Datenbasierte Sprechrhythmussteuerung: ein hybrides, neuronal-regelbasiertes Verfahren

Diane Hirschfeld, Oliver Jokisch, Matthias Eichner, Uwe Koloska

Institut für Technische Akustik
Technische Universität Dresden, D-01062 Dresden
mailto:kom@eakss1.et.tu-dresden.de

Mit der derzeit erreichten hohen segmentalen Sprachqualität des Dresdner TTS-Systems, werden nun verstärkt Mängel in den prosodischen, regelbasierten Verarbeitungsstufen wahrnehmbar. In diesem Artikel wird ein Ansatz zur Lautdauersteuerung vorgestellt, der zwei völlig verschiedene Ansätze – regelbasiert und lernend mit neuronalen Netzen – kombiniert. Die Parameter beider Verfahren wurden aus natürlichem Sprachmaterial des Inventarsprechers gewonnen.

Informelle Vortests ergaben deutliche qualitative Verbesserungen im Vergleich zu dem alten, lautorientierten Regelverfahren nach Klatt[7].

1 Motivation

Hintergrund für die vorgestellten Arbeiten zur Dauersteuerung im Dresdner Sprachvollsynthesesystem ist die Verbesserung von Natürlichkeit und Verständlichkeit der synthetischen Sprache durch die Analyse großer Mengen natürlichen Sprachmaterials. Dabei kann ausgenutzt werden, daß die Quellsprecher für die am Institut für Technische Akustik entwickelten Baustein-Inventare [4, 8] für weitere Sprachaufnahmen zur Verfügung stehen und somit eine *prosodische Personifizierung* der Synthesestimme erfolgen kann.

Es wurde ein mehrstufiges hybrides Konzept für die Dauersteuerung entwickelt, das einen hierarchischen Aufbau in Laut-, Silben- und Phrasenebene aufweist und die Kombination eines regelbasierten Verfahrens und eines künstlichen neuronalen Netzes zur Dauersteuerung [6] bietet (Abb. 3). Die Schnittstellen zwischen den alternativen Ansätzen sind so definiert, daß ein Datenaustausch möglich ist und sich das Endergebnis aus einer Kombination der Teilergebnisse verschiedener Verarbeitungsstufen ergibt. Auf diese Weise lassen sich die Vorteile des regelbasierten und des neuronalen Verfahrens bei einem Zuwachs an Gesamtqualität optimal ausnutzen.

2 Die Datenbasis und ihre Aufbereitung

Für die vorliegenden Untersuchungen wurde das natürliche Sprachmaterial mit Zusatzinformationen auf unterschiedlichen Beschreibungsebenen versehen. Dabei wurde besonderer Wert darauf gelegt, daß die Etikettierung anhand objektiver Merkmale erfolgt und das Material für verschiedene Aufgaben wiederverwendbar ist (Anwendung/Training lernender Verfahren und automatischer Etikettierer, für statistische Zwecke etc.).

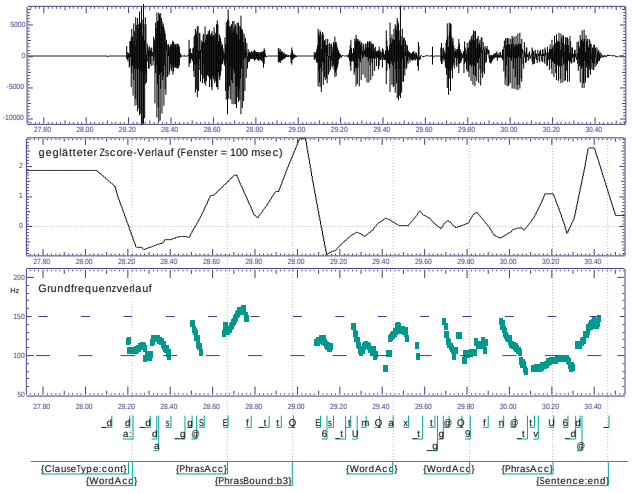

Abbildung 1: Die akustische Repräsentation des Satzfragmentes *„Da das Geschäft erst um acht geöffnet wurde. . . "* zusammen mit dem geglätteten *zscore*-Verlauf und der Grundfrequenz.

Um Bausteine für die konkatenative Synthese zu gewinnen, ist eine phonetische Ausgewogenheit des Materials sowie ein hoher Abdeckungsgrad der im deutschen gebräuchlichen Lautverbindungen wünschenswert. Andererseits machen die Forderungen nach Natürlichkeit im Sprechfluß der Einheiten, nach dem Auftreten prosodi-

scher Varianten und Effekte die Einbettung mindestens im Satzkontext notwendig. Diese Forderungen erfüllen die ca. 490 Sätze des PHONDAT1-Korpus (Satzkorpus).

Bestimmte prosodische Effekte sowie ein natürlicher für den Vorlesemodus typischer Sprechrhythmus erfordern die Aufnahme längerer, zusammenhängender Texte. Dafür wurden die zwei zum PHONDAT1-Korpus gehörigen Kurzgeschichten sowie das erste Kapitel des Buches *„Momo“* ausgewählt (Textkorpus, 344 Sätze mit 10.780 Segmenten).

2.1 Lautetiketten

Zur Markierung der Laute wird das SAMPA-Inventar verwendet, das um zusätzliche Symbole für Glottal-Stop/Glottalisierung, Pause sowie Störgeräusche/nicht zu verwendendes Material erweitert wurde. Plosive wurden in zwei Segmente unterteilt: Plosivpause und Burst mit Aspirationsphase. Die Etikettierung erfolgte möglichst objektiv anhand von Formanten als spektralen Merkmalen [3].

2.2 Prosodische Etikettierung

Prosodische Phrasen sind Wortgruppen, die inhaltlich zusammengehören und zwischen denen eine Sprechpause auftreten kann, aber nicht zwingend muß. Sie besitzen einen weitestgehend kontinuierlichen Intonationsverlauf, wobei an Phrasengrenzen oft ein Sprung (Deklinationsreset) in der Grundfrequenz zu beobachten ist.Sie sind Untereinheiten von Teilsätzen des Textes und können entweder durch syntaktische Satz-/Phrasengrenzen oder durch Akzentsilben begrenzt sein.

Für die Phrasensteuerung werden nur Intonations-Akzente (Wortakzent) als relevant angesehen, da für die Ausbildung von Grundfrequenzbewegungen auf einer Silbe auch eine Mindest-Dauer zur Verfügung stehen muß.

Innerhalb einer Phrase wurden Intonationsakzente als Wort- bzw. Phrasenakzent gelabelt. Akzente werden am Beginn des Silbenkerns der akzentuierten Silbe gelabelt. Der satzfinale Intonations-Akzent besitzt besonderes perzeptives Gewicht und ist meist auch Startpunkt der satz-/phrasenfinalen Intonationsmuster. Daher wird er als Phrasenakzent bezeichnet.

Als Hilfsmittel für die Etikettierung diente die Darstellung von Zeitfunktion, geglättetem Zscore-Verlauf, der Grundfrequenz, sowie der Lautgrenzen (Abb. 1).

2.3 Syntaktische Etikettierung

Die tatsächlich realisierte (und etikettierte) Lautfolge im fließenden Text weicht von der kanonischen Lautung meist erheblich ab. Automatische Silbentrennung und Markierung von Wortgrenzen wird somit innerhalb der Lautlabel zusätzlich erschwert. Daher wurde der Textkorpus manuell mit Satz-, Wort- und Silbengrenzen etikettiert.

Es wurden zwei alternative pragmatische Ansätze für die Silbendefinition verwendet:

- Der erste setzt die Silbengrenze unmittelbar an den Beginn jedes Vokals. Der Vorteil dieses Verfahrens liegt in der einfachen Automatisierbarkeit und dem geringen Aufwand.
- Der zweite Ansatz orientiert sich an phonologisch/akustischen Kriterien. Zunächst wird der morphologische Aufbau des Wortes berücksichtigt, indem Vor- und Nachsilben abgespalten werden und bei Komposita in der Wortfuge getrennt wird. Zur Positionierung der Silbengrenze in Konsonantclustern wird die akustische Trennbarkeit des Signals als Kriterium verwendet (z.B. in Plosivpausen / nach stimmlosen Frikativen). Zur Vermeidung offener Silben mit Kurzvokalen erfolgt eine gleichmäßige Aufteilung intervokalischer Einzelkonsonanten auf benachbarte Silben.

2.4 Datenanalyse

Für die Analyse der Lautdauern, Silbendauern und Phrasendauern wurden die prosodischen und syntaktischen Labelfiles auf die Lautgrenzen abgebildet und automatisch alle relevanten Informationen abgeleitet.

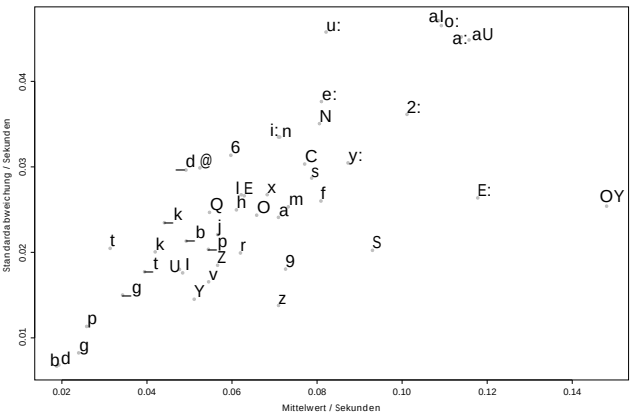

Abbildung 2: Die Standardabweichung in Bezug zum Mittelwert der Lautdauer für die Phonemklassen.

Auf *Lautebene* wurde das Verhältnis der Lautlänge zum Mittelwert aller Laute dieser Klasse, der CAMPBELLsche *zscore* [1]

$$zscore = \frac{Lautdauer - Mittelwert_{\text{Klasse}}}{Standardabweichung_{\text{Klasse}}} \qquad (1)$$

ausgewertet (Abb. 2).

Für die in Abschnitt 2.3 beschriebenen Silbenvarianten wurde je eine *Silbendatenbank* erzeugt, die folgende Informationen enthält: Index der Silbe im Wort, Index des Wortes im Satz, Index des Satzes in der Quelldatei, Filename, Lautstring der Silbe, Silbendauer, Silbenkern-Typ, Funktionswort, Akzenttyp, Phrasenposition und Wortposition (initial, medial, final), Lautzahl der Silbe und Position des Silbenkerns.

Die *Phrasendatenbank* enthält den Index des Teilsatzes in der Quelldatei, den Filenamen, Phrasendauer, Phrasentyp und Anzahl der Silben in der Phrase. Folgende Phrasentypen werden unterschieden: Satzbeginn-Wortakzent, Satzbeginn-Phrasenakzent, Wortakzent-Phrasenakzent, Wortakzent-Wortakzent sowie Phrasenakzent-Satzende.

2.5 Ergebnisse

Lautebene: Es besteht ein eindeutiger Zusammenhang zwischen den Phrasengrenzen und einem Anstieg des *zscore*-Verlaufs (vgl. Abb. 1 und [1]). Der Zusammenhang zwischen Akzentposition und *zscore*-Verlauf ist schwächer und daher in dieser Form nicht für die automatische Etikettierung zu verwenden.

Silbenebene: Die mittlere Silbendauer steht in linearem Zusammenhang mit der Anzahl der enthaltenen Phoneme. Weitere wesentliche Einflußfaktoren sind: Typ des Silbenkerns (Langvokal, Diphthong, Kurzvokal, Reduktionsvokal, silbischer Konsonant), Position der Silbe innerhalb des Satzes (final, nicht final), Akzentsituation (nicht akzentuiert, Wortakzent, Satzakzent) und Informationsgehalt/Auftretenshäufigkeit der Silbe. Problematisch ist die Trennung der verschiedenen Einflüsse, da die formulierten Analysekriterien nicht statistisch unabhängig voneinander sind, sondern sich in ihrer Wirkungsweise verstärken (Akzent und Informationsgehalt) oder kompensieren können (z.B. Kürzung aufgrund der Auftretenshäufigkeit versus finale Dehnung).

Phrasenebene: Die mittlere Phrasendauer steht in linearem Zusammenhang mit der Anzahl der Silben. Eine Modifikation dieser mittleren Dauer erfolgt durch phrasentyp-abhängige Einflußfaktoren.

3 Das System zur Rhythmussteuerung

3.1 Der mehrstufige, hybride Ansatz

Das Ziel einer hybriden datengetriebenen bzw. regelbasierten Rhythmussteuerung ist die Kombination von bewährten Wissenskomponenten mit der Fähigkeit, den Sprechrhythmus zu variieren und sogar sprecherindividuelle Merkmale nachzutrainieren. Die Strategie berücksichtigt vier Gesichtspunkte:

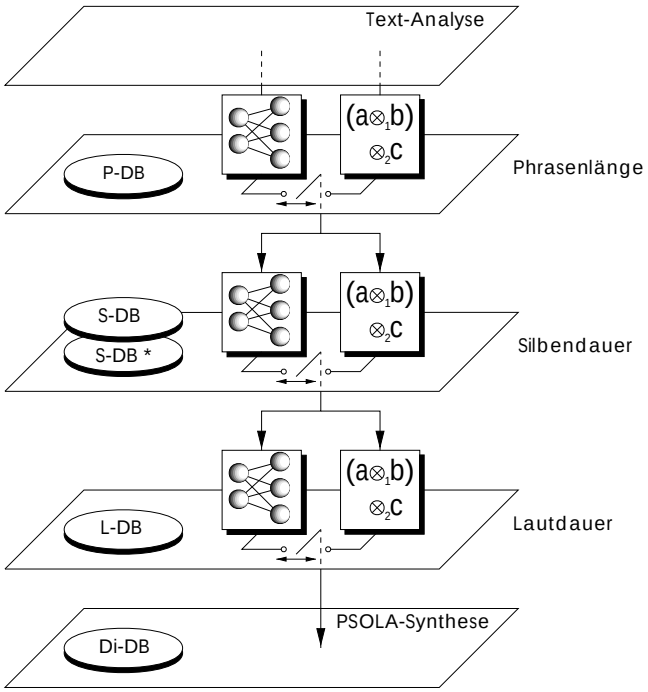

Abbildung 3: Konzept der hybriden Rhythmussteuerung

❶ Einteilung der Segmentdauersteuerung in die drei Repräsentationsebenen: Phrase, Silbe und Laut mit jeweils eigenen Datenknoten zum Training sowie zur Generierung der Zieldauern.

❷ Pro Ebene läuft alternativ ein neuronaler bzw. ein regelbasierter Algorithmus (unter Verwendung jeweils der gleichen Datenbasis).

❸ Die extrahierte prosodische, Silben- sowie Lautdatenbasis (P-DB, S-DB, L-DB) einschließlich der statistischen Parameter stammt von genau einem (variablen) Sprecher.

❹ Das Diphon-Inventar (Di-DB) für die akustische Synthese und die erwähnten P-DB, S-DB, L-DB basieren auf einem identischen Sprecher.

Die geeignete Kombination der Algorithmen soll durch perzeptive Testung gefunden werden. Das Konzept der hybriden Rhythmussteuerung ist in Abbildung 3 dargestellt.

3.2 Der neuronale Algorithmus

Der eingesetzte neuronale Algorithmus entspricht den in [6] vorgestellten Neu*Rosy*-Netzen (ELMAN-Typ). Dabei werden aus binär kodierten (linguistisch-phonetischen) Eingangsattributen direkt Stützwerte prosodischer Konturen (hier: relative Segmentdauern) trainiert und in

der Kannphase vorhergesagt. In Erweiterung zu dem Neu*Rosy*-Ansatz hängt die Eingangskodierung von der Verarbeitungsebene (Phrase, Silbe, Laut) ab. Neben einer silbenweisen Kodierung (1 Fokussilbe + 2 Vorgänger und Nachfolger) mit je 8 Attributen wird mit phrasen- und lautorientierten Attributvektoren und verschiedenen Netzgrößen experimentiert. Im Training wird das Datenmaterial gefenstert und je Lernschritt um ein Segment (Phrase, Silbe oder Laut) verschoben. Zwei verschiedene Silbeneinheiten (Onset-Nucleus-Coda und Nucleus-Coda-Folgeonset) können verarbeitet werden.

Unabhängig vom hybriden Mehrstufenkonzept zur Rhythmussteuerung steht jedoch das Design der konsistenten Phrasen-, Silben-, Laut- und Diphon-Datenbasen derzeit im Vordergrund.

3.3 Regelbasierte Dauersteuerung – Algorithmus und Implementierung

Der Algorithmus zur Lautdauerberechnung basiert auf einem 3-stufigem Modell (Abb. 3). In der ersten Stufe werden für alle prosodischen Phrasen innerhalb eines Satzes die Phrasendauern berechnet. Als zweites wird für jede Silbe die Silbendauer bestimmt. In der letzten Stufe erfolgt die Berechnung der Lautdauern bei gegebenen Silbendauern.

❶ Die Dauer einer Phrase wird in erster Linie durch die Silbenzahl bestimmt. Die mittlere Phrasendauer in Sekunden ergibt sich aus

$$\textit{Phrasenlänge} = 0{,}157 \cdot \textit{Silben} + 0{,}058. \qquad (2)$$

In Abhängigkeit vom Phrasentyp wird die mittlere Phrasendauer über Koeffizienten korrigiert.

❷ Als Grundlage für die Berechnung der Silbendauern wurden die Ergebnisse der statistischen Untersuchung über die Silbendauer in Abhängigkeit von der Lautanzahl, der Akzentuierung, dem Informationsgehalt, dem Typ des Silbenkerns, der Position der Silbe im Wort sowie die Position der Silbe in der Phrase verwendet. Diese Einflußfaktoren auf die Silbendauer wurden durch lineare Abhängigkeiten ausgedrückt. Die ermittelten Silbendauern wurden anschließend für jede Phrase aufsummiert, mit der in Stufe 1 ermittelten Phrasendauer durch lineare Dehnung bzw. Stauchung aller Silbendauern angepaßt.

❸ Die Berechnung der eigentlichen Lautdauern basiert auf den in ❷ berechneten Silbendauern. Dabei wird die unterschiedliche Elastizität der einzelnen Laute berücksichtigt. Es wird angenommen, daß alle Laute einer Silbe einer konstanten Dehnung k unterworfen sind [2]. Die Einzellautdauer ergibt sich aus:

$$\begin{aligned} &\textit{Lautdauer}_i = \exp(\mu_i + k\sigma_i) \qquad (3)\\ &n : \text{Anzahl der Laute in der Silbe} \end{aligned}$$

und steht in Zusammenhang mit der in ❷ ermittelten Silbendauer über die Summenformel:

$$\textit{Silbendauer} = \sum_{i=1}^{n} \exp(\mu_i + k\sigma_i) \qquad (4)$$

Über eine Iteration wird aus der Silbendauer, den Mittelwerten (μ_i) und Standardabweichungen (σ_i) der Lautdauern der Faktor k berechnet. Als Iterationsschrittweite wurde $\Delta k = 0,1$ gewählt. Die Einzellautdauern werden abschließend nach Formel 3 berechnet.

4 Ausblick

Erste informelle Hörtests zeigen, daß der vorgestellte Ansatz zu einer deutlichen Verbesserung der Sprachqualität führt. Die Entscheidung darüber, welche Stufe des hybriden Verfahrens welche Ebene der Verarbeitung übernimmt, wird derzeit noch über umfangreiche und aufwendige perzeptive Tests bestimmt. Zur Bewertung der Güte einer erzeugten Dauerstruktur ist ein objektives Maß notwendig; aufgrund der hohen Variabilität der Lautdauern erscheint jedoch ein Vergleich von synthetischem und natürlichem Sprechrhythmus wenig geeignet.

Literatur

[1] Campbell, N.: *Prosodic influence on segmental quality*. In *ESCA. Eurospeech '95*, pp. 1011–1014, 1995.

[2] Campbell, W. N. and S. D. Isard: *Segment durations in a syllable frame*. J. of Phonetics, 19:29–38, 1991.

[3] Hirschfeld, D.: *Variabilität und Stabilität segmentaler Merkmale unter dem Aspekt der konkatenativen Sprachsynthese – Vokale*. In: *Tagungsband Elektr. Sprachsignalverarbeitung*, S. 94–101, Berlin, 1996.

[4] Hirschfeld, D. und M. Eichner: *Dynamische Bausteinauswahl zu Synthese fließender Sprache*. In: Fellbaum, K. (Hrsg.): *Elektronische Sprachsignalverarbeitung*, Bd. 8. Konferenz, S. 177–183, Cottbus, 1997.

[5] Ihaka, R. and R. Gentleman: *R: A language for data analysis and graphics*. Journal of Computational and Graphical Statistics, 5(3):299–314, 1996.

[6] Jokisch, O. und M. Pescheck: *Neuronale Prosodiegenerierung – Einfluß der Trainingsdaten*. In: *Fortschritte der Akustik – DAGA 98*, Zürich, 1998. (im Druck).

[7] Klatt, D. H.: *Linguistic uses of segmental duration in english*. J. Acoustic Soc. Am., 59:1208–1221, 1976.

[8] Wunderlich, U.: *Sprachsynthesesystem im Zeitbereich*. Diplomarbeit, TU Dresden, 1993.