Prominenzbasierte Prosodieprädiktion für die Unit-Selection-Sprachsynthese

Andreas Windmann[1,2], Igor Jauk[3], Fabio Tamburini[4], Petra Wagner[2]

[1]CITEC, Universität Bielefeld
[2]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld
[3]Technische Fakultät, Universität Bielefeld
[4]Dipartimento di Studi Linguistici e Orientali, Università di Bologna
andreas.windmann@uni-bielefeld.de

Abstract: Der Beitrag stellt einen Prosodievorhersagealgorithmus für das Unit-Selection-Sprachsynthesesystem BOSS vor, der prosodische Strukturen nicht in Form akustischer Parameter wie Segmentdauern und Grundfrequenz, sondern direkt in der perzeptiven Domäne modelliert. Die prosodische Struktur einer Äußerung wird dabei als Muster von Werten wahrgenommener *Prominenz* repräsentiert. Für die Prominenzprädiktion verwenden wir einen regelbasierten Algorithmus, der das Prominenzmuster einer Äußerung auf Silbenebene basierend auf Wortbetonung, Vokalreduktion, Phrasenstruktur und Wortklasseninformation vorhersagt. In einer früheren Arbeit konnte gezeigt werden, dass die Vorhersagen dieses Algorithmus eine hohe Übereinstimmung mit der Prominenzwahrnehmung deutscher Muttersprachler erreichen. Um die Auswahl von Einheiten mit entsprechenden Prominenzwerten aus dem BOSS-Sprachkorpus zu gewährleisten, wurde dieses mittels eines akustikbasierten automatischen Annotationswerkzeugs mit kontinuierlichen Prominenzwerten auf Silbenebene etikettiert. Für die Einheitenauswahl werden die vorhergesagten Prominenzwerte linear auf den Wertebereich der Korpusannotation abgebildet und die Differenz zwischen vorhergesagter und tatsächlicher Prominenz einer Einheit zu ihrem Zielkostenvektor hinzugefügt. Präferenzurteile von Hörern aus einem paarweisen Vergleichstest zeigen, dass unser Ansatz die Prosodie der Sprachausgabe gegenüber der Standardausführung des Systems deutlich verbessert. Ein Transkriptionsexperiment mit semantisch unvorhersagbaren Sätzen ergab allerdings, dass dies bei ungenügender Korpusabdeckung zu Lasten der Verständlichkeit gehen kann. Da dies jedoch ein allgemeines Problem der Prosodieprädiktion in der Unit-Selection-Sprachsynthese ist, argumentieren wir, dass der prominenzbasierte Ansatz eine ernstzunehmende Alternative zu akustikbasierten Verfahren darstellt.

1 Einleitung

Für die Prosodieprädiktion in Unit-Selection-Sprachsynthesesystemen werden üblicherweise Algorithmen eingesetzt, die akustisch-prosodische Parameter des Sprachsignals wie Grundfrequenzkonturen und Segmentdauern vorhersagen. Wir verfolgen eine alternative Herangehensweise, bei der prosodische Strukturen nicht in der akustischen, sondern in der perzeptiven Domäne modelliert werden. Die prosodische Struktur einer Äußerung wird dabei in Form eines Musters von Prominenzwerten repräsentiert, die die graduell wahrgenommene relative Salienz einer Silbe oder einer größeren prosodisch relevanten Einheit im Äußerungskontext angeben [1]. Die Wahrnehmung von Prominenz wird zum einen durch den kumulativen Effekt der akustisch-prosodischen Parameter Grundfrequenz (F0), Dauer, Intensität und spektrale Energieverteilung bestimmt, wobei die Wichtigkeit der einzelnen Parameter sprachspezifisch variieren kann [2],[3],[4],[5]; andererseits haben jedoch auch linguistische Erwartungshaltungen von Hörern einen erheblichen Einfluss auf die Prominenzwahrnehmung [1],[5],[6],[7],[8]. Für die prominenzbasierte Prosodievorhersage im Rahmen der Unit-Selection-Sprachsynthese werden drei Komponenten benötigt: Ein Modul

für die Prädiktion von Prominenzmustern für den Eingabetext (1), Prominenzannotation des Synthesekorpus (2) und die Modellierung von Prominenz als Kostenfaktor (3). Prominenz ist in einigen älteren Studien als intermediärer Parameter verwendet worden, der zunächst aus dem Eingabetext vorhergesagt wird und dann als Basis für die Berechnung der individuellen akustisch-prosodischen Parameter des Sprachsignals dient [9],[10]. In [11] wird die Etikettierung eines Synthesekorpus mit Prominenzwerten auf einer vierstufigen Skala beschrieben, jedoch nicht auf die Vorhersage von Prominenz eingegangen. Ein komplett prominenzbasiertes System stellen [12] vor, legen jedoch eine andere Definition des Begriffs zugrunde, die im Wesentlichen die Wahrscheinlichkeit der Akzentuierung eines Wortes angibt. Unser Beitrag beschreibt schwerpunktmäßig die Anwendung der oben genannten Schritte (1) und (3) auf das Bonn Open Synthesis System (BOSS; [13]) und gibt einen Überblick über die Evaluation der Prominenzprädiktion.

2 Prominenzprädiktion

Für die Vorhersage von Prominenzmustern auf Äußerungsebene verwenden wir einen Algorithmus, dem die Implementierung eines metrisch-phonologischen Regelsatzes aus [1], basierend auf Überlegungen in [14] zugrunde liegt. Die Regeln ordnen den einzelnen Silben in einer Äußerung nach bestimmten linguistischen Kriterien *rhythmische Schläge* zu. Dadurch wird ein *metrisches Gitter* erzeugt, das das Prominenzmuster der Äußerung repräsentiert. Die relative Prominenz einer Silbe im Äußerungskontext lässt sich an der Zahl der rhythmischen Schläge ablesen, die ihr zugeordnet wurden. Der Algorithmus enthält die folgenden Regeln:

1. Jede Silbe erhält einen Basisschlag.
2. Jede Silbe, deren Nukleus kein silbischer Konsonant oder ein reduzierter Vokal ist, erhält einen zusätzlichen Schlag.
3. Die hauptbetonte Silbe in jedem Wort erhält eine Anzahl von Schlägen, abhängig von der Wortklasse des Wortes:
    1. Nomen, Eigennamen, Numerale: 5 Schläge
    2. Adjektive, Adverbien: 4 Schläge
    3. Lexikalische Verben, Pronomina: 3 Schläge
    4. Auxiliare, Affirmations- und Negationspartikeln: 2 Schläge
    5. Andere Wortklassen: 1 Schlag
4. Die hauptbetonte Silbe im letzten Nomen, Adjektiv oder Adverb innerhalb einer prosodischen Phrase sowie in jedem Verb, das nicht auf ein Nomen, Adjektiv oder Adverb folgt, erhält einen zusätzlichen Schlag (*Nuclear Stress Rule*).
5. Die hauptbetonte Silbe in einem äußerungsinitialen Funktionswort erhält einen zusätzlichen Schlag.
6. Jede zweite Silbe in Sequenzen von drei oder mehr Silben mit zwei Schlägen erhält einen zusätzlichen Schlag (*Euphonieregel*).

Regelbasierte Algorithmen kommen in der Sprachtechnologie nur noch selten zum Einsatz, da intuitiv gewonnene Regeln anfällig für Autosuggestion sind und maschinelle Lernverfahren sich im Allgemeinen als leistungsfähiger erwiesen haben. [1] konnte jedoch in einer umfangreichen Perzeptionsstudie zeigen, dass die Vorhersagen unseres Algorithmus eine hohe Übereinstimmung mit der Prominenzwahrnehmung deutscher Muttersprachler erzielen. Die vorgestellten Regeln stellen somit eine einfache, aber empirisch adäquate Lösung für die Modellierung von Prominenzmustern in pragmatisch neutralen Deklarativsätzen des Deutschen dar. Abbildung 1 zeigt ein Beipiel für ein metrisches Gitter, das durch Anwendung von allen sechs Regeln erzeugt wurde.

									x4
									X3.1
x5	**x4**			**X3.2**					**X3.1**
X3.3	**X3.3**			**X3.2**					**X3.1**
X3.3	**X3.3**			**X3.2**					**X3.1**
X3.3	**X3.3**		**X3.5**	**X3.2**		**x6**			**X3.1**
x2	**x2**		**x2**	**x2**	**x2**	**x2**	**x2**		**x2**
x1	**x1**	**x1**	**x1**	**x1**	**x1**	**x1**	**x1**	**x1**	**x1**
ɪç	fliː	ɡə	am	aɪn	ʊn	tsvan	tsɪçs	tən	maɪ
Ich	flie	ge	am	ein	und	zwan	zigs	ten	Mai
PRON	V		PREP	ADJ					N

Abbildung 1: Metrisches Gitter für die Äußerung "Ich fliege am einundzwanzigsten Mai". Die Zahlen geben an, von welcher Regel der zugeordnete Schlag erzeugt wurde.

3 Implementierung in BOSS

Der Prominenzvorhersage gehen in BOSS zwei weitere Verarbeitungsschritte voraus: Der *BOSS Client* empfängt zunächst die Texteingabe des Benutzers, führt eine Textvorverarbeitung durch und bestimmt, basierend auf der Interpunktion, die Grenzen von Intonationsphrasen. Der vorverarbeitete Text wird in einem XML-Dokument abgespeichert, das hierarchisch in <SATZ>- und <WORT>-Elemente gegliedert ist. Letztere sind als phraseninitial, -medial oder -final markiert. Das XML-Dokument wird im zweiten Verarbeitungsschritt an den *BOSS Server* weitergeleitet und durchläuft hier zunächst ein Transkriptionsmodul. Hier wird das XML-Dokument um die Silben-, Phon- und Halbphonebene erweitert, phonetisch transkribiert und auf den neu hinzugefügten Ebenen mit Wortbetonungsinformationen angereichert. Für diese Arbeitsschritte wird ein umfangreiches Aussprachelexikon für das Deutsche verwendet [15]; Ausnahmen werden durch eine Morphemliste des Deutschen und einen CART-basierten G2P-Algorithmus abgefangen.

Nach Beendigung des Transkriptionsprozesses wird das Äußerungs-XML-Dokument an das Prominenzprädiktionsmodul weitergegeben, das die oben eingeführten Regeln zunächst auf alle Elemente vom Typ <SILBE> anwendet. Mit Ausnahme der Wortklassen stehen alle benötigten Informationen zu diesem Zeitpunkt zur Verfügung: Das Transkriptionsmodul ermittelt die Silbengrenzen, stellt die phonetische Transkription bereit (Regel 2) und identifiziert hauptbetonte Silben (Regel 3,4,5); phrasenfinale und äußerungsinitiale Wörter (Regel 4,5) können mittels der im Client hinzugefügten Phrasierungs-Tags identifiziert werden. Für die Bereitstellung von Wortklasseninformation für die Anwendung der Regeln 3, 4 und 5 wird ein HMM-basierter POS-Tagger für das Deutsche verwendet [16]. Dieser wird direkt aus dem Prädiktionsmodul heraus aufgerufen und reichert die <WORT>-Elemente im aktuellen XML-Dokument mit POS-Tags an. Um die Prominenzwerte später in der Einheitenauswahl auf allen Ebenen berücksichtigen zu können, werden sie nach Beendigung der Prädiktion auf Silbenebene noch auf die Wort- und Phonebene propagiert. <WORT>-Elemente erhalten den Prominenzwert der prominentesten Silbe, die sie enthalten, <PHON>-Elemente erhalten den Prominenzwert der Silbe, deren Teil sie sind.

Die Einheitenauswahl in BOSS gliedert sich in zwei Teilschritte: Zunächst wird im Rahmen einer Vorauswahl ein Suchraum aus allen Einheiten aus dem Sprachkorpus generiert, die aufgrund ihrer segmentellen Struktur und, je nach Konfiguration auch anderer diskreter Eigenschaften, für die Synthese der Zieläußerung in Frage kommen. Im zweiten Schritt, der eigentlichen Einheitenauswahl, wird die optimale Sequenz von Einheiten für die Synthese der

Zieläußerung bestimmt, definiert als derjenige Pfad durch den Suchraum, der die Summe von Übergangs- und Einheitenkosten minimiert. Prominenz wird als Einheitenkostenfaktor berücksichtigt. Hierfür wurde das BOSS-Sprachkorpus mittels eines automatischen Annotationswerkzeugs [17] basierend auf Analysen der akustischen Prominenzkorrelate F0, Dauern, Intensität und spektrale Energieverteilung mit Prominenzwerten auf einer kontinuierlichen Skala zwischen 0 und 1 etikettiert [18]. Diese Annotation wurde zunächst auf Silbenebene durchgeführt; anschließend wurden die Prominenzwerte innerhalb der Korpusannotation nach der auch in der Prädiktion benutzten Methode auf die anderen Einheitenebenen propagiert. Ein Perzeptionstest auf einem Teil der Korpusdaten zeigte eine zufriedenstellende Übereinstimmung der Annotation mit wahrgenommenen Prominenzmustern [18]. Für die Einheitenauswahl werden die vorhergesagten Prominenzwerte linear auf den Wertebereich der Korpusannotation abgebildet und die Differenz zwischen vorhergesagter und tatsächlicher Prominenz einer Einheit zu ihrem Einheitenkostenvektor hinzugefügt. Somit kann das System eine Sequenz von Einheiten mit einem vorhergesagten Prominenzmuster auswählen. Abbildung 2 zeigt einen schematischen Überblick der BOSS-Systemarchitektur einschließlich der Prominenzprädiktion.

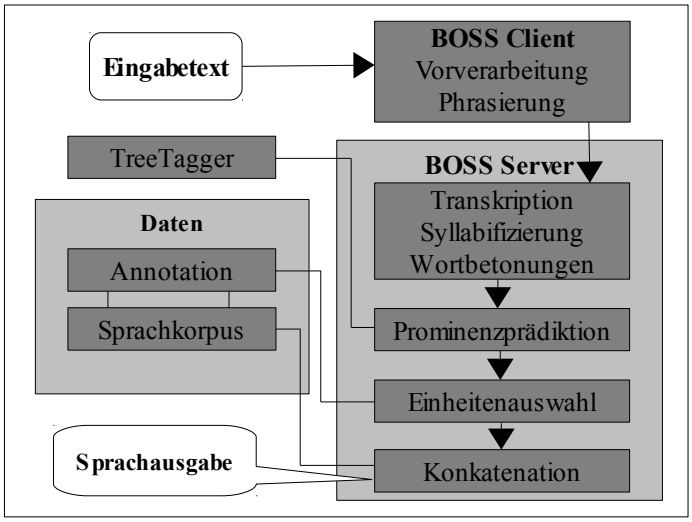

Abbildung 2: BOSS-Systemarchitektur mit Prominenzprädiktion.

4 Evaluation

4.1 Prosodie

4.1.1 Experimentaldesign

Mittels eines paarweisen Vergleichstests wurde überprüft, ob die Prominenzprädiktion die Prosodie der BOSS-Sprachausgabe verbessert. Dazu wurden zwei Listen mit jeweils zehn Satzpaaren synthetisiert. Jedes Paar bestand aus zwei Instanzen des gleichen Satzes, von denen eine mit der BOSS-Standardausführung (*Standard*), die andere mit der neuen Version des Systems inklusive der Prominenzprädiktion (*Prominenz*) erzeugt wurde. Zweck der beiden Listen war es, die Prominenzprädiktion sowohl *diagnostisch* als auch *global* zu testen.

Für die zehn Stimuluspaare in der *diagnostischen* Liste wurden mittels einer trial-and-error-Strategie Sätze gesucht, die in der *Standard*-Version von BOSS mit offensichtlichen Abweichungen vom erwarteten Prominenzmuster produziert wurden. In den meisten Fällen waren die Abweichungen auf die Auswahl von übermäßig prominenten Funktionswörtern

zurückzuführen. Um sicherzustellen, dass im Test tatsächlich die Prominenzmuster bewerten werden würden, wurden bei der Auswahl der Stimuli für die *diagnostische* Liste alle anderen Faktoren bestmöglich kontrolliert. Es wurden nur grammatische und inhaltlich sinnvolle Sätze verwendet, die von BOSS korrekt transkribiert und mit korrekten POS-Tags versehen und ausschließlich aus ganzen Wörtern zusammengesetzt wurden. Außerdem wurde darauf geachtet, dass die *Standard*-Stimuli in der *diagnostischen* Liste keine weiteren Auffälligkeiten außer dem abweichenden Prominenzmuster aufwiesen, und dass zumindest für die problematischen Einheiten jeweils eine substantielle Anzahl an Alternativen im Korpus zur Verfügung stand. Bei den Stimuli in der *globalen* Liste wurden dagegen nur Transkription und POS-Tagging kontrolliert.

Die Evaluation wurde mittels einer Web-Anwendung durchgeführt. Die Stimuluspaare wurden den Teilnehmern einzeln präsentiert, wobei beiden Stimuli jeweils ein Abspielknopf und ein Kontrollkästchen zugeordnet waren. Die Reihenfolge der Stimuluspaare sowie die Zuordnung von *Standard*- und *Prominenz*-Stimulus zum rechten bzw. linken Abspielknopf erfolgte randomisiert. Die Teilnehmer wurden instruiert, sich jeweils beide Versionen anzuhören und die besser klingende Version zu markieren.

4.1.2 Ergebnisse

An dem Experiment nahmen 105 Personen teil. 12 Personen gaben an, keine deutschen Muttersprachler zu sein, an Hörschäden zu leiden oder den Test in einer lauten Umgebung durchgeführt zu haben, und wurden deshalb nicht in die Auswertung aufgenommen. Die verbliebenen 93 Personen (36 m, 57 f) waren zwischen 19 und 60 Jahre alt (ø 27,8). 75 Versuchspersonen verfügten nach Selbstauskunft über Erfahrung in Linguistik oder Phonetik, 18 über Vorerfahrungen mit synthetischer Sprache. 44 Versuchspersonen benutzten eingebaute Lautsprecher, 26 Versuchspersonen benutzten Lautsprecherboxen, 13 Versuchspersonen benutzten Muschelkopfhörer und 10 Versuchspersonen benutzten Ohrhörer. Abbildung 3 zeigt die Ergebnisse des Experiments.

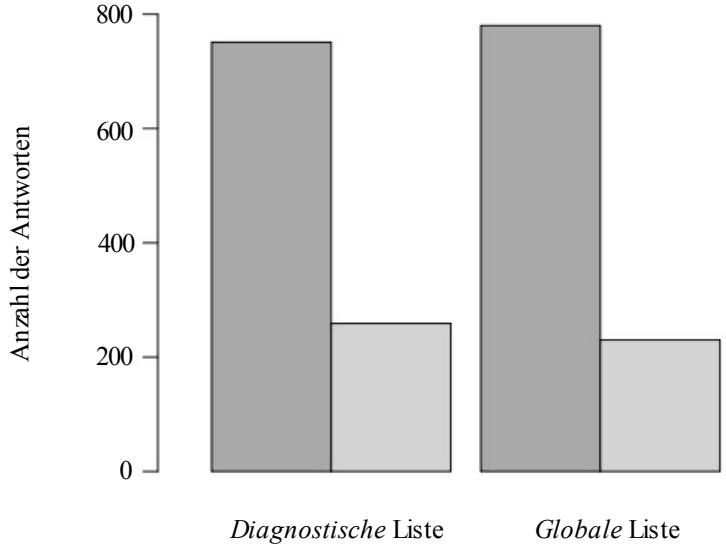

Abbildung 3: Anzahl der Antworten mit Präferenz für *Prominenz*- (dunkel) und *Standard*-Version (hell) in der *diagnostischen* und *globalen* Liste.

Die *Prominenz*-Stimuli wurden ihren *Standard*-Pendants sowohl in der *diagnostischen* ($\chi^2(1,930)=215.81$, $p<0.0001$) als auch in der *globalen* Liste ($\chi^2(1,930)=273.14$, $p<0.0001$) von der Mehrheit der Versuchspersonen vorgezogen. Mit zwei Ausnahmen zeigte sich dieses Muster auch bei den einzelnen Stimuluspaaren. Bei einem der *diagnostischen* Stimuluspaare war eine klare Präferenz für die *Standard*-Version zu beobachten, wohl aufgrund eines auffälligen Grundfrequenzsprunges in der *Prominenz*-Version. Bei einem der Stimuluspaare aus der *globalen* Liste wurden beide Versionen gleich oft bevorzugt. Keine der genannten Kontrollvariablen hatte einen nennenswerten Einfluss auf das Ergebnis. Damit lässt sich zusammenfassend festhalten, dass unsere Prominenzprädiktionsmethode das Potential hat, die Prosodie von synthetischer Sprache deutlich zu verbessern.

4.2 Verständlichkeit

4.2.1 Experimentaldesign

Um mögliche Auswirkungen der Prominenzprädiktion auf die Verständlichkeit der synthetischen Sprachausgabe zu testen, wurde ein Transkriptionsexperiment mit semantisch unvorhersagbaren Sätzen durchgeführt. Dazu wurde ein Set von 20 Stimuluspaaren synthetisiert, die jeweils aus zwei Instanzen des gleichen Satzes bestanden. Analog zum paarweisen Vergleichstest wurde jeweils ein Satz mit der BOSS-Standardausführung (*Standard*), der andere mit der neuen Version des Systems inklusive der Prominenzprädiktion (*Prominenz*) erzeugt. Die Sätze bestanden aus sechs oder sieben Wörtern und wurden zwecks besserer Vergleichbarkeit nach einem syntaktischen Template mit kleinen Variationen konstruiert. Für die Synthese dieser Sätze musste BOSS größtenteils auf Einzelphone zurückgreifen. Um Lerneffekte zu vermeiden und dennoch Vergleichbarkeit zwischen den Bedingungen zu gewährleisten, wurde das Testset für das Transkriptionsexperiment in zwei Subsets aufgeteilt, die jeweils die Hälfte der Stimuli aus jeder Bedingung und genau einen Satz aus jedem Stimuluspaar enthielten. Im Test wurde jede Versuchsperson zufällig einem der beiden Subsets zugeordnet. Das Transkriptionsexperiment wurde als Subtest innerhalb der selben Webapplikation implementiert wie der paarweise Vergleichstest, so dass die selben Personen an beiden Experimenten teilnahmen. Obwohl hier in der umgekehrten Reihenfolge beschrieben, durchliefen die Versuchspersonen das Transkriptionsexperiment zuerst, um Trainingsffekte zu vermeiden. Die Stimuli wurden separat, jeweils mit einem Abspielknopf und einem Textfeld präsentiert und konnten jeweils einmal abgespielt werden. Die Versuchspersonen wurden instruiert, sich den Stimulus anzuhören und ihn im Textfeld orthographisch zu transkribieren. Sie wurden darüber informiert, dass sie unsinnige Sätze hören würden und dass jeder einzelne Satz nur einmal abspielbar war.

4.2.2 Ergebnisse

Die beiden Subsets im Transkriptionsexperiment wurden 46 beziehungsweise 47 Versuchspersonen zugeordnet. Als Maß für die Verständlichkeit der Sätze wurde, wie von [19] vorgeschlagen, die Levenshtein-Distanz auf Wortebene zwischen den Transkriptionen der Versuchspersonen und den Referenztranskriptionen der Sätze berechnet. Interpunktion und Groß-/Kleinschreibung am Wortanfang wurden zuvor in den Referenztranskriptionen und den Transkriptionen der Versuchspersonen vereinheitlicht, um diese trivialen Faktoren als Fehlerquelle auszuschließen. Außerdem wurden die Levenshtein-Distanzen für die einzelnen Stimuli durch deren Silbenzahl dividiert, da sich diese zwischen den Stimuluspaaren zum Teil deutlich unterschied. Abbildung 4 zeigt die Mediane der normalisierten Levenshtein-Distanz für die *Standard*- und *Prominenz*-Stimuli in den beiden Subsets des Transkriptionsexperiments.

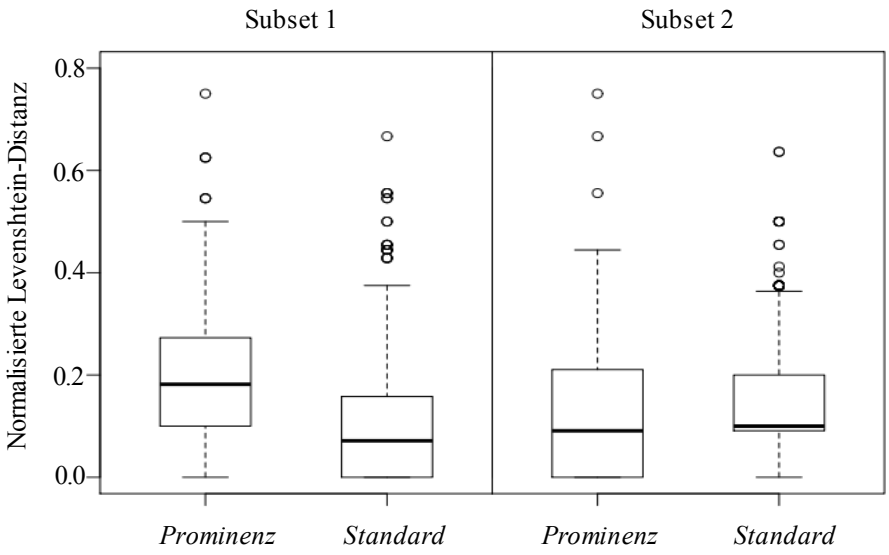

Abbildung 4: Mediane der normalisierten Levenshtein-Distanz für *Prominenz*- und *Standard*-Stimuli in den zwei Subsets.

Die Ergebnisse des Transkriptionsexperimentes wurden in R [20] mit *Linear Mixed Effects Models* analysiert. Es wurden ein schrittweises Modellauswahlverfahren sowie p-Wert-Schätzung nach [21] angewendet. Die Daten in Subset 1 werden durch ein Modell mit den festen Effekten *BOSS-Version* ($t(913)=-2.53$, $p<0.05$) und *Audioequipment* ($t(913)=-2.699$, $p<0.01$) am besten erklärt; dagegen war in Subset 2 kein Effekt des Faktors *BOSS-Version* festzustellen. Die Prominenzprädiktion verschlechterte somit die Verständlichkeit der Stimuli in Subset 1, nicht jedoch in Subset 2. In der Tendenz bestätigt dieses Ergebnis den bekannten Befund, dass Kontrolle der Prosodie auf Kosten der segmentalen Qualität von synthetischer Sprache gehen kann [12].

5 Diskussion und Ausblick

Unsere Ergebnisse zeigen, dass der prominenzbasierte Ansatz eine praktikable Methode für die Modellierung von Prosodie im Rahmen der Unit-Selection-Sprachsynthese darstellt. Der vorgestellte Vorhersagealgorithmus ist trotz seiner Einfachheit theoretisch wohlfundiert und empirisch adäquat und kommt im Gegensatz zu probabilistischen Verfahren ohne Merkmalsextraktion und handannotierte Trainingsdaten aus. Die sehr ermutigenden Ergebnisse des paarweisen Vergleichstests zeigen, dass unser Ansatz imstande ist, die Prosodie von synthetischer Sprache signifikant zu verbessern. Der weniger positive Ausgang des Transkriptionsexperiments macht allerdings deutlich, dass der Erfolg der Prominenzprädiktion von der Korpusabdeckung abhängig ist. Bei einer hohen Konkatenationsstellendichte können die zusätzlichen Kosten durch die Prominenzprädiktion unter Umständen zu einer weiteren Verschlechterung der Verständlichkeit führen. Dies war bei den Stimuli im Transkriptionsexperiment der Fall. Wie bereits erwähnt, sind die möglichen negativen Auswirkungen prosodischer Kontrolle auf die Verständlichkeit von synthetischer Sprache jedoch ein bekanntes Problem, das kaum zu vermeiden ist.

Mittelfristig wäre es wünschenswert, die Prominenzprädiktion auch auf andere Satztypen zu erweitern. Die Synthese von Fragesätzen ist ein solches Desiderat; im Kontext von Concept-to-Speech-Systemen wäre auch eine Anwendung auf Fokusstrukturen denkbar [10]. Derartige Erweiterungen würden einen komplexeren Regelsatz erfordern, für den der Einsatz maschineller Lernverfahren vermutlich sinnvoll ist. Eine zentrale Perspektive für weiterführende Arbeiten ist darüber hinaus ein direkter experimenteller Vergleich unseres Ansatzes mit konventionellen akustikbasierten Verfahren der Prosodieprädiktion.

Literatur

[1] Wagner, P.: *Vorhersage und Wahrnehmung Deutscher Betonungsmuster*. Dissertation, Universität Bonn, 2002.

[2] Fry, D.B.: *Experiments on the perception of stress.* In *Language and Speech 1*, S. 126-152, 1958.

[3] Fant, G. und A. Kruckenberg: *Preliminaries to the study of swedish prose and reading style*. In *STR-QPSR 2/1989*, KTH Stockholm, S. 1-83, 1989.

[4] Streefkerk, B.: *Prominence – acoustic and lexical/syntactic correlates*. Dissertation, Universität Amsterdam, 2002.

[5] Heuft, B., T. Portele, C. Widera, P. Wagner und M. Wolters: *Perceptual prominence*. In Sendlmeier, W. (ed.): *Speech and Signals*, S. 97-115, Frankfurt a.M., Hektor, 2000.

[6] Widera, C., T. Portele und M. Wolters: *Prediction of word prominence*. In *Proc. EUROSPEECH 1997*, Rhodos, S. 999-1003, 1997.

[7] Wagner, P.: *Great expectations – introspective vs. perceptual prominence ratings and their acoustic correlates*. In *Proc. Interspeech 2005*, Lissabon, S. 2381-2384, 2005.

[8] Arnold, D. und P. Wagner: *The influence of top-down expectations on the perception of syllable prominence*. In *Proc. ISCA Workshop on Experimental Linguistics*, Athen, S. 25-28, 2008.

[9] Ross, K. und M. Ostendorf: *Prediction of abstract prosodic labels for speech synthesis*. In *Computer Speech and Language 10,* S. 155-185, 1996.

[10] Portele, T. und B. Heuft: *Towards a prominence-based synthesis system*. In *Speech Communication 21*, S. 61-72, 1997.

[11] Wightman, C.W., A.K. Syrdal, G. Stemmer, A. Conkie und M. Beutnagel: P*erceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis*. In *Proc. ICSLP 2000*, Peking, S. 71-74, 2000.

[12] Strom, V., A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King und D. Jurafsky: *Modeling prominence and emphasis improves unit-selection synthesis*. In *Proc. Interspeech 2007*, Antwerpen, S. 1282-1285, 2007.

[13] Breuer, S. und W. Hess: *The Bonn Open Synthesis System 3.* In *International Journal of Speech Technology 13*, S. 75-84, 2010.

[14] Uhmann, S.: *Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie*. Tübingen, Niemeyer, 1991.

[15] Portele, T., J. Krämer und D. Stock.: *Symbolverarbeitung im Sprachsynthesesystem Hadifix*. In *Proc. 6. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, S. 97-104, Wolfenbüttel, 1995.

[16] Schmid, H.: *Improvements in part-of-speech tagging with an application to German*. In *Proc. ACL SIGDAT*, Dublin, S. 47-50, 1995.

[17] Tamburini, F. und P. Wagner.: *On automatic prominence detection for German*. In *Proc. Interspeech 2007*, Antwerpen, S. 1809-1812, 2007.

[18] Windmann, A., P. Wagner, F. Tamburini, D. Arnold und C. Oertel: *Automatic prominence annotation of a German speech synthesis corpus: towards prominence-based prosody generation for unit selection synthesis*. In *Proc. ISCA SSW7*, Nara, Japan, S. 377-382, 2010.

[19] Bunnell, H.T.: *Crafting small databases for unit selection TTS*: *effects on intelligibility*. In *Proc. ISCA SSW7*, Nara, Japan, S. 40-44, 2010.

[20] R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Wien, 2011. (www.R-project.org, 08.07.2011).

[21] Baayen, H.R.: *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge, CUP, 2008.