5.7 Sprachsynthesesysteme

Bernd Möbius

Sprachsynthese wird überall dort eingesetzt, wo die Ausgabe von Information nur oder vorzugsweise auf akustischem sprachlichem Weg erfolgen kann. Derzeit wird die Sprachsynthese zunehmend in Auskunftsystemen (siehe Unterkapitel 5.13 und 5.10) eingesetzt. Hier sind die Anwendungsmöglichkeiten vielfältig: Navigationssysteme, Verkehrsmeldungen, Reiseauskünfte, Kinoprogramme, Börsenkurse, Webseiten, Email, und andere mehr. Insbesondere in der Mobiltelefon-Kommunikation, aber auch etwa im Auto, wo der Gesetzgeber oder die Vernunft des Fahrers eine Informationsausgabe auf einen Bildschirm untersagt, muss auf akustische Sprachausgabe zurückgegriffen werden. Klassische Anwendungen sind weiterhin der Computerarbeitsplatz für Blinde und Sehbehinderte oder die künstliche Stimme für Sprechbehinderte (siehe Unterkapitel 5.12).

Die übergreifende wissenschaftliche Theorie hinter der Sprachsynthese kann als ein funktionales Modell der menschlichen Sprachproduktion gelten. Unter diesem Aspekt kann die Ambition der Sprachsynthese als die Modellierung der wohl komplexesten kognitiven Fähigkeit des Menschen charakterisiert werden. So wenig perfekt dieses funktionale Modell ist, so wenig ist das Problem der optimalen Sprachsynthesequalität bislang gelöst.

5.7.1 Struktur eines TTS-Systems

Sprachsynthese (**text-to-speech, TTS**) kann als ein zweistufiger Prozess beschrieben werden. In einem ersten Schritt wird der Eingabetext linguistisch analysiert, und in einem zweiten Schritt wird die aus der Analyse resultierende linguistische Repräsentation in ein synthetisches Sprachsignal umgesetzt. Ein **Sprachsynthesesystem** (**TTS-System**) ist ein komplexes System, dessen Leistungsfähigkeit durch die Qualität der einzelnen Komponenten bestimmt wird, aus denen es besteht. Abbildung 5.10 zeigt die Hauptkomponenten, die in allen TTS-Systemen anzutreffen sind. Obwohl es durchaus Unterschiede in der Architektur verschiedener Systeme gibt, können sie im Allgemeinen auf die in der Abbildung gewählte „Pipeline“-Architektur zurückgeführt werden.

Infolge der nicht umkehrbaren Verarbeitungsrichtung lässt sich, anders als in einem Spracherkennungssystem (siehe Unterkapitel 5.8), eine lücken- oder fehlerhafte Verarbeitung durch eine TTS-Komponente nicht in einer späteren Komponente ergänzen oder korrigieren. Fehlanalysen pflanzen sich also durch das System fort und lösen oft Folgefehler aus. Da die in den Komponenten zum Einsatz kommenden linguistischen, phonetischen und akustischen Modelle nicht perfekt sind, führt die Verarbeitung im System zu einer zunehmenden Distanz der Qualität des synthetischen Sprachsignals zur Qualität der natürlichen Sprache. Auf die verschiedenen Verfahren zur Generierung des künstlichen Sprachsignals, d.h. insbesondere auf die zugrunde liegenden Modelle der Sprachproduktion und Artikulation, kann im Rahmen des vorliegenden Buches nicht eingegangen werden; diese Aspekte werden in *Dutoit* 1997 und *Sproat* 1998 eingehend behandelt.

Die folgenden Abschnitte konzentrieren sich auf diejenigen Komponenten und Verarbeitungsschritte in einem TTS-System, die aus dem schriftlichen Eingabetext eine linguistische Repräsentation herleiten und für die akustische Synthese bereitstellen.

5.7.2 Computerlinguistische TTS-Komponenten

Die in Abbildung 5.10 dargestellten Verarbeitungsblöcke bestehen üblicherweise aus mehreren, im Fall der linguistischen Textanalyse sogar oft aus einer Vielzahl von Modulen, von denen jedes einem wohldefinierten Teilproblem entspricht.

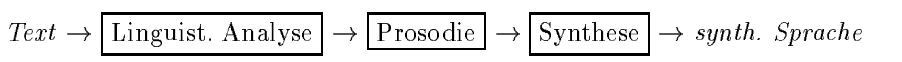

Abbildung 5.10: Hauptkomponenten von Sprachsynthesesystemen.

Zur Illustration der Komplexität der linguistischen Textanalyse soll der folgende Satz dienen:

Bei der Wahl am 12.3.1998 gewann Tony Blair ca. 52% der Wählerstimmen.

Welches Wissen muss ein Sprecher des Deutschen mitbringen, um ihn korrekt vorzulesen? Zunächst einmal muss er die Aussprache regulärer Wörter aus ihrer schriftlichen Form ableiten können. Dies setzt unter anderem die Kenntnis der internen Struktur von Wörtern voraus. So muss *Wählerstimmen* in die Komponenten *Wähler* und *Stimmen* zerlegt werden, um die Buchstabenfolge *st* korrekt als [ʃt] auszusprechen, im Unterschied etwa zu dem Wort *Erstimpfung*. Weiterhin sollte *Tony Blair* als ausländischer Name erkannt und idealerweise englisch ausgesprochen werden. Die Abkürzungen *ca.* und *%* sowie die Zahl *52* und das Datum *12.3.1998* schließlich müssen in reguläre Wortformen umgewandelt werden. Eine besondere Schwierigkeit ist, dass der orthographische Punkt beim ersten und zweiten Auftreten im Beispielsatz als Teil des Datums erkannt werden muss, im dritten Fall eine Abkürzung und im vierten Fall das Satzende markiert. Tatsächlich stellen sich dem Sprecher noch weitere Probleme, etwa die richtige Betonung von Wörtern und Silben sowie die Auswahl einer geeigneten Sprachmelodie oder Intonation.

Ein Problem der Textanalyse, das bislang nicht angesprochen wurde, ist die Zerlegung des Eingabetextes in Wörter. Dies ist selbst für das Deutsche, das Wörter in der Regel durch Leerzeichen voneinander trennt, keine triviale Aufgabe. So müssen die numerischen Ausdrücke *52%* und das Datum in mehrere separate Wörter expandiert werden. Wesentlich schwieriger verhält es sich in Sprachen wie dem Chinesischen oder Japanischen, in denen Wörter keine direkte orthographische Entsprechung haben und Wortgrenzen demnach auch nicht durch Leerzeichen markiert werden. Dennoch existieren in diesen Sprachen Wörter als lexikalische Einheiten, so dass die linguistische Analyse auch hier eine Wortsegmentierung vornehmen muss.

Die Expandierung von Symbolen wie *%* ist in einigen Sprachen ebenfalls komplexer als im Deutschen, wo es ausnahmslos als *Prozent* gesprochen wird. So ist etwa im Russischen eine Analyse des Satzzusammenhangs erforderlich, um die korrekte grammatische Form von *Prozent* zu ermitteln, da abhängig vom Kontext eine Vielzahl von Varianten möglich ist. Die Beispiele des russischen *%*, aber auch des Datums *12.3.1998*, zeigen, dass eine simple Vorverarbeitung oder Textnormalisierung, wie sie in manchen Systemen anzutreffen ist, unzureichend ist. Um Symbole, Abkürzungen und komplexe numerische Ausdrücke in die korrekten Wortformen zu expandieren, ist eine gründliche Analyse des Kontextes unumgänglich. Tag und Monat des Datums müssen als Ordinalzahlen ausgedrückt und in die mit der vorangehenden Präposition übereinstimmende grammatische Form (Dativ Singular) gesetzt werden: *am zwölften dritten*. Die Jahreszahl bedarf ebenfalls einer besonderen Behandlung: *neunzehnhundert achtundneunzig*, nicht *eintausend neunhundert achtundneunzig*.

Im Folgenden werden die aus computerlinguistischer Sicht wichtigsten Aspekte der Textanalyse näher betrachtet, und zwar die lexikalische und morphologische Analyse, wortübergreifende Sprachmodelle und die Ausspracheregeln. Schließlich wird ein einheitlicher Formalismus für die linguistische Repräsentation und für deren Implementierung im TTS-System vorgestellt.

Lexikalische Analyse

Die weitaus meisten TTS-Systeme verfügen über ein Lexikon, das zu jedem Eintrag Informationen über die Wortart und andere grammatische Kategorien und außerdem die Aussprache in Form einer phonetischen Transkription enthält. In vielen Fällen handelt es sich um ein Vollformenwörterbuch, d.h. es ist nicht nur jeweils die Grundform des Wortes aufgeführt, sondern auch die unterschiedlichen Wortformen.

Eleganter ist die Methode, für Wörter mit komplexer Flexionsmorphologie, im Deutschen also Nomina, Adjektive und Verben, Flexionsparadigmata oder Fortsetzungsklassen zu definieren und an jedem Wortstamm zu markieren, welches Paradigma zutrifft (siehe Unterkapitel 3.2). Die Expansion eines Wortstammes in alle legalen flektierten Wortformen kann dann automatisch und vollständig erfolgen.

Nichtflektierte und nicht abgeleitete Wortarten werden in einfachen Teillexika abgelegt. Spezielle Wortlisten lassen sich außerdem für Eigennamen, geographische Namen und ähnliche Kategorien sowie für die Expansion von Abkürzungen erstellen. Weiterhin verfügen TTS-Systeme oft über spezielle linguistische Modelle für die Behandlung von numerischen Ausdrücken. Als letzter Ausweg steht immer das Buchstabieren von Graphemsequenzen offen, die nicht weiter analysiert werden können.

Derivation und Komposition

Eine Besonderheit des Deutschen und einiger anderer Sprachen sind zusammengesetzte Wörter, also Komposita, wie *Wählerstimmen* in dem Beispielsatz. Die

Bildung von Komposita ist ausgesprochen produktiv: Sprecher des Deutschen können jederzeit neue Zusammensetzungen bilden. Dies hat zur Konsequenz, dass in nahezu jedem Text Wörter auftreten, die in keinem noch so umfangreichen Lexikon aufgelistet sind. Die linguistische Analyse muss daher in der Lage sein, Komposita und auch Derivationen in ihre Bestandteile zu zerlegen. Als Grundlage hierzu kann ein Modell der morphologischen Struktur von Wörtern und der Kombinierbarkeit von Morphemen dienen.

Ein solches Wortmodell könnte beispielsweise für *Wählerstimmen* folgende mehr oder weniger plausible Analysen liefern:

wähl [Vb-Stamm] + *erst* [Adj-Stamm] + *imme* [Nom-Stamm] + *n* [pl]
wähler [Vb-Stamm] + *st* [2per-sg] + *imme* [Nom-Stamm] + *n* [pl]
wähler [Nom-Stamm] + *stimme* [Nom-Stamm] + *n* [pl]

Die korrekte Lesart muss anhand von Wahrscheinlichkeiten, Kosten oder Auftretenshäufigkeiten in Korpora ermittelt werden, möglicherweise unterstützt durch eine Analyse des syntaktischen Kontextes.

Sprachmodelle und prosodische Analyse

Die lexikalische und morphologische Analyse liefert häufig alternative Lesarten, die erst durch lokale grammatische Sprachmodelle, die über die Wortgrenze hinaus den syntaktischen Kontext miteinbeziehen, disambiguiert werden können. Die häufigste Aufgabe für solche lokalen Grammatiken ist die Sicherstellung der syntaktischen Kongruenz (engl. *Agreement*) zwischen zusammengehörigen Wörtern.

Zu den wortübergreifenden Modellen gehören auch die syntaktische und prosodische Phrasierung und die Bestimmung des Satzmodus. Viele TTS-Systeme verfügen nur über Heuristiken, um diese Aufgaben zu bewältigen. Unter den TTS-Systemen für das Deutsche zeichnen sich das SVOX-System der ETH Zürich (*Traber* 1995) und das IMS-Festival-System der Universität Stuttgart (*IMS Festival* 2000) durch den Einsatz eines syntaktischen Parsers (siehe Unterkapitel 3.3) und Part-of-Speech-Taggers (siehe Unterkapitel 3.2) aus; die von diesen Modulen gelieferte Information bildet die Basis für die Festlegung von Phrasengrenzen und Akzenten.

Phonologische Analyse und Aussprache

In TTS-Systemen, die ein Vollformenwörterbuch verwenden, ist die Aussprache eines Wortes durch seine Transkription im Lexikon gegeben. Im Eingabetext auftretende Wörter, die nicht im Lexikon enthalten sind, werden durch Aussprachеregeln transkribiert. Solche Systeme zeichnen sich häufig durch eine Vielzahl von Ausnahmeregeln aus.

Eleganter ist hier ein Design der linguistischen Analysekomponente, die jedem Wort gerade so viel morphologische Annotation mitgibt, dass generische Aussprachеregeln eine zuverlässige Transkription liefern können. Bei im TTS-Lexikon vorhandenen Wörtern ist diese Information bereits gegeben, und für

„unbekannte“ Wörter liefert die Komposita- und Derivationsanalyse eine Granularität der Annotation, die der der bekannten Wörter äquivalent ist. Auf diese Weise werden Ausnahmeregeln weitestgehend überflüssig. Zur Aussprache eines Wortes gehört selbstverständlich nicht nur die Phonemfolge, sondern auch die Markierung der Silbenbetonung.

Im Deutschen hängt die Aussprache vorrangig von der morphologischen Struktur eines Wortes und erst danach von der Silbenstruktur ab. So wird in *Tonart* die Standard-Syllabifizierung (/to:-nart/) durch die Morphemgrenze außer Kraft gesetzt (/to:n+art/). Eine Syllabifizierung der ermittelten Phonemfolge muss dennoch vorgenommen werden, da die akustischen prosodischen Komponenten des TTS-Systems, also die Lautdauer- und Intonationsmodule, die Silbenstruktur als Eingabeinformation benötigen.

5.7.3 Ein einheitlicher Formalismus

Die Vielfalt der Probleme, die sich in den verschiedenen Sprachen im Zusammenhang mit der linguistischen Analyse stellen, scheint zunächst gegen eine generelle Lösung zu sprechen. Es ist jedoch möglich, die Problematik in einer abstrakteren Weise zu betrachten als in den angeführten Beispielen geschehen. Jedes Teilproblem kann als Transformation von einer Kette von Symbolen (konkret: Schriftsymbolen) in eine andere Kette von Symbolen (konkret: linguistische Analyse) beschrieben werden. So wird etwa die Buchstabenfolge *Wählerstimmen* in eine linguistische Repräsentation überführt, die nun auch Informationen über die Struktur des Wortes enthält: *wähler* [Nom-Stamm] + *stimme* [Nom-Stamm] + *n* [pl]. Auf vergleichbare Weise wird eine Folge von Schriftzeichen in einem chinesischen Satz in eine Darstellung überführt, die unter anderem Informationen über Wortgrenzen enthält.

Analog lässt sich auch der nächste Schritt im Rahmen der Textanalyse beschreiben, nämlich die Bestimmung der Aussprache von Wörtern. Dabei nutzen die Ausspracheregeln für eine bestimmte Sprache die aus der linguistischen Analyse gewonnenen Informationen und konvertieren die linguistische Repräsentation in eine Folge von Lautsymbolen. So ermöglicht erst die Information über die wortinterne Grenze vor *st* in *Wählerstimmen* die Bestimmung der korrekten Aussprache des Wortes.

Ein flexibles und zugleich mathematisch elegantes Modell, das die soeben skizzierte Konvertierung von Symbolketten erlaubt, beruht auf der Technologie der *Finite State Transducer* (FST, siehe Unterkapitel 2.2). Ein FST ist ein endlicher Automat, der eine Eingabe-Zeichenkette erkennt und daraus eine Ausgabe-Zeichenkette erzeugt. Ein solcher Automat enthält eine endliche Anzahl von Zuständen; für jeden dieser Zustände bestimmt eine Tabelle, zu welchen anderen Zuständen Übergänge möglich sind, und zwar in Abhängigkeit davon, welche Eingabesymbole gerade verarbeitet werden. Die Tabelle bestimmt auch, welche Symbole daraufhin ausgegeben werden.

Ein einfacher Transducer mitsamt seiner Übergangsfunktion als Tabelle sind im Beitrag über Automatentheorie (2.2) dargestellt. Der dort gezeigte Transducer enthält nur zwei Zustände; es ist einleuchtend, dass ein Transducer, der

eine komplexe Aufgabe wie die linguistische Analyse in einem Sprachsynthesesystem übernehmen soll, über eine sehr viel größere Zahl (typischerweise einige hunderttausend) von Zuständen verfügt.

Die linguistische Analysekomponente im multilingualen TTS-System der Bell Labs (*Sproat* 1998) ist vollständig nach diesen Prinzipien konstruiert und verarbeitet viele der Phänomene und Probleme, die in den einzelnen Sprachen im Rahmen der Textanalyse auftreten, einschließlich der verschiedenen Schriftsysteme (lateinisch, kyrillisch, chinesisch, japanisch). Die hier skizzierte einheitliche Software-Architektur für multilinguale Sprachsynthese ermöglicht eine vergleichsweise einfache Erweiterung auf neue Sprachen, und ihre modulare Struktur erleichtert die Integration verbesserter Komponenten für bereits existierende Systeme. Die linguistische Analysekomponente der deutschen Version dieses Systems ist in *Möbius* 1999 detailliert beschrieben worden. Endliche Automaten werden auch im TTS-System SVOX der ETH Zürich eingesetzt (*Traber* 1995).

5.7.4 Perspektiven

In diesem Artikel wurde bislang von der meistverbreiteten und zugleich ambitioniertesten Zielrichtung der Sprachsynthese ausgegangen, der Sprachsynthese für unbeschränkte Texteingabe und für unbeschränkte Anwendungsdomänen – also dem klassischen Vorleseautomaten. Synthetische Sprache kann jedoch aus recht unterschiedlichen Eingabeinformationen erzeugt werden. Die Eingabe kann maschinenlesbarer Text sein oder ein strukturiertes Dokument oder mit speziellen Steuerzeichen annotierter Text (es gibt eigens für die Sprachausgabe entwickelte *Markup Languages*) oder auch semantische Konzepte.

Unbeschränkte textbasierte Sprachsynthese stellt hier ein Extrem in einem Quasi-Kontinuum von Szenarien dar. Am anderen Ende des Kontinuums stehen Sprachausgabesysteme, die ein kleines Inventar abgespeicherter Sprachbausteine (z.B. Systemprompts oder wiederkehrende Phrasen) neu kombinieren und wiedergeben. Solche auf *canned speech* oder *sliced speech* basierende Systeme sind nur in strikt definierten und geschlossenen Anwendungsdomänen einsetzbar. Sie erfordern keine ernsthafte computerlinguistische Verarbeitung und sollen daher hier auch nicht weiter diskutiert werden.

TTS-Systeme müssen eine sehr große, ja unbegrenzte Anzahl möglicher Eingabesätze verarbeiten können. Sie benötigen hierzu linguistische und prosodische Modelle sowie ein akustisches Inventar, das die synthetische Sprachausgabe für eine solche Texteingabe in einer Qualität ermöglicht, die für die Benutzer des Systems akzeptabel ist. TTS-Systeme bieten so die größtmögliche Flexibilität, für die jedoch ein hoher Preis in Form reduzierter Natürlichkeit der Sprachausgabe zu zahlen ist.

Hingegen ermöglicht die **konzeptbasierte Sprachsynthese** (**concept-to-speech, CTS**), üblicherweise integriert in ein Dialog- oder Übersetzungssystem (siehe Unterkapitel 5.10, 5.13 und 5.14), die Generierung synthetischer Sprache auf der Grundlage pragmatischen, semantischen und Diskurs–Wissens. Der Vorteil gegenüber einem TTS-System ist, dass die sprachgenerierende Komponente des CTS-Systems „weiß“, was sie sagen will, ja sogar, wie es gesagt werden soll.

Sie weiß es, weil sie eine vollständige linguistische Repräsentation des Satzes selbst generiert. Die zugrunde liegende Struktur ist bekannt, die intendierte Interpretation ist möglicherweise verfügbar, und die entsprechende syntaktische Struktur ist ebenfalls bekannt.

In einem CTS-System ist der Umweg über eine Textgenerierung nicht nur unnötig, sondern hinderlich. Orthographischer Text ist eine stark verarmte Repräsentation der Sprache. Es wäre kontraproduktiv, zunächst eine vollständige linguistische Repräsentation einer sprachlichen Äußerung in orthographischen Text zu konvertieren, nur um dann aus diesem Text wieder eine linguistische Struktur zu berechnen, die gegenüber der ursprünglichen Struktur defizitär sein muss.

Da in den Schriftsystemen der meisten Sprachen die prosodischen Strukturen allenfalls rudimentär (Satzmodus und Phrasierung durch Interpunktion) wiedergegeben werden, erwartet man sich von einem CTS-System eine signifikante Verbesserung gerade der prosodischen Qualität der synthetischen Sprache. Allerdings ist die Beziehung zum einen zwischen der symbolischen Repräsentation der Intonation und ihrer akustischen Realisierung durch Grundfrequenzkonturen und zum anderen zwischen der symbolischen Repräsentation der Intonation und der Bedeutung, die sie ausdrücken soll, selbst noch Forschungsgegenstand. Die computerlinguistische Erforschung und Modellierung der Schnittstellen zwischen Pragmatik (siehe Unterkapitel 3.5), Semantik (3.4), Syntax (3.3) und der Prosodie kann somit entscheidend zu einer Verbesserung der synthetischen Sprachqualität in Dialogsystemen und anderen Anwendungen beitragen.

5.7.5 Literaturhinweise

Die beiden aktuellen Standardwerke zur Sprachsynthese, *Dutoit* 1997 und *Sproat* 1998, ergänzen sich in vielerlei Hinsicht optimal, auch aus der Sicht des Computerlinguisten. Beide Bücher behandeln alle Komponenten von TTS-Systemen, setzen dabei aber unterschiedliche Schwerpunkte: *Dutoit* 1997 bietet die vollständigste Darstellung der Signalverarbeitungsmethoden für die Sprachsynthese, während *Sproat* 1998 eine Fülle von linguistischen Textanalyseproblemen, und elegante Lösungen, für eine ganze Reihe von Sprachen bereithält. Einsatzmöglichkeiten für *Finite State Transducer* in der Sprachtechnologie (Synthese und Erkennung) werden von *Mohri* 1997 besprochen; dieser Artikel setzt allerdings fortgeschrittene Kenntnisse in der Automatentheorie voraus.