Automatische Prominenzetikettierung einer Datenbank für die korpusbasierte Sprachsynthese

Petra S. Wagner, Stefan Breuer, Karlheinz Stöber
Institut für Kommunikationsforschung und Phonetik, Universität Bonn

Zusammenfassung

Die Methode der korpusbasierten Sprachsynthese [9] benötigt große, segmental und prosodisch annotierte Datenbanken für eine optimale Einheitenauswahl.[1] Es wird überprüft, inwieweit eine Etikettierung von *perzeptiver Silbenprominenz* auf der Basis von CART-Bäumen automatisiert werden kann. Die Ergebnisse sind vielversprechend, sofern auf die Maße *Akzentuierbarkeit des Silbenkerns* (keine Schwa-Silbe), *Silbendauer*, sowie *Wortart* (*part of speech = POS*), *Wortbetonung* und *Vorhandensein eines F0-Maximums* zurückgegriffen werden kann.

1. Motivation

Für eine natürlich klingende korpusbasierte *Unit Selection*-Synthese spielt nicht nur die segmentale, sondern auch die prosodische Struktur der Einheit eine große Rolle. Die perzeptive Silbenprominenz ist ein prosodisches Maß, welches die wahrgenommene Betonungsstärke widerspiegelt und in traditionellen Syntheseansätzen bereits erfolgreich eingesetzt wurde [8]. Das Wissen um die perzeptive Prominenz ist hilfreich bei der Erzeugung natürlich klingender Wortbetonung sowie prosodischer Foki [11]. Angesichts der Größe der Synthesekorpora und dem Aufwand manueller Etikettierungen ist eine Automatisierung in diesem Bereich nahezu erforderlich. Bisherige Untersuchungen [7] haben hohe Korrelationen ($cc > 0.8$) zwischen vorhergesagten und handetikettierten Werten ergeben, sofern auf linguistische wie akustische Parameter für eine Vorhersage zurückgegriffen werden konnte. Vergleichbar hohe Korrelationen konnten zwischen Prominenzbeurteilungen verschiedener Hörer gemessen werden [5].

2. Synthesekorpus

Das Sprachmaterial des Synthesekorpus [12] stammt von einer der Sprecherinnen der Bonner Prosodischen Datenbank (BPD) [4] und besteht aus knapp 4 Stunden gelesener Sprache. Das Material wurde auf Wortbasis manuell segmentiert. Anhand der Wortgrenzen wurde anschließend eine automatische Phonem- und Silbensegmentierung sowie die Zuweisung der Wortakzente vorgenommen. Dies geschah anhand der Transkriptionen innerhalb der offiziellen Wortliste des Verbmobilprojekts [6]. Ein Tagging der Wortklassen wurde halbautomatisch mit den Wortklassen des Verbmobil-Generierungsmoduls durchgeführt. Zusätzlich erfolgte eine automatische Annotierung der F0-Maxima auf der Basis des Prosodieerkenners von [1].

3. Methode

Als Trainingsmethode wurde das CART-basierte Programm *wagon* der *Edinburgh Speech Tools Library* [10] gewählt. CART-Bäume haben den Vorteil, daß sie hierarchisch aufgebaut sind, Interpretationen der wichtigsten Einflußfaktoren auf die erfolgreiche Vorhersage zulassen und sowohl numerische wie auch symbolische Eingaben verarbeiten können. Die Anwendbarkeit der Methode bzgl. phonetischer Fragestellungen wurde u.a. in [2] gezeigt.

Als Trainingsmaterial stand die BPD zur Verfügung. Diese Datenbank besteht aus 266 von drei Sprechern gelesenen Sätzen und drei Geschichten und ist linguistisch und akustisch etikettiert. Die Etikettierung der perzeptiven Silbenprominenz wurde von drei Hörern auf einer Skala von 0-31 nach der Methode von [3] vorgenommen. Die annotierten Wortklassen unterscheiden sich von denen der Synthesedatenbank. Um eine Vergleichbarkeit der Wortklassen zu gewährleisten, wurde eine Abbildung auf die Wortklassen der Synthesedatenbank durchgeführt.

Die CART-Bäume wurden so trainiert, daß die Parameter, welche die Daten am besten hinsichtlich der Prominenzwerte ausdifferenzieren können, im Baum an höherer Stelle auftauchen.

80% der BPD wurden für das Training verwendet, 20% als Testmenge. Als weitere Testmenge wurden 400 Äußerungen des Synthesekorpus ebenfalls handetikettiert. Anschließend wurden Korrelationskoeffizienten zwischen vorhergesagter und wahrgenommener Prominenz berechnet.

Da das Synthesekorpus nicht so detailliert annotiert ist wie die BPD, können die CART-Bäume auch nicht so viele Trainingsmerkmale heranziehen wie in den vorherigen Untersuchungen. Daher wird geprüft, welche bzw. wieviele Parameter für eine erfolgreiche Vorhersage essentiell sind.

4. Ergebnisse

Ein CART-Baum, der auf einer minimalistischen Merkmalsmenge (Silbendauer, Akzentuierbarkeit, Wortklasse) traininert wurde, führte bereits zu einer annähernd hohen Korrelation auf den Testdaten der BPD, schnitt aber auf

[1] Die hier vorgestellte Forschung wurde vom Bundesministerium für Bildung und Forschung im Rahmen des Projekts Verbmobil, Förderungsnummer 01 IV 101 G, finanziert. Wir danken Hannes Pirker (ÖFAI Wien) für die freundliche Herausgabe und Anpassung des verwendeten Werkzeugs zur Handetikettierung von Prominenzwerten sowie den Mitarbeitern am Lehrstuhl für Mustererkennung der Universität Erlangen-Nürnberg für die Annotierung der F0-Maxima.

	Dauer, Schwa, POS	Dauer, Schwa, POS, Wort-betonung	Dauer,Schwa, POS, Wortbe-tonung, F0-Maximum
BPD	0.67	0.78	0.82
Synthese-korpus	0.60	0.65	0.73

Tabelle 1: Korrelationen zwischen vorhergesagten und wahrgenommenen Prominenzwerten mit unterschiedlichen Vorhersageparametern auf beiden Testdaten

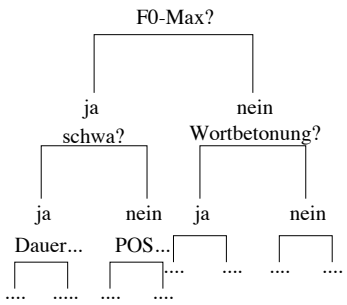

Abbildung 1: Die Spitze des CART-Baumes, der F0-Maxima berücksichtigt

dem neuen Korpus eher unbefriedigend ab. Die Hinzunahme der Wortbetonung in die Menge der Trainingsparameter führte bereits zu besseren Ergebnissen auf beiden Testmengen. Information über das Vorhandensein von F0-Maxima ist aber essentiell, um bei der Vorhersage für das Synthesekorpus zu Ergebnissen zu gelangen, die mit den Hörerurteilen vergleichbar sind.

5. Diskussion

Die Korrelationen zwischen vorhergesagten und wahrgenommenen Prominenzen waren durchgängig höher auf der Testmenge der Trainingsdatenbank. Dies kann mehrere Gründe haben. Hauptgrund ist sicherlich der bekannte Umstand, daß CART-Klassifikationen immer sehr speziell auf das Trainingskorpus zugeschnitten sind und ein Problem mit der Generalisierbarkeit ihrer Vorhersagen haben. Weiterhin sind die Testdaten nur von einer Person etikettiert worden, so daß idiosynkratische Höreindrücke nicht durch Bildung des Medians kompensiert werden konnten. Außerdem sind sämtliche verwendete akustische wie linguistische Merkmalsannotationen im Synthesekorpus automatisch erstellt worden und somit fehlerbehafteter als im Ausgangskorpus. Ein weiterer Faktor für Fehler in der Klassifikation ist die Verwendung ursprünglich unterschiedlicher Wortklassen in Trainings- und Vorhersagekorpus. Hier wurde ein Mapping vorgenommen, welches eine weitere Ungenauigkeit zwischen beiden Mengen zur Folge hat. Wie in vorherigen Untersuchungen konnte einmal mehr der Zusammenhang zwischen perzeptiver Prominenz und dem Vorliegen eines F0-Maximums gezeigt werden. In solchen Fällen, wo kein F0-Maximum vorliegt, sind Informationen zur Wortbetonung ausschlaggebend für eine Diskriminierung verschiedener Prominenzklassen. Dauer, Akzentuierbarkeit und Wortart sind wichtiger für eine Feinabstimmung der Prominenzstufen.

6. Schlußfolgerung

Eine erfolgreiche automatische Etikettierung perzeptiver Silbenprominenz ist bereits dann möglich, wenn auf nur wenige akustische wie linguistische Information zurückgegriffen werden kann. Steht die Information über das Vorhandensein eines F0-Maximums auf der Silbe zur Verfügung, gelangen die Vorhersagen in Bereiche, die mit Beurteilungen menschlicher Hörer vergleichbar sind.

Literatur

1. Anton Batliner, Volker Warncke, Elmar Nöth, Jan Buckow, Richard Huber, and Matthias Nutt. How to label accent position in spontaneous speech automatically with the help of syntactic-prosodic boundary labels. VERBMOBIL-Report 228, Universität Erlangen-Nürnberg, August 1998.
2. Stefan Breuer. Zur Regelhaftigkeit segmentaler Reduktion in deutschen Funktionswörtern. Magisterarbeit, Universität Bonn, 1999.
3. Gunnar Fant and Anita Kruckenberg. Preliminaries to the study of Swedish prose reading and reading style. *STL–QPSR 2/1989*, pages 1—68, 1989.
4. B. Heuft, T. Portele, F. Höfer, J. Krämer, H. Meyer, M. Rauth, and G. Sonntag. Parametric description of f0–contours in a prosodic database. In *Proceedings of the International Conference on Spoken Language Processing, Stockholm*, volume 2, pages 378—381, 1995.
5. Barbara Heuft. *Eine prominenzbasierte Methode zur Prosodieanalyse und -synthese*, volume 2 of *Computer Studies in Language and Speech*. Peter Lang, 1999.
6. Harald Lüngen, Karsten Ehlebracht, Dafydd Gibbon, und Ana Paula Quirino Simoes. Morphologie in VERBMOBIL Phase II. VERBMOBIL Report 366, Universität Bielefeld, 1998.
7. Thomas Portele. Just concatenation — a corpus–based approach and its limits. In *Selected Papers of the 3rd Speech Synthesis Workshop at Jenolan Caves*. (Working Title), erscheint.
8. Thomas Portele and Barbara Heuft. Towards a prominence–based synthesis system. *Speech Communication*, pages 61—72, 1997.
9. Karlheinz Stöber, Thomas Portele, Petra Wagner, and Wolfgang Hess. Synthesis by word concatenation. In *Proceedings of EUROSPEECH'99*, Budapest, Hungary, 1999.
10. Paul Taylor, Richard Caley, Alan W. Black, and Simon King. Edinburgh speech tools library–system documentation. `http://www.cstr.ed.ac.uk/projects/speechtools/manual-1.2.0`, June 1999. Edition 1.2.0.
11. Petra Wagner. The synthesis of German contrastive focus. In *Proceedings of ICPhS 99, San Francisco*, 1999.
12. Petra Wagner, Felicitas Haas, Karlheinz Stöber, and Jörg Helbig. Multilinguale korpusbasierte Sprachsynthese auf der Basis domänenspezifischen Ausgangsmaterials. In *10. Konferenz Elektronische Sprachsignalverarbeitung '99*, Görlitz, Germany, 1999.