_m \cos(m\omega_0 t + \varphi_m) + \dots \quad (3.7)$$

Jedes Glied der Reihe lässt sich gemäß (3.4) in einen Kosinus- und einen Sinusanteil zerlegen; Kosinus- und Sinusanteil lassen sich (unter bestimmten mathematischen Voraussetzungen, die aber bei realen Signalen in der Regel erfüllt sind) getrennt zusammenfassen,

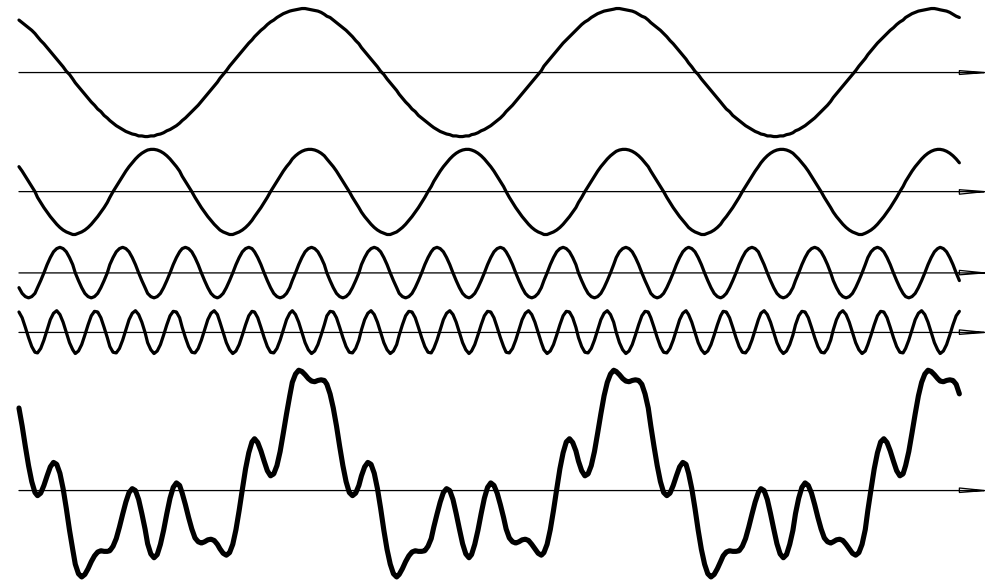


Bild 3.2. Beispiel für einen Ton, d.h., eine nichtsinusförmige periodische Zeitfunktion, und seine Zusammensetzung aus Sinoidalschwingungen (*Teilschwingungen*), deren Frequenzen ganzzahlige Vielfache der Grundfrequenz des Tones sind (hier im Beispiel ist der Ton zusammengesetzt aus der 1., 2., 5. und 8. Teilschwingung)

$$x_{Pg}(t) = \frac{a_0}{2} + a_{1g} \cos \omega_0 t + a_{2g} \cos 2\omega_0 t + \dots + a_{mg} \cos m\omega_0 t + \dots ;$$

$$x_{Pu}(t) = a_{1u} \sin \omega_0 t + a_{2u} \sin 2\omega_0 t + \dots + a_{mu} \sin m\omega_0 t + \dots \quad (3.8a,b)$$

Wie sich zeigt, ist $x_{Pg}(t)$ der *gerade* Anteil der Funktion $x_p(t)$; $x_{Pu}(t)$ stellt den *ungeraden* Anteil dar.

Die Darstellung einer periodischen (Zeit-)Funktion mit Hilfe der Fourierreihe ist dann nützlich, wenn es gelingt, die Koeffizienten a_{ig} und a_{iu} bzw. a_i und φ_i aus dem Signal $x_p(t)$ zu gewinnen. Dies wird von der Fourierreihe geleistet.

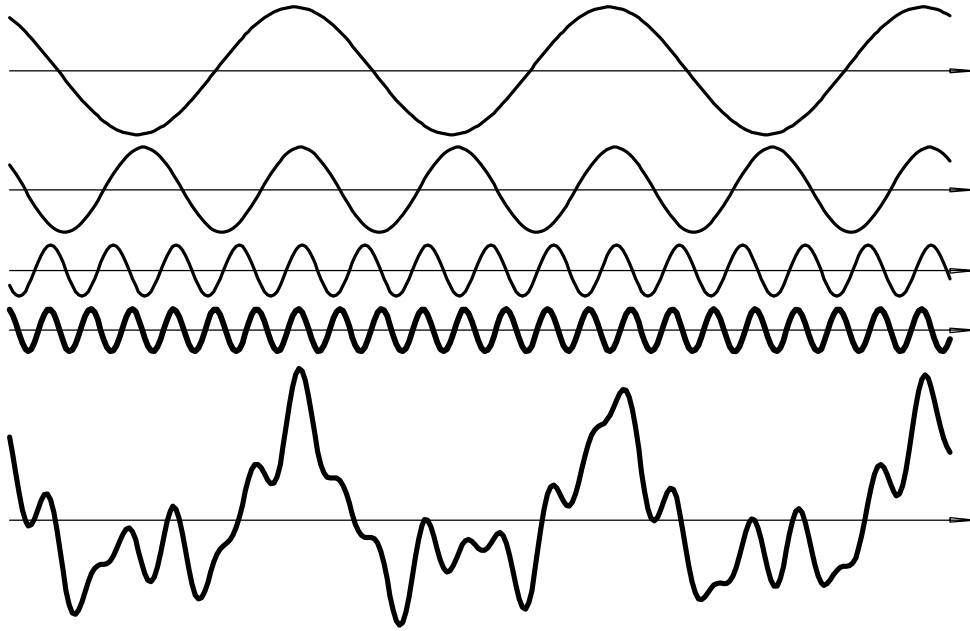


Bild 3.3. Eine durch Addition mehrerer Sinoidalschwingungen resultierende Zeitfunktion wird nur dann periodisch, wenn Bedingung (3.6) erfüllt ist. Wenn auch nur eine Teilschwingung "aus der Reihe tanzt", wird die Zeitfunktion aperiodisch (hier im Beispiel die dick eingezeichnete Teilschwingung, deren Frequenz ungefähr das 7,5-fache der Frequenz der 1. Teilschwingung beträgt)

3.1.2 Orthogonalitätsbeziehung und Fourierkoeffizienten

Mit Hilfe einer einfachen Beziehung wird zunächst gezeigt, dass es gelingt, eine einzelne Teilschwingung

$$x_{iP}(t) = a_i \cos i\omega_0 t + a_i \sin i\omega_0 t$$

vom Rest des Signals $x_P(t)$ zu isolieren. Es lässt sich leicht nachweisen, dass gilt

$$\int_0^{2\pi} \cos kx \cos mx \, dx = \begin{cases} \pi & k = m \\ 0 & k \neq m \end{cases}; \quad (3.9a)$$

$$\int_0^{2\pi} \sin kx \sin mx \, dx = \begin{cases} \pi & k = m \\ 0 & k \neq m \end{cases}; \quad (3.9b)$$

$$\int_0^{2\pi} \cos kx \sin mx \, dx = 0. \quad (3.9c)$$

Diese Beziehung bezeichnen wir als *Orthogonalitätsbeziehung* der trigonometrischen Funktionen. Wichtig ist, dass sich die Integrale stets über eine ganze Zahl von Perioden *beider* Integranden erstrecken. Dann und nur dann verschwinden die Integrale bis auf die eine Ausnahme $k=m$ in (3.9a,b).

Eine entsprechende Beziehung existiert bei digitalen Signalen; sie wird als *Summenorthogonalität* bezeichnet und lautet

$$\sum_{n=0}^{N-1} \cos \frac{k \cdot 2\pi n}{N} \cos \frac{m \cdot 2\pi n}{N} = \begin{cases} N/2 & k = m \\ 0 & k \neq m \end{cases} \quad (3.10a)$$

$$\sum_{n=0}^{N-1} \sin \frac{k \cdot 2\pi n}{N} \sin \frac{m \cdot 2\pi n}{N} = \begin{cases} N/2 & k = m \\ 0 & k \neq m \end{cases} \quad (3.10b)$$

$$\sum_{n=0}^{N-1} \cos \frac{k \cdot 2\pi n}{N} \sin \frac{m \cdot 2\pi n}{N} = 0. \quad (3.10c)$$

Bild 3.5 zeigt diese Orthogonalitätsbeziehung an einem Beispiel.

Mit Hilfe der in (3.9, 3.10) gezeigten Orthogonalitätsbeziehung gelingt es nun, aus einem beliebigen mit t_0 (bzw. N) periodischen Signal die Koeffizienten a_{ig} und a_{iu} zu gewinnen; hierbei erhalten wir im analogen Fall

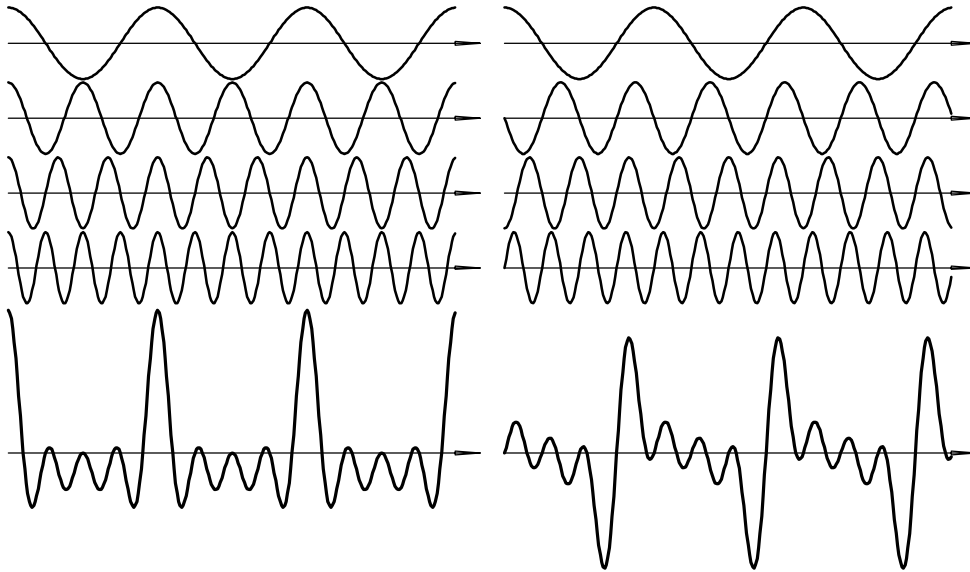


Bild 3.4. Phasenverschiebungen einzelner Teilschwingungen bewirken eine Änderung der Kurvenform der resultierenden Schwingung. Dabei kann sich auch der Spitzenwert ändern. Die Energie der Schwingung ändert sich dagegen nicht, wenn nur die Phasen, nicht aber die Amplituden der Teilschwingungen geändert werden. (Links) 1. bis 4. Teilschwingung gleicher Amplitude als Kosinusschwingungen mit Phase Null; (rechts) ebenso, nur ist jede Teilschwingung gegenüber der vorangehenden um 90° phasenverschoben

$$a_{mg} = \frac{2}{t_0} \int_0^{t_0} x_P(t) \cos m\omega_0 t dt, \quad m = 1(1) \dots; \quad (3.11a)$$

$$a_{mu} = \frac{2}{t_0} \int_0^{t_0} x_P(t) \sin m\omega_0 t dt, \quad m = 1(1) \dots; \quad (3.11b)$$

$$a_0 = \frac{2}{t_0} \int_0^{t_0} x_P(t) dt. \quad (3.11c)$$

sowie bei digitalen Signalen, wo der Einfachheit halber das Eingangssignal als $x(n)$ bezeichnet werden soll

$$a_{mg} = \frac{2}{N} \sum_{n=0}^{N-1} x(n) \cos \frac{m \cdot 2\pi n}{N}, \quad m = 0(1) N-1; \quad (3.12a)$$

$$a_{mu} = \frac{2}{N} \sum_{n=1}^{N-1} x(n) \sin \frac{m \cdot 2\pi n}{N}, \quad m = 0(1) N-1; \quad a_{0u} = 0. \quad (3.12b)$$

Auf den Beweis von (3.11) im allgemeinen Fall, bei dem die Konvergenzbedingungen für die Reihe betrachtet werden müssen, wird hier verzichtet.

3.1.3 Exkurs: Was benötigen wir an Mathematik in diesem Kapitel?

Aus der Schulmathematik wird gebraucht: Kenntnisse in Differentialrechnung (Ableitung von Funktionen, besonders von transzendenten, sowie die wichtigsten Differentiationsregeln) und Grundkenntnisse in Integralrechnung (Berechnung bestimmter Integrale reeller Funktionen). Zu bemerken ist allerdings, dass sich bei der Behandlung *digitaler* Signale die meisten Integrale auf Summen reduzieren.

Über die Schulmathematik hinausgehende Kenntnisse (werden in der Übung zu Grundlagen der Sprachsignalverarbeitung vermittelt) benötigen wir aus drei Bereichen: 1) komplexe Zahlen, 2) Reihenrechnung, 3) Grundkenntnisse über die komplexe Exponentialfunktion.

Es ist nun keineswegs so, dass wir diese drei Gebiete zur Gänze kennen müssen – dies würde Lehrveranstaltungen zur Mathematik von mehreren Semestern erfordern. Kennen müssen wir die Grunddefinitionen und in den meisten Fällen ein paar elementare Sätze. Hierzu nachfolgend ein paar Details.

1) Komplexe Zahlen. Wir benötigen die Rechnung mit komplexen Zahlen, um Schwingungen mit Hilfe der komplexen Exponentialfunktion beschreiben zu können. Diese Beschreibungsweise ist, sobald man sie einmal kennt, die bei weitem eleganteste, universell einsetzbar und wird von Physikern und Ingenieuren ausnahmslos verwendet.

Wichtigstes Prinzip bei der Rechnung mit komplexen Zahlen ist das Prinzip der *Permanenz von Regeln und Operatoren* bei der Erweiterung des Zahlensystems. In anderen Worten: bei der Erweiterung des Zahlensystems bleiben die Rechenregeln unverändert, und das bisherige Zahlensystem wird ein Sonder-

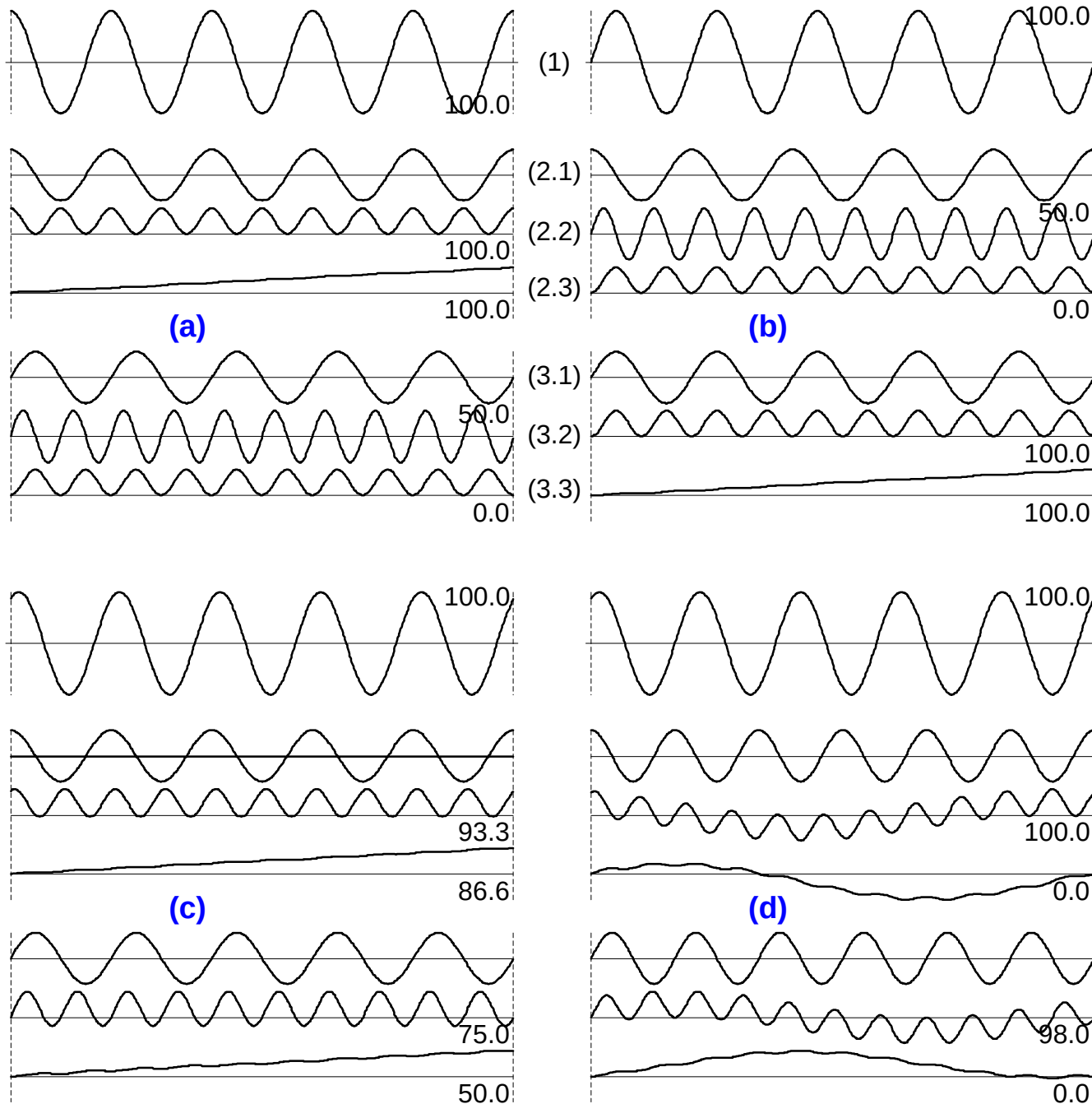


Bild 3.5a-d. Beispiel für die (Summen-)Orthogonalität der trigonometrischen Funktionen. In jedem Teilbild sind (von oben nach unten) gezeichnet: (1) Eingangssignal $x_p(n) = A \cos(k \cdot 2\pi n / N + \varphi)$ mit Angabe der Amplitude A ; (2.1) Testschwingung: gerader Anteil $g(n) = \cos(m \cdot 2\pi n / N)$; (2.2) Produkt $p_g(n) = x_p(n) \cdot g(n)$ normiert mit Angabe des Spitzenwertes; (2.3) Summe $(2/N) \cdot \sum p_g(i)$ von Beginn des Analyseintervalls bis zum laufenden Punkt $i = n$ mit Angabe des Endwertes für $n = N - 1$; (3) wie (2), nur mit dem ungeraden Anteil $u(n)$ der Testschwingung. Im Bild gilt stets $k = 5$, $N = 1024$, $A = 100$. (a) $x_p(n)$ gerade, d.h. $\varphi = 0$; $m = k = 5$; dementsprechend ergibt sich A aus dem geraden Anteil in Zeile (2.3); (b) $x_p(n)$ ungerade, d.h. $\varphi = -\pi/2$; $m = k = 5$; dementsprechend ergibt sich A aus dem geraden Anteil in Zeile (3.3); (c) $\varphi = -\pi/6$ (entspricht einer Phasenverschiebung in Zeitrichtung von 30°) – damit ist $x_p(n)$ weder gerade noch ungerade; $m = k = 5$; dementsprechend ergibt sich ein von Null verschiedener Beitrag im geraden wie auch im ungeraden Teil; die Amplitude A lässt sich hieraus durch Anwendung von (3.5) berechnen; (d) $x_p(n)$ wie in (c), jedoch $m = 6 \neq k$; damit werden gerader und ungerader Anteil gemäß (3.10a-c) gleichermaßen Null

fall des erweiterten. Beim schrittweisen Übergang von den natürlichen Zahlen auf die komplexen Zahlen gilt das wie folgt:

- Ausgangspunkt ist die Menge der natürlichen Zahlen: 1, 2, 3, Mit natürlichen Zahlen lassen sich die Grundrechenarten Addition und Multiplikation sowie ihre Umkehrungen Subtraktion und Division definieren und ausführen.
- Bei der Subtraktion können Ergebnisse herauskommen, die mit den natürlichen Zahlen nicht mehr zu beschreiben sind: wir erweitern um Null und die negativen Zahlen und erhalten die Menge der ganzen Zahlen; die natürlichen Zahlen sind eine Teilmenge hiervon. Die Rechenregeln gelten unverändert weiter; die Multiplikationsregel wird erweitert: minus mal minus gibt plus, Zahl mal null gibt null. Division durch Null ist nicht möglich.
- Bei der Division natürlicher oder ganzer Zahlen können Ergebnisse herauskommen, die mit ganzen Zahlen nicht mehr zu beschreiben sind. Wir erweitern um die gebrochenen Zahlen und erhalten die Menge der rationalen Zahlen; die ganzen Zahlen sind eine Teilmenge hiervon. Die Rechenregeln gelten unverändert weiter.
- Wir nehmen als neue Grundrechenart das Potenzieren hinzu: Potenz gleich Basis hoch Exponent. Ist die Basis rational und der Exponent ganzzahlig, so bleiben alle Potenzen im Rahmen der rationalen Zahlen. Bei den beiden Umkehrungen *Radizierung* (Berechnung der Basis aus Potenz und Exponent) und *Logarithmierung* (Berechnung des Exponenten aus Potenz und Basis) können Ergebnisse herauskommen, die nicht mit rationalen Zahlen zu beschreiben sind (z.B. $\sqrt{2}$). Wir erweitern das Zahlensystem um die irrationalen Zahlen und erhalten die Menge der reellen Zahlen; die rationalen sind eine Teilmenge hiervon. Die Rechenregeln gelten unverändert weiter; Radizieren und Logarithmieren werden (zumindest für positive Zahlen) möglich. Auch Potenzieren mit nicht ganzzahligen Exponenten wird möglich.
- Mit den reellen Zahlen ist die Zahlengerade komplett. Jede Zahl ist geometrisch darstellbar als Strecke vom Nullpunkt der Zahlengeraden bis zu dem Punkt, der durch die Zahl markiert ist.
- Die Quadratwurzel negativer Zahlen kann durch eine reelle Zahl nicht angegeben werden. Das Grundproblem hierbei ist: welchen Wert besitzt $\sqrt{-1}$? Das Problem wird gelöst, indem das Zahlensystem um die *imaginären* Zahlen erweitert wird; wir erhalten damit das System der komplexen Zahlen. Die reellen Zahlen sind eine Teilmenge hiervon.

- Die Umstellung auf komplexe Zahlen ist eine größere. Aus der eindimensionalen Zahlengeraden wird die zweidimensionale Zahlenebene; sie wird beschrieben durch ein kartesisches Koordinatensystem, dessen Abszisse durch die bisherige Zahlengerade der reellen Zahlen (*reelle Achse*) und dessen Ordinate durch die neue Zahlengerade der imaginären Zahlen gebildet wird (*imaginäre Achse*). Jede komplexe Zahl ist geometrisch darstellbar als Vektor vom Ursprung des Koordinatensystems bis zu dem Punkt in der Zahlenebene, der durch die Zahl markiert ist.
- Dementsprechend werden die Rechenregeln erweitert: Addition und Subtraktion folgen den Gesetzen der Vektorrechnung. Wichtig ist die alternative Darstellung im Polarkoordinatensystem durch Betrag und (Phasen-)Winkel. Multiplikation, Division und auch Potenzieren laufen in dieser Darstellung ab; z.B. bedeutet Multiplikation zweier komplexer Zahlen die Multiplikation der Beträge und die Addition der Winkel. Die Rechenoperationen für reelle Zahlen folgen genau den gleichen (erweiterten) Gesetzmäßigkeiten. Im Rahmen der komplexen Zahlen lassen sich alle Probleme der Grundrechenarten und ihrer Umkehrungen lösen; einer nochmaligen Erweiterung des Zahlensystems bedarf es nicht mehr.
- Was müssen wir von komplexen Zahlen wissen: Definition; Real- und Imaginärteil; Darstellung durch Betrag und Winkel; und die Tatsache, dass eine Vergrößerung des Winkels um $k \cdot 2\pi$ (k ganzzahlig) den Wert einer komplexen Zahl nicht verändert; die vier Grundrechenarten; konjugiert komplexe Zahlen. Potenzieren, Radizieren und Logarithmieren kann auf Grundprobleme beschränkt bleiben (etwa Potenzieren komplexer Zahlen mit ganzzahligen reellen Exponenten; Ziehen von Quadratwurzeln im Komplexen). Weitergehende Probleme werden im Zusammenhang mit der komplexen Exponentialfunktion erklärt.

2) *Reihenrechnung*. Bei digitalen Signalen haben wir es häufig mit Summen zu tun (etwa über Signalwerte). Um solche Summen behandeln zu können, bedarf es der Rechnung mit Reihen. In der Mathematik ist die Reihenrechnung jedoch ein ganzer Wissenschaftszweig, von dem wir nur einige wenige Grundkenntnisse brauchen. Diese sind:

- Definition: Eine Reihe ist eine Summe von Gliedern (Summanden). Jedes Glied (außer dem ersten) geht durch ein bestimmtes Bildungsgesetz aus dem vorangehenden Glied hervor. Wir kennen endliche und unendliche Reihen; eine unendliche Reihe hat unendlich viele Glieder. Beispiel für eine unendliche Reihe ist die unendliche geometrische Reihe

$$s = 1 + q + q^2 + q^3 + \dots + q^k + \dots = \sum_{n=0}^{\infty} q^n.$$

- **Konvergenz:** Eine unendliche Reihe ist konvergent, wenn die Summe einen endlichen Wert ergibt. Wir müssen dies im Einzelfall nachweisen. Bei unserer geometrischen Reihe ist dies leicht möglich. Eine endliche geometrische Reihe lässt sich nach der Summenformel berechnen,

$$s_k = \sum_{n=0}^k q^n = \frac{1 - q^{k+1}}{1 - q};$$

ist nun $|q| < 1$, so kann die Zahl k der Glieder gegen Unendlich streben; wir erhalten dann

$$s = \sum_{n=0}^{\infty} q^n = \lim_{k \rightarrow \infty} \frac{1 - q^{k+1}}{1 - q} = \frac{1}{1 - q}; \quad |q| < 1.$$

Der Bereich, hier $|q| < 1$, in dem die Reihe konvergiert, wird *Konvergenzbereich* genannt. Eine Reihe ist absolut konvergent, wenn sich jedes Glied durch seinen Betrag ersetzen lässt und die Reihe trotzdem noch konvergiert (ist bei der geometrischen Reihe der Fall).

- **Majorantenprinzip:** Eine Reihe ist konvergent, wenn sich zu ihr eine andere konvergente Reihe finden lässt, deren sämtliche Glieder ab einem bestimmten Index (betragsmäßig) gleich oder größer sind als die entsprechenden Glieder der untersuchten Reihe. – Für viele unserer Probleme in der Signalverarbeitung lässt sich eine konvergente geometrische Reihe als Majorante finden. Konvergenz und Majorantenprinzip gelten gleichermaßen für Reihen mit reellen und Reihen mit komplexwertigen Gliedern.
- Sinnvoll ist es, auch die grundlegenden Definitionen von Potenzreihen zu kennen, insbesondere der Taylorsche Reihe. Diese wird aber hier noch nicht unmittelbar benötigt.
- Für die Spektralanalyse unmittelbar notwendig ist die Kenntnis der harmonischen Reihe (Fourierreihe). Diese Reihe wird in diesem Kapitel behandelt.

3) *Komplexe Exponentialfunktion.* Die komplexe Exponentialfunktion (also die Exponentialfunktion zur Basis e mit komplexem Exponenten) erlaubt eine einheitliche und elegante Darstellung ungedämpfter, exponentiell gedämpfter und exponentiell anklingender Schwingungen mit einer einzigen Funktion und

führt zudem Kosinus und Sinus auf die Exponentialfunktion zurück. Hierzu sollten wir uns folgendes vergegenwärtigen.

- Das mathematische Gebiet der *Funktionentheorie*, auf das wir uns hier begeben, behandelt Funktionen komplexwertiger Veränderlicher. Eine solche Funktion $y=f(x)$ mit x und y in der Regel komplex können wir graphisch nicht darstellen, weil die Darstellung vier Dimensionen benötigen würde. Das hindert uns aber nicht daran, mit solchen Funktionen wie mit reellen Funktionen zu rechnen.
- Die Theorie der komplexen Exponentialfunktion ist selbst sehr kompliziert und sprengt den Rahmen dieser Vorlesung. Kennen müssen wir die Eulersche Formel und die Tatsache, dass die üblichen Gesetze der Potenzrechnung auch im Komplexen gelten. Dies wird im folgenden Abschnitt erklärt.

3.1.4 Übergang zur komplexen Exponentialfunktion. Die diskrete Fouriertransformation (DFT)

Wir beschränken uns im folgenden auf den Fall digitaler Signale, bei denen die Betrachtungen einfacher sind. Der (3.8) entsprechende Ansatz würde hier für den geraden Anteil lauten

$$x_{Pg}(n) = \frac{a_0}{2} + a_{1g} \cos \frac{2\pi n}{N} + a_{2g} \cos \frac{2 \cdot 2\pi n}{N} + \dots + a_{N/2,g} \cos \frac{(N/2) \cdot 2\pi n}{N}; \quad (3.13a)$$

und für den ungeraden Anteil

$$x_{Pu}(n) = a_{1u} \sin \frac{2\pi n}{N} + a_{2u} \sin \frac{2 \cdot 2\pi n}{N} + \dots + a_{N/2,u} \sin \frac{(N/2) \cdot 2\pi n}{N}; \quad (3.13b)$$

wegen der Abtastung ist die Reihe, wie wir noch sehen werden, auf den Bereich $m=0$ (1) $N/2$ beschränkt, d.h., die Glieder $m=0$ (1) $N/2$ beschreiben das Signal $x_p(n)$ vollständig.

Ein alternativer Ansatz verwendet die komplexe Exponentialfunktion und gibt die Möglichkeit, den geraden und den ungeraden Anteil der Fourierreihe zusammenzufassen. Die Euler'sche Formel

$$\exp(jx) = e^{jx} = \cos x + j \sin x; \quad j = \sqrt{-1} \quad (3.14)$$

verbindet die komplexe Exponentialfunktion und die trigonometrischen Funktionen; hieraus lassen sich Kosinus und Sinus leicht wiedergewinnen:

$$\cos x = \frac{e^{jx} + e^{-jx}}{2} ; \quad \sin x = \frac{e^{jx} - e^{-jx}}{2j} . \quad (3.15a,b)$$

Der mathematische Nachweis der Euler'schen Formel kann hier nicht erbracht werden. Sie wird nachgewiesen durch die Entwicklung der drei beteiligten Funktionen in Potenzreihen und den Nachweis der Übereinstimmung der Potenzreihen für die linke und die rechte Seite von (3.14). Wir können uns die Formel höchstens etwas plausibel machen, indem wir beispielsweise beide Seiten zweimal nach x ableiten (das geht für komplexe Funktionen genau so wie für reelle) und sehen, dass die Formel auch für die zweiten Ableitungen gilt.

Mit (3.15a,b) haben wir nun ein Mittel in der Hand, Schwingungen sehr elegant mit der Exponentialfunktion auszudrücken. Wie wir eingangs gesehen haben, resultiert eine Sinoidalschwingung aus der gleichförmigen Kreisbewegung. Bewegt sich ein Punkt auf dem Umfang eines Einheitskreises, so führt seine Projektion auf die Abszisse des zugehörigen Koordinatensystems eine Kosinusschwingung aus (Bild 3.1). Die Projektion ist nun eine mathematisch nicht einfach zu beschreibende Operation. Betrachten wir aber unseren Einheitskreis von Bild 3.1 als in der komplexen Zahlenebene liegend, so beschreibt die Euler'sche Formel genau alle Punkte auf dem Einheitskreis, und unser laufender Punkt auf dem Einheitskreis ist beschrieben durch

$$P = e^{j\alpha} = e^{j\omega t} = \cos \omega t + j \sin \omega t . \quad (3.16)$$

Wir definieren zu diesem Punkt den konjugiert komplexen Punkt

$$P^* = e^{-j\alpha} = e^{-j\omega t} = \cos \omega t - j \sin \omega t , \quad (3.17)$$

der ebenfalls stets auf dem Einheitskreis liegt und sich dort mit entgegengesetzter Winkelgeschwindigkeit bewegt wie P . Die Summe beider Punkte ergibt dann eine Kosinusschwingung (mit der Amplitude 2):

$$P + P^* = e^{j\omega t} + e^{-j\omega t} = 2 \cos \omega t \quad (3.18)$$

und haben hiermit die Kosinusschwingung auf eine Summe von zwei komplexen Werten rückgeführt, die beide eine gleichförmige Kreisbewegung in entge-

gegengesetzter Richtung beschreiben. Die Summe ist mathematisch sehr viel einfacher zu traktieren als eine Projektion.

Gedämpfte und *anklingende* Schwingungen lassen sich beschreiben, indem wir dem Exponenten einen Realteil hinzufügen:

$$\exp(at + j\omega t) = \exp(at) \cdot \exp(j\omega t) ; \quad (3.19)$$

indem wir die Gesetze der Potenzrechnung anwenden, können wir die Exponentialfunktion mit komplexem Exponenten in ein Produkt aus Exponentialfunktion mit reellem Exponenten und Exponentialfunktion mit imaginärem Exponenten zerlegen; der letztere Faktor beschreibt die Schwingung, während der erste eine anklingende ($\alpha > 0$) bzw. exponentiell gedämpfte ($\alpha < 0$) beschreibt; für die ungedämpfte Schwingung verschwindet der Realteil im Exponenten.

Damit sind Sinus und Kosinus auf die Exponentialfunktion zurückgeführt, und gedämpfte wie ungedämpfte und anklingende Schwingungen lassen sich mit der gleichen Funktion elegant darstellen.

Wir setzen nun (3.15a,b) in (3.13) ein,

$$\begin{aligned} a_{mg} \cos \frac{2\pi n m}{N} &= a_{mg} \cdot \left[\exp \frac{2\pi j n m}{N} + \exp \frac{-2\pi j n m}{N} \right] / 2 \\ &= a_{mg} \cdot \left[\exp \frac{-2\pi j n m}{N} + \exp \frac{2\pi j n m}{N} \right] / 2 \\ &= a_{mg} \cdot \left[\exp \frac{-2\pi j n m}{N} + \exp \frac{-2\pi j n \cdot (-m)}{N} \right] / 2 ; \end{aligned}$$

$$\begin{aligned} a_{mu} \sin \frac{2\pi n m}{N} &= a_{mu} \cdot \left[\exp \frac{2\pi j n m}{N} - \exp \frac{-2\pi j n m}{N} \right] / 2j \\ &= -j \cdot a_{mu} \cdot \left[\exp \frac{2\pi j n m}{N} - \exp \frac{-2\pi j n m}{N} \right] / 2 \\ &= j \cdot a_{mu} \cdot \left[\exp \frac{-2\pi j n m}{N} - \exp \frac{2\pi j n m}{N} \right] / 2 \\ &= j \cdot a_{mu} \cdot \left[\exp \frac{-2\pi j n m}{N} - \exp \frac{-2\pi j n \cdot (-m)}{N} \right] / 2 ; \end{aligned}$$

gleichzeitig definieren wir durch Aufspaltung des Klammerausdrucks neue Koeffizienten $X_m = X(m)$,

$$a_{mg} \cos \frac{2\pi n m}{N} = X_g(m) \exp \frac{-2\pi j n m}{N} + X_g(-m) \exp \frac{-2\pi j n \cdot (-m)}{N},$$

$$X_g(m) = X_g(-m); \quad (3.15c)$$

$$a_{mu} \sin \frac{2\pi n m}{N} = jX_u(m) \exp \frac{-2\pi j n m}{N} + X_u(-m) \exp \frac{-2\pi j n \cdot (-m)}{N},$$

$$X_u(m) = -X_u(-m); \quad (3.15d)$$

$$\text{entsprechend } \frac{a_0}{2} = X(0) = X_g(0). \quad (3.15e)$$

Beim Ansatz von (3.13a,b) mit der komplexen Exponentialfunktion gemäß (3.15c-e) läuft der Index m im digitalen Fall damit nicht mehr von 0 bis $+N/2$, sondern von $-N/2$ bis $+N/2$, wenn die Summe nach den Argumenten der Exponentialfunktion geordnet wird.

Da die trigonometrischen Funktionen Kosinus und Sinus – und damit auch die komplexe Exponentialfunktion mit imaginärem Argument – mit 2π periodisch sind, ergibt sich im digitalen Fall – und nur dort – eine Periodizität der verwendeten Abtastwerte der Exponentialfunktion mit N ; hier gilt beispielsweise

$$\cos \frac{(m+kN) \cdot 2\pi n}{N} = \cos \frac{m \cdot 2\pi n}{N}, \quad k \text{ ganzzahlig}, \quad (3.15f)$$

so dass der Index m beim Ansatz nach (3.15c-e) ebenso gut von 0 bis $N-1$ laufen kann. Diese letztere Darstellung wird in der Signalverarbeitung bevorzugt. Die Beziehung (3.15f) begegnet uns an anderer Stelle im Zusammenhang mit dem Abtasttheorem² wieder.

Hieraus ergibt sich nach Umrechnung und Umbenennung der spektralen Koeffizienten a_m die *diskrete Fouriertransformation*,

$$X(m) := \sum_{n=0}^{N-1} x(n) \exp \frac{-2\pi j n m}{N}; \quad (3.20)$$

$$x(n) := \frac{1}{N} \sum_{m=0}^{N-1} X(m) \exp \frac{2\pi j n m}{N}. \quad (3.21)$$

Hierbei stellt (3.20) die eigentliche diskrete Fouriertransformation dar; (3.21) definiert die *inverse diskrete Fouriertransformation* (IDFT), die das Spektrum $X(m)$ in das Eingangssignal $x(n)$ rückwandelt. Auf den Beweis von (3.21) soll hier verzichtet werden; die Probe wird gemacht, indem wir (3.21) in (3.20) einsetzen.

Unter einer *Transformation* verstehen wir hier die Substitution der unabhängigen Variablen der Zeitfunktion $x(n)$ – wobei der Index n für die Zeit steht – durch eine andere unabhängige Variable, die eine physikalisch von der Zeit verschiedene Größe, die Frequenz darstellt; in der digitalen Darstellung steht hierfür der Index m . Durch die diskrete Fouriertransformation wird also die finite Zeitfunktion $x(n)$, $n=0$ (1) $N-1$ in das zugehörige Spektrum $X(m)$, $m=0$ (1) $N-1$ umgewandelt und umgekehrt.

Entsprechend dem Ansatz (3.20, 3.21) ergeben sich die Koeffizienten des Spektrums in *komplexer* Form. Auch das Zeitsignal $x(n)$ kann – zumindest formal – komplexe Werte annehmen. Reale Signale sind aber stets reell. Ist $x(n)$ reell, so entspricht der Realteil des Spektrums dem geraden Anteil, der Imaginärteil dem ungeraden Anteil von $x(n)$:

$$x(n) = x_g(n) + x_u(n) \rightarrow X(m) = X_g(m) + jX_u(m). \quad (3.22)$$

Wir bezeichnen $X(m)$ im folgenden als das *komplexe Spektrum* oder auch einfach als das *Spektrum* von $x(n)$. Entsprechend (3.5) können wir auch ein *Betrags-* bzw. *Amplitudenspektrum* definieren:

$$|X(m)| := \sqrt{X_g^2(m) + X_u^2(m)}; \quad (3.23)$$

ebenso das *Phasenspektrum*, das aber in der Sprachsignalverarbeitung selten benötigt wird:

² Vgl. Kap. 1 dieses Skriptums. Genauer betrachtet und hergeleitet wird das Abtasttheorem in der Vorlesung *Signalverarbeitung 1*, Kapitel 1.

$$\varphi(m) := \arctan \frac{X_u(m)}{X_g(m)}. \quad (3.24)$$

Von Bedeutung ist neben dem Amplitudenspektrum auch das *Leistungsspektrum*, das wie folgt definiert ist:

$$P(m) := X_g^2(m) + X_u^2(m) = |X(m)|^2; \quad (3.25)$$

Insbesondere für die graphische Darstellung, wo es fast ausschließlich verwendet wird, aber auch für die Verarbeitung von Sprachsignalen ist das *logarithmierte* Leistungs- oder Amplitudenspektrum von Bedeutung; es wird meist in Dezibel (dB) angegeben und entspricht den für akustische Messung üblichen *Pegelwerten* (vgl. Abschnitt 2.2.1),

$$L(m) := 20 \lg |X(m)| = 10 \lg P(m). \quad (3.26)$$

Von allen diesen Darstellungen erlaubt nur das komplexe Spektrum die Wiedergewinnung der Zeitfunktion aus der spektralen Darstellung via (3.21).

3.1.5 Einige besondere Zeitfunktionen und ihre Spektren

Für die weiteren Betrachtungen von Interesse sind einige besondere Zeitfunktionen, obwohl sie mit Sprachsignalen direkt nichts zu tun haben. Es handelt sich um die Rechteckfunktion, den Sägezahn, die Dreiecksfunktion und die Impulsfunktion.

Bild 3.6 zeigt Signalverlauf und Spektren (erste 50 Teilschwingungen) für diese vier Zeitfunktionen. Durch Anwendung von (3.11a,b) lässt sich im analogen Fall für die beiden Zeitfunktionen, die Sprünge enthalten (d.h., Sägezahn und Rechteck) leicht zeigen, dass die Amplituden der Teilschwingungen nach höheren Frequenzen hin mit $1/m$ abfallen, wobei das Spektrum der Rechteckschwingung nur ungerade Teilschwingungen enthält. Das Spektrum der Dreieckschwingung fällt mit $1/m^2$ ab. (Wegen der Abtastung gelten diese Zusammenhänge für digitale Signale nur näherungsweise; die Abweichungen vergrößern sich mit steigender Frequenz.) Das Spektrum der Impulsfunktion (Impulsbreite 1 Abtastwert) ist für alle Teilschwingungen konstant.

Für Zeitfunktionen, die Impulse, Sprünge oder Ecken enthalten, konvergiert die Fourierreihe verhältnismäßig schlecht. Hierzu zeigt Bild 3.7 Signalverlauf und Spektren einer Annäherung dieser vier Zeitfunktionen durch die Teilschwingungen für $m \leq 40$.

3.2 Nichtharmonische Spektralanalyse

Die Orthogonalitätsbeziehung (3.9, 3.10) für die trigonometrischen Funktionen gilt nur für *ungedämpfte* Schwingungen und den Fall, dass sowohl für das Eingangssignal $x(n)$ als auch für die Testfunktionen die Integration bzw. Summierung über eine *ganze Anzahl von Perioden* erfolgt.

In der Sprachsignalverarbeitung haben wir es jedoch zum einen nur mit annähernd periodischen Signalen zu tun, zum anderen sind die Schwingungen innerhalb einer Grundperiode in der Regel gedämpft. Zudem ist die Grundperiode selbst oftmals zum Zeitpunkt der Spektralanalyse eines Sprachsignals noch gar nicht ermittelt. Für viele Aufgaben der Sprachsignalverarbeitung ist daher eine Spektralanalyse ohne Kenntnis der Grundperiode des Sprachsignals erforderlich. Eine solche Analyse ist dann sicher bezüglich des Sprachsignals *nichtharmonisch*. Um die harmonische Analyse bzw. die diskrete Fouriertransformation trotzdem anwenden zu können, müssen wir uns überlegen, wie sie sich verhält, wenn gedämpfte Schwingungen auftreten oder eine nicht ganze Zahl von Perioden analysiert wird. Hierbei sei zuerst auf den Fall der gedämpften Schwingung eingegangen.

Die mathematische Herleitung ist im Einzelfall kompliziert und erscheint hier zudem unnötig, da sich aus der Phänomenologie gut erschließen lässt, welche Artefakte durch diese Verallgemeinerung auftreten und wie ihre Auswirkungen abgeschätzt bzw. reduziert werden können.

3.2.1 Spektraltransformation gedämpfter Schwingungen

Die harmonische Analyse approximiert jedes periodische Signal mit *ungedämpften* Schwingungen und erzwingt die Periodizität des Signals mit der Periodendauer gleich dem Analyseintervall t_0 bzw. N . Weist die interne Signalstruktur eine Periodizität auf, die kein ganzzahliger Teiler von N ist, oder treten gedämpfte Schwingungen auf, so äußert sich dies im Spektrum daran, dass sich jetzt zusätzliche Spektrallinien von Null

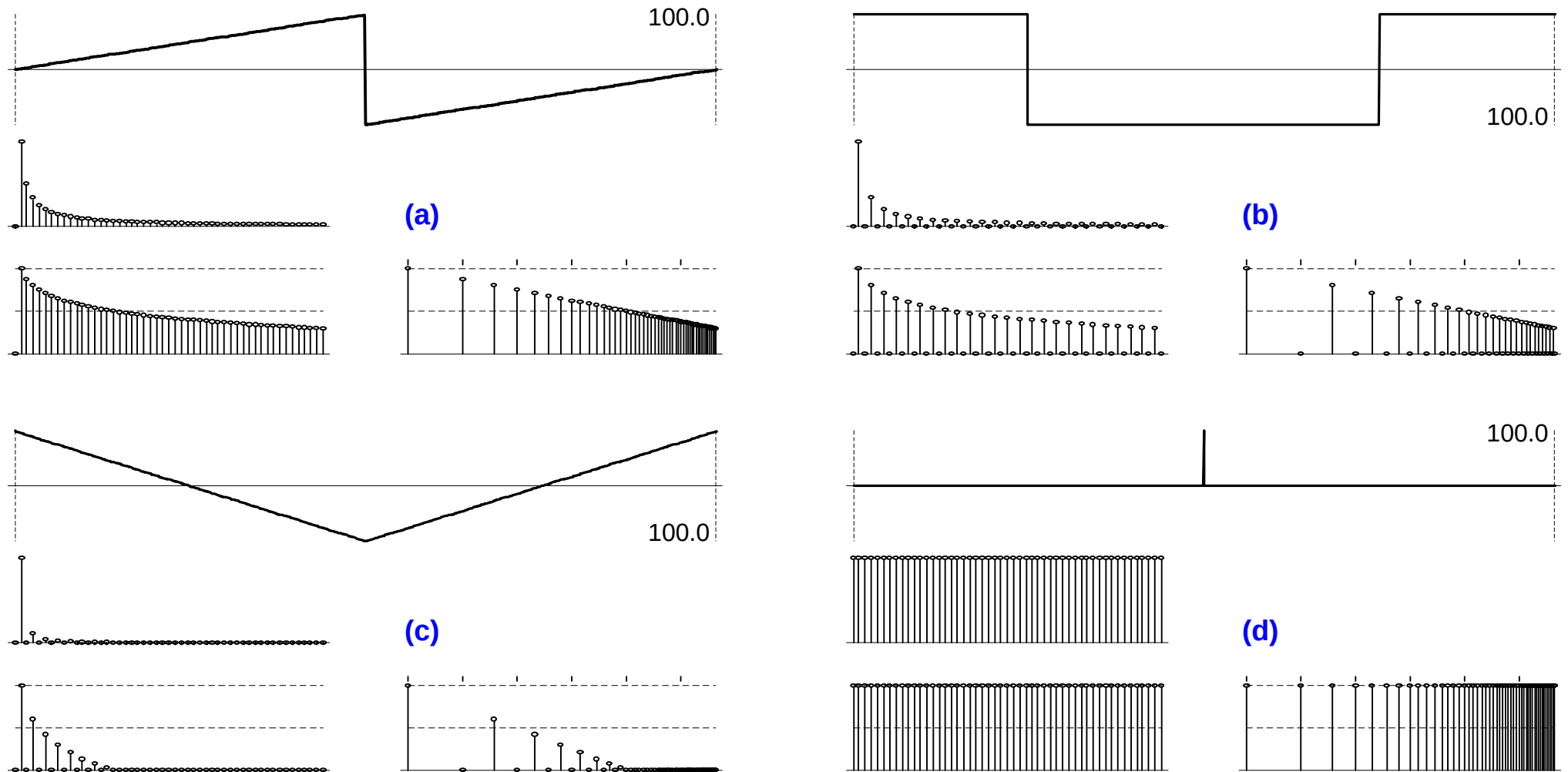


Bild 3.6a-d. Sägezahn (a), Rechteck [mit Tastverhältnis 1:1, d.h., gleicher Länge der positiven und negativen Signalabschnitte] (b), Dreieck [mit gleichem Betrag der Steigung für steigende und fallende Flanke] (c) und Impuls [Breite: 1 Abtastwert; aus Gründen der Darstellung in der Mitte des Bildes] (d). Darstellung in jedem Teilbild: (oben) Signal; 1 Periode, Amplitude 100, Zahl der Abtastwerte $N=1260$; (Mitte) Amplitudenspektrum; (links unten) logarithmiertes Amplitudenspektrum (12 dB/Skalenteil), für beide Spektren gilt $m=0$ (1) 50; (rechts unten) logarithmiertes Amplitudenspektrum mit logarithmischer Darstellung der Frequenzachse, $m=1$ (1) 50; die senkrechten Striche über dem Spektrum geben an, an welchen Stellen sich die Frequenz jeweils verdoppelt (Oktavabstand). Für die Impulsfunktion bleibt das Spektrum mit steigender Frequenz konstant; für Sägezahn und Rechteck, die beide Sprünge enthalten, fällt es nach höheren Frequenzen hin mit 6 dB/Oktave (entspricht $1/m$), für die Dreiecksfunktion mit 12 dB/Oktave (entspricht $1/m^2$). Das Spektrum enthält mehr als 50 Teilschwingungen, doch nur die ersten 50 sind dargestellt

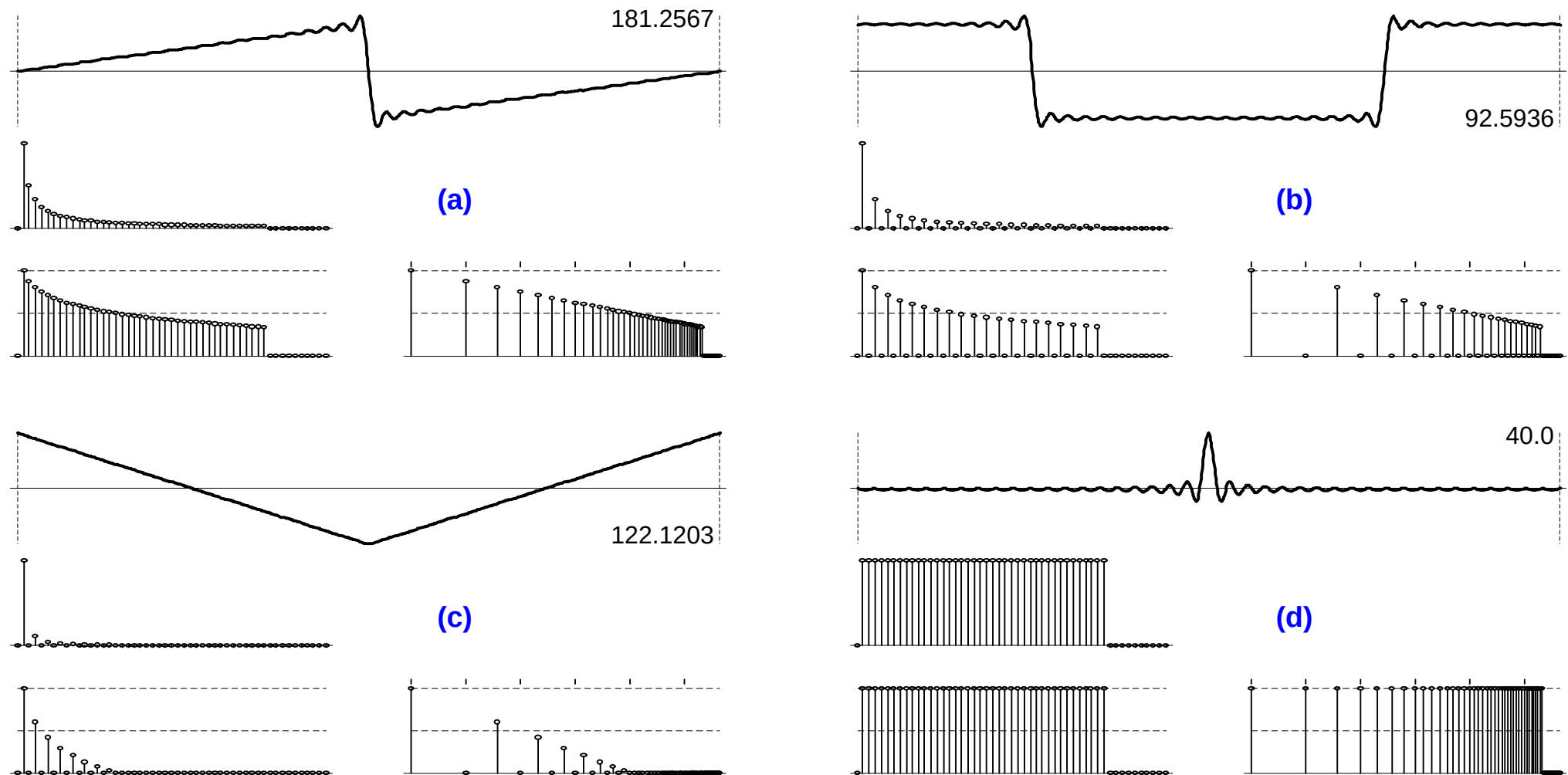


Bild 3.7a-d. Sägezahn, Rechteck, Dreieck und Impuls, jeweils approximiert durch die ersten 40 Teilschwingungen in richtiger Amplitude und Phasenlage. Darstellung: wie Bild 3.6. Die Amplitude der ersten Teilschwingung beträgt 1 für den Impuls, 100 für die übrigen Signale. Aus Gründen der Darstellung wurde die Impulsfunktion so approximiert, dass der Impuls in der Bildmitte entsteht

verschieden ergeben. Die erzwungene Periodizität mit N bewirkt zudem, dass am Rand des Intervalls Sprünge bzw. Ecken im Signal auftreten. Bild 3.8 zeigt einige Beispiele. Zu beachten sind insbesondere die Teilbilder 3.8b und 3.8c. In Bild 3.8b wird eine gedämpfte Kosinusschwingung transformiert. Bei hohen Frequenzen ergibt sich ein spektrales Verhalten, das

dem der Rechteck- oder Sägezahnschwingung gleicht. Der zugehörige Sprung tritt an der Grenze des Analyseintervalls auf; durch die Dämpfung setzt sich die Zeitfunktion über die Intervallgrenze nicht stetig fort. Entsprechendes gilt für Bild 3.8c, wo an der Intervallgrenze auf diese Weise eine Ecke auftritt.

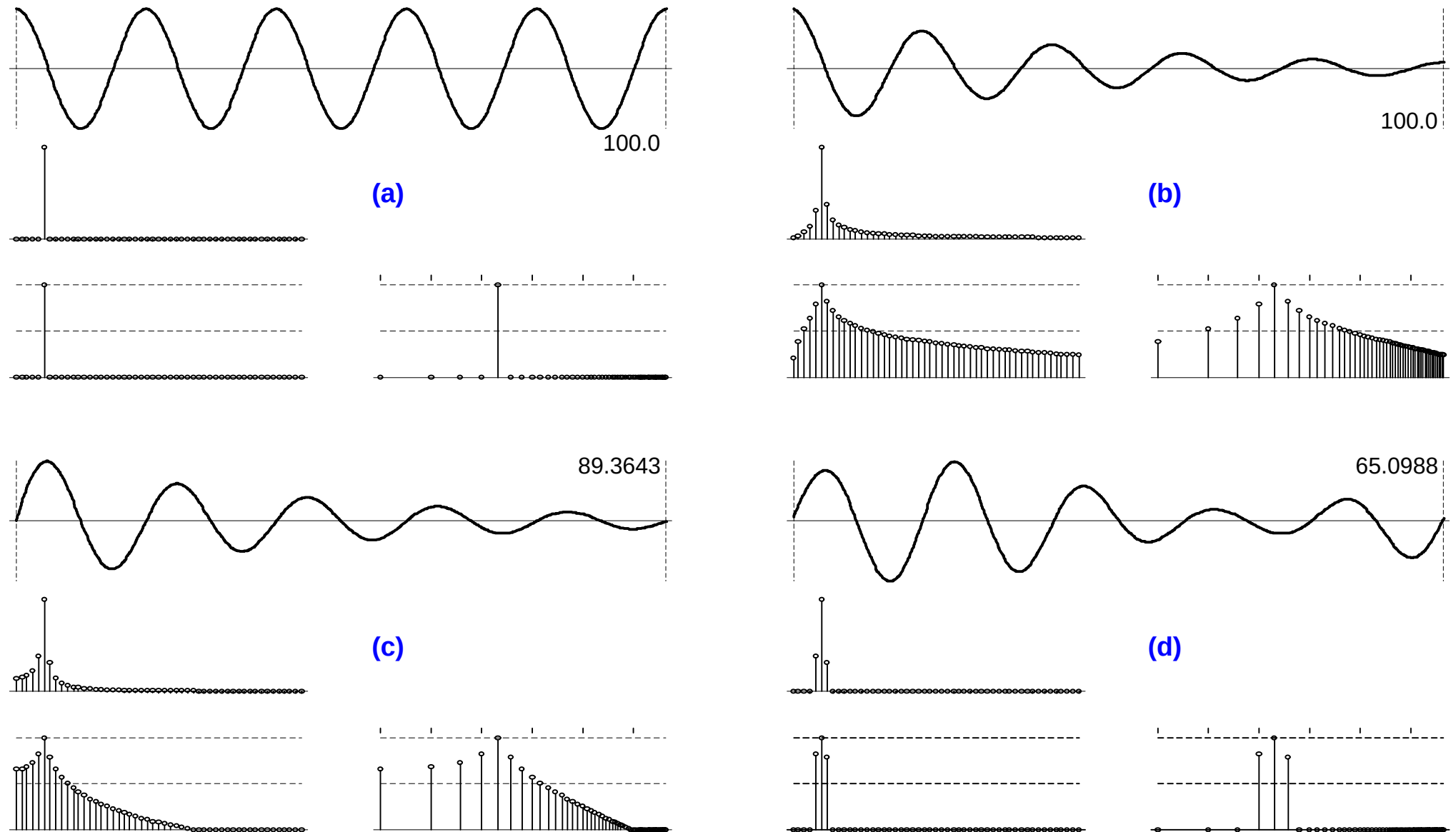


Bild 3.8a-d. Spektralanalyse gedämpfter Schwingungen. Die Darstellung stimmt mit der von Bild 3.7 überein. (a) Ungedämpfte Schwingung ($k=5$) zum Vergleich; (b) gedämpfte Kosinusschwingung; (c) gedämpfte Sinusschwingung – beide Schwingungen sind am Ende des Analyseintervalls auf 10% ihrer ursprünglichen Amplitude abgeklungen (d) drei Harmonische ($m=4, 5, 6$) mit Amplitude und Phase wie in (c); man sieht deutlich, dass die drei Teilschwingungen eine gedämpfte Schwingung zu synthetisieren beginnen

3.2.2 Nichtharmonische Spektralanalyse

Umfasst die Schwingung $x(n)$ – der Einfachheit halber in diesem Abschnitt wieder als rein sinusförmig angenommen – innerhalb des Analyseintervalls $n=0$ (1) $N-1$ keine ganze Anzahl von Perioden, so treten am Rand des Analyseintervalls zwangsläufig Sprungstellen, zumindest aber Ecken auf. Das Verhalten nach hohen Frequenzen hin wird sich also nach der Art der Unstetigkeit am Rand des Intervalls richten.

Die Schwingung selbst wird im wesentlichen durch die Spektrallinien in der Nachbarschaft des nun nicht mehr ganzzahligen Wertes k charakterisiert. Die Spektralanalyse mit Hilfe der DFT erzwingt nach wie vor die Periodizität mit N ; deshalb kann das Signal $x(n)$ nur durch Schwingungen mit ganzzahligen Werten von m dargestellt werden. Beispielsweise ergeben sich für $k=5,5$ (Bild 3.9b,c) die Spektralwerte für $m=5$ und $m=6$ als die energiereichsten.

Der Gesamtverlauf des Spektrums ist dem einer gedämpften Schwingung in gewisser Weise ähnlich. Es fällt auf, dass ein Großteil des Bildes durch die Unstetigkeit an den Intervallgrenzen, also durch ein Artefakt dominiert wird.

3.2.3 Gewichtung des Eingangssignals

Um die Auswirkung der Unstetigkeit eines beliebigen Eingangssignals an den Intervallgrenzen zu mildern, greift man in der Regel nach dem Mittel der *Gewichtung* des Signals vor Anwendung der DFT. Hier wird das Signal über die gesamte Länge des Analyseintervalls mit einer *Gewichtungs-* oder *Fensterfunktion* multipliziert, die in Intervallmitte den Wert 1 annimmt und nach den Intervallgrenzen hin sanft gegen Null oder einen kleinen, von Null verschiedenen Wert abfällt. Die Form des Fensters ist

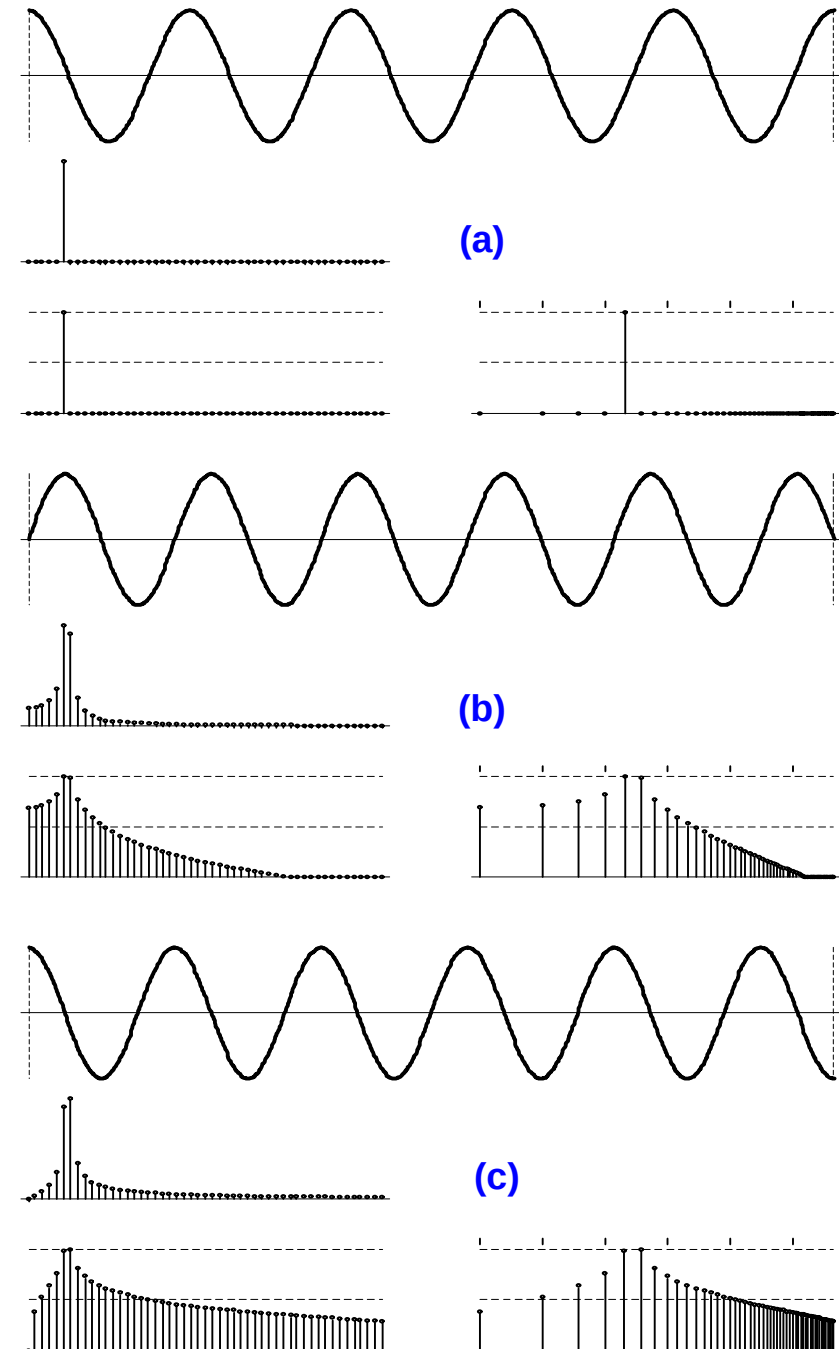


Bild 3.9a-c (nebenstehend). Nichtharmonische Spektralanalyse. Die Darstellung stimmt mit der von Bild 3.7 überein. (a) Schwingung mit ganzer Zahl von Perioden im Analyseintervall ($k=5$) zum Vergleich; (b) Kosinusschwingung mit 5,5 Perioden; (c) Sinusschwingung mit 5,5 Perioden. In den Bildern (b) und (c) ist der Einfluss des Sprunges bzw. der Ecke am Rand des Analyseintervalls deutlich zu sehen

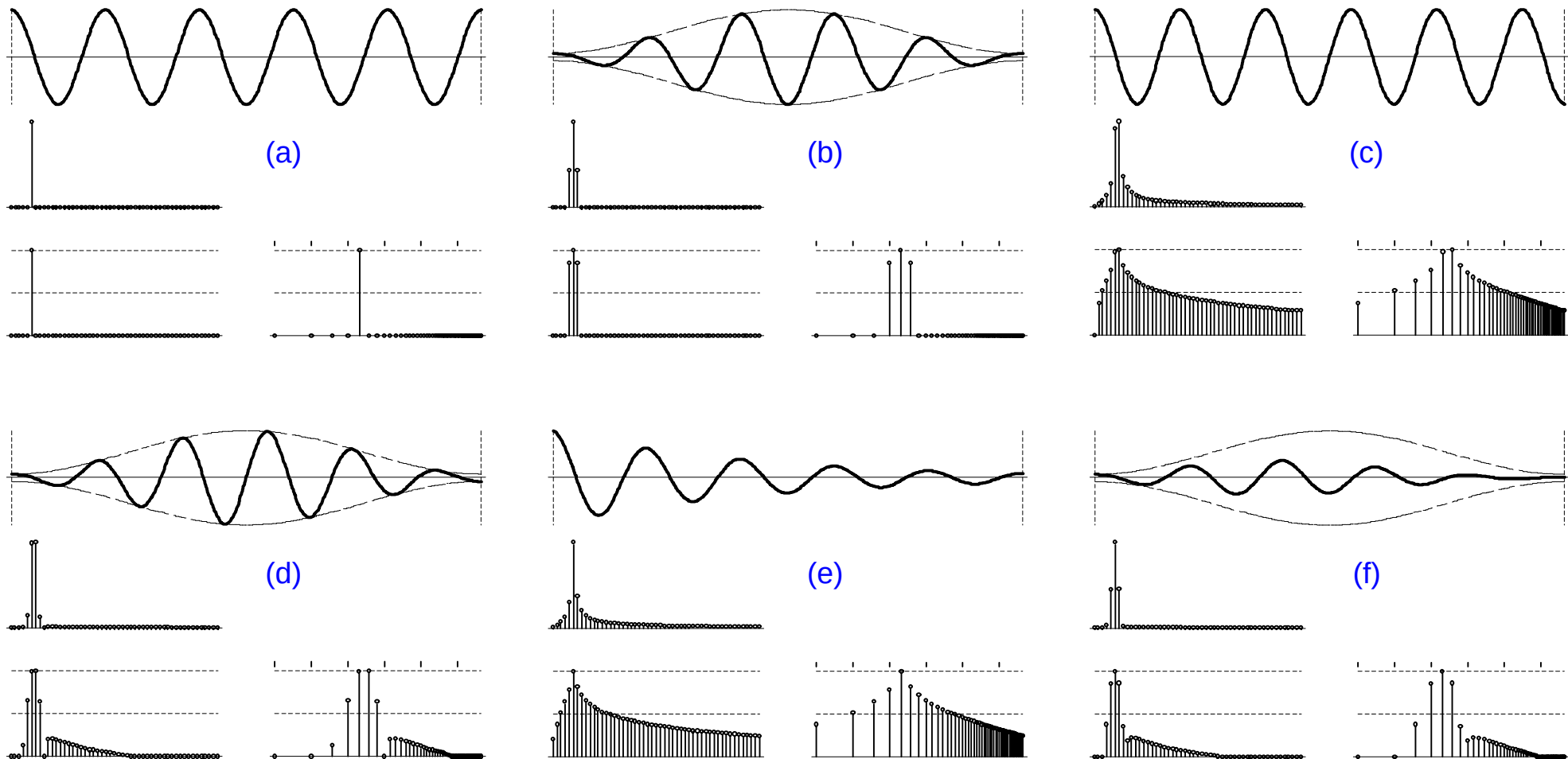


Bild 3.10a-f. Spektralanalyse von Signalen mit und ohne Gewichtung. (a, c, e) Ohne Gewichtung, (b, d, f) Gewichtung mit Hammingfenster. (a, b) Kosinusschwingung mit 5 Perioden; (c, d) Kosinusschwingung mit 5,5 Perioden; (e, f) Gedämpfte Schwingung (wie in Bild 3.8b)

mehr oder minder willkürlich. In der Sprachsignalverarbeitung sehr verbreitet ist das *Hamming-Fenster*; dieses ist definiert durch

$$\begin{aligned}
 w_H(k) &= 0,08 + 0,92 \cos^2\left(\frac{\pi(k-K)}{N}\right) \\
 &= 0,54 + 0,46 \cos \frac{2\pi(k-K)}{N} ; \quad K = \frac{N}{2} .
 \end{aligned}
 \tag{3.27}$$

Bild 3.10 zeigt einige Beispiele. Wie deutlich zu sehen, wird in der Spektralfunktion das Artefakt – die Auswirkung von Sprüngen und Ecken an den Intervallgrenzen – durch ein anderes Artefakt ersetzt, nämlich die Auswirkung der Fensterfunktion. Diese bewirkt eine Art künstlicher Dämpfung des Signals, die sich im Spektrum – gleichgültig, ob die ursprüngliche Schwingung gedämpft ist oder nicht – als Verbreiterung des Hauptmaximums zeigt (Bild 3.10b, f). Wesentlich geringer

sind jedoch die Unterschiede zwischen den Spektren eines Signals mit ganzzahligem und nicht ganzzahligem Wert von k (ohne Gewichtung: Bild 3.10a, c; mit Gewichtung: Bild 3.10b, d). Dieser Effekt ist durchaus erwünscht. Soll mit den Mitteln der harmonischen Analyse, also der DFT, nichtharmonische Analyse eines Sprachsignals betrieben werden, so ist es nicht wünschenswert, wenn die Form des Spektrums stark davon abhängt, ob eine einzelne Teilschwingung innerhalb des Analyseintervalls zufällig mit einer ganzen Zahl von Perioden vertreten ist oder nicht. Ausgesprochen ungünstig wirkt sich die Gewichtung auf die spektrale Darstellung gedämpfter Schwingungen aus; hier entstehen Verzerrungen derart, dass sich aus dem Spektrum auf die Dämpfung des Signals, sofern diese gering ist, kaum mehr Rückschlüsse ziehen lassen.

3.3 Kurzzeitspektren von Sprachsignalen

3.3.1 Definition, spektrale Auflösung, Transformationsintervall

Das Verfahren der diskreten Fouriertransformation [und die zugehörige algorithmische Variante, die *schnelle Fouriertransformation* (FFT), deren Ausführung gegenüber der direkten Anwendung von (3.20, 3.21) um ein Vielfaches schneller ist] bildet die Grundlage für die digitale Kurzzeitspektralanalyse von Sprachsignalen.

Die diskrete Fouriertransformation ist für zeitbegrenzte Signale definiert. Sie bietet sich daher direkt für die Kurzzeitanalyse an. Dort entspricht sie dem *periodischen* Ansatz, da sie – wie bereits gezeigt – die Zeitfunktion von Haus aus als außerhalb des Transformationsintervalls periodisch fortgesetzt betrachtet. Die Zahl N der transformierten Signalwerte ("Länge" der DFT bzw. *Transformationsintervall*) ist größer oder gleich der durch die Kurzzeitanalyse vorgegebenen Fensterlänge K . Ist sie größer, so muss das zu transformierende Signal $x_T(n)$ entsprechend mit Nullwerten aufgefüllt werden:

$$x_T(k) := \begin{cases} x_K(k, q) & k = 0(1)K-1; \text{ entspricht } n = q-K+1(1)q \\ 0 & k = K(1)N-1. \end{cases} \quad (3.28)$$

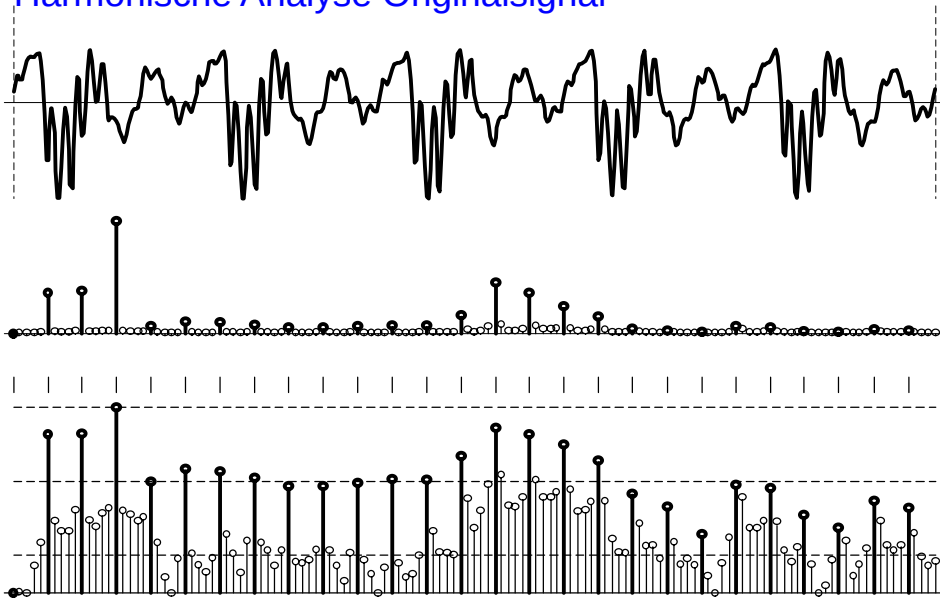
Die spektrale Auflösung, d.h., der Abstand benachbarter Abtastwerte im Spektrum ergibt sich aus der Abtastfrequenz F des Signals sowie der Länge N des Transformationsintervalls:

$$\Delta f = F / N. \quad (3.29)$$

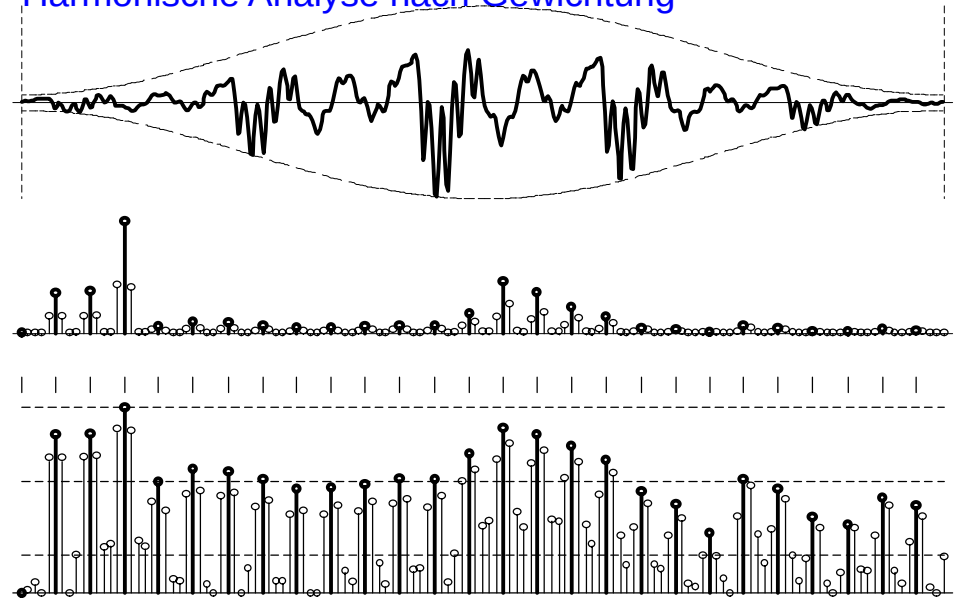
Hierbei ist Δf der Abstand zweier Spektralwerte, F die Abtastfrequenz, beides in Hz. Die spektrale Auflösung ist damit durch die Länge N des Transformationsintervalls gegeben. Die im Spektrum enthaltene

Bild 3.11. (*nächste Seite*) Harmonische und nichtharmonische Analyse von Sprachsignalen, Beispiel 1. Signal: 5 Perioden des Vokals [ε], Sprecher WGH (männlich); Abtastfrequenz 16 kHz; 536 Messwerte; Grundfrequenz 149,3 Hz. Alle Darstellungen zeigen oben das Signal (in den beiden oberen Teilbildern gegenüber den beiden unteren zeitlich gespreizt) im gesamten Transformationsintervall, das in den beiden unteren Teilbildern durch Nullwerte von 536 auf 2048 Werte verlängert ist, in der Mitte das Amplitudenspektrum und unten das logarithmierte Amplitudenspektrum (20 dB/Skalenteil). Die senkrechten Linien zwischen den beiden spektralen Darstellungen markieren die Vielfachen der Grundfrequenz des Signals. Das Spektrum umfasst einen Bereich von 0 bis 4 kHz. (Links oben) Harmonische Analyse Originalsignal. Bei dieser strengen harmonischen Analyse mit $N=136$ ergibt sich ein ausgeprägtes Maximum bei jedem 5. spektralen Abtastwert. Dass die Werte dazwischen nicht Null werden, liegt daran, dass das Signal nur quasiperiodisch ist. Diese Zwischenwerte repräsentieren also den nichtharmonischen Anteil des Signals (also die Perturbationen); in diesem Bild beträgt der nichtharmonische Anteil an der Gesamtenergie des Signals 1,9%. (Rechts oben) Gleiche Analyse nach Gewichtung mit einem Hammingfenster. Die Maxima bei den Vielfachen der Grundfrequenz sind nunmehr verbreitert. Berechnet man auch hier einen "nichtharmonischen Anteil", so beträgt dieser nunmehr rund 27% der Gesamtenergie. (Links unten) Nichtharmonische Analyse ohne Gewichtung. In der spektralen Darstellung sind die Maxima bei den Vielfachen von F_0 deutlich zu sehen, werden aber teilweise durch zusätzliche Nebenmaxima verwischt, die durch den Sprung an den Rändern des Signalintervalls entstehen. (Rechts unten) Gleiche Analyse nach Gewichtung mit einem Hammingfenster. Die Maxima bei Vielfachen von F_0 sind etwas breiter, aber insgesamt gesehen besser ausgeprägt.

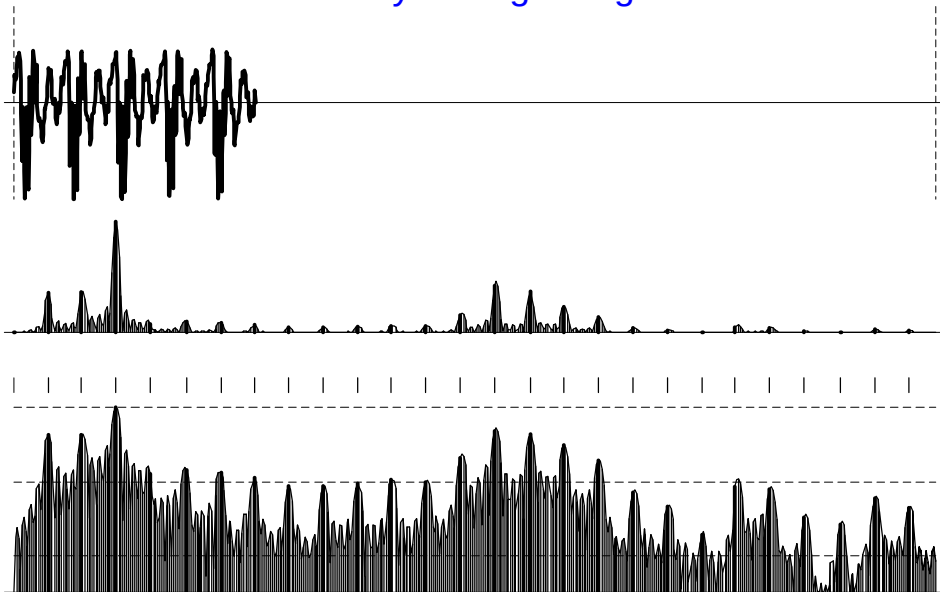
Harmonische Analyse Originalsignal



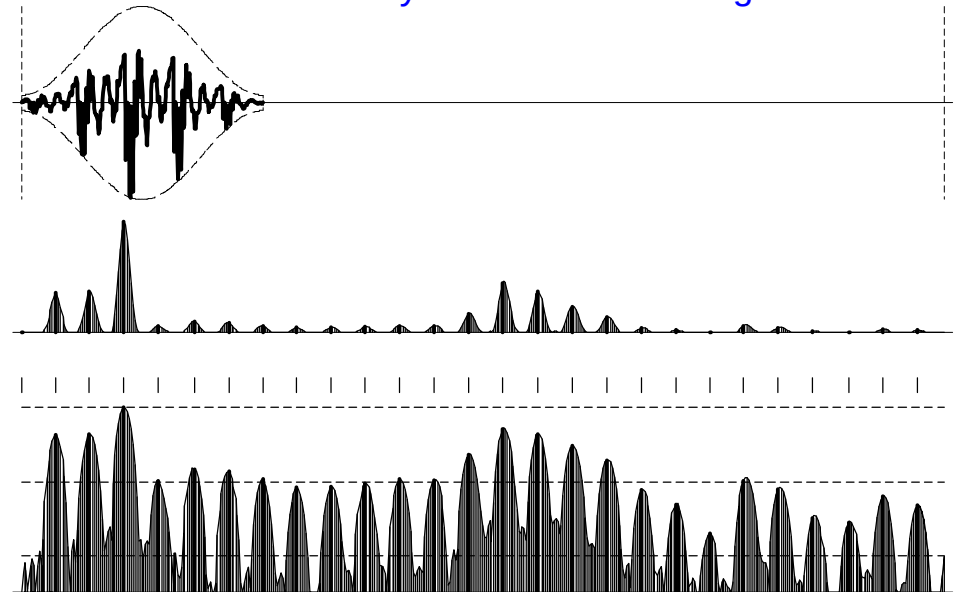
Harmonische Analyse nach Gewichtung



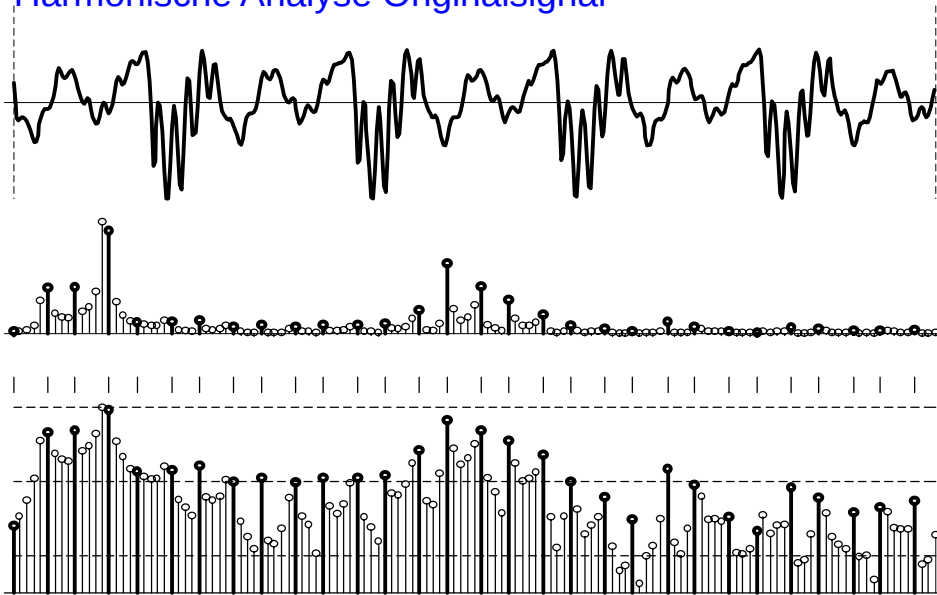
Nichtharmonische Analyse Originalsignal



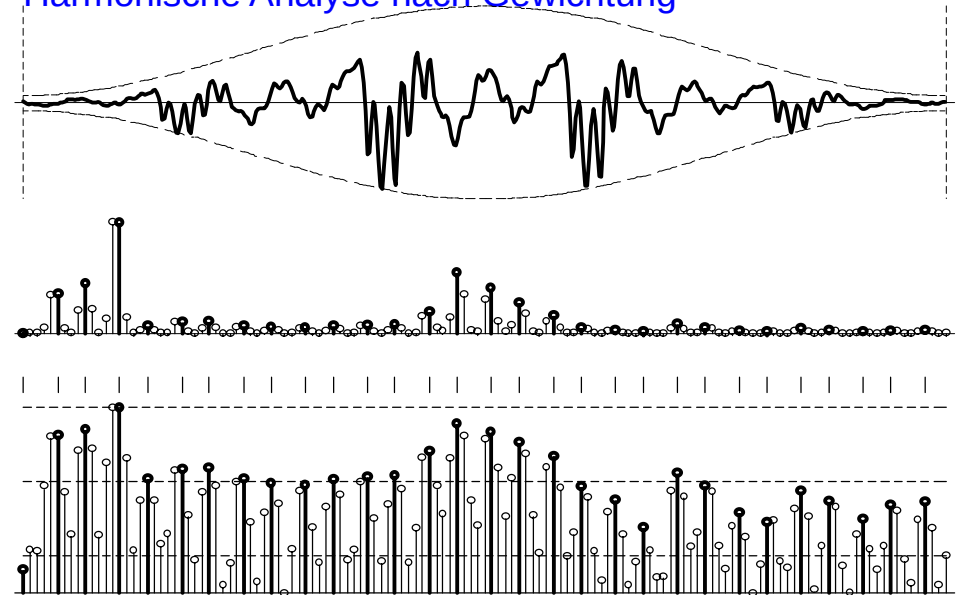
Nichtharmonische Analyse nach Gewichtung



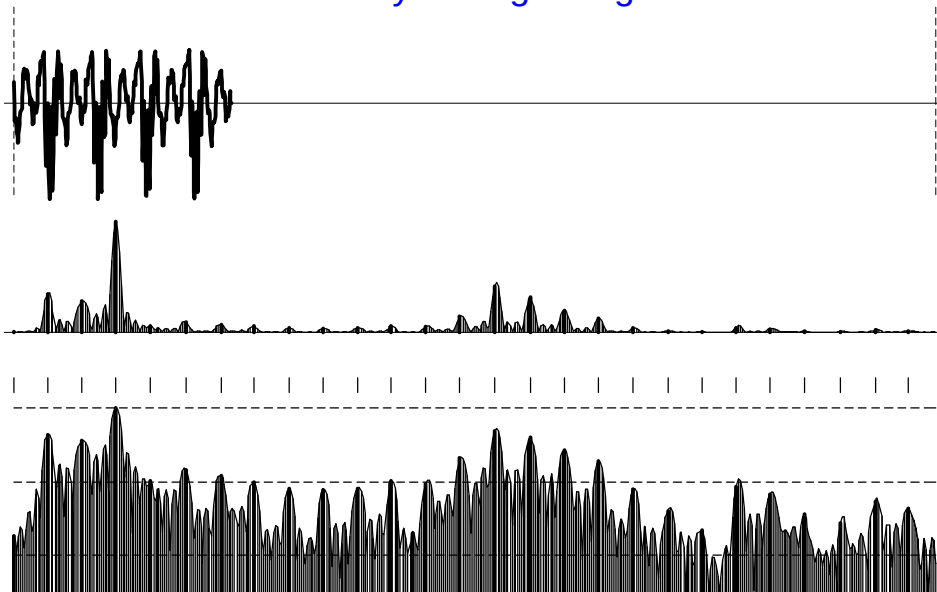
Harmonische Analyse Originalsignal



Harmonische Analyse nach Gewichtung



Nichtharmonische Analyse Originalsignal



Nichtharmonische Analyse nach Gewichtung

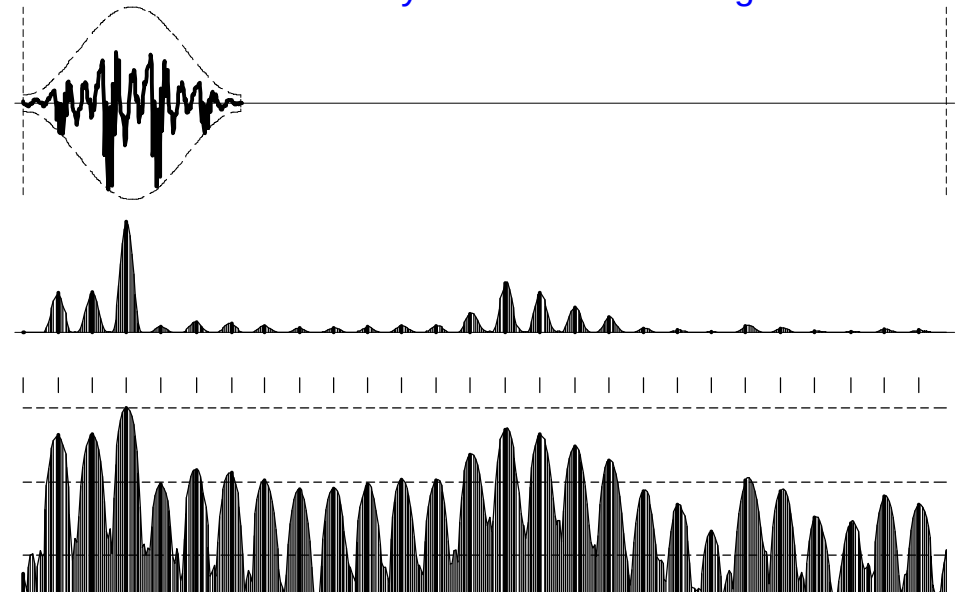


Bild 3.12. (vorige Seite) Harmonische und nichtharmonische Analyse von Sprachsignalen, Beispiel 1. Signal: 4,5 Perioden des Vokals [ε], Sprecher WGH (männlich); Abtastfrequenz 16 kHz; 536 Messwerte; Grundfrequenz 149,3 Hz. Darstellung wie in Bild 3.11. (Links oben) "Harmonische Analyse" Originalsignal. Hier ergibt sich gegenüber Bild 3.11 der markanteste Unterschied. Eine harmonische Struktur ist aus diesem Bild nicht mehr zu erkennen. (Rechts oben) Gleiche Analyse nach Gewichtung mit einem Hammingfenster. Hier treten wieder Maxima auf, die die Vielfachen der Grundfrequenz markieren. (Links unten) Nichtharmonische Analyse ohne Gewichtung; (rechts unten) gleiche Analyse nach Gewichtung mit einem Hammingfenster. Die Unterschiede in der Darstellung gegenüber dem gleichen Verfahren (mit 5 vollständigen Perioden) in Bild 3.11 sind hier gering geworden.

Information beschreibt allerdings das Zeitsignal bereits vollständig, wenn $N=K$ wird. Für $N>K$ ist das Spektrum lediglich überabgetastet.

Vor allem bezüglich der Darstellung bringt die Überabtastung des Spektrums Vorteile. Wie bereits aus Bild 3.10 ersichtlich, werden durch die Gewichtung des Zeitsignals Unterschiede zwischen harmonischer Analyse (d.h. Transformation eines Intervalls, das eine ganze Zahl von Grundperioden umfasst) und nichtharmonischer Spektralanalyse (alle anderen Fälle) weitgehend verwischt. Dieser Effekt verstärkt sich noch bei Überabtastung des Spektrums. Weiterhin kann durch die Überabta- stung die spektrale Auflösung unabhängig von der Länge K des Kurzzeit- analyseintervalls festgelegt werden. Da K sich im wesentlichen nach den Signaleigenschaften richtet (siehe Abschnitt 2.1), ist eine Koppelung der spektralen Auflösung an diesen Wert nicht wünschenswert.

3.3.2 Harmonische und nichtharmonische Spektralanalyse bei Sprachsignalen

Bild 3.11 und Bild 3.12 (vorangehende Seiten) zeigen ein Beispiel für die verschiedenen Anwendungsmöglichkeiten der Spektralanalyse bei Sprachsignalen. Bei nichtharmonischer Analyse (die Termini *harmonisch* bzw. *nichtharmonisch* beziehen sich, wohlgemerkt, auf die Grundperiode des Sprachsignals, nicht auf das Verfahren der Spektralanalyse als solches) spielt es nur eine geringe Rolle, ob sich eine ganze Zahl von Grundperioden im Messintervall befindet oder nicht; Gewichtung mit beispielsweise einem Hamming-Fenster ist in jedem Fall vorteilhaft. Um eine nichtharmonische Spektralanalyse durchzuführen, ist es also nicht notwendig, vorher die Grundperiode des Sprachsignals zu bestimmen.

Dem gegenüber steht die harmonische oder *grundperiodensynchrone* Analyse von Sprachsignalen. Sie bietet von allen Verfahren die genaue-

sten Ergebnisse, wenn das Transformationsintervall N genau gleich der Grundperiode T_0 (oder einem Vielfachen davon) ist. Bereits geringe Abweichungen des Intervalls N von T_0 führen zu groben Verzerrungen. Hier muss die Grundperiode also genau bekannt sein.

Beide Verfahren haben nebeneinander ihre Berechtigung: die nichtharmonische Analyse (bei stimmlosen Signalen kommt ohnehin nur sie in Frage) als allgemeines Verfahren zur raschen Untersuchung beliebiger Signale, die grundperiodensynchrone harmonische Analyse für Feinuntersuchungen, die sich auf das Intervall einer Grundperiode beschränken und Verzerrungen durch Gewichtung o.ä. nicht zulassen.

3.4 Das lineare Modell der Spracherzeugung in spektraler Darstellung

3.4.1 Systemtheoretische Vorbetrachtung: Beschreibung eines Systems im Spektralbereich

Gegeben sei, wie in Abschnitt 2.1 definiert, ein System mit der Rechen- vorschrift S , das ein Eingangssignal $u(n)$ in ein Ausgangssignal $y(n)$ über- führt,

$$y(n) = S\{u(k), k = \dots (1) n\}; \quad (2.7)$$

in dieser Form verbindet S das Ein- und das Ausgangssignal; genauer ge- sagt, wird für jeden Abtastzeitpunkt n der Abtastwert $y(n)$ des Ausgangs- signals durch eine Verknüpfung der Gegenwart [repräsentiert durch den Abtastwert $u(n)$] und der gesamten Vergangenheit des Eingangssignals mit Hilfe der Rechenvorschrift S berechnet. In dieser Form haben wir die Kausalität des Systems vorausgesetzt. Im folgenden wollen wir zu- sätzlich voraussetzen, dass das System linear sei.

Was mit Hilfe von (2.7) für die Signale $u(n)$ und $y(n)$ im Zeitbereich formuliert ist, lässt sich ebenso für die zugehörigen Spektren $U(m)$ und $Y(m)$ formulieren. Wir definieren hierzu

$$H(m) := \frac{Y(m)}{U(m)} \quad (3.30a)$$

als *Systemfunktion* im Spektralbereich (auch *Übertragungsfunktion* genannt) und erhalten damit

$$Y(m) = U(m) \cdot H(m); \quad (3.30b)$$

es lässt sich zeigen,³ dass damit ein lineares System vollständig beschrieben werden kann. Konkret bedeutet dies, dass ein lineares, durch (3.30) beschriebenes System in der Lage ist, die Amplitude und Phase jeder sinusoidalen Teilschwingung zu ändern, die das System durchläuft, *nicht aber deren Frequenz*. Der Übergang in den Spektralbereich bringt bei linearen Systemen damit eine entscheidende Vereinfachung der Systembeschreibung mit sich, da jeder Wert $Y(m)$ des Ausgangsspektrums nur noch von einem Wert $U(m)$ des Eingangsspektrums abhängt.

3.4.2 Beschreibung des linearen Modells der Spracherzeugung im Spektralbereich

Gemäß Abschnitt 2.4.1 wird das lineare Modell der Spracherzeugung für stimmhafte Sprachsignale (auf diese wollen wir uns hier beschränken) durch die 5 Komponenten Periodizität, Klangfarbe der Stimme, Amplitude, Vokaltrakt und Abstrahlung gebildet. Weist man diesen Komponenten entsprechend (3.30) Übertragungsfunktionen zu, die wir mit $P(m)$ für die Periodizität, $G(m)$ für die Klangfarbe, $V(m)$ für den Vokaltrakt und $R(m)$ für die Abstrahlung bezeichnen wollen, so ist das gesamte Modell (Bild 3.13) zu beschreiben als

$$\begin{aligned} S(m) &= P(m) \cdot G(m) \cdot A \cdot V(m) \cdot R(m) \\ &= P(m) G(m) V(m) R(m) A, \end{aligned} \quad (3.31)$$

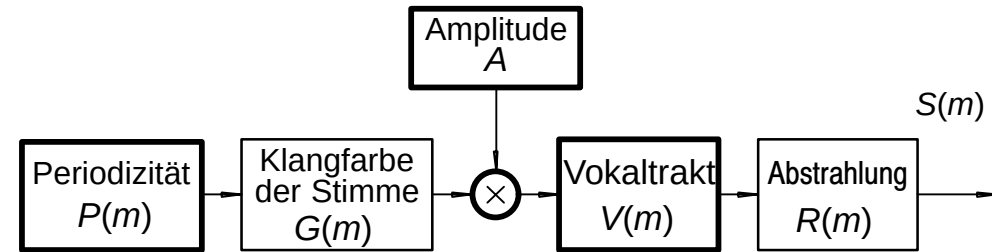


Bild 3.13. Das lineare Modell der Spracherzeugung für stimmhafte Signale in spektraler Darstellung

wobei $S(m)$ das Spektrum des Sprachsignals darstellt; A repräsentiert den nicht von der Frequenz abhängigen Faktor Amplitude.

Literatur

- Bronstein, I. N. / Semendjajew, K. A. (²⁵1991): *Taschenbuch der Mathematik* (Teubner, Stuttgart)
- Flanagan, James L. (²1972, ¹1965) *Speech analysis, synthesis, and perception* (Springer, Berlin; 2. Aufl.)
- Harris, Frederic J. (1978): "On the use of windows for harmonic analysis with the discrete Fourier transform." *Proc. IEEE* 66, 51...63
- Rabiner, Lawrence R. / Schafer, Ronald W. (1978): *Digital processing of speech signals*. Prentice-Hall Signal Proc. Series, ed. by A.V.Oppenheim (Prentice Hall, Englewood Cliffs, NJ)
- Rosen, Stuart / Howell, Peter (1991): *Signals and systems for speech and hearing* (Academic Press, London)
- Vary, Peter / Heute, Ulrich / Hess, Wolfgang (1998): *Digitale Sprachsignalverarbeitung* (Teubner, Stuttgart)

³ Siehe Vorlesung *Signalverarbeitung 1*, Kapitel 1. Dort wird das Übertragungsverhalten linearer digitaler Systeme allgemein beschrieben und eingehend diskutiert.

Grundlagen der Sprachsignalverarbeitung

4. Vom Spektrum zum Spektrogramm

Eine der wesentlichen Aufgaben der Sprachanalyse ist der Rückschluss aus dem Sprachsignal $s(n)$ auf die momentanen Eigenschaften des Vokaltraktes. Soll dies im Rahmen der Spektralanalyse durchgeführt werden, so ist es notwendig, die beiden Komponenten Anregung und Signalformung (Vokaltrakt) in der spektralen Darstellung zu trennen. Im einfachsten Fall wird dies durch eine Glättung des Spektrums erreicht (Abschnitt 4.1). Als besonders günstig hat sich hierzu das Cepstrum-Verfahren erwiesen (Abschnitt 4.2), das über diesen begrenzten Anwendungsfall hinaus allgemein große Bedeutung besitzt.

Wie keine andere Darstellung ist die Zeit-Frequenz-Darstellung als Visualisierungsmittel zur Sichtbarmachung der Eigenschaften von Sprachsignalen geeignet. Gebräuchliche Darstellungsformen sowie ihre Eigenschaften sind in Abschnitt 4.3 beschrieben.

4.1 Glättung von Kurzzeitspektren

Der Einfluss der periodischen Anregung $[P(m)]$ äußert sich durch die spektralen Maxima bei Vielfachen der Grundfrequenz F_0 , also durch multiplikative Überlagerung einer Welligkeit in Form äquidistanter Maxima im Spektrum, deren Distanz der Grundfrequenz F_0 entspricht. Der Einfluss des Vokaltraktes $[V(m)]$ zeigt sich durch die globale Einhüllende des Spektrums, die man beispielsweise durch Verbinden der Extremwerte in den Spektren der Bilder 3.11 und 3.12 aus Kap. 3 (jeweils rechts unten) erhält. Eine Aufteilung ist am ehesten derart möglich, dass die Kom-

ponente $P(m)$ der periodischen Anregung (siehe Bild 3.13) von den übrigen Komponenten des Modells abgetrennt wird. Hierbei existieren im wesentlichen drei Möglichkeiten.

1) Die unerwünschten Auswirkungen der periodischen Anregung werden im Spektrum nachträglich unterdrückt. Dies erreichen wir durch eine geeignete Glättung.

2) Durch Wahl einer Fensterlänge $K < T_0$ bei der Berechnung des Spektrums wird die Information über die periodische Anregung bereits durch die Kurzzeitanalyse unterdrückt; im Kurzzeitsignal ist keine Information mehr darüber enthalten. Für die reine Visualisierung in Form des Spektrogramms (siehe Abschnitt 4.3) wird diese Methode gern verwendet; das einzelne Spektrum wird jedoch wegen des kurzen Fensters so anfällig gegen Artefakte, dass es für die maschinelle Weiterverarbeitung in der Regel ungeeignet ist.

3) Das Spektrum wird so weiterverarbeitet, dass es gelingt, die Information über die periodische Anregung abzutrennen, d.h., herauszufiltern. Hier besitzt das *Cepstrum*-Verfahren besondere Bedeutung (Abschnitt 4.2).

Direkte Glättung eines Spektrums erfolgt durch "Verschmieren" der Maxima. Dies ist eine – im Leistungs- oder Amplitudenspektrum direkt durchgeführte – gewichtete Mittelung des Spektrums mit einer Glättungsfunktion, die derart erfolgt, dass im Abstand F_0 benachbarte spektrale Maxima zusammenfallen. Die Methode ist nicht sehr zuverlässig, besonders wenn Maxima mit stark unterschiedlicher Amplitude nebeneinanderliegen. Das Beispiel in Bild 4.1 macht die Probleme deutlich.

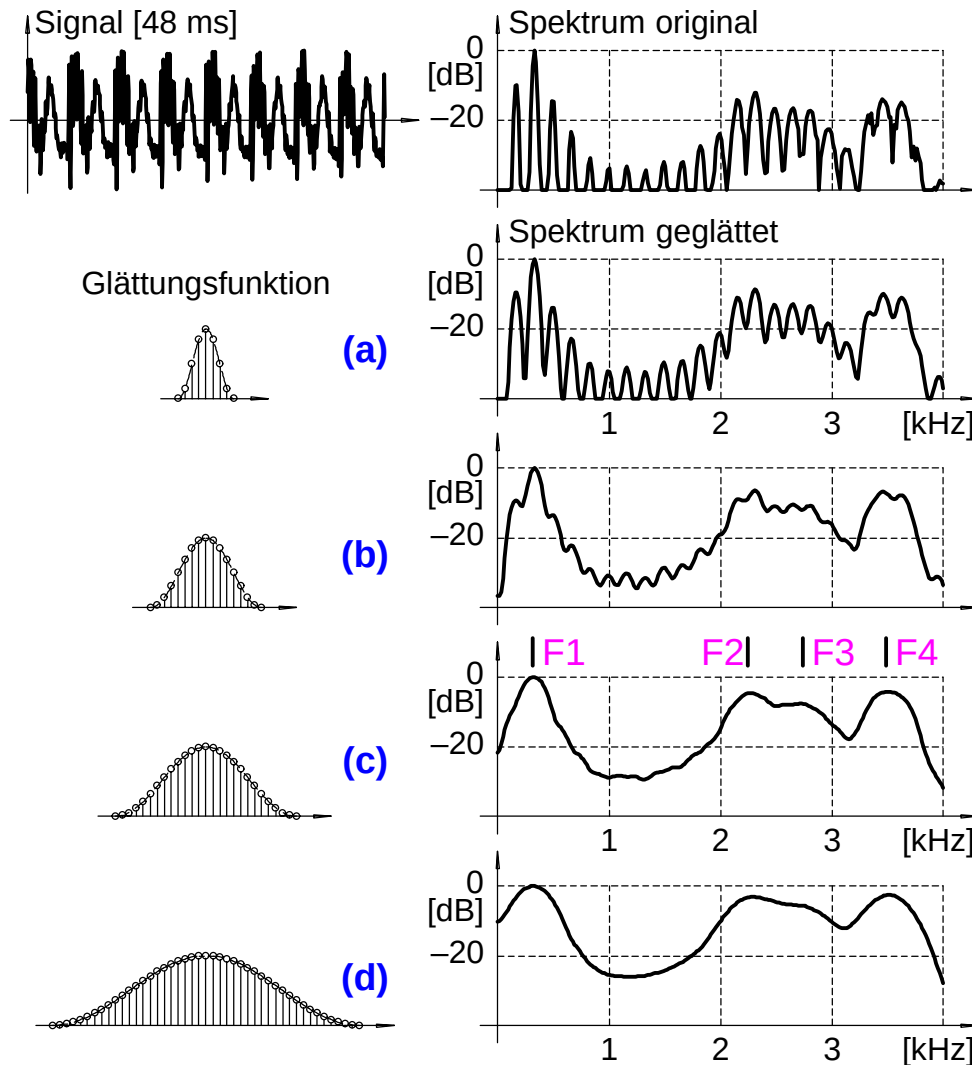


Bild 4.1. Glättung von Kurzzeitspektren durch gewichtete Mittelung. (Links oben) Sprachsignal, Vokal [i:], 48 ms; (rechts oben) ungeglättetes Kurzzeitspektrum; (links) verschiedene Glättungsfunktionen; (rechts) zugehörige geglättete Spektren (Frequenzbereich 0-4 kHz). Geglättet wurden die Amplitudenspektren; die Darstellung im Bild erfolgt im logarithmischen Maßstab. (a), (b) Glättung zu schwach; (c) Glättung angemessen; (F1...4) Formanten; (d) Glättung zu stark

4.2 Glättung von Spektren mit Hilfe des Cepstrum-Verfahrens

Durchgesetzt hat sich zur Glättung von Spektren ein Verfahren, das als *Cepstrum*-Verfahren bezeichnet wird (Bogert et al., 1963; das Wort "Cepstrum" ist durch Permutation der Buchstaben der englischen Bezeichnung "spectrum" entstanden). Es führt die Glättung durch Abwandlung eines Filterungsverfahrens durch.

Betrachten wir zum besseren Einstieg kurzzeitig das Spektrum als ein gewöhnliches in seiner unabhängigen Variablen begrenztes Signal (d.h. so, als ob es ein Zeitsignal wäre). Dann bedeutet Glättung die Tiefpassfilterung dieses Signals derart, dass die vorhandene kurzweilige Schwingung mit "Periode" F_0 beseitigt wird und die Einhüllende zurückbleibt. Ist F_0 bekannt, so kann anstelle des Tiefpasses auch eine Bandsperre verwendet werden, d.h., ein Filter, das nur die unerwünschte Schwingung und ihre unmittelbare Umgebung sperrt, alles andere aber durchlässt.

Die Schwingung mit der "Periode" F_0 ist nun im komplexen Spektrum wie auch im Leistungs- oder Amplitudenspektrum nicht ohne weiteres zu beseitigen, da sie mit der Einhüllenden gemäß (3.31) multiplikativ verknüpft ist:

$$S(m) = H(m) \cdot P(m); \quad (4.1)$$

wobei unter $H(m)$ alle Komponenten des linearen Modells außer $P(m)$ zusammengefasst seien. Das Problem ist beseitigt, wenn durch Logarithmieren des Spektrums die multiplikative Verknüpfung in eine additive umgewandelt wird:

$$\log S(m) = \log H(m) + \log P(m). \quad (4.2)$$

In dieser Darstellung, die in gleicher Weise für das Leistungs- oder Amplitudenspektrum gilt, kann der Anteil, der auf $P(m)$ entfällt, einfach von den übrigen Anteilen getrennt werden, indem man ihn durch Tiefpass oder Bandsperre wegfiltert. Gelten muss dabei allerdings (zumindest bei Verwendung eines Tiefpasses) die gleiche Voraussetzung wie bei direkter Glättung, nämlich, dass die Anteile von $H(m)$ an der Spektralfunktion "niederfrequenter" sind; in anderen Worten, dass benachbarte

Formanten und Antiformanten weiter auseinanderliegen, als die Grundfrequenz F_0 beträgt. Diese Voraussetzung gilt insbesondere bei weiblichen Stimmen oftmals nicht!

Der spektrale Tiefpass wird am ehesten realisiert durch die Rücktransformation des logarithmischen Leistungs- bzw. Amplitudenspektrums in den Zeitbereich. Man erhält damit das *Cepstrum*, dessen unabhängige Variable d (engl. *quefrency*; vor einer direkten Übertragung dieses Terminus ins Deutsche wird gewarnt!) nichts anderes als eine Zeiteinheit darstellt, die der *Verzögerung* bezüglich eines fiktiven Zeitnullpunktes entspricht und deshalb künftig auch so genannt werden soll. Im Cepstrum bleibt die additive Verknüpfung der einzelnen Komponenten des linearen Modells erhalten:

$$s(d) = h(d) + p(d). \quad (4.3)$$

Den gewünschten spektralen Tiefpass erhält man durch Abschneiden des Cepstrums vor der Grundperiode T_0 und anschließende erneute Transformation in den Spektralbereich. Bild 4.2 zeigt hierzu ein Beispiel.

In Kap. 3, Abschnitt 3.1.4 wurde festgestellt, dass nur die Rücktransformation des *unverzerrten* komplexen Spektrums in den Zeitbereich wieder auf das ursprüngliche Signal führt. Was zur Bildung des Cepstrums rücktransformiert wird, ist jedoch das logarithmierte Amplituden- oder Leistungsspektrum. Die nichtlineare Verzerrung im Frequenzbereich (durch die Betragsbildung und die anschließende Logarithmierung) führt also bei der Rücktransformation zu solch drastischen Unterschieden, wie sie beispielsweise zwischen dem Signal und seinem Cepstrum bestehen.

Mit der Anwendung als Verfahren zur Glättung von Spektren ist das Potential des Cepstrums keineswegs erschöpft. Wie sich beispielsweise in Bild 4.2 zeigt, äußert sich die periodische Anregung im Cepstrum in Form eines scharfen Maximums bei der Verzögerung $d=T_0/T$. Diese Eigenschaft ist für das Cepstrum charakteristisch (aber nicht nur auf dieses beschränkt) und macht das Verfahren auch zur Bestimmung der Grundperiode des Signals geeignet. Darüber hinaus lässt sich das Cepstrum grundsätzlich immer dann einsetzen, wenn ein lineares System untersucht wird, das aus mehreren in Reihe geschalteten Komponenten be-

steht (wie unser lineares Modell der Spracherzeugung). Die oben implizit definierte Aufgabe des Rückschlusses aus einem (beobachtbaren) Signal $s(n)$, repräsentiert durch sein Spektrum $S(m)$ auf eine Systemkomponente $H(m)$ oder $P(m)$, die im Grunde nicht beobachtbar ist, über die aber gewisse (theoretische oder empirische) Vorstellungen existieren, ist eine allgemeine Aufgabe der Signalverarbeitung, die sich häufig mit Hilfe des Cepstrums elegant lösen lässt, weil dieses die im Frequenzbereich multiplikative Verknüpfung der Systemkomponenten durch die Logarithmierung in eine additive umwandelt.

Die Tatsache, dass im Cepstrum das signifikante Maximum, das durch die periodische Anregung verursacht wird, stets bei $dT=T_0$ auftaucht, lässt sich durch eine einfache Betrachtung plausibel machen (Bild 4.3). Der komplexe (natürliche) Logarithmus trennt Betrag und Phase,

$$\begin{aligned} \ln S(m) &= \ln (|S(m)| e^{j\varphi(m)}) = \ln |S(m)| + \ln (e^{j\varphi(m)}) \\ &= \ln |S(m)| + j\varphi(m). \end{aligned} \quad (4.4)$$

Der Betragsanteil des komplexen Spektrums, also das Amplitudenspektrum, geht bei der Logarithmierung in den Realteil, das Phasenspektrum in den Imaginärteil des komplexen logarithmierten Spektrums über. Wird nun nur der Realteil weiterverarbeitet, so wird die Phaseninformation unterdrückt, d.h., die Phase aller Teilschwingungen wird bei der Rücktransformation Null. Die Phase ist aber für die Positionierung der einzelnen Teilschwingungen im Zeitsignal verantwortlich. Wenn wir insbesondere, wie in Bild 4.3 dargestellt, eine Folge äquidistanter Impulse betrachten, so ist deren komplexes Spektrum wiederum eine Impulsfolge (allerdings mit komplexen Werten). Das zugehörige Betragsspektrum ist damit ebenfalls eine Impulsfolge, die auch die Logarithmierung als solche übersteht. Damit wird auch das Cepstrum wieder eine Impulsfolge darstellen; darüber hinaus werden die Abstände aufeinanderfolgender Impulse (vorausgesetzt, die Impulse waren im Zeitsignal äquidistant) im Cepstrum mit denen im ursprünglichen Zeitsignal übereinstimmen. Die Unterdrückung der Phaseninformation bei der Berechnung des Cepstrums sorgt jedoch dafür, dass im Unterschied zum Zeitsignal (in dem die genaue Position der Impulse nicht festgelegt ist) immer

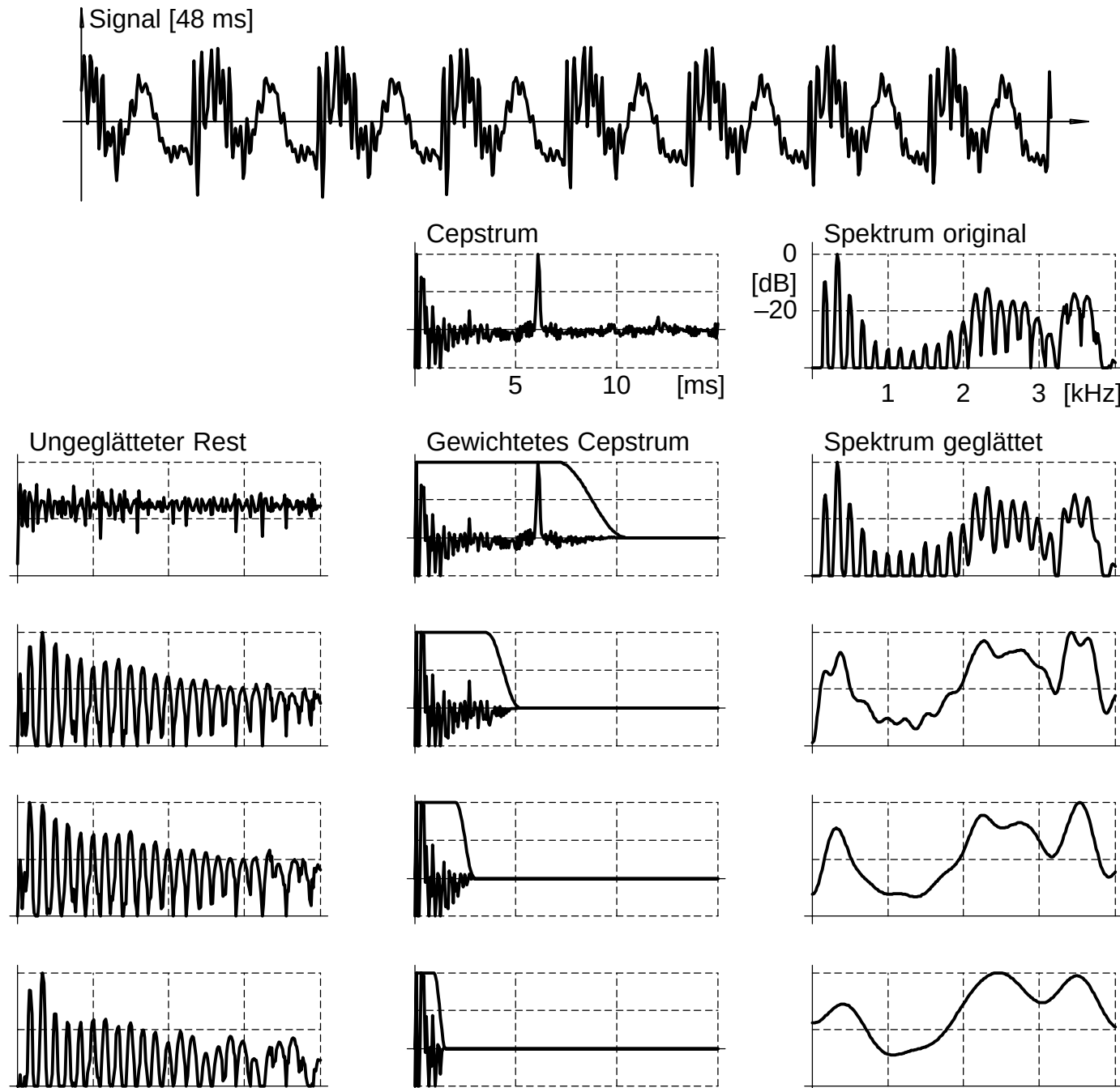


Bild 4.2. Glättung von Kurzzeitspektren mit Cepstrum: Trennung des Beitrages von Anregung (Periodizität) und Signalformung. In der Spalte "Cepstrum" ist jeweils der Anteil gezeichnet, der nach Gewichtung im Cepstrum-Bereich zur Rücktransformation für das geglättete Spektrum verwendet wird; die Spalte "Ungeglätteter Rest" geht aus der Rücktransformation des verbleibenden Anteils des Cepstrums hervor. Die Wirksamkeit des Verfahrens hängt entscheidend von der Wahl der Fensterfunktion ab, mit der das Cepstrum gewichtet wird. Im Beispiel wird die optimale Wirkung bei einer Fensterlänge von 2.5 ms erreicht (zweites Bild von unten). Signal: Vokal [i:], 48 ms, Sprecher WGH (männlich), Frequenzbereich für alle Spektren: 0–4 kHz. Die Anregung äußert sich im Cepstrum vor allem durch die scharfe Spitze bei der Grundperiode $dT=T_0$ (hier ungefähr 7 ms); (T) Abtastintervall

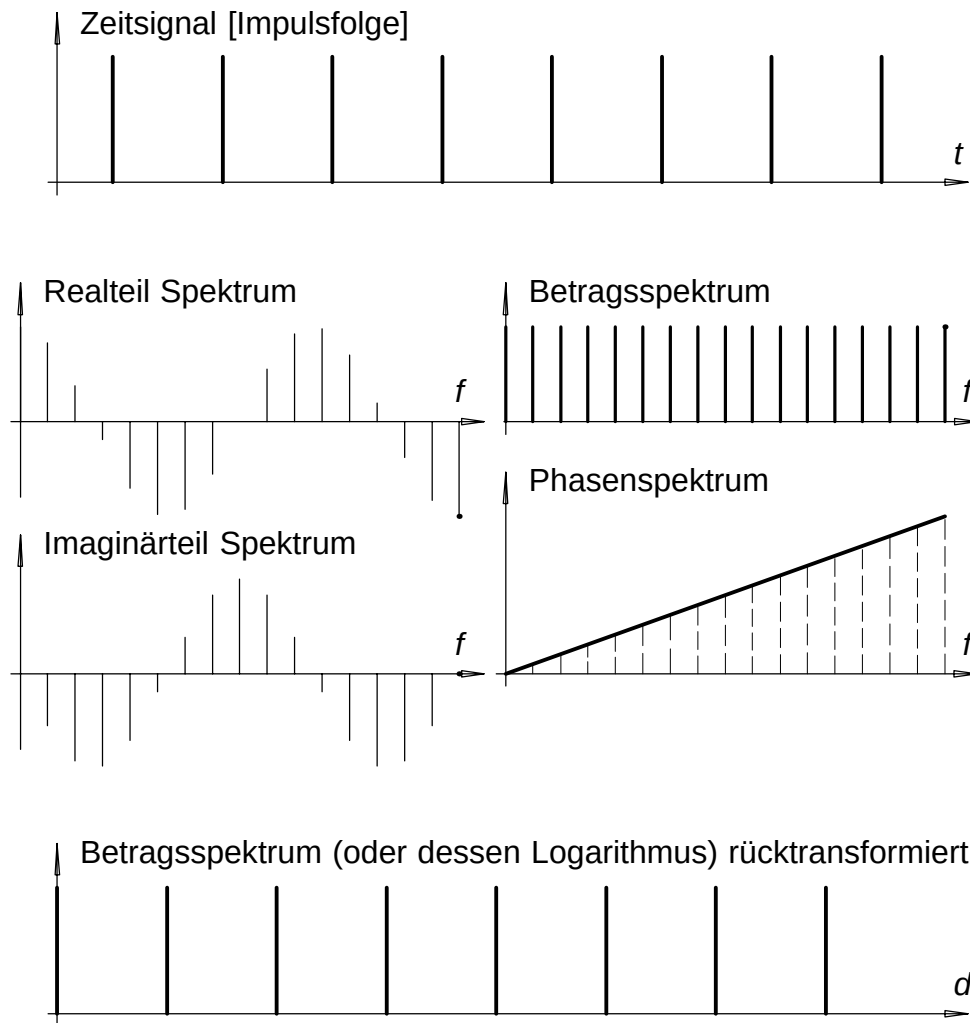


Bild 4.3. Zur Berechnung des Cepstrums einer Impulsfolge. Das Spektrum einer Impulsfolge ist wiederum eine Impulsfolge (dies ergibt sich zwangsläufig aus den Gesetzen der harmonischen Analyse); damit wird auch das Cepstrum eine Impulsfolge. Wird nun im Spektrum nur der Betragsanteil weiterverarbeitet und der Phasenanteil unterdrückt, so besitzen alle Teilschwingungen im Cepstrum die Phase Null; somit liegt ein Impuls der Impulsfolge im Cepstrum zwangsläufig an der Stelle $d=0$

ein Impuls an der Stelle $d=0$ liegt; damit muss der nächstfolgende in den Verzögerungspunkt $d=T_0/T$ fallen.

4.3 Kombinierte Zeit-Frequenz-Darstellung. Spektrogramm und Lautgebirge

Die spektrale Darstellung eines zeitveränderlichen Signals, wie es das Sprachsignal ist, erfolgt als eine Folge von Kurzzeitspektren über der Zeit. Dient die Zeit-Frequenz-Darstellung der Visualisierung, so werden die Abstände zwischen den einzelnen Stützstellen im Zeitbereich bewusst kurz gewählt, um in beiden Bereichen grundsätzlich eine pseudoanaloge graphische Darstellung zu ermöglichen. Das bedeutet allerdings, dass die Parameterabtastrate – wie auch in den meisten Fällen die spektrale Auflösung – absichtlich höher als notwendig gewählt wird, um eine dem Auge gerechte Darstellung zu sichern.

Es ist allerdings nicht möglich, durch Wahl einer hohen spektralen Abtastrate und eines kleinen Parameterabtastintervalls im Frequenz- und im Zeitbereich *gleichmaßen* eine beliebig feine Detailauflösung zu erreichen. Eine gute Auflösung in Zeitrichtung bringt eine schlechte spektrale Auflösung mit sich und umgekehrt, so dass hier eine natürliche Grenze existiert. Mit dieser Frage beschäftigt sich Abschnitt 4.3.1.

Wie jede Darstellung eines zeitveränderlichen Signals nach erfolgter Kurzzeitanalyse ergibt sich die Zeit-Frequenz-Darstellung einer Folge von Kurzzeitspektren als Funktion zweier Veränderlicher, d.h. der Zeit (in Form der Folge von Parameterabtastintervallen an den Stützstellen, an denen die einzelnen Spektren berechnet werden) und der Frequenz (als unabhängiger Veränderlicher des einzelnen Spektrums). Die Darstellung wird damit eigentlich dreidimensional. Für die Visualisierung auf Papier oder Bildschirm muss eine pseudo-dreidimensionale Darstellung gewählt werden. Hierbei hat sich vor allem das *Spektrogramm* (Sonagramm) durchgesetzt, das auf eine Entwicklung von Koenig et al. (1946) zurückgeht. Daneben findet sich gelegentlich auch die Darstellung in Form eines *Lautgebirges*. Mit beiden Darstellungsformen wird sich Abschnitt 4.3.2 befassen.

4.3.1 Spektrale und zeitliche Auflösung oder das Zeit-Bandbreite-Gesetz der Nachrichtentechnik

Aus der Anschauung können wir uns rasch plausibel machen, dass ein Zeitsignal, das innerhalb eines sehr kurzen Zeitintervalls abklingt, ein breitbandiges Spektrum besitzt, d.h., ein Spektrum, in dem zahlreiche Komponenten von Null verschieden sind. Zwei wohlbekannte Extrembeispiele hierzu: Eine ungedämpfte Sinoidalschwingung besitzt eine einzige Frequenz, also ein Spektrum, bei dem nur eine Komponente von Null verschieden ist. Ein Impuls, also ein Zeitsignal, bei dem nur ein einziger Abtastwert von Null verschieden ist, besitzt spektrale Anteile bei sämtlichen Frequenzen.

Diese Feststellungen legen nahe, dass sich die Reziprozität zwischen Zeit und Frequenz nicht nur in den physikalischen Maßen selbst äußert [man beachte, dass die Frequenz die physikalische Dimension $(\text{Zeit})^{-1}$ besitzt!], sondern auch in der Begrenztheit bzw. Ausdehnung von Vorgängen (d.h., Signalen und Spektren) in den beiden Bereichen.

In Verbindung mit Resonanzen und spektralen Maxima ist in der Nachrichtentechnik die Definition der *Bandbreite* geläufig. Die Bandbreite einer Resonanz ist definiert als das Frequenzintervall Δf um die Resonanz herum, bei dem die Energie auf die Hälfte des Maximalwertes abgeklungen ist; dies entspricht einer Veränderung der relativen Amplitude auf den Wert $1/\sqrt{2}$ und des Pegels um -3 dB, jeweils bezogen auf den Maximalwert an der Resonanzstelle. Bild 4.4 verdeutlicht den Zusammenhang und stellt auch fest, daß zwischen der Bandbreite einer Resonanz im Spektrum und der Dämpfung der zugehörigen Schwingung im Zeitbereich ein reziproker Zusammenhang besteht: je länger die Zeitspanne, bis eine gedämpfte Schwingung abklingt, desto geringer die Bandbreite der Resonanz. Die Dauer einer gedämpften Schwingung, hier definiert als die Zeit, die die Schwingung benötigt, um auf einen bestimmten Bruchteil ihrer Anfangsamplitude abzuklingen, und die Bandbreite der Resonanz im Spektrum sind also umgekehrt proportional, oder, verbal formuliert:

Zeitliche Bandbreite mal spektrale Bandbreite ergibt eine Konstante. (4.5)

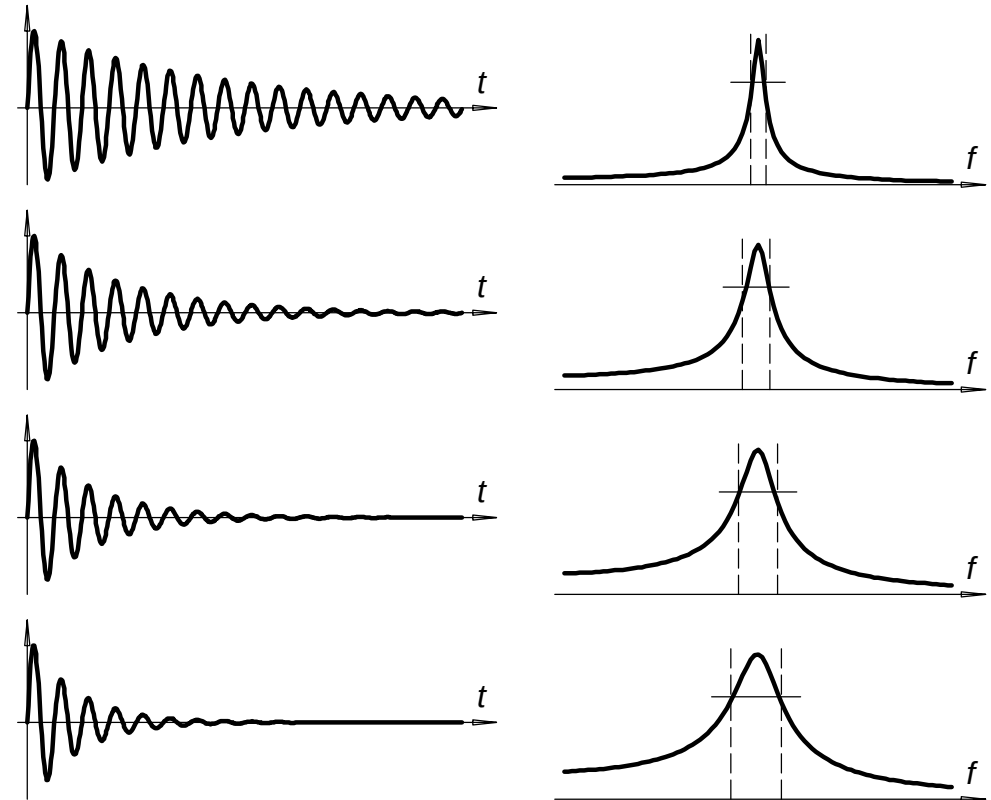


Bild 4.4. Zur Beziehung zwischen Dämpfung einer Schwingung und Bandbreite der zugehörigen Resonanz im Spektrum. Aufgezeichnet sind vier verschieden stark gedämpfte Schwingungen gleicher Frequenz und ein Ausschnitt des jeweiligen Amplitudenspektrums um die Resonanzstelle. Die Bandbreite ist definiert durch die beiden Frequenzen um die Resonanzstelle, bei denen die Amplitude im Vergleich zum Maximum um 3 dB abgesunken ist

Diese Beziehung ist auch bekannt als das *Zeit-Bandbreite-Gesetz* der Nachrichtentechnik. Marko (1956) hat diesen Zusammenhang für eine spezielle Art von Impulsen untersucht, deren Signalform definiert ist durch

$$s(t) = \hat{s} \exp(-at^2), \quad (4.6a)$$

also die Form der Gauß'schen Fehlerfunktion (Glockenkurve) besitzt. Ein derartiges Signal hat (im Analogen) ein Spektrum der Form

$$S(f) = \hat{S} \exp(-\beta f^2), \quad (4.6b)$$

so dass die Darstellung im Zeit- und im Frequenzbereich durch die gleiche mathematische Funktion erfolgt. In diesem Fall lässt sich die im Frequenzbereich vorgenommene – und zunächst auf diesen beschränkte – Definition der Bandbreite unverändert auch auf das Zeitsignal übertragen. Wird die zeitliche Bandbreite in Sekunden und die spektrale in Hertz gemessen, so ergibt sich für diesen Spezialfall und darüber hinaus allgemein (Marko, 1956)

$$\Delta t \cdot \Delta f \geq 1. \quad (4.7)$$

Bild 4.5 soll diesen Zusammenhang verdeutlichen.

Für die Darstellung zeitveränderlicher Signale in einer Zeit-Frequenz-Repräsentation bedeutet dieses Gesetz eine Unschärferelation, die es nicht gestattet, die zeitliche Auflösung und die spektrale Auflösung gleichzeitig beliebig fein zu machen. Konkret bedeutet dies: wird die Fensterlänge K (vgl. Abschnitt 2.2) bei der Spektralanalyse klein gewählt, so kann eine detailreiche Abbildung entlang der Zeitachse erwartet werden, die aber mit einer wesentlich ärmeren Auflösung in spektraler Richtung einhergeht. Ist K groß, so wird die zeitliche Variation des Signals infolge der größeren zeitlichen Bandbreite nur schlecht abgebildet, während im Spektralbereich eine sehr detailreiche Abbildung erfolgt.

Dies betrifft insbesondere die Darstellung der periodischen Anregung $P(m)$. Ist $K < T_0$, so gelangt die Information über $P(m)$ nicht mehr in den Frequenzbereich, da eine harmonische Struktur im Spektrum nur dann auftreten kann, wenn das Zeitsignal im Kurzzeitfenster der Länge K Periodizität erkennen lässt; dies ist wiederum nur dann möglich, wenn das Intervall mehr als eine (besser: mehr als zwei) komplette Grundperiode(n) enthält. Ist K geringer, d.h., umfasst ein Messintervall deutlich weniger als eine Grundperiode, so lässt sich die periodische Struktur entlang der Zeitachse sichtbar machen. Ist K hingegen deutlich größer als T_0 , so wird jedes Einzelspektrum die aus Kap. 3 gewohnte harmoni-

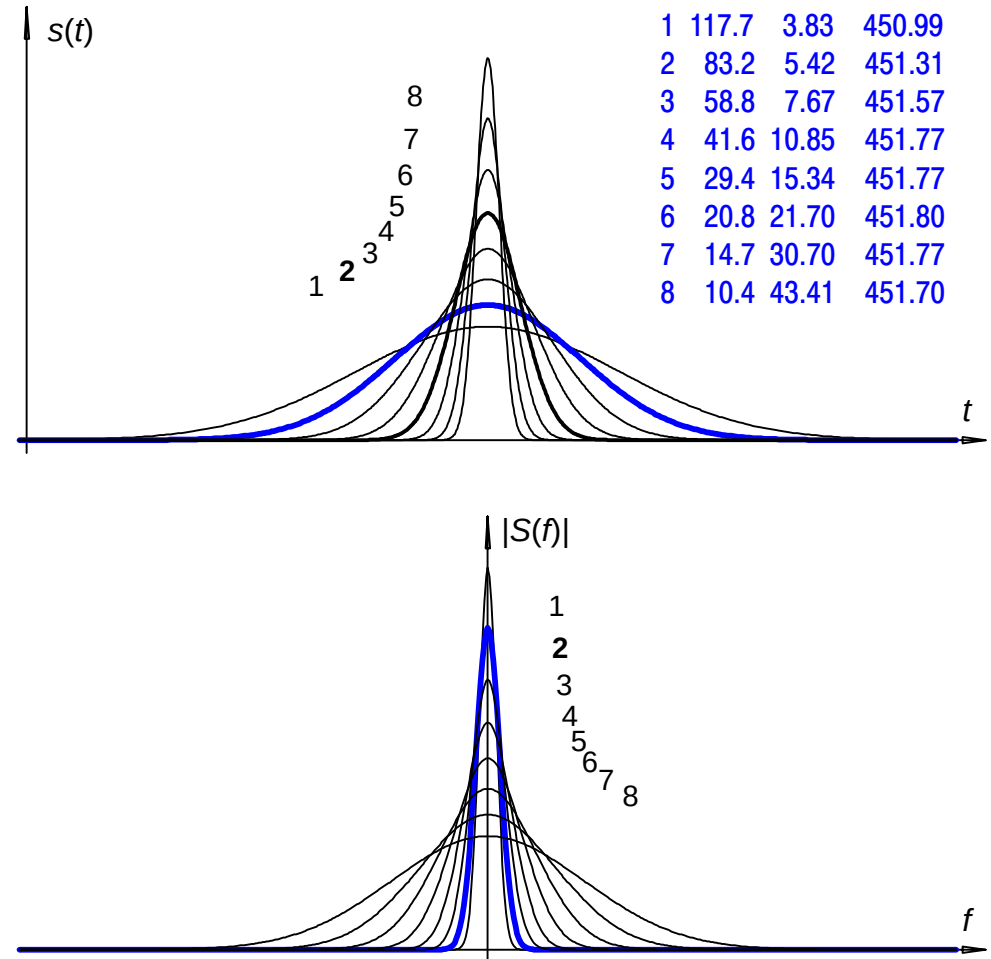


Bild 4.5. Zum Zeit-Bandbreite-Gesetz. Dargestellt sind insgesamt 8 Impulse, deren Zeitverlauf der Gauß'schen Glockenkurve entspricht, für verschiedene Werte des Parameters a in (4.6), sowie die zugehörigen Amplitudenspektren. Die Tabelle liest sich wie folgt (von links nach rechts): Index des Impulses bzw. des Spektrums, Parameter a , Parameter β (beide gemessen in willkürlichen Einheiten, die einem zeitlichen bzw. spektralen Abtastwert entsprechen), Produkt von a und β . Das Produkt bleibt mit guter Annäherung konstant; dass sein Wert nicht 1 beträgt, ist auf die Willkür bei der Wahl der Einheiten zurückzuführen. Zur Verdeutlichung ist die Kurve 2 blau hervorgehoben gezeichnet

sche Struktur mit den äquidistanten spektralen Maxima besitzen, während es nicht mehr gelingt, aus dieser Darstellung die einzelnen Perioden herauszulesen. Die periodische Anregung kann also je nach Wahl der Fensterlänge K entlang der Zeitachse oder entlang der Frequenzachse sichtbar gemacht werden, niemals jedoch entlang beiden Achsen gleichzeitig.

Das Zeit-Bandbreite-Gesetz spezifiziert eine untere Schranke für das Zeit-Bandbreite-Produkt. Es ist jederzeit möglich, eine Zeit-Frequenz-Darstellung mit schlechterer Gesamtauflösung zu gewinnen. Genau dies tut eine Darstellung, wenn sie von *geglätteten* Spektren ausgeht: nach der Glättung ist die Information über die periodische Anregung in keinem der beiden Bereiche mehr präsent.

4.3.2 Visualisierungsformen

Die beiden hauptsächlichen Visualisierungsformen von Zeit-Frequenz-Darstellungen zeitveränderlicher Signale sind *Spektrogramm* und *Lautgebirge*, wobei sich in der Sprachsignalverarbeitung, wenn es mehr um die Visualisierung als um die maschinelle Weiterverarbeitung geht, das Spektrogramm durchgesetzt hat.

Das Spektrogramm geht auf eine Entwicklung von Koenig et al. (1946) zurück und entstammt somit der Analogtechnik. Zeit und Frequenz bilden die beiden Achsen eines kartesischen Koordinatensystems; die abhängige Veränderliche (Amplitude bzw. Pegel) wird durch einen Helligkeitsparameter (Schwärzungsgrad des Papiers bzw. Helligkeit des Bildschirms) dargestellt. Wir unterscheiden zwei Formen: das *Breitbandspektrogramm* und das *Schmalbandspektrogramm*. Hierbei werden die einzelnen Spektren durchweg *nicht* geglättet; der entscheidende Parameter ist die Länge K des Signalabschnitts, der den Ausgangspunkt für die Berechnung der einzelnen Spektren bildet.

Beim Breitbandspektrogramm wird K (wesentlich) kleiner als T_0 gewählt. Damit kann über die Grundfrequenz des Sprachsignals keine Information mehr in das einzelne Spektrum gelangen; entsprechend dem Zeit-Bandbreite-Gesetz geht die Information über die periodische Anregung jedoch nicht verloren, sondern schlägt sich in der zeitlichen Dar-

stellung nieder. Hier sind die einzelnen Perioden wie senkrechte Striche markiert: da alle Formanten punktuell zum Zeitpunkt des Glottalverschlusses angeregt werden, ergeben sich zu Beginn jeder Periode die höchsten Signalamplituden; damit erhalten wir bei den Spektren, die einen Periodenbeginn enthalten, insgesamt höhere Energie, also in der visuellen Darstellung einen höheren Schwärzungsgrad. Dies ist in dem Beispiel in Bild 4.6 deutlich zu sehen. Das Breitbandspektrogramm erlaubt die manuelle Messung von Formantfrequenzen sowie zeitlicher Parameter (Lautdauer, Pausendauer etc.). Es ist in der Phonetik und in den Philologien das wichtigste experimentelle Hilfsmittel zur (manuellen) Untersuchung gesprochener Sprache.

Da die einzelnen Spektren nicht geglättet sind und K sehr kurz ist, hängt die Form des Einzelspektrums stark von der Lage des Fensters relativ zum Beginn der nächstgelegenen Grundperiode ab. Die einzelnen Spektren sind daher für die maschinelle Weiterverarbeitung (z.B. für eine automatische Formantanalyse) in der Regel ungeeignet.

Beim Schmalbandspektrogramm wird K so gewählt, dass das Fenster mehrere Grundperioden umfasst. Das Einzelspektrum erhält somit die Form, wie wir sie in Kapitel 3 kennengelernt haben. Die Information über die periodische Anregung findet sich nun in den einzelnen Spektren wieder; Grenzen einzelner Perioden lassen sich nicht mehr darstellen. Da nun die einzelnen Teilschwingungen im Spektrum (und damit im Spektrogramm) erscheinen, ist auch eine Formantmessung nicht mehr möglich. Schmalbandspektrogramme dienen somit einzig dem Zweck, den Grundfrequenzverlauf sichtbar und manuellen Messungen zugänglich zu machen. Bild 4.7 zeigt ein Beispiel.

Ein Spezialfall der Visualisierung ist das sogenannte *Lautgebirge*. Hier wird eine pseudodreidimensionale Darstellung in Form einer geometrischen Projektion gewählt. In der Ebene der Zeichnung liegen Frequenz und Amplitude (bzw. Pegel), während die Zeitachse senkrecht dazu (also in der Projektion schräg nach "hinten") verläuft. Im Gegensatz zum Spektrogramm ist hier eine quantitativ nachvollziehbare Darstellung auch der abhängigen Veränderlichen des Spektrogramms (d.h., Amplitude bzw. Pegel) möglich.

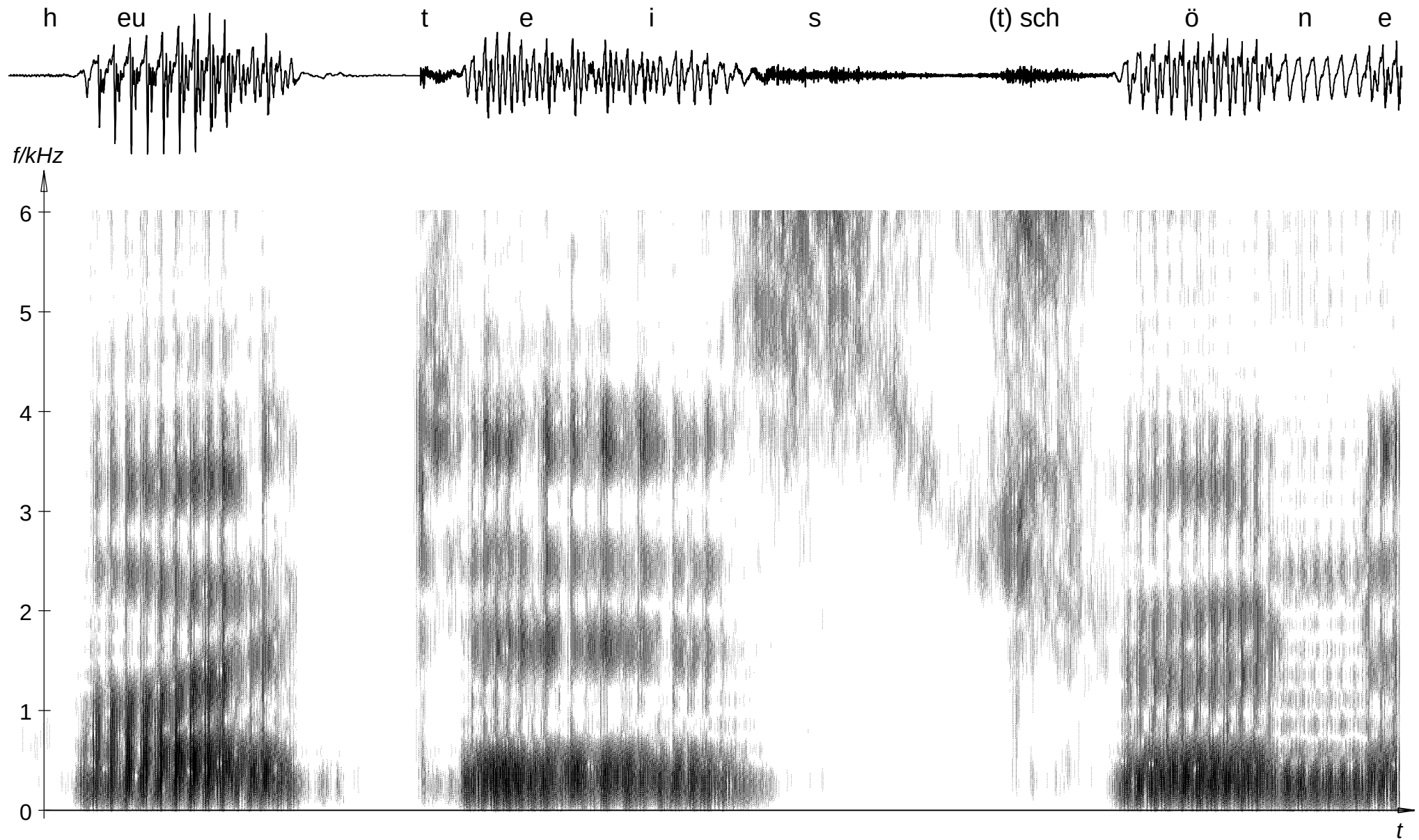


Bild 4.6. Breitbandspektrogramm eines Sprachsignals. Sprachliche Äußerung: Beginn des Satzes "Heute ist schönes Frühlingswetter", etwa 750 ms. Fensterlänge K : 3 ms, also Bandbreite etwa 400 Hz; Hammingfenster. Die Periodengrenzen lassen sich in Zeitrichtung als vertikale Striche verfolgen

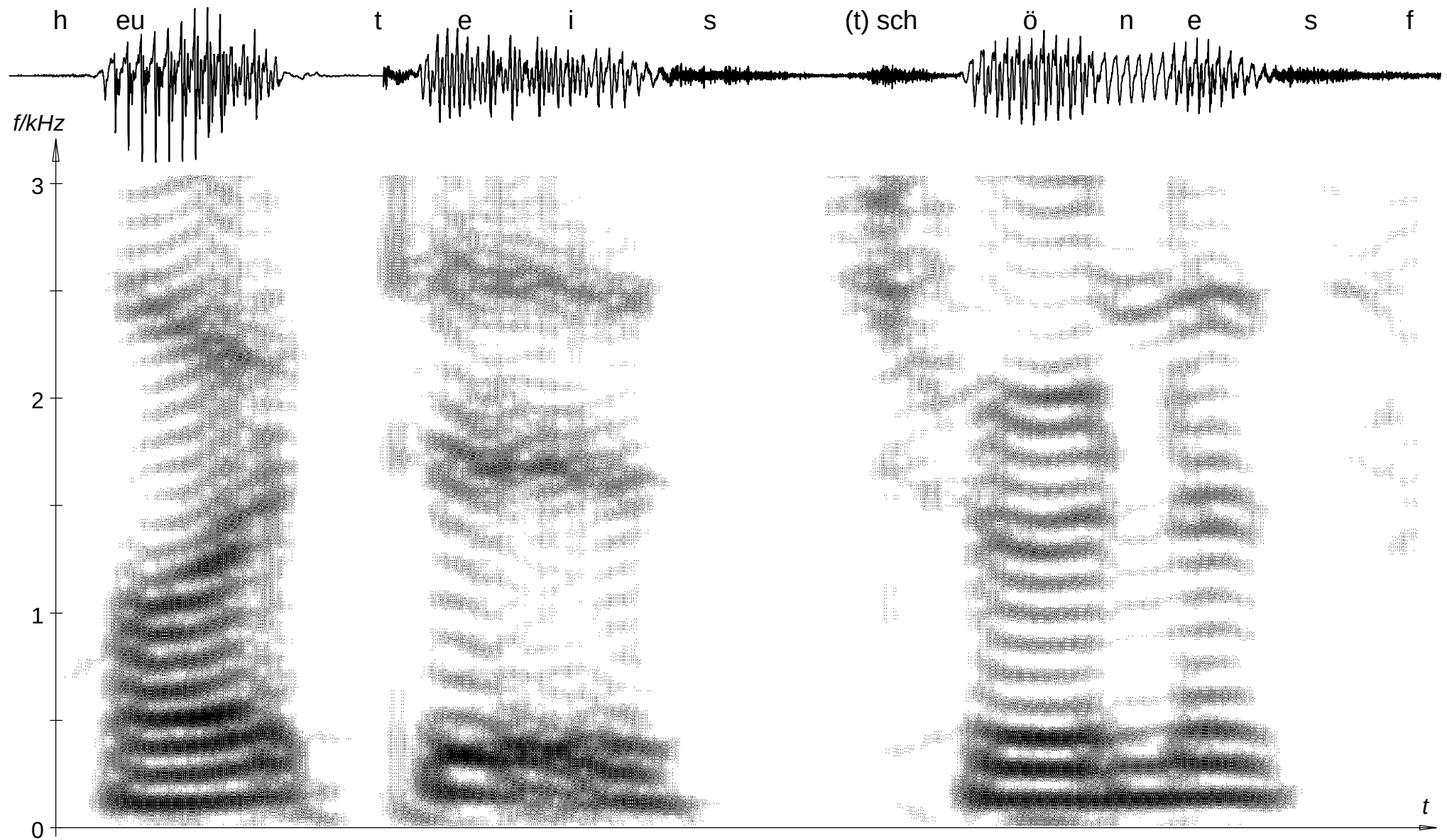


Bild 4.7. Schmalbandspektrogramm eines Sprachsignals. Sprachliche Äußerung: wie Bild 4.6, etwa 850 ms. Fensterlänge K : 30 ms, also Bandbreite etwa 40 Hz; Hammingfenster. Die periodische Anregung äußert sich in Form einer harmonischen Struktur; dies ist deutlich daran zu sehen, daß eine solche Struktur während der Pausen sowie der stimmlosen Abschnitte nicht auftritt. Ebenfalls keine harmonische Struktur ist zu beobachten während der zwei Grundperioden im zweiten stimmhaften Signalabschnitt, während derer die Stimmbandschwingung unregelmäßig wird

Grundlage dieser Darstellung bilden im Gegensatz zum Spektrogramm möglichst konsistente, geglättete Spektren. Lautgebirge enthalten somit keine Information über die periodische Anregung (es sei denn, man verzichtet auf die Glättung, was grundsätzlich möglich ist). Artefakte, wie sie bei schlechter Vorverarbeitung durch die veränderliche Lage der Stützstellen relativ zu den Grundperioden auftreten können, würden das Bild erheblich stören. Bild 4.8 zeigt ein Beispiel.

Als Visualisierungsmittel hat sich in der Sprachsignalverarbeitung – im Gegensatz zu einigen anderen Zweigen der Signalverarbeitung – das Lautgebirge gegenüber dem Spektrogramm nicht durchsetzen können. Für die maschinelle Weiterverarbeitung von Sprachsignalen, wenn eine Zeit-Frequenz-Darstellung gewählt wird, geht man jedoch durchweg von dieser Repräsentation aus.

Literatur

- Bogert B.P., Healy M.J.R., Tukey J.W. (1963): "The quefrency alalysis [sic!] of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking." *Proceedings of the Symposium on Time Series Analysis*, ed. by M.Rosenblatt, 209–243 (Wiley, New York)
- Childers D.G., Skinner D.P., Kemerait R.C. (1977): "The cepstrum: a guide to processing." *Proc. IEEE* 65, 1428–1443
- Flanagan J.L. (²1972) *Speech analysis, synthesis, and perception* (Springer, Berlin)
- Harris F.J. (1978): "On the use of windows for harmonic analysis with the discrete Fourier transform." *Proc. IEEE* 66, 51...63
- Koenig W., Dunn H.K., Lacy L.Y. (1946): "The sound spectrograph." *J. Acoust. Soc. Am.* 18, 19–49
- Marko H. (1956): "Die Reziprozität von Zeit und Frequenz in der Nachrichtentechnik." *Nachrichtentechn. Z.* 9, 222–228 und 266–271
- Noll A.M. (1964): "Short time spectrum and 'cepstrum' techniques for vocal pitch detection." *J. Acoust. Soc. Am.* 36, 296–302
- Rabiner L.R., Schafer R.W. (1978): *Digital processing of speech signals*. Prentice-Hall Signal Proc. Series, ed. by A.V.Oppenheim (Prentice Hall, Englewood Cliffs, NJ)
- Vary, Peter / Heute, Ulrich / Hess, Wolfgang (1998): *Digitale Sprachsignalverarbeitung* (Teubner, Stuttgart)

Bild 4.8. Zeit-Frequenz-Darstellung als Lautgebirge. Äußerung: wie Bild 4.6; dargestellt ist ein Zeitintervall von etwa 0,625 s. Fensterlänge K im Zeitsignal: 30 ms; Glättung der Spektren mit Cepstrum; Fensterlänge im Cepstrum: 2,5 ms. Zeitlicher Abstand aufeinanderfolgender Spektren: 5 ms. Dargestellter Dynamikbereich: 50 dB

