
Automatische Erkennung von Akzentuierungen und Phrasierungen in Sprachsynthesekorpora

Diplomarbeit im Fachbereich Informatik der Universität Hamburg

Timo Baumann <mail@timobaumann.de>

2009-01-07, 23:52

Kurzfassung der Diplomarbeit

Prosodieannotierung, die symbolische Kennzeichnung des Sprechverlaufs, ist ein grundlegender Schritt der Aufbereitung von Sprachsynthesedaten. Sie dient zum einen der datengetriebenen Modellierung der Prosodie des aufgenommenen Sprechers einschließlich idiolektaler Eigenschaften. Außerdem erlaubt sie modernen Sprachsynthesystemen, Akzentuierungsgrad und Junktur bei der Einheitenauswahl mitzuberechnen, was die Natürlichkeit des Resultats erhöhen kann.

Die hier zusammengefasste Arbeit beschäftigt sich mit automatischer Prosodieannotierung mittels Methoden des maschinellen Lernens. Für jede Silbe soll entschieden werden, ob sie akzentuiert ist oder nicht. Für Wortgrenzen wird entschieden, ob eine Phrasengrenze (ggfs. welcher Art) vorliegt oder nicht. Dafür werden Klassifizierer trainiert und auf unabhängigem Testmaterial evaluiert. Auf eine weitergehende Prosodiemodellierung (die zum Beispiel die Ergebnisse der Klassifizierung im Zusammenhang integriert) wird bewusst verzichtet um so die Möglichkeit zu bewahren, solche Modelle mit den hier ermittelten Ergebnissen zu überprüfen.

Der Begriff der Prosodie wird vor dem Hintergrund der Differenzierung in Phonologie und Phonetik betrachtet. Auf phonologischer Ebene werden die drei Teilsysteme Akzent, Junktur und Intonation angenommen. Der Akzent bildet ein (binäres) Silbenmerkmal und ist lexikalisch festgelegt. Die Junktur beschreibt den Grad der Zusammengehörigkeit benachbarter Äußerungsteile und hilft damit bei der Dekodierung und Gliederung des Sprechstromes. Die Intonation dient der Bedeutungsauswahl und -differenzierung durch die Zuordnung einer Sprechmelodie. Die Intonation selektiert zu realisierende Akzente und versieht sie mit Tönen, die die Sprechmelodie beschreiben. Zudem stützen sogenannte Phrasierungstöne die Gliederungsfunktion der Junktur. Die phonologischen Teilsysteme haben einen gemeinsamen Einfluss auf die phonetischen Phänomenbereiche Akzentuierung und Phrasierung, die sich wiederum auf die phonetisch-prosodischen Merkmale Intensität, Melodieverlauf, Dauer und Pause auswirken, deren vielfältige akustische Korrelate schließlich im Sprachschall hör- und messbar sind. Im praktischen Teil der Arbeit werden die akustischen (und weitere) Merkmale genutzt um den Akzentuierungs- und Phrasierungsstatus von Silben bzw. Wortgrenzen zu bestimmen.

Als Datengrundlage dieser Arbeit dienen mehrere prosodieannotierte Sprachsynthesekorpora, die jeweils eine größere Anzahl von Äußerungen eines oder mehrerer Sprecher umfassen. Die Heterogenität sowohl der zugrundeliegenden theoretischen Annahmen bei der prosodischen Annotierung als auch in der praktischen Realisierung der Korpora (Textauswahl, Sprecherauswahl inklusive Dialekt, automatische oder manuelle Segmentierung) wird sowohl qualitativ als auch quantitativ dargelegt und anschließend die Annotierung soweit möglich vereinheitlicht. Alle Korpora müssen zunächst silbifiziert werden, d. h. in die segmentale Annotierung werden automatisiert Silbengrenzen eingetragen. Außerdem muss bei einigen Korpora die Zuordnung des Sprechtextes zum ursprünglichen Schrifttext wiederhergestellt werden um auch textbasierende Werkzeuge zur Merkmalsbestimmung nutzen zu können.

Eine Vielzahl von Merkmalen wird für die Akzentuierungserkennung (segmentale, silbische, wortbasierte, Dauern, Tonhöhen und -verläufe) sowie für die Phrasierungserkennung (wortbasierte, syntaktische, Pausen, Dehnung, Tonhöhen und -verläufe) extrahiert und in einer Merkmalsauswahl auf ihre Tauglichkeit untersucht. Für die Akzentuierungs- als auch Phrasierungserkennung werden sowohl sprecherabhängige als auch sprecher- und korpusübergreifende Experimente durchgeführt. Es zeigt sich, dass die Ergebnisse für sprecherabhängige Akzentuierungen zwischen den Korpora ähnlich sind (F-Maß: 70–80 %) und für sprecherübergreifende (F-Maß: 70 %) sowie sprecher- und korpusübergreifende Experimente (F-Maß: 65 %) die Leistung nur geringfügig abfällt. Für die Phrasierungserkennung gilt dies hingegen nicht: Bei der sprecherabhängigen Erkennung unterscheiden sich die

Ergebnisse stark und spiegeln damit Unterschiede in der Annotierung der Korpora wider, wie die sprecherübergreifende, aber korpusabhängige Erkennung zeigt; die korpus- und sprecherübergreifende Erkennung erreicht folglich sehr schlechte Ergebnisse. Im Umkehrschluss kann gefolgert werden, dass die Akzentuierungsannotierung in den verwendeten Korpora hinreichend einheitlich ist und sich auch sprecherübergreifend erzeugen lässt, während bei der Phrasierungsannotierung noch Lücken in der Standardisierung bestehen, bevor eine automatische sprecher- und korpusübergreifende Annotierung möglich wird.

Im letzten Schritt der Arbeit wird ein bisher unannotiertes Korpus automatisch prosodieannotiert und anschließend zum Training der sprecherabhängigen prosodischen Modelle in einem Sprachsynthesesystem verwendet. Die Auswertung in einem Perzeptionsexperiment zeigt im Vergleich zum Baseline-System mit einem sprecherfremden Prosodiemodell, eine schlechte Leistung. Die ausführliche Fehleranalyse wirft mehrere Punkte auf, ausgehend von der grundsätzlich schlechten Phrasierungsannotierung über das Training des prosodischen Modells des TTS-Systems (bei dem nur textuelle Merkmale verwendet werden können, wohingegen die Klassifizierer gerade solche Merkmale nur wenig nutzen), bis hin zu nicht-trainierter, regelbasierter Einfügung von Pausen durch das TTS-System, welche auf die Modelle des Baseline-Systems optimiert ist und sich im neuen System deutlich negativ auswirkt.