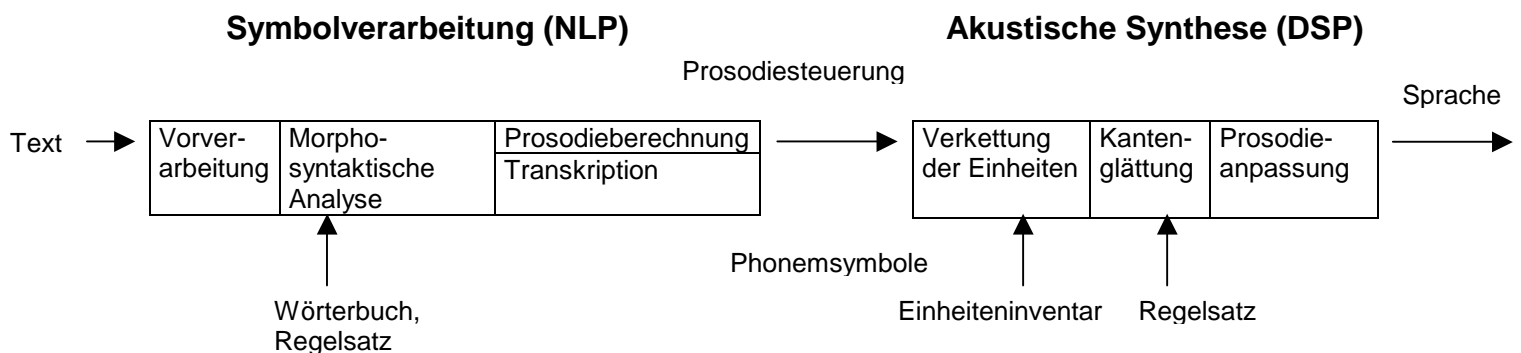


A-21 Sprachsynthese

Spezifikationen für moderne Synthesizerprodukte

Grundsätzlich bestehen alle Sprachsynthesizer aus mindestens zwei Komponenten: der Symbolverarbeitung, bei der eine Generierung der Prosodie (Sprechmelodie, Sprechrhythmus, insbesondere Intonation, Lautdauer und Lautintensität) sowie eine phonetische Transkription erfolgt und der akustischen Synthese, bei der das eigentliche Sprachsignal generiert wird. Diese beiden großen Teilbereiche werden auch als NLP (Natural Language Processing) und DSP (Digital Speech Processing) bezeichnet.



Symbolverarbeitung (NLP)

Der Text der synthetisiert werden soll wird zuerst durch die Symbolverarbeitung in eine phonetische Darstellung umgewandelt. Dazu wird noch eine Prosodiebeschreibung produziert. Zur Erfüllung dieser Aufgabe ist eigentlich eine vollständige syntaktisch-morphologische Analyse des eingegebenen Textes notwendig. Auch wenn dies erfüllt ist, kann eine natürlich wirkende Prosodie eigentlich nur durch Wissen über den semantischen und pragmatischen Gehalt des Eingabetextes generiert werden. Dies betrifft im speziellen die automatische Simulation emotionaler Sprechweise.

Bei den meisten TTS-Systemen (Text-to-Speech-Systeme) ist nicht einmal ein syntaktischer Parser implementiert. Der Grund dafür ist, dass die Systeme ja in Echtzeit laufen sollen und die natürliche Sprache eine hohe Komplexität besitzt.

Im Gegensatz zur früheren sequentiellen Abarbeitung von Verarbeitungsschritten, verwenden Sprachsynthesysteme heute komplexe Datenstrukturen, welche eine Erfassung der Information aus vielen verschiedenen Modulen erlaubt.

Im NLP-Modul kann man die Verarbeitung in mehrere Schritte untergliedern:

In der Vorverarbeitung werden Abkürzungen, Ziffern und Sonderzeichen in orthographische Form gebracht. Grosse Probleme gibt es hier vor allem in der Vieldeutigkeit von Satzzeichen, Zahlenangaben und der Aussprache von Abkürzungen. Aus dem Text entsteht eine Kette von Wörtern in orthographischer Form, welche zu Sätzen gruppiert sind. Auch Umformatierungen finden hier statt, um etwa einen Tabelleninhalt verständlich vorlesen zu können

Die aus dem ersten Schritt erhaltene Wortliste wird, falls keine Analyse der Syntax erfolgt, mindestens basierend auf einem Morphemlexikon in Funktions- und Inhaltswörter unterteilt. Die Funktionswörter (z.B. Partikel, Präpositionen, Konjunktionen) sind meist in Lexika aufgeführt, da sie in der Regel nicht allzu viele sind. Die potentiell unendlich vielen Inhaltswörter (z.B. Substantive, Verben, Adjektive, Adverbien), werden einer morphologischen Analyse unterzogen.

Die Wortebenenanalyse liefert sehr viele Falschklassifikationen, weshalb eine kontextuelle Analyse notwendig ist.

Jetzt findet die phonetische Transkription statt. Besonders schwierig ist dabei die Transkription von Eigennamen. Meistens werden dafür HMM (Hidden Markov Models), Entscheidungsbaumverfahren oder selten künstliche neuronale Netze verwendet. Nach der Klassifizierung der Wörter nach grammatischer Funktion, Stellung im Satz und eventuell semantischem Gehalt, erfolgt die Berechnung der Prosodieparameter. Eine natürlich wirkende Prosodie erhöht die Verständlichkeit eines Satzes. Zum Beispiel der Satz: „Er sah den Mann mit einem Fernrohr“ kann auf mehrere Arten verstanden werden, wenn die Betonung nicht klar ist.

Akustische Synthese (DSP)

Die akustische Synthese generiert aus der lautlichen und prosodischen Beschreibung das Sprachsignal. Dabei wird die Kette von phonetischen Symbolen in einen kontinuierlichen Datenstrom von Signalabstastwerten oder -parametern überführt. Je nach Synthesemethode werden entweder aus Parametern Sprachsignale berechnet und/oder Glättungsregeln auf die Übergangsstellen der Einheiten angewendet. Zusätzlich wird zur Anpassung der Prosodie das Ausgabesignal manipuliert bzw. die Parameter angepasst.

Die akustische Synthese kann in zwei Hauptansätze unterteilt werden: Regelbasierte – und datenbasierte Synthese. Bei der Regelbasierten Synthese ist das phonetische Wissen wie Sprache erzeugt wird, explizit in Regeln gespeichert. Um einen kontinuierlichen Sprachfluss zu erzeugen, wird zwischen den Werten einzelner Parameter unter Berücksichtigung bestimmter Regeln interpoliert.

Datenbasierte Systeme generieren Sprache aus Grundeinheiten, die als Inventar im System vorliegen und miteinander verknüpft werden. Ein Grossteil des phonetischen Wissens ist implizit in den Einheiten kodiert, weshalb die resultierende Sprachqualität zunächst höher ist. Bei diesem Ansatz ist die Wahl der Einheit entscheidend. Als Einheit werden Silben, Halbsilben, Diphone, Phone oder Phonemkluster verwendet. Je länger die gewählten Segmente sind, desto weniger Schnittstellen gibt es in der resultierenden Sprachausgabe. Allerdings wächst bei gross gewählten Segmenten auch die Anzahl die man abspeichern muss. Somit muss ein Kompromiss gefunden werden. Denkbar wäre auch das Abspeichern der Einheiten in verschiedenen emotionalen Zuständen wie angespannt, gelassen oder wütend. Noch vor einigen Jahren waren die regelbasierten Systeme in der Überzahl, da sie weniger Speicherplatz benötigen. Inzwischen wurden sie aber von den datenbasierten System überholt, weil diese eine höhere Qualität garantieren.

Literatur

T. Dutoit. *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers, 1997
 O’Shaughnessy. *Speech Communication: Human and Machine*. Addison-Wesley, 1987.