Kapitel 2

Sprachverarbeitung

2.1 Spracherzeugung

Sprache wird durch Veränderung des von der Lunge kommenden Luftstroms erzeugt. Zunächst passiert der Luftstrom am Eingang der Luftröhre den Kehlkopf (medizinisch *Larynx*) mit den Stimmbändern. Stehen die Stimmbänder dicht beieinander, dann werden sie durch den Luftstrom zu periodischen Schwingungen angeregt. Es entsteht ein stimmhafter Laut. Die Frequenz der Schwingungen (Sprachgrundfrequenz, *fundamental frequency*) liegt im Mittel bei Männern um 100 Hz und bei Frauen um 180 Hz. Die Variation der Sprachgrundfrequenz ist zusammen mit Lautstärkeänderungen wesentlich für die Sprachmelodie einer Äußerung. Ist der Abstand zwischen den Stimmbändern groß, so kommt es zu keinen Schwingungen sondern nur zu Turbulenzen. Dann spricht man von stimmlosen Lauten.

Die weitere Lautbildung erfolgt im Vokaltrakt, d. h. dem durch Rachen, Mundhöhle und Nasenraum gebildeten Raum. Durch Absenken des Gaumensegels wird bei Bedarf die Verbindung zum Nasenraum geöffnet. Der Luftstrom kann in diesem Fall auch durch die Nasenhöhle austreten (Nasallaute). Je nach Art der Luftströmung lassen sich zwei Lautklassen unterscheiden:

- Vokale: die Luft strömt relativ ungehindert durch den Vokaltrakt. Im Wesentlichen abhängig von der Position von Zunge und Lippen bilden sich unterschiedliche Resonanzen aus. Die markanten Resonanzen werden als Formanten bezeichnet.

- Konsonanten: der Luftstrom wird durch eine Verengung gestört oder sogar vorübergehend vollkommen unterbrochen. Je nach Verengungsstelle ergeben sich unterschiedliche Laute. Konsonanten können stimmhaft (Beispiel [p]) oder stimmlos ([b]) sein.

Weitere Beispiele für Konsonanten sind die Zahnverschlusslaute [t d]. Hier berührt die Zungenspitze die oberen Schneidezähne oder die Zahnfächer (Alveolen).

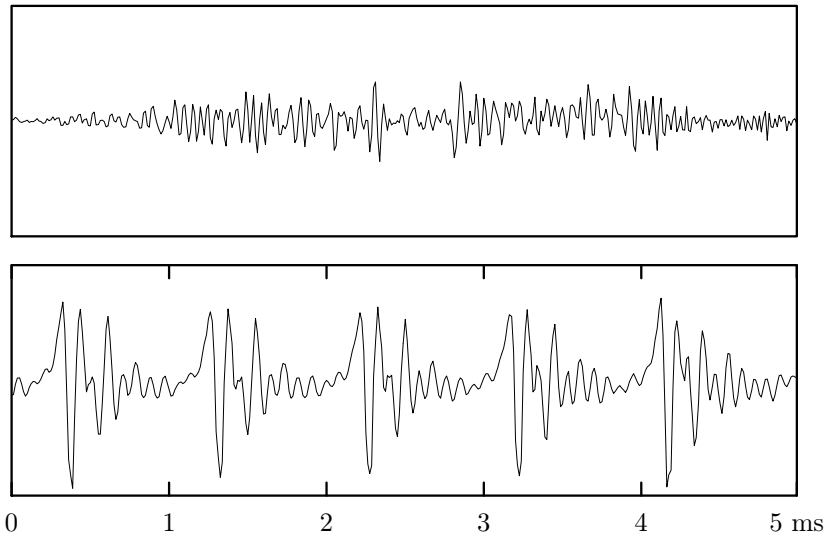

Abbildung 2.1: Abschnitte aus den Lauten tz und ei einer Äußerung des Wortes *Zwei* (Abtastrate 8 kHz)

Liegt die Verengung weiter hinten am Gaumen, so entstehen die Hintergaumenverschlußlaute [k g]. Bei den Plosiven [p b] sind die Lippen zunächst vollständig geschlossen. Im Mundraum wird ein Druck aufgebaut. Beim Öffnen der Lippen entsteht dann der Laut.

Bild 2.1 illustriert den unterschiedlichen Charakter verschiedener Laute. Dargestellt sind zwei jeweils 5 ms lange Abschnitte aus einer Äußerung des Wortes *Zwei*. In dem Abschnitt aus dem Laut ei erkennt man deutlich die Periodizität als Folge der stimmhaften Anregung. Der Abstand der aufeinander folgender Anregungen beträgt etwa 1 ms. Daraus ergibt sich die Sprachgrundfrequenz zu circa 100 Hz.

Die aus den beiden Signalabschnitten abgeleiteten spektralen Darstellungen sind in Bild 2.2 wiedergegeben. Es handelt sich genauer gesagt um die logarithmierten Betragsspektren. Diese Darstellung zeigt den Anteil der einzelnen Frequenzen in dem Signalabschnitt. Das Spektrum des stimmlosen Lautes ist insgesamt gesehen relativ flach. Demgegenüber kann man im stimmhaften Fall die Erhöhungen durch die Formanten deutlich erkennen. Darüber hinaus ist der Verlauf durch die Grundfrequenz bei 100 Hz und die dazu gehörenden Obertöne bei ganzzahligen Vielfachen geprägt.

Um einen Laut zu erzeugen, ist ein wohl koordiniertes Zusammenspiel der Sprechwerkzeuge erforderlich. Eine komplette Äußerung als Abfolge von Lauten bedingt darüber hinaus dynamische Lautübergänge. Die Bewegungsabläufe der Artikulatoren unterliegen dabei dem Prinzip der *Ökonomie des Artikulationsauf-*

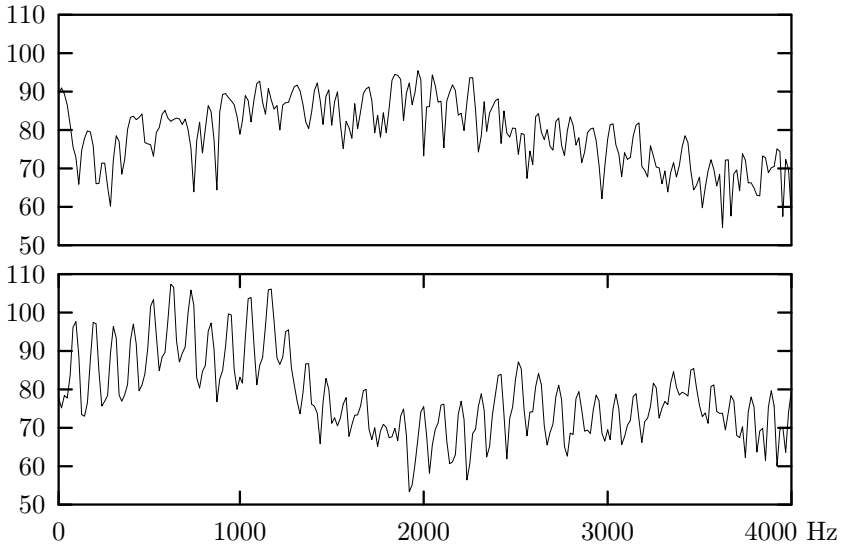

Abbildung 2.2: Frequenzdarstellung mit logarithmierten Betragsspektren der beiden Signalabschnitte aus Bild 2.1

wandes. Die Bewegungen von z. B. der Zunge werden so ausgeführt, dass der Gesamtaufwand minimal wird. Dazu werden die einzelnen Laute derart artikuliert, dass der Übergang an den Verbindungsstellen möglichst glatt wird. Als Randbedingung gilt natürlich, dass der Laut noch als solcher erkennbar bleibt. Die Realisierung eines Lautes richtet sich dementsprechend nach den vorhergehenden und nachfolgenden Lauten (Koartikulation).

Sprechen hat eine große Ähnlichkeit mit anderen Bewegungsabläufen. Man kann daher durchaus das Sprechen eines Satzes mit einer Aktion wie z. B. dem Werfen eines Balls vergleichen. Aus dieser Analogie lassen sich wichtige Konsequenzen ableiten:

- Jeder Mensch hat eine individuelle Sprechweise. Die Art, wie er oder sie spricht, ist einerseits von den anatomischen Voraussetzungen und andererseits von der erlernten Technik bestimmt.

- Sprachliche Äußerungen sind nicht beliebig exakt reproduzierbar. Wie andere Bewegungsabläufe hängen sie beispielsweise von Stimmung, Müdigkeit und Gesundheitszustand ab. Da die Muskeln nicht mit beliebiger Genauigkeit gesteuert werden können, kommt noch ein mehr oder weniger zufälliges Moment hinzu. Selbst ein trainierter Basketball-Profi kann Freiwürfe nicht beliebig reproduzieren.

2.1.1 Quelle-Filter-Modell

Die Beschreibung der Spracherzeugung soll einen Eindruck von der Komplexität der zugrundeliegenden Abläufe vermitteln. Eine vollständige Modellierung ist überaus schwierig. Glücklicherweise kann man in der Sprachverarbeitung auch mit einem vergleichsweise einfachen Modell bereits viel erreichen. Das so genannte Quelle-Filter-Modell geht zurück auf Arbeiten von Gunar Fant in den 60er Jahren [Fan60]. Die Annahmen sind:

- Das Modell besteht aus unabhängigen Komponenten.
- Die Anregung wird durch eine Quelle, umschaltbar zwischen periodischem Signal (stimmhaft) und rauschartigem Signal (stimmlos), geliefert.
- Das Frequenzverhalten des Vokaltraktes wird durch ein digitales Filter beschrieben.
- Ein weiteres Filter modelliert die Schallabstrahlung am Mund.

Trotz der starken Vereinfachungen erlaubt das Modell in vielen Fällen eine hinreichend gute Beschreibung der Spracherzeugung. Gleichzeitig ist es gerade wegen seiner Einfachheit gut handhabbar. In Systemen zur Spracherkennung konzentriert man sich häufig auf die Schätzung der Parameter des digitalen Filters. Dazu wurde eine ganze Reihe von effizienten Verfahren entwickelt.

In den allermeisten Anwendungen werden Allpol-Filter eingesetzt. Ein solches Filter kann als Modell für eine Röhre mit Segmenten von unterschiedlichem Durchmesser interpretiert werden (akustisches Röhrenmodell). Das Modell von konzentrischen Röhren mit harter Innenseite ist nur eine grobe Annäherung an den Vokaltrakt. Nichtsdestoweniger erweist es sich in vielen Anwendungsfällen als hilfreich.

2.2 Einheiten

Eine sprachliche Äußerung – geschrieben oder gesprochen – besteht aus kleineren Einheiten. Eine große und weitgehend selbständige Einheit ist der Satz. In geschriebener Sprache wird ein Satz durch ein Satzzeichen beendet. Weitere Satzzeichen verdeutlichen die innere Struktur eines Satzes wie beispielsweise die Aufteilung in Haupt- und Nebensatz. Die Grenzen eines gesprochenen Satzes sind durch den Intonationsverlauf und eine abschließende Pause markiert. Fehlen die Satzzeichen, so ist das Verständnis wesentlich erschwert. Der Satz *Bei mir geht Mittwoch nicht aber Donnerstag* als Beispiel ist mehrdeutig.

Ein Satz besteht aus einem (Beispiel: *Ja.*) oder mehreren Wörtern. Ein Wort ist die kleinste funktionale Einheit. Je nach Merkmal unterscheidet man verschiedene Wortarten. Hauptwörter (Substantive) beispielsweise stehen für konkrete

Gegenstände, Personen, abstrakte Ideen oder Handlungen. Verben beschreiben Zustände oder Handlungen. Verknüpfungen werden durch Konjunktionen wie *und*, *oder*, *dass*, ... ausgedrückt. Ein Wort kann, je nach Gebrauch, unterschiedliche Wortarten annehmen. So zeigt das Beispiel 2 aus Abschnitt 1.2, dass das Wort *time* sowohl Verb als auch Substantiv sein kann.

Viele Wörter haben eine innere Struktur. Häufig werden Wörter aus kleineren Einheiten zusammengesetzt (*Fachhochschule*). Vorsilben verändern die Bedeutung eines Wortes (*unsicher*). Je nach Zeitform werden verschiedene Endungen angefügt. Diese innere Struktur der Wörter ist Gegenstand der Morphologie.

Nach der Sprechweise lassen sich Wörter in Silben aufteilen. Im Gegensatz zu den Schreibsilben, die nur bei der Trennung eine Rolle spielen, spricht man genauer von Sprechsilben. Silben sind relativ eigenständige Einheiten. Eine Silbe besteht aus einem Vokal oder Diphthong (Doppellaut wie z. B. ei oder au) als Kern. Vor und nach dem Kern können Konstanten stehen. Einige Beispiele für die Aufteilung in Silben sind:

- *Haus*: eine Silbe
- *Ufer*: U - fer
- *schaufeln*: schau - feln
- *Kannen*: Kan - nen

Die Aussprache einer Silbe ist weitgehend unabhängig von der vorausgehenden und der nachfolgenden Silbe. Eine Silbe selbst kann wiederum in 2 Halbsilben aufgeteilt werden: in eine Anfangshalbsilbe von der Silbengrenze bis zur Silbenmitte und eine Endhalbsilbe von der Silbenmitte bis zum Silbenende.

Die kleinste sprachliche Einheit ist ein Laut (Phon). Laute werden in Lautschrift mit speziellen Symbolen dargestellt. Damit kann man einem Wörterbuch in einer fremden Sprache entnehmen, wie einzelne Wörter ausgesprochen werden. Ähnlich wie ein Buchstabe in verschiedenen Fonts unterschiedlich aussieht, kann ein Laut unterschiedlich klingen. Der Klang hängt von dem Kontext (d. h. den umgebenden Lauten), dem Sprecher, gegebenenfalls seinem Dialekt oder Akzent etc. ab.

Für einen geübten Hörer sind viele tausend verschiedene Laute unterscheidbar. Eine derart feine Auflösung ist aber für das Verstehen nicht notwendig. Es reicht aus, Laute so gut zu erkennen, dass die Bedeutung des gesprochenen Wortes klar wird. Solange Lautunterschiede keine Bedeutungsunterschiede bewirken, können die Laute als zu einer Klasse gehörig angesehen werden. Eine solche Klasse bildet ein Phonem. Ein Phonem ist eine Gruppe von Lauten (Phonen), die ähnlich klingen und niemals einen Bedeutungsunterschied bewirken. Die Phone innerhalb der Gruppe bezeichnet man als Allophone. Wenn zwei Phone zu einem Phonem gehören, so gibt es keine ansonsten identische Wörter, deren Bedeutung sich

durch Austausch der beiden Phone ändert. Kann man umgekehrt mindestens ein Wortpaar mit unterschiedlicher Bedeutung (Minimalpaar) angeben, so handelt es sich dementsprechend um zwei Phoneme. Die Laute [r] und [l] als Beispiel sind unterschiedliche Phoneme, da es z. B. das Paar *Ratte* und *Latte* gibt.

Auch die Dauer eines Lautes kann einen Bedeutungsunterschied bewirken. Als Beispiel sind – zumindest im Deutschen – kurzes und langes [a] unterschiedliche Phoneme, wie die Paare *Bann* [ban] und *Bahn* [ba:n] oder *Ratte* und *Rate* belegen. Da die Einteilung an die Bedeutung der Wörter geknüpft ist, hängt das Inventar an Phonemen vom betrachteten Wortschatz ab. Ein Standardbeispiel dazu ist die Zeichenfolge ch, zu der zwei Laute gehören: der Ich-Laut [ç] und der Ach-Laut [x]. In den allermeisten Wörtern kann nur eine der beiden Formen verwendet werden. Verwendet man nur diese Wörter, so können beide Laute zu einem Phonem zusammengefasst werden. Ein – zugegebenermaßen etwas konstruiertes – Gegenbeispiel ist das Paar *Kuhchen* (kleine Kuh) und *Kuchen*. Will man derartige Fälle auch abdecken, muss man die beiden Laute doch als zwei verschiedene Phoneme betrachten.

In seltenen Fällen wird durch die Betonung der Silben eines Wortes ein Bedeutungsunterschied ausgedrückt. Der Satz

Er wollte das Schild umfahren

kann je nach Betonung des Wortes *umfahren* zwei Bedeutungen haben:

1. Er wollte dem Schild ausweichen (Betonung auf der zweiten Silbe)
2. Er wollte das Schild zerstören (Betonung auf der ersten Silbe)

Aus diesen Betrachtungen wird deutlich, dass die Definition eines optimalen Inventars von Lauteinheiten schwierig ist. Die Entwickler von Systemen zur Spracherkennung gehen bei der Auswahl des Inventars oftmals recht pragmatisch vor. Das Ziel ist eine möglichst gute Modellierung der wesentlichen Laute. Daher können durchaus bei einem als „phonembasiert“ bezeichneten Erkenner die Einheiten von den im obigen Sinne definierten Phonemen abweichen. Eine Zusammenstellung der beschriebenen Einheiten enthält Tabelle 2.1. Zusätzlich ist die Größe des jeweiligen Inventars für Deutsch eingetragen.

Tabelle 2.1: Sprachliche Einheiten

Einheit	Inventar (D)
Satz	unbegrenzt
Wort	≈ 500000
Silbe	5000
Halbsilbe	2000
Phonem	≤ 50