

Prädiktion von Intonationsverläufen durch „Unit Selection“*

Kerstin Sehnert und Fred Englert

Zusammenfassung

Der Verlauf der Grundfrequenz (F_0) ist eines der wichtigsten prosodischen Merkmale. Die bisher entwickelten – stochastischen oder regelbasierten – Modelle zur Vorhersage von F_0 -Konturen für die automatische Sprachausgabe erscheinen noch immer verbesserungswürdig. Eine neue Möglichkeit der Modellierung ergibt sich aus der Anwendung der „Unit Selection“-Methode als eine nach syntaktischen und phonetischen Kriterien optimierte Auswahl von F_0 -Konturen aus einer Datenbank. In diesem Beitrag werden die Architektur eines derartigen Basissystems für Frage-sätze, Messungen zur Ähnlichkeit von vorhergesagten und natürlichen F_0 -Konturen und die Ergebnisse erster Hörtests vorgestellt.

6.1. Einleitung

Automatische Sprachausgabe spielt in digitalen Medien eine zunehmend wichtige Rolle. Mit der inzwischen realisierbaren Sprachverständlichkeit ist das Erzeugen synthetischer Sprache zu einer Standardtechnik in sogenannten „Interactive-Voice-Response-Systemen“ geworden. Eines der noch immer nicht zufriedenstellend gelösten Probleme ist die Prädiktion von natürlich klingenden und allgemein akzeptablen F_0 -Konturen (Sprechmelodie) für beliebige Texte. Zu den wichtigsten Ansätzen in diesem Bereich zählen neben rein regelbasierten Modellen (Kohler, 1997) auch solche, in denen F_0 -Verläufe durch die „Analyse-durch-Synthese-Methode“ (Möbius et al., 1993) gewonnen werden, sowie Verfahren zur Generierung der Sprechmelodie durch künstliche neuronale Netze (Traber, 1992).

Durch die Verfügbarkeit großer, annotierter Sprachdatenbanken sowie leistungsfähiger Computersysteme wurde für die Sprachausgabe in den letzten Jahren die Technik der „Unit Selection“ ermöglicht, die sogenannte „Nonuniform Units“ für die konkatenative Sprachsynthese verwendet.¹ Die Sprachsynthese resultiert unter diesem Ansatz unmittelbar aus einer Konkatenation von aufgezeichneten Elementen, die in einer linguistisch und phonetisch annotierten Datenbank zusammengefasst sind. Durch die optimale Auswahl möglichst großer Einheiten – Wörtern, Phrasen oder ganzen Sätzen – kann bei einer hinreichend großen und gut strukturierten Datenbank eine sehr gute Synthesequalität erreicht werden. Diese Qualität resultiert nicht nur aus dem

* Erschienen in: *Proceedings der GLDV-Frühjahrstagung 2001*, Henning Lobin (Hrsg.), Universität Gießen, 28.–30. März 2001, Seite 59–68. <http://www.uni-giessen.de/fb09/ascl/gldv2001/>

¹ <http://www.itl.atr.co.jp/chatr/>

Playback der verketteten Einheiten, bei dem oft keine weitere Signalverarbeitung eingesetzt wird, sondern auch aus den bereits in den Einheiten enthaltenen prosodischen Merkmalen.

Problematisch bei diesem Ansatz sind einerseits die hohen Anforderungen an Rechen- und Speicherkapazität, zum anderen der zum Teil drastische Abfall der Synthesequalität beim Zurückgreifen auf kleinere Konkatenationseinheiten. Durch eine Kombination von Diphon-Synthese auf lautlicher Ebene und der Anwendung der „Unit Selection“-Methode im Bereich der Prosodie soll hier ein datengesteuerter Ansatz zur Prädiktion intonatorischer Parameter für deutsche Fragesätze vorgestellt werden.

Die Annahme eines Zusammenhangs zwischen formaler Satzstruktur und Intonation und die Vermutung, dass Sätze oder Satzteile, die bzgl. ihrer strukturellen Merkmale eine geringe Distanz aufweisen, auch mit ähnlichen F_0 -Konturen realisiert werden, waren Argumente für die Wahl der „Unit Selection“-Methode. Diese strukturellen Merkmale können syntaktischer, semantischer und auch phonetisch-phonologischer Natur sein. Ein Überblick zu linguistischen Theorien, die einen Zusammenhang zwischen syntaktischer Organisation und prosodischer Struktur annehmen, wird in Abschnitt 6.2 gegeben.

Die Auswahl der Merkmale zur Textbeschreibung sollte so getroffen werden, dass die wichtigsten Einflussgrößen auf die Intonation einer lautsprachlichen Realisierung berücksichtigt werden. In dieser Untersuchung wurde auf relativ robuste Merkmale zurückgegriffen, die mit geringerem Aufwand aus Texten gewonnen werden können. Als grundlegende Syntheseinheit wurde die Silbe ausgewählt. Jede der in einem Korpus enthaltenen Silben wurde durch eine Menge von Merkmalen charakterisiert, die sich sowohl auf die Eigenschaften der betreffenden Silbe selbst, als auch auf die sie umgebenden Einheiten beziehen. Zusätzlich war jede Silbe mit einem Vektor von sechs F_0 -Werten versehen. Diese wurden vorab durch Messungen aus dem Korpus gewonnen, durch Vektorquantisierung zusammengefasst und den jeweiligen Silben zugeteilt. Abschnitt 6.3 beschreibt die für das Basissystem getroffene Merkmalsauswahl sowie die Zusammenfassung eines Inventars von F_0 -Konturen und beschreibenden Merkmalen in einer Datenbank.

Um eine F_0 -Kontur für einen zu synthetisierenden Text vorhersagen zu können, wird dieser in Einheiten geeigneter Größe zerlegt. Daraufhin werden in der Datenbank möglichst lange Ketten von Einheiten gesucht, die in bezug auf die formalen Merkmale eine möglichst geringe Distanz zum Text aufweisen. Diese Suche wird effizient mittels Dynamischer Programmierung (DP) durchgeführt, einem algorithmischen Konzept, das zur Analyse von mehrstufigen Entscheidungsprozessen dient. Bei Anwendung dieses Prinzips im „Unit Selection“-Verfahren sind die Entscheidungen mit der Auswahl einzelner Konkatenationseinheiten assoziiert. Zwei Funktionen, welche die Distanzen für die Einheiten selbst und für die Übergänge zwischen den Einheiten repräsentieren, gehen in die Berechnungen der Gesamtdistanz für eine Silbenfolge ein. Die global optimale Lösung entspricht einer Folge von Datenbankeinheiten für eine konkrete Äußerung. Dabei entstehen Elemente unterschiedlicher Größe, aus denen sich die endgültige Synthesesequenz zusammensetzt. Die Organisation der Datenbank für eine effiziente Suche mittels DP und die Prinzipien der Distanzberechnung werden in Abschnitt 6.4 erläutert.

Ein Problem bei der Prädiktion prosodischer Parameter ist die Bewertung der Resultate. Als beste Möglichkeit für eine objektive Bewertung scheint zunächst ein numerischer Vergleich mit einer Zielgröße angebracht zu sein. Diese Zielgröße, im vorliegenden Fall die F_0 -Kontur für eine Äußerung, weist aber für verschiedene Realisierungen eines Textes häufig eine starke Variabilität auf. Hinzu kommt, dass eine numerische Bewertung nicht notwendigerweise mit der Bewertung durch Hörer übereinstimmt. Der Versuch einer Bewertung der vorhergesagten Intonationskonturen durch numerische Maße und Hörtests wird in den Abschnitten 6.5 und 6.6 beschrieben.

6.2. Formale Struktur und Intonation

Die Möglichkeit, intonatorische Eigenschaften als Funktion von syntaktischen Beziehungen zwischen Wörtern und Phrasen zu beschreiben, ist von großer Bedeutung für die Text-to-Speech Synthese, da der Umgang mit syntaktischen Strukturen gegenüber automatischen Analysen der Semantik und Pragmatik relativ unproblematisch ist. Als Motivation für die Wahl des „Unit Selection“-Ansatzes zur Prädiktion von Intonationsverläufen gilt die Annahme eines unmittelbaren Zusammenhangs zwischen formaler Struktur und F_0 -Verlauf. Demnach sollten strukturell ähnliche Äußerungen – oder Teile von Äußerungen – durch sich gleichende intonatorische Muster gekennzeichnet sein, die als typisch für eine bestimmte Struktur gelten. Intonatorische Phrasierung bezieht sich auf eine strukturelle Repräsentation syntaktischer oder semantischer Information, welche durch die lineare Abfolge der Wörter repräsentiert wird, während die zugrunde liegenden syntaktischen Strukturen hierarchisch organisiert sind.² Deutlich wird der Zusammenhang zwischen formaler Struktur und Intonation beispielsweise bei der Unterscheidung verschiedener Fragesatztypen (insbesondere Ergänzungs- und Entscheidungsfragesatz). Unterschiede können sowohl durch charakteristische formale Merkmale (wie etwa An- und Abwesenheit finiter Verben, *w*-Phrasen oder Modalpartikeln, sowie Stimmmerkmale dieser Kategorien), als auch durch typische intonatorische Eigenschaften (meist F_0 -Kontur über der Hauptakzentsilbe und Offset) gekennzeichnet werden. Untersuchungen dieser Zusammenhänge finden sich vorwiegend in Arbeiten zur intonatorischen Kennzeichnung von Satzmodi.³

Auch in anderen linguistischen Disziplinen beschäftigt man sich mit der Beziehung von Satzstruktur und Prosodie (insbesondere Grundfrequenz): So gehen beispielsweise Chomsky und Halle (1968) und Jackendoff (1972) (zit. in Caelen-Haumont, 1994) davon aus, dass intonatorische Strukturen Hinweise auf zugrundeliegende syntaktische Gliederungen liefern. Die hauptsächliche Funktion der Intonation besteht in der Reorganisation syntaktischer Strukturen. Selkirk (1986) betrachtet die Bestimmung von prosodischen Domänen in der Satzphonologie. Sie versucht, die lokalen Domänen, innerhalb derer phonologische Prozesse operieren, aus syntaktischen Konstituenten abzuleiten. Damit wird ihr System sprachübergreifend anwendbar. Im Rahmen einer Text-to-Speech Prosodiesynthese definiert Aubergé (1992) ein Lexikon mit Intonationskonturen. Sie stützt sich dabei auf die These, dass ein sog. „Rendezvous-Punkt“ zwischen intonatorischen und syntaktischen Strukturen existiert. Jede linguistische Einheit (Satz, Phrase usw.), in die sich eine Äußerung untergliedern lässt, entspricht einer intonatorischen Einheit und wird durch eine spezifische Intonationskontur charakterisiert.

Die Auswahl der formalen Kennzeichen, welche die Datenbankeinheiten bezüglich ihrer intonatorischen Eigenschaften charakterisieren, wurde so getroffen, dass die Merkmale mit relativ geringem Aufwand aus der Textbasis zu extrahieren waren. Deshalb beziehen sich diese ausschließlich auf die oberflächliche Satzstruktur.

² Aus diesem Grund bezieht sich die Beschreibung „Formale Struktur“ in den folgenden Ausführungen auf die durch die lineare Wortfolge repräsentierte Satzstruktur.

³ Siehe dazu z. B. Barliner (1989), Oppenrieder (1998), Luukko-Vinchenzo (1998) oder Reis und Rosengren (1991).

6.3. Annotation der Datenbank

Als Material zur Erstellung der Datenbank dienten die PHONDAT Signaldateienkorpora⁴. Signaldateien zu jeweils 90 Fragesätzen zweier Sprecher und die dazu gehörigen auf Lautebene segmentierten und etikettierten Labeldateien wurden verwendet. Unter Verarbeitung der in den Labeldateien vorhandenen symbolischen Informationen, wie Transkription von Standardlautung und tatsächlich Produziertem, Betonungszeichen, Wortgrenzen, Satzzeichen und Zeitangaben, wurde das Sprachmaterial annotiert. Zur Wortartenanalyse diente das morphologische Analyseprogramm MORPHY⁵. Da Silben als primäre Konkatenationseinheiten galten, wurde für jede in der Datenbank enthaltene Silbe ein Merkmalsvektor erstellt. Damit nicht nur beste Entsprechungen für eine konkrete Silbe, sondern auch möglichst lange Folgen von Silben ermittelt werden konnten, mussten sowohl Merkmale, welche die Silbe selbst betrafen, als auch Informationen über Kontext und benachbarte Einheiten im Merkmalsvektor festgehalten werden. Jeder Vektor beinhaltete 14 Merkmale (s. Tab. 6.1) die für die Ermittlung optimaler Entsprechungen relevant waren. Da angenommen wurde, die Verteilung der Lautsegmente innerhalb der Silbe sei für die F_0 auf Silbenebene bedeutsam, wurden Silbenstruktur und Reduktionsgrad angegeben. Durch die Kombination von Silbenposition im Wort und Silbenzahl pro Wort ließ sich eine Aussage über die Silbenstruktur eines Wortes treffen. Des weiteren ergab sich eine Darstellung der Akzentstrukturen aus Markierungen der Akzente für aktuelle und umgebende Silben. Eine grobe oberflächenstrukturell-syntaktische Beschreibung resultierte aus der Betrachtung der Wortarten in Verbindung mit satzbezogenen Positionsmerkmalen.

Neben formalen Merkmalen wurden Informationen über den F_0 -Verlauf jeder Silbe registriert. Die F_0 -Messungen erfolgten unter Zuhilfenahme des Analyseprogramms PRAAT⁶. Jeder Silbe wurde eine F_0 -Kontur bestehend aus sechs Werten zugeordnet. Um eine normierte Repräsentation der Konturen zu erreichen, wurde die mittlere F_0 des jeweiligen Sprechers von allen Werten in der Datenbank abgezogen. Die auf diesem Wege erhaltenen F_0 -Konturen wurden durch Vektorquantisierung zusammengefasst. Diese Methode zur Datenkompression bildet eine Menge von Eingabevektoren auf eine weniger große Repräsentantenmenge ab. Die dabei entstehenden Referenzvektoren werden in einem Codebuch gespeichert. Im konkreten Fall wurden die F_0 -Konturen zu 16 Codebuchvektoren zusammengefasst (s. Abb. 6.1).

Für die weitere Verarbeitung standen ausschließlich die Codebuch-Indices zur Verfügung. War nach Auswahl der optimalen Konkatenationseinheiten eine Folge von Silben für eine Eingabesequenz gefunden, wurden die zugehörigen F_0 -Konturen aus dem Codebuch entnommen und zusammengesetzt.

6.4. Auswahl der Konkatenationseinheiten

Für einen aus N Silben bestehend Eingabetext sollte eine optimale Auswahl aus M in der Datenbank enthaltenen Silben gefunden werden. Als optimal wurde eine Folge von Silben angesehen, deren Pfad durch ein Gitter die geringsten „Kosten“ verursachte. Die Knoten des Gitters repräsentierten dabei die möglichen Verknüpfungen von Silben aus dem Eingabetext mit Silben aus

⁴ <http://www.ipds.uni-kiel.de/forschung/kielcorpus.de.html>

⁵ <http://www-psycho.uni-paderborn.de/lezius/titel.html>

⁶ <http://www.fon.hum.uva.nl/praat/>

	Merkmal	Darstellung
1	Silbenstruktur	Repräsentation der Laute innerhalb einer Silbe mit Unterscheidung von stimmlosen Konsonanten, stimmhaften Konsonanten und Vokalen
2	Reduktionen	Anzahl der Abweichungen von der Standardlautung
3–5	Akzentuierung der aktuellen, vorhergehenden und nachfolgenden Silbe	Binäres Merkmal (nimmt die Werte 0 für „nein“ und 1 für „ja“ an)
6	Position der Silbe im Wort	Wortinitial, wortmedial oder wortfinal
7–9	Wortart des aktuellen, vorhergehenden und nachfolgenden Wortes	Substantiv, Eigennamen, Finite Verbform, Infinitiv, Partizip, Imperativ, Auxiliar, Modalverb, erw. Infinitiv mit zu, unbestimmter Artikel, bestimmter Artikel, Adjektiv, Demonstrativ-, Relativ-, Possesiv-, Indefinit-, Interrogativ-, Personal-, Reflexivpronomen, Adverb, Konjunktion, Präposition, Sonderklasse zu, Verbzusatz, Interjektion, Zahlwort, Zahl oder Abkürzung
10	Silbenzahl pro Wort	Zählung der Silben
11–13	Position des aktuellen, vorhergehenden und nachfolgenden Wortes	Satzinitial, satzmedial oder satzfinal
14	Satztyp	Einteilung in V1, V2 oder andere bezüglich der Verbstellung

Tabelle 6.1.: Annotierte Merkmale

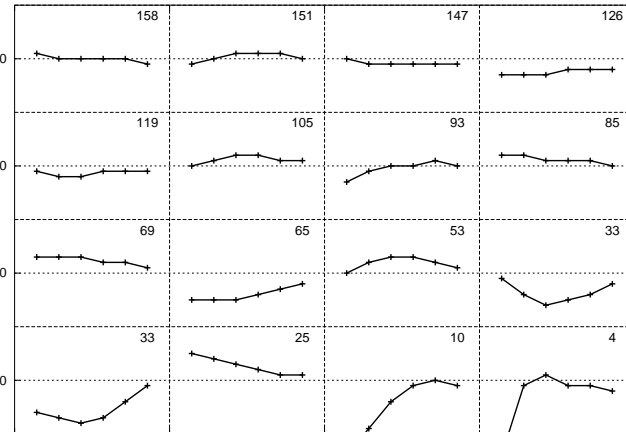


Abbildung 6.1.: Die 16 Codebuchvektoren der silbenbasierten F_0 -Konturen für die Sprecherin; Darstellung in Halbtönen, bezogen auf die mittlere F_0 . Für jede Kontur ist die Auftrittshäufigkeit vermerkt.

der Datenbank (s. Abb. 6.2). Wurde keine weitere Vorauswahl getroffen, entstand ein Gitter mit $M * N$ Knoten.

Ein Pfad bestand aus einer Folge $(i_0, j_0), (i_1, j_1), \dots, (i_N, j_N)$ von Indexpaaren, mit $1 \leq i \leq N$ und $1 \leq j \leq M$. Der Knoten mit dem Index 0 markierte einen externen Startpunkt. Die Kosten für einen Pfad wurden durch zwei Funktionen beschrieben. Die *Übergangskosten*

$$K_A = f_A[(i_k, j_k)|(i_{k-1}, j_{k-1})]$$

wurden für je zwei auf einem Pfad folgende Knoten berücksichtigt. Dabei galt $K_A = 0$ für Silbenpaare, die in der Datenbank – und damit auch in natürlichen Äußerungen – innerhalb eines Satzes direkt aufeinander folgten. Sonst galt $K_A = 1$. Die *Knotenkosten*

$$K_B = f_B(i, j)$$

wurden aus der Differenz der Merkmale von i -ter Eingangsilbe und j -ter Auswahlilbe berechnet. Hierbei wurde eine Normierung angewandt, so dass $K_B = 0$ für Silben mit identischen Merkmalen und $K_B = 1$ für Silben mit komplementären Merkmalen galt. Die Kosten für einen Gesamtpfad

$$K = \sum_{k=1}^N K_{A_k} + K_{B_k}$$

ergaben sich durch die Kombination von Übergangs- und Knotenkosten. Der optimale Pfad wurde effizient mittels Dynamischer Programmierung nach dem „Bellman-Optimality-Principle“ (Deller et al., 1993) berechnet.

6.5. Ergebnisse für unterschiedlich gewichtete Selektionen

Es wurde angenommen, dass einige der formalen Merkmale bezüglich des Auffindens ähnlicher F_0 -Konturen wichtiger waren, als andere. Der Grad der Relevanz einzelner Merkmale konnte

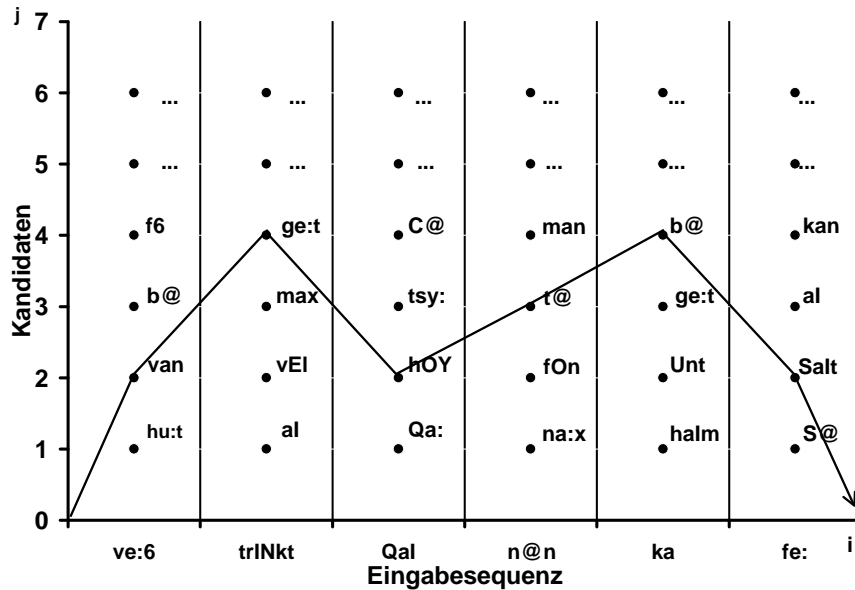


Abbildung 6.2.: Optimaler Pfad durch ein mit Silben gefülltes Gitter für die Eingabesequenz „Wer trinkt einen Kaffee?“.

durch Gewichtungseinstellungen für die Auswahl der Konkatinationseinheiten ausgedrückt werden. Für den Vergleich der Merkmalsvektoren von Eingangs- und Zielsilbe konnten bestimmte Merkmale einen größeren oder kleineren Beitrag zu K_A liefern. Um erste Gewichtungseinstellungen vorzunehmen, wurden die Merkmale in Gruppen zusammengefasst: Merkmale der Gruppe 1, die sich auf strukturelle Eigenschaften der Silbe bezogen, Merkmale der Gruppe 2, die auf Wortebene operierten und Merkmale der Gruppe 3, die Eigenschaften des ganzen Satzes beschrieben. Die Übergangskosten K_B wurden für $K_B \neq 0^7$ verändert. Je nach Gewichtung von K_A und K_B veränderte sich die Auswahl der Konkatinationseinheiten und somit die resultierende F_0 -Kontur. Zur Bewertung dieser Konturen wurde der numerische Vergleich mit einer Kontur vorgenommen, die für den betreffenden Satz tatsächlich realisiert worden war. Da sich der Korrelationskoeffizient im Hörtest (Abschnitt 6.6) und in weiteren Arbeiten zur Evaluierung von F_0 -Konturen (Hermes, 1998) als das numerische Maß erwies, welches dem Höreindruck am nächsten kommt, galt das Korrelationsmaß als Beurteilungskriterium.

K_B wurde bei konstantem K_A im Bereich zwischen 0,1 und 1,0 mit einer Schrittweite von 0,1 variiert, wobei sich folgendes Resultat ergab: Tendenziell wurde der Korrelationskoeffizient geringer, je größer die Kosten für eine unnatürliche Verkettung waren. Offensichtlich verursachte die Konkatination größerer Einheitenfolgen aus der Datenbank keine Verbesserung der Linearen Korrelation (und damit evtl. der Synthesequalität). Die besten Korrelationswerte⁸ lagen für die Sprecherin bei $K_B = 0,3$, für den Sprecher bei $K_B = 0,1$.

Die Gewichtung von K_A wurde für die drei genannten Merkmalsgruppen mit $K_B = 0,2$, $K_B = 0,5$ und $K_B = 0,8$ durchgeführt. Nach Auswertungen der Ergebnisse für beide Sprecher wurde deutlich, dass die besten Korrelationswerte bei Hervorhebung von Merkmalen der Grup-

⁷ Verbindung zweier Silben, die in der Datenbank nicht aufeinander folgten.

⁸ 0,27 und 0,26

pe 1 (dazu zählten Silbenstruktur, Reduktionen, Akzentmerkmale und Silbenposition im Wort) erreicht wurden. Durch Variation der Kosten ließen sich Anhaltspunkte zum Auffinden optimaler Gewichtungsverhältnisse gewinnen.

6.6. Hörtest

Um die F_0 -Verläufe als Intonationskonturen hörbar zu machen, diente eine Diphon-Synthese als lautsprachliche Basis. Dafür wurde aus dem Input für die Synthese⁹ eine automatische, normative Transkription erstellt. Angaben über die Dauer einzelner Laute wurden aus der Gesamtdauer der ausgewählten Silben abgeleitet. Transkription, Dauer und Grundfrequenz bildeten den Input für eine MBROLA Sprachsynthese¹⁰.

Im Rahmen des Hörtests bewerteten 53 Testpersonen 45 Satzpaare, wobei ein Paar aus zwei bezüglich der F_0 -Kontur verschiedenen Realisierungen eines Fragesatzes bestand.¹¹ Anhand der Testauswertungen wurden drei unterschiedliche Aspekte untersucht: Zum einen sollte durch den Vergleich von Hörerurteil und numerischen Maßen das Maß gefunden werden, welches dem auditiven Urteil der Testpersonen am nächsten kam, um für die Weiterentwicklung des Systems auf aufwendige Hörtests verzichten zu können. Des weitern sollte festgestellt werden, ob Hörer die Auswahl längerer Ketten innerhalb der zusammengesetzten Intonationskonturen bevorzugten. Ferner fand eine Gegenüberstellung mit einem – auf Künstlichen Neuronalen Netzen basierenden – System zur Prädiktion von F_0 -Konturen und segmentaler Dauer (Englert, 1999) statt.

Nach Auswertung der Hörerurteile konnte festgestellt werden, dass das Korrelationsmaß in 48,5% der Fälle mit dem auditiven Urteil übereinstimmte, wobei die Übereinstimmungen bezüglich anderer Maße (darunter auch der quadratische Fehler) zwischen 40% und 45% lagen.¹² Obwohl dieses Resultat nicht als eindeutig zu bezeichnen ist, zeigte es eine Tendenz, auf die sich auch die Wahl des Korrelationsmaßes zur numerischen Beurteilung der Intonationskonturen in Abschnitt 6.5 stützte.

Bei den Untersuchungen zu Präferenzen unterschiedlicher Einheitenlängen in den Ausgabesequenzen stellte sich heraus, dass in 52,5% der Fälle die Version des Fragesatzes bevorzugt wurde, deren Bestandteile mit $K_B = 0,2$ ausgewählt worden waren (im Gegensatz zu 34,3% für $K_B = 0,8$ bei 13,1% nicht beurteilten Fragesätzen). Das bedeutet, es schien den Höreindruck positiv zu beeinflussen, dass die Einzelteile der Gesamtkette in bezug auf ihre Merkmale denen der Eingabesequenz ähnlicher waren, obwohl es mehr Schnittstellen gab, an denen einzelne kleinere Konturen zusammengefügt wurden.

Für die Gegenüberstellung von „Unit Selection“-Ansatz und Synthese mit Künstlichen Neuronalen Netzen ergab sich folgendes Resultat: Die Intonationskonturen, die durch „Unit Selection“ erzeugt worden waren, wurden in 58,3% der Fälle bevorzugt; in 33,7% der Entscheidungen wurden die durch Neuronale Netze generierten F_0 -Konturen als besser beurteilt. Die durch „Unit Selection“ hergestellten Konturen zeigten wesentlich größere Bewegungen in den F_0 -Verläufen und wurden von den Probanden möglicherweise aus diesem Grund favorisiert. Offensichtlich wurden größere Sprünge zwischen aufeinanderfolgenden F_0 -Werten, die durch fehlende Anglei-

⁹ Fragesätze aus dem PHONDAT Korpus, die nicht in der Datenbank vorhanden waren.

¹⁰ <http://tcts.fpms.ac.de/synthesis/>

¹¹ Für jedes Satzpaar sollte angegeben werden, welche der beiden Versionen bevorzugt wird. Kriterien zur Beurteilung wurden den Hörern nicht vorgegeben.

¹² Für jeweils 13,1% der Satzpaare konnte von den Hörern kein Urteil abgegeben werden.

chung der Teilkonturen oder nicht optimales Datenbankdesign entstanden sein könnten, zugunsten größerer F_0 -Bewegungen toleriert.

6.7. Zusammenfassung

Im vorliegenden Beitrag wurde der Versuch beschrieben, intonatorische Parameter für deutsche Fragesätze mittels „Unit Selection“ zu erzeugen. Intonationskonturen für deutsche Fragesätze wurden durch die Konkatenation möglichst großer Einheiten aus einer annotierten Datenbank erzeugt, wobei Silben als grundlegende Konkatenationseinheiten fungierten. Als zentrale Probleme wurden die Wahl der Annotationsmerkmale und das Auswahlverfahren zur Suche einer optimalen Entsprechung für eine Eingabesequenz erläutert.

Zur Beurteilung der resultierenden Intonationskonturen wurde ein Hörtest durchgeführt, in dessen Rahmen sich die Lineare Korrelation als das numerische Maß erwies, welches dem Höreindruck zur Qualitätsbeurteilung einer F_0 -Kontur am nächsten kam. Anhand dieses Korrelationsmaßes wurde der Versuch unternommen, Anhaltspunkte für die Optimierung des Auswahlverfahrens zu finden. Als Resultat konnte festgehalten werden, dass Merkmale, die sich auf die primäre Konkatenationseinheit Silbe beziehen, zum Erlangen eines besseren Ergebnisses stärker gewichtet werden sollten als Merkmale, die größere Einheiten beschreiben. Das Ergebnis eines Vergleichs mit Intonationskonturen, die durch Künstliche Neuronale Netze generiert wurden, zeigte eine Präferenz der Probanden für die durch „Unit Selection“ erzeugten Konturen.

Im Ganzen ist das hier vorgestellte System eine erste Realisierung, die an unterschiedlichen Stellen verbessert werden könnte: Mögliche Ansatzpunkte sind effizientes Korpusdesign (das speziell auf die hier angewandte Technik ausgerichtet sein könnte), weitere Optimierungen des Selektionsvorgangs und Nachbearbeitungen der Konturelemente bei der Konkatenation einzelner Teilkonturen.

Literaturverzeichnis

- AUBERGÉ, V. (1992): “Developing a structured lexicon for synthesis of prosody”. In: *Talking Machines*, herausgegeben von Bailly, G. und Benoit, C., Elsevier, S. 247–307.
- BATLINER, A. (1989): “Eine Frage ist eine Frage ist keine Frage. Perzeptionsexperimente zum Fragemodus im Deutschen”. In: *Zur Intonation von Modus und Fokus im Deutschen*, herausgegeben von Altmann, H.; Batliner, A. und Oppenrieder, W., Tübingen: Niemeyer, S. 87–109.
- CAELEN-HAUMONT, G. (1994): “Semantic and Pragmatic Prediction of Prosodic Structures”. In: *Fundamentals of Speech Synthesis and Speech Recognition*, herausgegeben von Keller, E., Chichester: John Wiley & Sons, S. 271–293.
- CHOMSKY, N. UND HALLE, M. (1968): *The sound pattern of English*. New York: Harper & Row.
- DELLER, J. R.; PROAKIS, J. G. UND HANSEN, J. H. (1993): *Discrete-Time Processing of Speech Signals*. New York: Macmillan. S. 623 ff.
- ENGLERT, F. (1999): “We’ve got rhythm (but no melody). An experiment with basic input parameters for prosody networks”. In: *Papers in Phonetics and Linguistics*, herausgegeben von Wodarz, H. W., Frankfurt am Main, S. 1–10.
- HERMES, D. J. (1998): “Measuring the Perceptual Similarity of Pitch Contours”. *Journal of Speech, Language and Hearing Research* 41: S. 73–82.

JACKENDOFF, R. (1972): *Semantic Interpretation in Generative Grammar*. Cambridge: MIT Press. Zitiert nach Caelen-Haumont (1994).

KOHLER, K. J. (1997): "Parametric Control of Prosodic Variables by Symbolic Input in TTS Synthesis". In: *Progress in Speech Synthesis*, herausgegeben von van Santen et al., J. P. H., New York: Springer, S. 495–475.

LUUKKO-VINCENZO, L. (1998): *Formen von Fragen und Funktionen von Fragesätzen*. Tübingen: Niemeyer.

MÖBIUS, B.; PÄTZOLD, M. UND HESS, W. (1993): "Analysis and synthesis of German f_0 contours by means of Fujisaki's model". *Speech Communication* 13 (1): S. 53–61.

OPPENRIEDER, W. (1998): "Intonatorische Kennzeichnung von Satzmodi". In: *Intonationsforschungen*, Tübingen: Niemeyer, Nummer 200 in Linguistische Arbeiten.

REIS, M. UND ROSENGREN, I. (Herausgeber) (1991): *Fragesätze und Fragen*. Tübingen: Niemeyer.

SELKIRK, E. (1986): "On derived domains in sentence phonology". In: *Phonology Yearbook*, Nummer 3, S. 371–405.

TRABER, C. (1992): " f_0 generation with a database of natural f_0 patterns and with a neuronal network". In: *Talking Machines*, herausgegeben von Bailly, G. und Benoit, C., Elsevier, S. 287–304.