Universität zu Köln
Institut für Linguistik
Abteilung Sprachliche Informationsverarbeitung

Hausarbeit

Sprachsynthese

Seminar: Computerlinguistische Grundlagen

SoSe 2007

Dozent: Jürgen Hermes

Oskar Henry Solich
Karl-Jaspers-Str. 71
51377 Leverkusen
oskar@solich.de

Inhaltsverzeichnis

1 Einleitung

Sprachsynthese ist die Möglichkeit, Sprache zu erzeugen, auch wenn kein Sprecher vorhanden ist. Es finden sich zahlreiche Anwendungsgebiete, in denen dies durchaus von Vorteil ist:

So kann man beispielsweise *Informationen über das Telefon* abfragen, die nur in schriftlicher Form vorliegen, etwa Informationen zu Kulturveranstaltungen wie Theater oder Kino, E-Mails, Faxe, oder auch Datenbankeinträge verschiedenster Art. Ein nicht zu vernachlässigendes Gebiet sind *Hilfen für Sprachgestörte und Blinde*, wie etwa sprechende Browser oder Sprache erzeugende Tastaturen. Bei *Mess- und Kontrollsystemen* kann es ebenfalls von Vorteil sein, wenn die Informationen lautlich wiedergegeben werden, da sie so schneller abgerufen werden können und man sich nicht auf das Display konzentrieren muss. *Übersetzungs- und Lernprogramme* können gut durch ein Dialogsystem unterstützt werden, in dem nicht nur Schrift, sondern auch die Aussprache gelernt wird. Außerdem eignen sich Sprachsynthesesysteme auch zur *Erforschung der Sprache*.

Am häufigsten anzutreffen sind die sogenannten TTS-Systeme (text-to-speech), bei denen der zu synthetisierende Text bereits in schriftlicher Form vorliegen (oder entsprechend erzeugt) werden muss. Wie ein solches System funktioniert, werde ich im Laufe meiner Arbeit erläutern.

Daneben gibt es noch die *CTS-Systeme (concept-to-speech)*.

„Als eine künftige Anwendung der Sprachsynthese zeichnet sich ihr Einsatz in Sprachdialogsystemen, Auskunftssystemen und automatischen Ansagediensten ab [...]. Sofern nicht vorgefertigte Texte verwendet werden, wird die Sprachsynthese bei dieser Anwendung von einer Sprachgenerierungsstufe angesteuert. Deren Eingangsinformation besteht in einer abstrakten semantischen Repräsentation des auszugebenden Inhalts. Bei herkömmlichen Realisierungen wandelt die Generierungsstufe diese in natürlichsprachlichen Text, der wiederum an die Sprachsynthese weitergereicht wird.“ (Hess, 2002: S. 28)

Dies geht allerdings auf Kosten des Sprachsignals, weshalb inzwischen entsprechende *CTS-Systeme (concept to speech)* entwickelt werden. Da sich diese aber noch in der Entwicklungsphase befinden, werde ich im Verlauf meiner Arbeit nicht näher darauf eingehen.

2 Geschichte der Sprachsynthese

Christian Gottlieb Kratzenstein (1723-1795) und Wolfgang von Kempelen (1734-1804) waren die ersten Wissenschaftler, die Versuche unternommen hatten, Sprachlaute mechanisch zu erzeugen. Kratzenstein entwickelte im Zuge eines Wettbewerbs an der Petersburger Akademie eine Maschine, die mit Orgelpfeifen und daran angeschlossenen Resonanzröhren Vokale erzeugen konnte.

Nach dem Studium der menschlichen Sprechorgane entwickelte von Kempelen diese Maschine weiter und veröffentlichte 1791 die Schrift 'Mechanismus der menschlichen Sprache nebst der Beschreibung einer sprechenden Maschine'.

Mit seiner Sprechmaschine konnte von Kempelen nicht nur Vokale, sondern auch einige Konsonanten erzeugen und damit einfache Wörter wiedergeben.

„Im 19. Jhdt. wurden zwar einige weitere Maschinen ähnlicher Art konstruiert, aber grundsätzliche Neuerungen auf dem Gebiet der Sprachsynthese sind für dieses Jahrhundert eigentlich nicht zu verzeichnen. Erwähnenswert ist aber das von Joseph Faber in 1835 vorgestellte Gerät, das im Vergleich mit Kempelens Maschine insofern einen Fortschritt darstellte, als es auch eine Zunge und einen formveränderlichen Rachenraum hatte und außerdem zur Synthese von Gesang geeignet war. Sein Blasebalg wurde über ein Fußpedal getrieben, und die sonstige Bedienung erfolgte über eine Klaviatur.“ (Traunmüller, 1997)

Erst etwa ein Jahrhundert später gab es mit dem VODER (Voice Operation DEmonstratoR) von Homer Dudley einen entscheidenden Fortschritt. Es war das erste Gerät, das Sprachschall elektronisch erzeugen konnte und war zum ersten Mal 1939 auf der Weltausstellung in New York zu sehen. Der Nachteil an dieser Maschine war jedoch, dass eine lange Übungszeit nötig war, um sie bedienen zu können.

Parallel zu Dudleys VODER entwickelte Frank Cooper in den Haskins Labs den *Pattern Playback,* den er 1950 fertigstellte. Dieses Gerät konnte Spektrogramme[1] lesen und in Ton umwandeln. Diese konnten zuvor mit einem Sonagraphen aufgenommen oder per Hand selbst gezeichnet werden.

Mit Verbreitung der Computer wurde die Entwicklung der Sprachsynthese zunehmend mit der elektronischen Datenverarbeitung verknüpft. Während die Geräte, die zuvor gebaut worden waren nur zu Forschungs- und Veranschaulichungszwecken entwickelt worden waren, konnten die neueren Entwicklungen auch praktisch genutzt werden.

1 Das Spektrogramm, auch Sonagramm genannt, ist ein Diagramm, das Frequenz, Intensität und Zeit im Verhältnis zueinander darstellt.

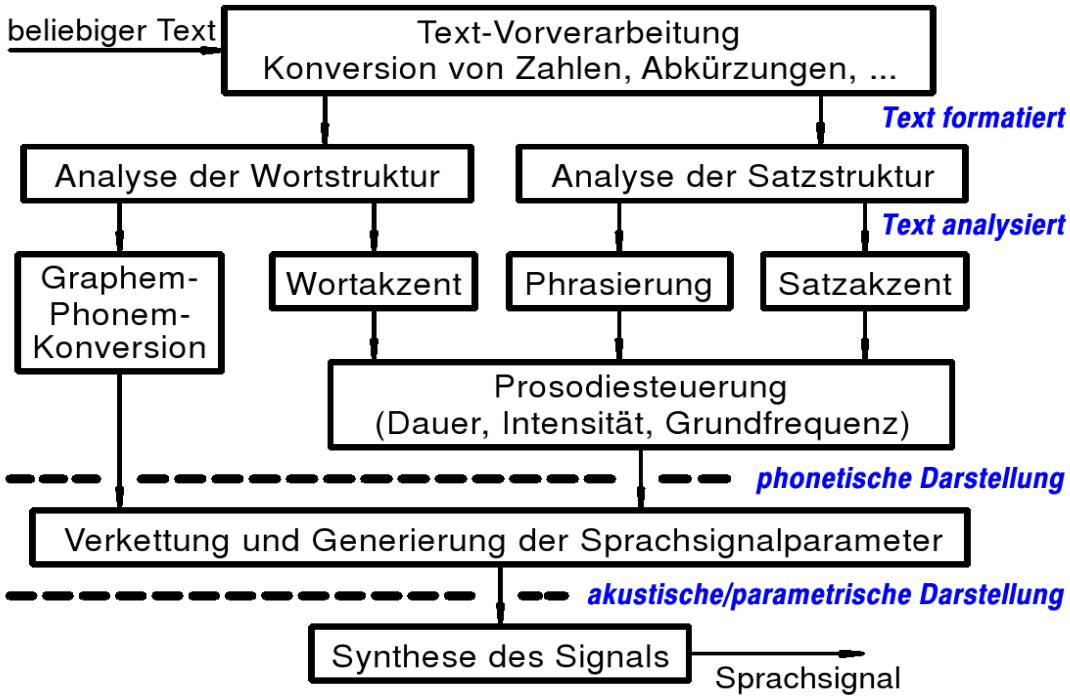

Bild 3.1. Blockdiagramm eines TTS-Systems

(Hess, 2002: S. 1)

Ein Text-To-Speech-System (TTS-System) besteht im wesentlichen aus drei Teilen, der *Symbolverarbeitung*, der *Verkettung* und der akustischen *Synthese,* in der Graphik durch die gestrichelten Linien voneinander getrennt.

Im ersten Teil müssen zunächst Zahlen und Abkürzungen erkannt und umgewandelt werden. Einige bessere Systeme filtern auch fremdsprachliche Ausdrücke und Namen heraus, um diese separat zu behandeln. Anschließend werden die Wörter in eine Lautschrift umgesetzt und Wortakzente gesetzt. Während die Akzentuierung der Wörter bei einigen Sprachen, wie dem Polnischen, regelmäßig ist, muss sie bei deutschen Sprachsynthesesystemen dem Lexikon entnommen werden. In einem weiteren Schritt, der parallel zur Umwandlung der Buchstaben in Lautschriftzeichen erfolgen kann, wird eine Analyse der Satzstruktur vorgenommen. Das heißt, der Text wird in Sätze und Phrasen eingeteilt und die Satzakzente ermittelt. Zusammen mit den Wortakzenten wird die Prosodie, also die Satzmelodie bestimmt, die aus Dauer, Intensität und Grundfrequenz besteht.

Die Informationen aus der Symbolverarbeitung müssen in einem zweiten Teil miteinander verkettet werden, damit sie schließlich im dritten Teil in Ton umgesetzt, also akustisch synthetisiert zu werden.

3.1 NLP (natural language processing)

Das *NLP* ist eins von zwei Modulen, das sich in modernen Sprachsynthesesystemen findet und umfasst die Symbolverarbeitung und die Verkettung. Es beinhaltet eine *morphosyntaktische Analyse*, die die Identifizierung von Wortklassen zur Aufgabe hat und den Input-Satz in Phrasen aufgliedert.

Ein *LTS (letter-to-sound)* und ein *Prosodie-Generator* liefern die Folge der zu sprechenden Phoneme, wie auch deren Dauer und Intonation.

Eine *morphosyntaktische Analyse* besteht aus einem morphologischen Analysemodul, dem Teil, das für jedes Wort sämtliche möglichen Wortklassen und die entsprechenden Aussprachemöglichkeiten auswählt. Flektierte und derivatisierte Wörter, sowie Kompositionen werden in ihre einzelnen Morphe gespalten.

Außerdem enthält die Analyse auch ein Modul für die kontextuelle Analyse, das Wörter in ihrem Zusammenhang betrachtet, was zu einer Verkleinerung der Liste der möglichen Wortklassekategorien für jedes Wort führt.

Das letzte Modul der Analyse ist ein Syntax-Prosodie-Parser, der die Wörter in einer Phrasenstruktur hierarchisch anordnet, was für die Intonation sehr wichtig ist.

Das *LTS* ist für die automatische Ermittlung der Lautübertragung des Textes verantwortlich. Auf den ersten Blick erscheint diese Aufgabe so einfach wie das Nachschlagen in einem Wörterbuch. Bei näherer Betrachtung stellt man jedoch fest, dass die meisten Wörter in einer gesprochenen Rede in Variationen vorkommen, die in keinem Aussprachewörterbuch zu finden sind.

So enthalten solche Wörterbücher lediglich Wortwurzeln und zählen nicht explizit alle morphologischen Variationen auf die sich z. B. durch Numerus oder Konjugation ergeben. Mit diesen Variationen beschäftigt sich die Morphophonologie, ein Teilbereich der Phonologie.

Außerdem finden sich Wörter mit mehreren Einträgen im Wörterbuch, beziehungsweise zu unterschiedlichen morphologischen Komponenten, oft mit jeweils unterschiedlicher Aussprache. In deutschen Texten kommt noch erschwerend hinzu, dass sich Wörter teilweise nur durch die Betonung unterscheiden können (z. B. Ténor vs. Tenór)

Wörter im Satz werden anders ausgesprochen als isolierte. Dies kann sowohl die Wortgrenzen bei Koartikulation betreffen (z. B. an Gas), als auch innerhalb des Wortes, dort vor allem Betonung und Rhythmik.

Und nicht alle Wörter können im Wörterbuch gefunden werden, dies gilt vor allem für Neologismen und Namen.

Für automatische LTS müssen Wege gefunden werden, wie diese Probleme gelöst werden können. Oft werden hier *wörterbuchbasierte (dictionary-based)* oder *regelbasierte (rule-based)* Strategien genutzt.

Wörterbuchbasierte Lösungsansätze versuchen, möglichst viel phonologisches Wissens in einem Lexikon zu speichern. Eintragungen werden dabei oft auf Morpheme eingeschränkt. Die Aussprache der Lemmata wird erklärt, indem sie durch Flektions-, Derivations- und morphonemische Regeln zusammengesetzt werden, die beschreiben, wie die Laute der morphemischen Bestandteile geändert werden, wenn sie in Wörter kombiniert werden. Morpheme, die nicht im Lexikon gefunden werden können, werden durch Regeln übertragen. Nach einer ersten phonemischen Übertragung eines jeden Wortes, werden die Laute nachbearbeitet, um der Koartikulation zu berücksichtigen.

Eine deutlich andere Strategie wird bei den regelbasierten Lösungsansätzen verfolgt. Hierbei wird versucht, die meisten der phonologischen Merkmale aus dem Wörterbuch in Buchstabe-zu-Ton- oder Graphem-zu-Phonem-Regeln umzusetzen. In diesem Fall werden nur jene Wörter gespeichert, die sich nicht durch Regeln darstellen lassen, darunter fallen viele Fremdwörter und Namen. Sie werden in einem Ausnahmelexikon gespeichert.

Zum Schluss dieses Abschnitts will ich noch auf den *Prosodie-Generator* eingehen. Der Ausdruck Prosodie bezieht sich auf bestimmte Eigenschaften des Sprachsignals, die mit hörbaren Änderungen in Tonhöhe, Lautstärke und Silbenlänge zusammenhängen. Prosodische Eigenschaften haben spezielle Funktionen in der Sprache. Der offensichtlichste Effekt der Prosodie ist der des Fokus. Es gibt bestimmte Ereignisse, die eine Silbe innerhalb der Äußerung verändern, und indirekt das Wort oder die syntaktische Gruppe, zu der es gehört, als wichtiger oder neuer Bestandteil in der Bedeutung dieser Äußerung hervorheben.

Der Prosodie-Generator sorgt dafür, dass die oben genannten Eigenschaften auch beachtet werden können.

3.2 DSP *(digital speech processing)*

Das *DSP* ist das zweite Modul, das sich in modernen Sprachsynthesesystem findet. Hier wird das Sprachsignal erzeugt, was auf drei unterschiedliche Arten umgesetzt werden kann.

Bei der *Konkatenationssynthese (Non-Uniform unit selection)* wurden zunächst komplette Wörter und Wortteile aufgenommen, die miteinander verkettet werden. Der Vorteil an dieser Art der Synthese ist, dass sie sich sehr natürlich anhört. Allerdings benötigt die Datenbank, in der die Aufnahmen gespeichert werden, sehr viel Speicher, weshalb sie nur für zweckgebundene Anwendungen geeignet sind.

Dagegen braucht die *Diphon-Synthese* deutlich weniger Speicher, da nur Diphone aufgenommen werden. Diphone sind Lautabschnitte, die von der Mitte eines Lautes bis zur Mitte eines anderen Lautes gehen. Der Grund, dass man nicht einfach die Laute selbst aufnimmt, liegt darin, dass jeder Laut von seiner Umgebung beeinflusst wird. Ohne diese Beeinflussung hört sich ein Text abgehackt und unverständlich an. Nachteil an der Diphonsynthese ist allerdings, dass sie sich unnatürlich anhört. Viele Systeme enthalten Elemente beider oben genannter Synthesearten.

Außerdem gibt es noch die *Formant-Synthese*, die zur Zeit allerdings nur zu Forschungszwecken verwendet wird. Bei dieser Synthese existieren überhaupt keine Aufnahmen, die Wörter werden allein mit Hilfe der Vokal- und Nasalformanten erzeugt. Diese Formanten sind Resonanzfrequenzen des Vokaltraktes, die lautspezifisch sind. Plosive, und im abgeschwächten Maße auch Frikative, erzeugen sogenannte Formantabbiegungen, die die Formanten 'beugen', womit die bereits beschriebene Beeinflussung der Laute untereinander erzeugt werden kann. Die sehr synthetisch klingende und schwer verständliche Stimme, die dabei erzeugt wird, ist für den praktischen Gebrauch allerdings noch zu schlecht, auch wenn darin möglicherweise die Zukunft der Sprachsynthese liegen könnte. Zur Forschung in der Phonetik ist diese Syntheseart allerdings bestens geeignet, weil sie die Erkenntnisse dieser wissenschaftlichen Disziplin direkt lautlich umsetzen kann.

3.3 Beispiel

Im Folgenden mache ich den Vorgang der Synthese an einem fiktiven Beispiel deutlich.

Beispielsatz:

'Eine alte Dame kauft 10 Gürtelschnallen für ca. 2€.'

Zunächst erfolgt die Vorverarbeitung. Dabei werden Zahlen, Sonderzeichen und Abkürzungen umgewandelt, um später besser verarbeitet zu werden.

'Eine alte Dame kauft zehn Gürtelschnallen für circa zwei Euro.'

Nun erfolgt zunächst eine morphologische Analyse. In unserem Beispiel haben wir dabei keine Doppeldeutigkeiten.

'Ein-e alt-e Dame kauf-t zehn Gürtel-schnalle-n für circa zwei Euro.'

Um die morphosyntaktische Analyse abzuschließen muss der Satz nun geparst werden. Die kontextuelle Analyse kann man in diesem einfache Beispiel auslassen, da sie hier nichts bewirken würde.

'(Ein-e alt-e Dame) (kauf-t ((zehn Gürtel-schnalle-n) (für (circa (zwei Euro))))).'

Bei einem lexikonbasiertem System würde nun für jedes Morphem im Lexikon nachgesehen werden, wie das jeweilige Wort aussieht und anschließend über Regeln verändert werden, während bei einem regelbasierten System nur das Wort 'circa' im Lexikon zu finden wäre, da sich die Aussprache der anderen Morpheme herleiten läßt.

lexikonbasiert:[2]

'?ajn @ '?alt_t @ 'da:m@ 'k_tawf t_t 'tse:n 'gYR\t_t@l 'Snal@ n fy:6 'tsIR\k_ta 'tsvaj '?OjR\o

Eine Regel würde an dieser Stelle sein können [t_t ts] wird zu [ts]:

'?ajn @ '?alt_t @ 'da:m@ 'k_tawf 'tse:n 'gYR\t_t@l 'Snal@ n fy:6 'tsIR\k_ta 'tsvaj '?OjR\o

Da die regelbasierte Variante den Rahmen etwas sprengen würde, übergehe ich diese und mache direkt weiter mit dem Prosodie-Generator. Dieser prüft anhand der Satzstruktur, wie die Akzentuierung aussehen muss. Bei Betrachtung eben dieser Struktur fällt auf, dass sich der Satz bereits mit zwei Prosodieregeln wesentlich verbessern lässt. Die erste wäre eine Erhöhung der Tonhöhe und leichte Anhebung der Silbenlänge der jeweils ersten Silbe in den beiden Hauptphrasen ([?ajn] und [k_tawf]) und die zweite wäre eine Absenkung der Tonhöhe auf das letzte Wort ([?OjR\o]).

2 Im Folgenden verwende ich als phonetische Zeichen das SAMPA, einem an das IPA angelehnte Alphabet, das in der Sprachsynthese häufig verwendet wird, da es anders als das IPA keine Sonderzeichen benötigt.

4 Fazit

Der Beginn der Sprachsynthese liegt im 18 Jahrhundert und hatte ursprünglich nur den Zweck, der Forschung und der Unterhaltung zu dienen. Erst mit der Erfindung des Computers wurde sie auch für andere Zwecke nutzbar und hat heute ein breites Anwendungsgebiet. Dennoch steht die Entwicklung erst am Anfang. Nach heutiger Auffassung beinhaltet die Sprachsynthese mehrere Module, die teilweise jedoch sehr unterschiedlich realisiert werden. Für den praktischen Nutzen kann zur Zeit nur das TTS-System genutzt werden, dem bereits Aufnahmen in einer Datenbank vorliegen, bei dem der Text bearbeitet und geparst werden muss um letztendlich in gesprochene Sprache umgewandelt werden kann. Besonders komplex gestaltet sich dabei die Prosodie, die im Fachgebiet der Phonetik und Phonologie angesiedelt ist. Auch die Entwicklung der letzten Jahrzehnte hat gezeigt, dass die Sprachsynthese vor allem von Phonetikern weiterentwickelt wurde und wohl auch in Zukunft wird.

Bibliographie

BLACK, Alan W., Kevin A. LENZO. 2007 *Building Synthetic Voices.* http://festvox.org/bsv/ 18. Sep. 2007

BREIDBACH, Günter. 1985. *Zur Sprachsynthese von deutschstämmigem Schrifttext mit Hilfe von Phonemklustern und dem LPC-Spracherzeugungsmodell*. Berlin: Papyrus-Druck GmbH.

BRINCKMANN, Caren. 2004. *Improving Prosody Prediction for Speech Synthesis*. (Phonus, 10. 2006). Saarbrücken.

BURKHARDT, Felix. 2007. *Deutsche Sprachsynthese*. http://ttssamples.syntheticspeech.de/deutsch/index.html 18. Sep. 2007

CARSTENSEN, K. U., C. Ebert, u. a. 2004. *Computerlinguistik und Sprachtechnologie.* Heidelberg, Berlin: Spektrum.

HESS, Wolfgang. 2002. *Systeme der akustischen Mensch-Maschine-Kommunikation.* http://www.ikp.uni-bonn.de/dt/lehre/materialien/sammk/sam_3.pdf 18. Sep. 2007

MITKOV, R. 2003. *The Oxford Handbook of Computational Linguistics.* Oxford, New York: Oxford University Press.

MÜLLER, Bernd S. (Hrsg.). 1985. *Sprachsynthese*. (Germanistische Linguistik, 79-80). Hildesheim: Georg Olms Verlag.

TRAUNMÜLLER, Hartmut. 1997. *Geschichte der Sprachsynthese.* http://www.ling.su.se/staff/hartmut/kempln.htm 18. Sep. 2007

2003. *Sprachsynthese*. http://www.logox.de/sprachsynthese.php 18. Sep. 2007