Datenbasierte und linguistisch interpretierbare Intonationsmodellierung

Uwe Reichel

München 2010

Datenbasierte und linguistisch interpretierbare Intonationsmodellierung

Uwe Reichel

Dissertation
an der Fakultät für Sprach- und Literaturwissenschaft
der Ludwig–Maximilians–Universität
München

vorgelegt von
Uwe Reichel
aus München

München, den 18.03.2010

Erstgutachter: Prof. Dr. Jonathan Harrington
Zweitgutachter: PD Dr.Ing. Florian Schiel
Tag der mündlichen Prüfung: 19.07.2010

Danksagung

Diese Arbeit entstand an der Ludwig-Maximilians-Universität München am Institut für Phonetik und Sprachverarbeitung. Mein besonderer Dank gilt dem Institutsvorsitzenden und meinem Doktorvater Prof. Dr. Jonathan Harrington für seine fortwährende Betreuung und Unterstützung. Durch Einführung eines regelmäßigen Doktorandentreffens und dadurch, dass er auch darüber hinaus bei fachlichen Fragen stets ansprechbar war, hat er ein ideales Umfeld geschaffen, in dem diese Arbeit entstehen konnte.

Weiter möchte ich Prof. em. Dr. Hans Tillmann danken, der mir den Weg zur Forschungsgemeinschaft der Sprachsynthese eröffnete und damit zu fruchtbarem fachlichen Austausch mit Vertretern dieses Gebiets, was für diese Arbeit nur förderlich war.

Sehr inspirierend waren stets die fachlichen Diskussionen mit Hartmut Pfitzinger, von dessen großer Erfahrung in phonetischen wie sprachtechnologischen Bereichen ich profitieren durfte, und der mich entscheidend dazu motivierte, phonetische Forschung mit Sprachtechnologie zu verbinden.

Weiter möchte ich Florian Schiel danken für die vielen hilfreichen Kommentare zu Vorträgen im Institutsrahmen im Zusammenhang mit dieser Arbeit.

Großer Dank gebührt auch Felicitas Kleber, Claudia Kuzla und Katalin Mády für ihre wertvollen Kommentare zur experimentellen Untersuchung der Intonationswahrnehmung. Katalin Mády möchte ich außerdem danken für ihren unermüdlichen Einsatz beim Korrekturlesen.

English Summary

In this thesis a data-driven and linguistically interpretable intonation model for the automatic analysis and synthesis of fundamental frequency (F0) contours was developed.

The intonation model The model can be characterised as parametric, contour-based, and superpositional. F0 contours are treated as a superposition of global and local components. These components are anchored in a hierarchic prosodic structure defined by global and local segments which correspond roughly to intonation phrases and accent groups respectively. The stylisation of the F0 contours is carried out as follows: Within each global segment a linear F0 base contour is fitted. After the subtraction of this global baseline a third order polynomial is fitted to the F0 residual within each local segment. Subsequently, a symbolic description of the intonation inventory in form of global and local contour classes is derived by polynomial coefficient clustering. On the phonetic level, linear regression models adjust these abstract units to the respective prosodic context.

As to the parametric and contour-based description, the model stands in the tradition of Fujisaki (1987), Möhler (1998b) and Taylor (2000). As to superposition, it stands in the tradition of Fujisaki (1987). As in Möhler und Conkie (1998) stylisation parameter clustering is carried out. Regarding the following aspects the approach chosen here provides additional benefit to intonation research: (1) The requirements for data preprocessing are comparably low. F0 stylisation was carried out in F0 sections at syllable nuclei, rendering an exact syllable segmentation unnecessary. The extraction of the prosodic structure is restricted to prosodic phrase boundaries guided by signal pauses, punctuation and part-of-speech information. Pitch accent localisation and classification is not needed. Due to this a complete automation of the preprocessing steps with acceptable quality is achieved, so that there is no need for a manual data preparation by experts. This property allows for a fast adaptation of the model to new speech data and avoids inconsistencies caused by incomplete inter-labeller agreement. Due to the partly text-based definition of prosodic structure, automatic preprocessing includes a signal-text alignment needed for subsequent linguistic interpretation. (2) In contrast to the more complex stylisation functions of the models mentioned above, the polynomial stylisation chosen in this study guarantees an analytic approximation and thus a biunique relation between the F0 to be modelled and its abstraction. This property is essential to partition the F0 stylisations into intonation classes based on their contour similarity as well as for later linguistic interpretation. At the same time the chosen polynomial order is capable of capturing

F0-coded prominence and boundary behaviour.

Linguistic interpretation The linguistic interpretability of local contour classes was examined for the concepts *significance, informational novelty*, and *utterance finality*. The approach chosen here can be described as follows: first, by automatic linguistic corpus analyses hypotheses about possible relations between contour classes and linguistic concepts are generated. These hypotheses are subsequently tested by perception experiments. By these means a systematic linguistic anchoring of the model was achieved in form of a decision tree to predict the linguistically appropriate contour class. The adequacy of its predictions was assured by a further perception test.

Conclusion It has been shown, that it is possible to build a perceptually acceptable and linguistically interpretable representation of intonation in a purely data-driven manner. This bottom-up approach guarantees consistency and easy adaptability of the model to new data. Due to its simultaneous signal proximity and linguistic anchoring, it covers the entire chain from text to signal and therefore can be used for intonation analysis and generation on a linguistic as well as on a phonetic-acoustic level. It is qualified for employment in speech technology applications as well as in phonetic fundamental research to automatically analyse raw speech data.

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

Ziel Ziel dieser Arbeit ist die Entwicklung eines Intonationsmodells, das folgenden Kriterien genügt:

- **Datenbasiertheit:** Das Modell soll automatisch aus Sprachdaten ableitbar sein, so dass auf manuelle Etikettierung verzichtet werden kann.
- **Interpretierbarkeit:** Es wird eine intonatorische Beschreibung angestrebt, die weitestmöglich linguistisch interpretierbar ist.
- **Anwendbarkeit:** Die sich aus den ersten beiden Kriterien ergebende Signalnähe und linguistische Verankerung soll das Modell unmittelbar zur maschinellen Analyse und Synthese von F0-Konturen qualifizieren.

Gliederung Im Forschungsüberblick in **Teil I** werden nach Behandlung wichtiger Aspekte der Intonation und prosodischen Struktur die bis zum jetzigen Zeitpunkt einflussreichsten Intonationsmodelle kategorisiert und vorgestellt (Kapitel 2 und 3). Anhand der präsentierten Modelle wird im Anschluss beschrieben, wie aus der F0-Kontur eine abstrakte Repräsentation gewonnen werden kann (Kapitel 4), wie sich diese abstrakte Repräsentation linguistisch interpretieren lässt (Kapitel 5) und wie umgekehrt aus der Repräsentation wieder eine konkrete F0-Kontur generiert werden kann (Kapitel 6). In Kapitel 7 folgt eine Diskussion der Modelle im Hinblick auf wesentliche Anforderungen an die Intonationsmodellierung.

Teil II hat die Entwicklung und Anwendung des in dieser Arbeit entwickelten PKS-Intonationsmodells zum Thema. P steht hierbei für parametrisch, K für konturbasiert und S für superpositional. Nach einer grundlegenden Vorstellung der Modellarchitektur (Kapitel 8) sowie der Trainingsdaten und deren Vorverarbeitung (Kapitel 9) folgen Modelldetails zur Überführung der F0-Konturen in eine Intonationsrepräsentation und umgekehrt (Kapitel 10). Die mathematischen und perzeptiven Evaluierungsergebnisse des Modells werden in Kapitel 11 zusammengefasst.

Teil III beinhaltet die linguistische Interpretation der modellgegebenen Intonationsrepräsentation in Hinblick auf Semantik (Kapitel 14) sowie Diskurs (Kapitel 15 und 16).

Hierbei werden anhand von statistischen Korpusanalysen Hypothesen über die Funktionen der Intonationseinheiten aufgestellt und mittels Perzeptionsexperimenten überprüft. Die gewonnenen Befunde dienen der Entwicklung eines linguistischen Modells zur Intonationsvorhersage. Dieses PKS-EB-Modell, das in Form eines Entscheidungsbaums (EB) vorliegt, wird seinerseits perzeptiv evaluiert (Kapitel 17).

Die Teile II und III schließen jeweils mit einer Diskussion zu Methodik, Resultaten und noch offenen Fragen (Kapitel 12 und 18).

Im abschließenden **Teil IV** wird nach einer knappen Zusammenfassung der entscheidenden Ergebnisse dieser Arbeit kurz auf weitere potentielle Einsatzbereiche des PKS-Modells eingegangen.

Teil I

Forschungsüberblick

Überblick Zunächst wird der Untersuchungsgegenstand *Intonation* eingehend beleuchtet als Teilbereich der Prosodie, der sich akustisch hauptsächlich als Grundfrequenzverlauf manifestiert. Betrachtet werden hierbei auch Aspekte der Intonationswahrnehmung. Im Anschluss werden Dichotomien zur Charakterisierung von Ansätzen der Intonationsmodellierung vorgeschlagen und einige bekannte Intonationsmodelle unter Bezugnahme auf dieser Dichotomien vorgestellt. Bezogen auf die vorgeschlagenen Unterscheidungskriterien und die vorgestellten Modelle werden daraufhin Verfahren zur Analyse und Synthese von Grundfrequenzkonturen beschrieben sowie Befunde der linguistischen Interpretation der Analyseergebnisse. Der Forschungsüberblick schließt mit einer Diskussion über die Anforderungen an ein Intonationsmodell, die zugleich die Basis legt für das in dieser Arbeit entwickelte und im Teil II präsentierte Modell.

Kapitel 2

Aspekte der Intonation

Intonation umfasst als Teil der Prosodie die melodischen Aspekte einer lautsprachlichen Äußerung.

2.1 Intonation und Prosodie

Prosodie bezeichnet alle *suprasegmentellen* Eigenschaften lautsprachlicher Äußerungen, worunter alle Phänomene verstanden werden, deren Wirkungsbereich größer ist als ein einzelnes Lautsegment. Die traditionelle Beschränkung der Prosodie auf ausschließlich linguistisch und paralinguistisch relevante Suprasegmentalia kann mittlerweile wohl in Anbetracht der Ausweitung phonetischer Untersuchungen auf Zusammenhänge zwischen prosodischen und extralinguistischen Phänomenen (Harrington et al., 2007) aufgegeben werden.[1]

Suprasegmentalia lassen sich nach Lehiste (1970) linguistisch in folgende Bereiche unterteilen:

- Quantität: distinktive Länge sprachlicher Einheiten,
- Intensität: Betonung,
- Intonation: Sprechmelodie.

Unter Intonation werden also die melodischen Aspekte der Prosodie verstanden. Zwischen den genannten Teilbereichen bestehen wechselseitige Abhängigkeiten, nicht zuletzt deshalb, weil ihnen teilweise dieselben akustischen und artikulatorischen Parameter zugrundeliegen:

- Dauer, Sprechgeschwindigkeit, Rhythmus,
- Energie,

[1] Als paralinguistisch gelten hierbei beispielsweise Emotion und Sprechstil, als extralinguistisch Alter und Geschlecht.

- Grundfrequenz-(F0)-Verlauf,
- artikulatorischer Aufwand.

So können sich betonte Silben (siehe Abschnitt 2.3.2) durch höhere zeitliche Ausdehnung, markantere F0-Bewegungen und durch erhöhten artikulatorischen Aufwand auszeichnen.

2.2 Intonation und Grundfrequenz

Die Grundfrequenz einer lautsprachlichen Äußerung als akustisches Hauptkorrelat der Intonation wird durch diverse Einflussfaktoren auf unterschiedlichen Ebenen bestimmt.

2.2.1 Segmentale Ebene

Mikroprosodie Mikroprosodie setzt sich zusammen aus *intrinsischer* und *kointrinsischer* F0. Intrinsische F0 bezeichnet die Lautabhängigkeit segmentinterner durchschnittlicher F0-Werte. So weisen hohe Vokale sprachunabhängig eine gegenüber tiefen Vokalen erhöhte F0 auf (Di Cristo, 1985; Whalen und Levitt, 1995). Weiter beeinflusst auch der Lautkontext die segmentale F0, was als kointrinsische F0 bezeichnet wird. So führen benachbarte stimmlose Konsonanten in vokalischen Segmenten gegenüber stimmhaften Konsonanten zu einer Erhöhung der Grundfrequenz, sowohl in CV-Sequenzen (Lehiste, 1970) als auch in VC-Sequenzen (Kohler, 1982). Die Erhöhung erstreckt sich hierbei über das gesamte vokalische Segment (Löfquist, 1975; Reichel und Winkelmann, 2010).

In der Mehrzahl der phonetischen Studien wird davon ausgegangen, dass mikroprosodische Effekte nicht willentlich vom Sprecher gesteuert, aber vom Hörer wahrgenommen werden und dabei beispielsweise als akustischer Cue für Stimmhaftigkeit (Kohler, 1982) dienen.

Eine Erschwernis bei der Isolierung dieser mikroprosodischen Effekte besteht in ihrer zusätzlichen Abhängigkeit von der Makroprosodie. So treten beispielsweise in betonten und äußerungsinitialen Silben mikroprosodische Unterschiede zwischen Vokalen deutlicher zu Tage (Silverman, 1984).

Ein umfassender Überblick über phonetische Befunde hierzu findet sich in Di Cristo und Hirst (1986).

Trunkierung, Kompression Weitere auf Segmentebene zu findende Einflussfaktoren auf den F0-Verlauf sind die Phänomene *Trunkierung* und *Kompression* (Grønnum, 1990), mit denen Strategien bezeichnet werden, wie der Sprecher den Intonationsverlauf über stimmlose Segmente vervollständigt. Im Falle der Trunkierung bricht die F0-Bewegung zum stimmlosen Segment ab (Erikson und Alstermark, 1972; Grabe, 1998), im Fall der Kompression wird sie so gestaucht, dass sie vor Ende des Stimmtons abgeschlossen werden kann. Nach aktuellem Forschungsstand ist die bevorzugte Wahl einer der beiden Strategien sprachabhängig (Rathcke, 2008).

Spektrale Unterstützung der Intonation F0-Verläufe werden auf Segmentebene spektral unterstützt. So konnte beispielsweise Niebuhr (2009) bei hohen steigenden F0-Konturen gegenüber tiefen fallenden in Frikativen höhere *Centers of Gravity*[2] feststellen, und in Vokalen eine Absenkung des ersten sowie eine Erhöhung des zweiten Formanten, was auf eine geschlossenere und frontiertere Vokalproduktion bei hohen F0-Werten schließen lässt. Diese Cues spielen auch eine wichtige Rolle bei der Tonhöhenwahrnehmung über längere stimmlose Passagen oder geflüsterte Äußerungen (Higashikawa und Minifie, 1999). Stoll (1984) konnte eine positive Korrelation zwischen F2 und wahrgenommener Tonhöhe feststellen, und Traunmüller (1987) identifizierte als Korrelat der wahrgenommenen Tonhöhe in geflüsterten Vokalen die sogenannte *sibilant pitch F2'* als Mittelwert des zweiten und höheren Formanten.

2.2.2 Silben- und lexikalische Ebene

Tonsprachen In Tonsprachen wie beispielsweise dem Mandarin oder dem Vietnamesischen werden den Silben phonologisch distinktive Töne, sogenannte *Toneme* zugeordnet. Charakterisieren lassen sie sich durch Bewegungsmuster (*Konturtöne*) oder durch zu erreichende Zieltonhöhen (*Registertöne*).

Tonakzentsprachen In Tonakzentsprachen wie dem Schwedischen und Norwegischen erstrecken sich bedeutungsunterscheidende F0-Verläufe über ganze Wörter.

Intonationssprachen Die übrigen Sprachen, in denen weder Toneme noch Tonakzente auftreten, werden als Intonationssprachen bezeichnet. Diesem Sprachtyp lässt sich beispielsweise auch das Deutsche zuordnen.

2.2.3 Phrasen-, Satz- und Äußerungsebene

Oberhalb der lexikalischen Ebene dient der F0-Verlauf der Phrasierung von Äußerungen, also dem Zusammenfassen inhaltlich zusammengehöriger Abschnitte zu Intonationsphrasen sowie der Hervorhebung relevanter Segmente.

Weiter macht der F0-Verlauf den Satzmodus kenntlich und codiert, ob eine Äußerung fortgesetzt (progredienter, d. h. nicht absinkender Verlauf) oder abgeschlossen wird (finaler, in Aussagesätzen absinkender Verlauf).

Globale Aspekte Der globale Verlauf der Sprechmelodie lässt sich hierbei unter anderem anhand der folgenden Kenngrößen charakterisieren:

- *Register:* Die Verwendung dieses Begriffs in der Intonationsforschung ist sehr variabel. Eine Zusammenfassung unterschiedlicher Definitionen findet sich in Rietveld und Vermillion (2003), demnach sich Register im Wesentlichen definieren lässt (a)

[2] *Center of Gravity*: Gewichteter Frequenz-Mittelwert im Amplitudenspektrum.

als Abstand eines F0-Abschnitts zu einer Referenzfrequenz, beispielsweise dem gemessenen F0-Minimum eines Sprechers (Ladd, 1992) oder (b) einem nach unten und oben begrenzter Frequenzbereich, der durch den Abstand der Begrenzungslinien (der *Baseline* und der *Topline*) charakterisiert ist (Connell und Ladd, 1990). Im letzteren Sinne legt das Register für Zeitabschnitte einer Äußerung den Frequenzbereich fest, in dem sich lokale F0-Bewegungen abspielen können.[3]

- *Downtrend:* Tendenz, dass die F0 im zeitlichen Verlauf sinkt; Downtrend lässt sich unterteilen in:
  - *Deklination:* Fallen von Baseline und Topline im zeitlichen Verlauf (Pike, 1945). Die Topline fällt in Deklarativ-Äußerungen tendenziell stärker (Cohen et al., 1982; Ladd, 1984), was dazu führt, dass F0-Gipfel im Laufe einer Äußerungseinheit zunehmend flacher werden.
  - *Downstep:* tiefere Realisierung von F0-Gipfeln in Abhängigkeit des vorausgehenden tonalen Kontexts – Ein Phänomen, was zunächst für westafrikanische Sprachen beschrieben (Welmers, 1959; Stewart, 1965) und dann auf andere Sprachen übertragen wurde (Pierrehumbert, 1980).
  - *Final lowering:* überdurchschnittlich starke Absenkung des letzten Akzents (Liberman und Pierrehumbert, 1984).
- *Inklination:* Anstieg von Base- und/oder Topline, beispielsweise bei Alternativ- und deklarativ formulierte Fragen im Niederländischen (Haan, 2001).
- *Pitch Reset:* Neujustierung des Registers (de Pijper und Sandermann, 1994) nach vorangegangener Deklination oder Inklination.

Für die Downtrend-Phänomene wird im Wesentlichen der im Laufe einer Äußerungseinheit nachlassende subglottale Druck verantwortlich gemacht (Collier, 1975; Titze, 1989b; Strik und Boves, 1995). Einige Studien verweisen auch auf eine Beteiligung der laryngalen Muskulatur (Öhman, 1968; Fujisaki, 1991). Das Zustandekommen des Downtrends wird teils als passive Reaktion auf den innerhalb eines Atemzyklusses fallenden subglottalen Druck erklärt, und teils als aktiv gesteuertes Instrument zur linguistischen Codierung (Ohala, 1990). Für letztere wird die Sprechatmung zur Steuerung des subglottalen Druckverlaufs verantwortlich gemacht (Strik und Boves, 1995) sowie die laryngale Muskelaktivität (Ohala, 1990).

Lokale Aspekte Intonatorisch relevante lokale F0-Bewegungen sind mit akzentuierten Silben oder Grenzen zwischen Äußerungseinheiten verbunden und basieren phonatorisch im Wesentlichen auf der Aktivität des Cricothyroid-Muskels (Collier, 1975). Sie werden in Abschnitt 2.3 genauer behandelt.

[3] Zusätzliche Verwendung findet der Begriff Register im Zusammenhang mit der Beschreibung von Phonationstypen (Laver, 1980), wo in Abhängigkeit des Schwingungsverhaltens der Stimmlippen zwischen Modal-, Falsetto und Strohbassregister unterschieden wird.

2.2.4 Para- und extralinguistische Ebene

Paralinguistik umfasst Faktoren wie Emotion und Sprechstil, Extralinguistik Faktoren wie Alter und Geschlecht. Befunde zu Zusammenhängen zwischen Emotion und Intonation finden sich unter anderem bei Uldall (1960) und Tischer (1993), und zwischen Sprechstil und Intonation bei Blaauw (1995) sowie Hirschberg (2000). Mit Auswirkungen des Alterns auf F0 befassen sich beispielsweise Linville (2001) sowie Xue und Deliyski (2001) und mit geschlechtsabhängiger F0 Carlson (1981) und Titze (1989a).

2.2.5 Intonationsbegriff in dieser Arbeit

Diese Arbeit beschränkt sich auf die Modellierung derjenigen Aspekte der Intonation, die auf Phrasen-, Satz- und Äußerungsebene anzusiedeln sind. Para- und Extralinguistik werden also ebenso ausgeklammert wie segmentale Effekte und F0-Muster im Kontext von Ton- und Intonationssprachen.

2.3 Intonationsverankerung: Prosodische Struktur

Die prosodische Struktur einer Äußerung dient ihrer Gliederung dahingehend, dass die enthaltene Information in verarbeitbaren Einheiten übermittelt wird und die wichtigsten Inhalte hervorgehoben werden. Die Struktur lässt sich festmachen an Phrasengrenzen und Akzenten, an denen die Intonationskontur verankert wird.

2.3.1 Prosodische Phrasengrenzen

Prosodische Phrasengrenzen zerlegen eine Äußerung in Einheiten, innerhalb derer die Intonation einer Äußerung beschrieben werden kann. Diese Einheiten werden in der Literatur aus diskursanalytischen oder intonationsphonologischen Betrachtungswinkeln behandelt. Im ersten Fall liegt der Schwerpunkt auf der Eigenschaft dieser Einheiten, inhaltlich zusammengehörige Äußerungsteile zusammenzufassen (*sense units* nach Selkirk, 1984), im zweiten Fall auf ihrer Eigenschaft als Domäne zur Ausbildung von Intonationskonturen. In diesem Zusammenhang werden die Segmente als Intonationsphrasen oder intermediäre Phrasen bezeichnet. In Abschnitt 3.2 wird darauf genauer eingegangen.

Phonetische Korrelate Prosodische Phrasengrenzen werden im Wesentlichen durch die folgenden akustischen Grenzsignale markiert:

- Pausen (Swerts und Geluykens, 1994),
- Grenztöne (Brown et al., 1980), die den Melodieverlauf unmittelbar vor der Grenze bestimmen (vgl. Abschnitt 3.2). Sie dienen der Codierung von Satzmodus sowie von Äußerungsende beziehungsweise -fortführung. Im Deutschen markiert wie in vielen anderen Sprachen ein *progredienter* nicht-fallender Intonationsverlauf eine Fortsetzung einer Äußerung und (zumindest bei Deklarativsätzen) ein *terminaler* fallender

Intonationsverlauf deren Ende. Dialoguntersuchungen (beispielsweise von Politikerinterviews) haben ergeben, dass ein unkonventioneller Gebrauch dieser Grenztöne das Gelingen des Dialogs beeinträchtigen kann (Beattie et al., 1982).

- Diskontinuierlicher Verlauf der Grundfrequenz (de Pijper und Sandermann, 1994), zumeist als *Pitch Reset*, die Rücksetzung des etwa durch Deklination im Laufe einer Phrase modifizierten Registers.

- *Präfinale Längung* (Wightman et al., 1992), worunter die Längung von Silben am Phrasenende zu verstehen ist.

- Reduzierung grenzübergreifender koartikulatorischer Effekte (Cho, 2004; Kuzla, 2009).

Perzeptionsexperimente mit delexikalisierten Stimuli (de Pijper und Sandermann, 1994) haben ergeben, dass diese akustischen Merkmale auch unabhängig von lexikalischer, syntaktischer und semantischer Information als Grenzsignale interpretiert werden.

2.3.2 Akzente

Akzentuierung bezeichnet die Hervorhebung linguistischer Einheiten, was perzeptiv zu einer Erhöhung ihrer *Prominenz* (Auffälligkeit) führt. Im Kontext der prosodischen Strukturierung ist vor allem die Akzentuierung auf *Phrasenebene* interessant, die von der Akzentuierung auf *Wortebene* (auch *Wortbetonung* genannt) abzugrenzen ist. Im Folgenden werden die Begriffe Akzent und Akzentuierung stets im Zusammenhang mit der Phrasenebene verwendet.

Phonetische Korrelate Akzentuierung lässt sich anhand der folgenden akustischen Parameter festmachen:

- Dauer,

- Grundfrequenz (und deren Verlauf),

- Intensität,

- spektrale Zusammensetzung von Lauten.

Die Abhängigkeit der Akzentuierung von Dauer, F0 und Intensität wurde unter anderem in Experimenten von Fry (1955, 1958) untersucht. Akzentuierung geht demnach einher mit einer Längung und Intensitätserhöhung der betroffenen Silbe sowie mit einer F0-Änderung über dem Silbenkern.

Frys experimentelles Design, das in der Untersuchung von Einwort-Stimuli bestand, war allerdings nicht dazu geeignet, die akustischen Korrelate von Akzenten und Wortbetonung auseinanderzuhalten. Für letztere wurde nach Untersuchung nicht akzentuierter Wörter im Deutschen im Wesentlichen eine längere Silbendauer (Dogil, 1995) festgestellt,

und im Niederländischen eine Änderung der spektralen Balance dahingehend, dass die Intensitätserhöhung nicht über das gesamte Spektrum, sondern nur in dessen mittleren Bereich zu beobachten ist (de Sluijter und van Heuven, 1996).

Die verkürzte Vokaldauer in unakzentuierten Silben kann gegenüber akzentuierten Vokalen eine Änderung der Vokalqualität bewirken, da die artikulatorische Zielkonfiguration nicht erreicht wird (*artikulatorischer undershoot*), was zur Zentralisierung der Vokale mit entsprechender Veränderung ihrer spektralen Charakteristik führt (Lindblom, 1963).

Die Prominenzverhältnisse mehrerer aufeinanderfolgender Akzente sind von deren Position in der Äußerung abhängig. Perzeptionsexperimente mit delexikalisierten Stimuli (Terken, 1991, 1994) ergaben, dass ein in der Äußerung weiter hinten liegender Akzent mit niedrigerer Tonhöhe realisiert werden muss, um als gleich prominent empfunden zu werden wie ein Akzent weiter vorne, eine gleiche Tonhöhe hat dagegen eine relative Erhöhung der Prominenz des hinteren Akzents zur Folge. Verantwortlich für diesen Effekt ist die Deklinationserwartung des Hörers.

2.3.3 Assoziation und Alinierung

Wie bei Ladd (1996) präzisiert, ist bei der Verankerung der Intonationskontur in der prosodischen Struktur zu unterscheiden zwischen Assoziation und Alinierung. Assoziation bedeutet die wechselseitige Zuordnung von Einheiten der segmentalen Ebene, strukturgebenden Ereignissen (Akzente und Phrasengrenzen) und intonatorischen Ereignissen (zum Beispiel ein F0-Anstieg).

Welche Einheiten auf der segmentalen Ebene zur prosodischen Assoziation herangezogen werden, ist sprach- und theorieabhängig. So halten im Englischen als sogenannte *Tontragende Einheiten* (*tone bearing units TBU*) Vokale (Goldsmith, 1976), Silben (Pierrehumbert, 1980) und metrische Füße (Beckman und Pierrehumbert, 1986b) her, während im Japanischen Moren als TBUs angenommen werden (Pierrehumbert und Beckman, 1988).

Unter Alinierung versteht man das genaue zeitliche Zusammenspiel der Ereignisse auf den unterschiedlichen Beschreibungsebenen. Ein Beispiel hierfür ist die häufig beobachtete Verzögerung des F0-Gipfels gegenüber dem Silbenkern (*peak delay*). Das Ausmaß der Verzögerung ist sprach- und dialektabhängig, beispielsweise stellten Atterer und Ladd (2004) einen größeren *Delay* für das Süddeutsche gegenüber dem Norddeutschen fest. Zudem ergab sich in diversen Studien eine größere Verzögerung bei nicht phrasenfinalen Akzenten gegenüber phrasenfinalen (zum Beispiel Silverman und Pierrehumbert (1990) für das Amerikanische Englisch und Mücke et al. (2006) für das Deutsche), sowie eine positive Korrelation mit der Länge des Reims der akzentuierten Silbe (van Santen und Hirschberg, 1994; Rietveld und Gussenhoven, 1995).

2.4 Sprachabhängigkeit der Intonation

Intonatorische Unterschiede zwischen Sprachen lassen sich nach Ladd (1996) unterteilen in

- *systemische Unterschiede* im intonatorischen Inventar,
- *phonotaktische Unterschiede* in der Aufeinanderfolge intonatorischer Einheiten und ihrer Beziehung zur segmentalen Ebene,
- *realisatorische Unterschiede* hinsichtlich der phonetischen Realisierung, wie die im vorangehenden Abschnitt besprochenen Dialektunterschiede in der Alinierung von F0-Gipfeln.
- *semantische Unterschiede* bei der linguistischen Interpretation der Intonation.

Peters (2006) diskutiert die Herangehensweise bei der Ermittlung solcher Unterschiede und stellt vergleichende Analysen zu Dialekten des Deutschen vor.

2.5 Perzeption der Intonation

In diesem Abschnitt sollen Erkenntnisse über grundlegende Aspekte der Intonationswahrnehmung zusammengetragen werden, die bei der Entwicklung einer geeigneten Repräsentation von F0-Verläufen nützlich sind: geeignete psychoakustische Maße der Tonhöhe, Beschränkungen des perzeptiven Systems bei der Tonhöhenwahrnehmung sowie die Perzeption von Intonationskonturen.

Auf perzeptive Urteile höherer Ebene zur linguistischen Bedeutung von Intonation wird an entsprechenden Stellen zu Intonationsmodellen und linguistischer Interpretation von Intonationskonturen (Kapitel 3 und 5) eingegangen.

2.5.1 Tonhöhenwahrnehmung

Akustische Cues Für die Tonhöhenwahrnehmung (engl. *pitch*) sind die Grundschwingung mit der Frequenz F0 sowie vor allem die dritte bis sechste Harmonische entscheidend (Ritsma, 1967), anhand derer die Tonhöhe über den größten gemeinsamen Teiler auch dann rekonstruiert werden kann, wenn die Grundschwingung gar nicht im Signal vorhanden ist. Periphere Erklärungsansätze zu dieser sogenannten *virtuellen Tonhöhe* verweisen auf die neben der Ortscodierung existierende zeitliche Codierung der Tonhöhe (Wever, 1930), derzufolge sich die Periodendauer des akustischen Signals, die sich ja bei fehlendem Grundton nicht ändert, in der Periodendauer des neuronalen Entladungsmusters wiederfindet. Zentralnervöse Erklärungsansätze wie in Terhardt (1979) sehen die F0-Rekonstruktion beispielsweise als Mustervervollständigungsprozess. Ein Überblick über Theorien hierzu ist in Terhardt (1998) zu finden.

Interpolation Das menschliche Gehör ist in der Lage, den F0-Verlauf über kurze Signalpausen (kleiner 200 ms), wie sie in stimmlosen Abschnitten des Sprachsignals auftreten, zu interpolieren (Nooteboom et al., 1978).

Perzeptiv motivierte F0-Maße Bei der perzeptiven Beurteilung der Äquivalenz zweier F0-Konturen spielen weniger absolute F0-Werte als vom Register abstrahierte F0-Verhältnisse eine Rolle. Zur Veranschaulichung: perzeptiv äquivalent zu einer F0-Bewegung von 100 auf 110 Hz ist ein F0-Verlauf von 200 auf 220 Hz (und nicht auf 210 Hz). Diesem Sachverhalt trägt nicht die absolute Hertz-Skala, wohl aber Verhältnisskalen Rechnung.

Hermes und van Gestel (1991) ließen Versuchspersonen die F0-Höhe von Akzenten so anpassen, dass ihre Prominenz als äquivalent zu Referenzstimuli anderen Registers empfunden wurde. Die Äquivalenzurteile konnten am besten auf einer *Equivalent-Rectangular-Bandwidth*-Skala (ERB; Moore und Glasberg, 1996) nachgestellt werden.

Bei Nolan (2003) mussten Versuchspersonen Intonationsmuster von männlichen und weiblichen Sprechern reproduzieren. Hier erwiesen sich die Halbton- (HT) und die ERB-Skala mit den geringsten Abweichungen zwischen Original- und reproduzierten Konturen als am geeignetsten zur Messung der empfundenen Äquivalenz.

2.5.2 Beschränkungen des perzeptiven Systems

Perzeptive Sensitivität bei lautsprachlichen Stimuli

Die Übertragung psychoakustischer Befunde anhand einfacher Stimuli wie Sinustönen auf die Tonhöhenwahrnehmung lautsprachlicher Stimuli ist problematisch, da allgemein gilt: je komplexer das Signal, desto weniger sensitiv das perzeptive System gegenüber F0. Systematisch wurde die Verschlechterung der Tonhöhenwahrnehmung bei Lautsprache beispielsweise in Abhängigkeit zeitlich variabler spektraler Charakteristik untersucht (t'Hart et al., 1990).

Absolute Schwellen

F0-Bereich und Stimulusdauer Eine Tonhöhenwahrnehmung ist in einem F0-Bereich ab 40 Hz möglich und verschlechtert sich deutlich ab etwa 4000 Hz (Henning, 1966). Für eine stabile Tonhöhenbestimmung ist eine Präsentationsdauer von mindestens 6 Perioden der Grundschwingung nötig (Doughty und Garner, 1948).

Tonhöhenänderung (*glissando threshold*) Die Wahrnehmungsschwelle g für Tonhöhenänderungen wird in Hz/s oder HT/s gemessen. Sergeant und Harris (1962) fanden für Tonglissandi dauerabhängige Schwellen zwischen 1 Hz/s bei Stimulusdauern von 10 s und 150 Hz/s bei 100 ms Darbietungszeit. Nach t'Hart et al. lässt sich diese Schwelle für sprachliche Stimuli nach der Formel $g = \frac{0.16}{T^2}$ berechnen, wobei T für die Stimulusdauer (in s) steht.

Unterschiedsschwellen

Statische Töne Bei statischen Tönen kann das menschliche Gehör etwa 640 Frequenzen unterscheiden. Die Unterschiedsschwelle (*just noticable difference JND*) ist abhängig von Frequenz sowie – bei kurzen (< 100 ms) oder leisen (< 20 phon) Darbietungen – von Dauer und Lautstärke der präsentierten Stimuluspaare. Bis zu 1 kHz liegt die JND bei etwa 3 Hz, darüber steigt sie progressiv an (Ritsma, 1965; Nordmark, 1968).

Die Befunde zu sprachlichen Stimuli (synthetisierte Vokale) variieren stark hinsichtlich der gefundenen Unterschiedsschwellen zwischen 2 und 7 Hz (Flanagan und Saslow, 1958; Isačenko und Schädlich, 1970; Rossi, 1971).

Tonhöhenänderung (*differential glissando threshold*) Hier geht es um die Beurteilung, ob zwei Glissandi dieselbe oder unterschiedliche Tonhöhenänderungen aufweisen. Gemessen wird diese Schwelle als Quotient der F0-Änderungen. Psychoakustische Experimente hierzu wurden von Pollack (1968); Nabelek und Hirsh (1969) unternommen. Bei lautsprachlichen Stimuli stellte Klatt (1973) für isolierte Vokale eine Schwelle von $\frac{g_1}{g_2} > 1.7$ fest, und t'Hart et al. für interkonsonantische Vokale eine Schwelle von 2, wobei g_1 und g_2 für die zu vergleichenden F0-Änderungen in den vokalischen Segmenten stehen.

Modellierung

Auf Grundlage der beschriebenen Beschränkungen des Wahrnehmungsapparats bei der Verarbeitung von F0-Verläufen entwickelten d'Alessandro und Mertens (1995) ein in Abschnitt 4.3.5 genauer vorgestelltes F0-Stilisierungsverfahren.

2.5.3 Wahrnehmung von Intonationskonturen

Konturen vs. Töne

House (1990) postulierte ausgehend von den oben beschriebenen Befunden zur abnehmenden Empfindlichkeit der Tonhöhenwahrnehmung bei steigender Komplexität der Stimuli in seiner *Tonal Movement Coding*-Hypothese, dass Intonation in eher stationären Sprachsignalen (z. B. Vokal-Stimuli) in Form von tonalen Bewegungen, also **Konturen** perzipiert wird, während bei wachsender Zeitveränderlichkeit des Signals (z. B. in Vokal-Plosiv-Vokal-Sequenzen) statt kompletter Konturen nur noch Sequenzen von **Tönen** wahrgenommen werden können. In ABX-Experimenten zur intonatorischen Zuordnung ließ sich diese Hypothese bestätigen: waren Vokalstimuli zu vergleichen, erfolgte die Zuordnung zu den Ankerstimuli konturgeleitet, während die Zuordnung bei Vokal-Konsonant-Vokal-Folgen auf dem Vergleich der Tonhöhen an den Stimulusrändern basierte.

Gleichheits- und Ähnlichkeitswahrnehmung

Die Untersuchung der Ähnlichkeitswahrnehmung von Intonationskonturen spielt eine wichtige Rolle in der Intonationsmodellierung (t'Hart et al., 1990, vgl. Abschnitt 3.9).

So fanden t'Hart et al., dass F0-Konturen durch eine perzeptiv nicht unterscheidbare Sequenz von Geradenstücken ersetzt werden können.

Korrelationen zwischen objektiv-mathematischen Distanzmaßen von F0-Konturen und zumeist ordinal gemessenen empfundenen Distanzen erreichten Werte bis etwa 0.7 (Hermes, 1998; Clark und Dusterhoff, 1999). Reichel et al. (2009) konnten feststellen, dass Ähnlichkeitsurteile bei wiederholter Darbietung derselben Konturpaare relativ konsistent gegeben werden und trainierten auf Grundlage dieser Urteile neuronale Netze zur Vorhersage der empfundenen Distanz von F0-Konturen auf 1-Silbern.

Kategoriale Wahrnehmung der Intonation

Die Messung der empfundenen Ähnlichkeit wird hochgradig erschwert durch Nonlinearitäten zwischen akustischem Kontinuum und Perzept. So konnte beispielsweise Kohler (1987) kategoriale Wahrnehmung bei kontinuierlich variierten Alinierungen zwischen F0-Gipfel und Kern der akzentuierten Silbe feststellen. Die Kategoriale Wahrnehmung im Sinne eines scharfen Kategorieübergangs sowie hoher Diskriminationsfähigkeit im Übergangsbereich ergab sich hierbei zwischen *frühem* und *mittlerem Gipfel*. *Früher Gipfel* bedeutet eine Vorverlagerung des F0-Maximums vor den Kern der akzentuierten Silbe, *mittlerer Gipfel* die Gleichzeitigkeit von Gipfel und Kern. Die Perzeption des Kontinuums zwischen *mittlerem* zu *spätem Gipfel* (Gipfel zeitlich nach dem Silbenkern) war dagegen gradueller Natur.

Der Identifikationstest wurde in indirekter Form durchgeführt und bestand in der Aufgabe, die Angemessenheit einer intonatorisch systematisch variierten Zieläußerung *Sie hat ja gelogen* im Kontext des Satzes *Jetzt verstehe ich das erst* zu bewerten. Dieser Kontext implizierte, dass die Zieläußerung neue Information trug.

Da ein früher Gipfel im Gegensatz zum mittleren und späten zur Codierung neuer Information als unangemessen beurteilt wurde, konnte ihm die Diskursfunktion *Gegebene Information* zugewiesen werden. *Mittlere* und *späte* Gipfel erhielten nach Kohler (1987, 1991) die Diskursfunktionen *Neu* und *Überraschend Neu*. Hierauf wird in Abschnitt 5.3 noch eingegangen.

Das Auftreten kategorialer Wahrnehmung von Intonationskonturen wird allerdings durch einer Vielzahl späterer Befunde in Frage gestellt:

- Ein Definitionskriterium kategorialer Wahrnehmung ist eine hohe Übereinstimmung zwischen der anhand der Identifikationsergebnisse vorhergesagten und der im Diskriminationstest empirisch ermittelten Diskriminationsfähigkeit. In den wenigen Studien, in denen neben Kohler (1987) überhaupt ein Diskriminationstest durchgeführt wurde, stellte sich häufig der Zusammenhang zwischen vorhergesagter und empirisch ermittelter Diskriminationsfähigkeit als nur sehr gering heraus. So beispielsweise in Ladd und Morton (1997), die bei der Beurteilung der Prominenz in Abhängigkeit der Höhe des F0-Gipfels zwar scharfe Kategoriengrenzen im Identifikationstest aber keinen daraus vorherzusagenden Verlauf der Diskriminationsfähigkeit feststellen konnten.

- Kategoriale Wahrnehmung ist abhängig vom Stimulus-Design. Niebuhr (2007a) stellte beispielsweise fest, dass Stimuli mit höherer Dynamik (schnelleren F0- und Intensitätsverläufen) in weit stärkerem Ausmaß scharfe Kategoriengrenzen hervorrufen als es Stimuli mit niedrigerer Dynamik tun.

- Kategoriale Wahrnehmung lässt sich als Artefakt der Fragestellung verstehen, wie Thomassen (1993) und Schouten et al. (2003) beim Vergleich von in unterschiedlichen Konstanzverfahren ermittelten Diskriminationsergebnissen feststellten. So führten Verfahren, die auch eine Kategorisierung der Stimuli beinhalten (*ABX*- und *2IFC*-Design[4]) zu einer Bestätigung der im Identifikationstest vorhergesagten Diskriminationsfähigkeit, während dies Verfahren ohne implizite Kategorisierung (*4IAX*[5])weit weniger oder gar nicht taten.

Im Hinblick auf indirekte Identifikationstest-Designs wie in Kohler (1987), die auf Beurteilung der Angemessenheit einer Kontur im Diskurs beruhen, besteht die Gefahr einer Abhängigkeit der Ergebnisse vom Grad der Vereinbarkeit der gewählten Diskurskategorien. So sind die Diskurskategorien *Neue* und *Überraschend neue Information* weniger unvereinbar als *Neue* und *Gegebene Information*. Dies könnte auch ein Grund dafür sein, dass im ersten Fall graduelle und im zweiten Fall kategoriale Wahrnehmung der zugehörigen Konturen festgestellt wurde.

[4] *ABX: "Ist Stimulus X gleich Kategorie A oder B", 2IFC (Two-Intervall-Forced-Choice): "Ordnen Sie die Stimuli X und Y im Hinblick auf ihre Ähnlichkeit zu Kategorie A"*

[5] *4IAX (Four-Interval-AX): "Welches der beiden präsentierten Stimuluspaare enthält unterschiedliche Stimuli?"*

Kapitel 3

Intonationsmodelle

3.1 Unterteilungskriterien

Die in den nächsten Abschnitten vorgestellten Intonationsmodelle lassen sich anhand der folgenden Kenngrößen unterteilen:

- Einheiten der F0-Abstrahierung: ton- vs. konturbasiert,
- Beschreibung der Einheiten: symbolisch vs. parametrisch,
- Gewinnung der Einheiten: perzeptiv vs. objektiv-mathematisch,
- Anordnung der Einheiten: einschichtig vs. superpositional.

3.1.1 Einheiten der F0-Abstrahierung: ton- vs. konturbasiert

Tonbasierte Abstrahierung In tonbasierten Modellen sind die intonationsphonologisch relevanten Einheiten F0-Zielpunkte, also Töne. Dieser Ansatz fußt auf der Tradition des Amerikanischen Strukturalismus mit Vertretern wie Pike (1945) und Wells (1945). In dieser Schule wurde das *Vier-Ebenen-Modell* entwickelt, das als bedeutungsunterscheidende Intonationsbausteine ("*pitch phonemes*") vier F0-Niveaus annimmt (*low, mid, high, overhigh*), die an bestimmten prosodisch relevanten Äußerungsstellen auftreten. Seine Fortsetzung fand dieser Ansatz in Tonsequenzmodellen (TSM), die in Abschnitt 3.2 beschrieben werden. Die F0-Kontur ergibt sich hier also aus einer Abfolge dieser Akzenten und Phrasengrenzen zugeordneten Zielpunkte (vgl. Abbildung 3.1). Daraus ergibt sich eine *unterspezifizierte* Repräsentation des F0-Verlaufs verbunden mit der Annahme, dass die F0-Konturen zwischen den Tönen hinreichend genau durch Interpolation abgeleitet werden können. Als Rechtfertigung für diese ökonomische aber zugleich stark abstrahierende Darstellungsform können Perzeptionsexperimente wie von Isačenko und Schädlich (1964) herangezogen werden, in dem kurze Äußerungen mit *Aussage-, Frage-, Kontrast-* und *nonfinaler* Intonation mit abstrahiertem F0-Verlauf resynthetisiert wurden. Trotz dieser Abstrahierung, die in der Reduzierung der Intonation auf eine diskontinuierliche

Sequenz zweier Frequenzniveaus bestand, waren die Versuchspersonen in der Lage, den Stimuli die intendierten Intonationskategorien zuzuordnen.

Konturbasierte Abstrahierung In konturbasierten Ansätzen hingegen sind die phonologisch relevanten Einheiten nicht die Ton-Targets sondern F0-Bewegungen, Diese Sichtweise steht in der Tradition Bolingers (1951), der den amerikanischen Strukturalisten entgegenhielt, dass ihr 4-Ebenen-System nichtexistente intonatorische Kontraste vorhersage und sich zugleich nicht zur Beschreibung diverser existierender Intonationsmuster eigne. Weiter steht dieser Ansatz in der Tradition der Britischen Schule (Halliday, 1967a), die in Palmer (1922) ihren Ausgang nimmt und Intonation mittels dynamischer, also konturbezogener Merkmale wie *steigend* und *fallend* beschreibt.

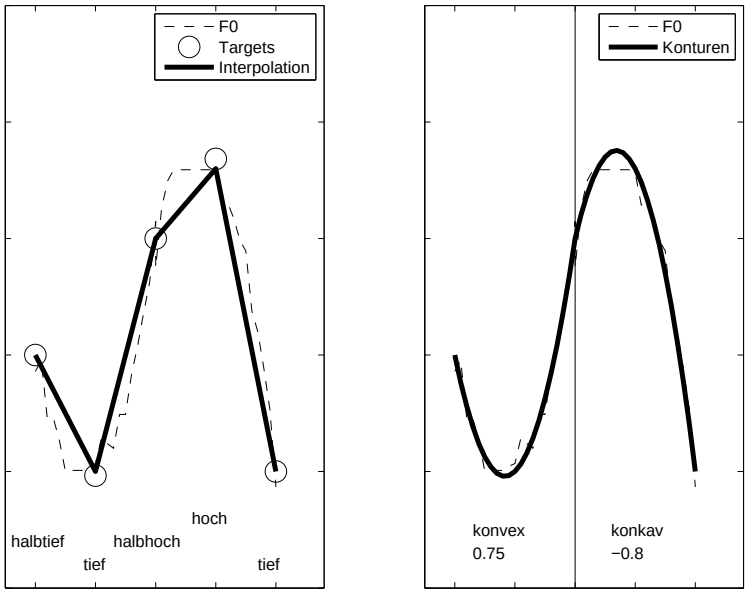

Abbildung 3.1: Ton- vs. konturbasierte F0-Abstrahierung mit symbolischer vs. parametrischer Beschreibung der Einheiten. **Links:** F0-Verlauf als Abfolge symbolisch etikettierter Targets. **Rechts:** F0-Verlauf als Abfolge von Konturen mit symbolischen Etiketten bzw. Krümmungskoeffizienten der Stilisierungsparabeln.

3.1.2 Beschreibung der Einheiten: symbolisch vs. parametrisch

Symbolische Beschreibung Die in Abbildung 3.1 dargestellte F0-Kontur lässt sich sowohl symbolisch als auch parametrisch beschreiben. Eine symbolische Beschreibung besteht in der Etikettierung der Kontur mit diskreten Labels aus einem endlichen Inventar. Im Falle der tonbasierten Abstrahierung kann das beispielsweise eine Abfolge von F0-Ebenen sein wie: *halbtief, tief, halbhoch, hoch, tief*, im Falle der konturbasierten Abstrahierung eine Abfolge formbeschreibender Symbole wie *konvex, konkav*.

Parametrische Beschreibung Parametrisch lassen sich die Abschnitte des F0-Verlaufs beispielsweise in Form der Krümmungskoeffizienten der dem Verlauf angepassten Parabeln repräsentieren.

3.1.3 Gewinnung der Einheiten: perzeptiv vs. mathematisch-objektiv

Perzeptive Gewinnung Perzeptiv motivierte F0-Modellierung basiert auf der Befragung von Versuchspersonen beziehungsweise in der prosodischen Etikettierung von Signalen durch Experten. Ersteres Vorgehen dient dem Erwerb perzeptiv-phonetischen Wissens. Im letzteren Fall sind die Freiheitsgrade zur Festlegung der Einheiten bereits durch eine Theorie *top-down* vorgegeben, ein Ansatz, der den gezielten Einsatz linguistischen und phonetischen Vorwissens erlaubt und somit die linguistische Verankerung der Intonationsbeschreibung, also ihre grundsätzliche Interpretierbarkeit aus Blickwinkeln der Semantik, Diskursanalyse etc. sicherstellen kann.

Mathematische Gewinnung Beim mathematischen Ansatz steht anstelle der Befragung von Versuchspersonen oder Experten die automatisierte Beschreibung von F0-Verläufen als Funktionen der Zeit. In die Modellierung können Vorwissen oder Annahmen über das Zustandekommen der Konturen mit einfließen. Dieses Vorwissen kann linguistische, phonetische oder physiologische Constraints für die F0-Verläufe umfassen. Bei einer reinen *Bottom-up-Orientierung*, also einer Modellierung, die unter Verzicht auf Vorwissen allein an der Oberflächenbeschaffenheit der F0-Kontur ausgerichtet ist, lässt sich deren linguistische Interpretierbarkeit erst post hoc ermitteln.

3.1.4 Anordnung der Einheiten: einschichtig vs. superpositional

Während einschichtige Beschreibungen die F0-Kontur im Frequenzbereich nicht weiter zerlegen, nehmen superpositionale Ansätze eben solche Zerlegungen der F0-Kontur in mehrere Komponenten vor (vgl. Abbildung 3.2). Die Originalkonturen sind also repräsentiert als (beispielsweise additive oder multiplikative) Verknüpfung ihrer Teilkomponenten. Es bieten sich Zerlegungen in globale und lokale F0-Bewegungen an, wobei die globalen Bewegungen mit Sprecherspezifika sowie größeren prosodischen Einheiten wie Intonationsphrasen assoziiert werden können, und die lokalen Bewegungen mit kleineren prosodischen Einheiten wie beispielsweise akzentuierten Silben oder Akzentgruppen (bestehend aus einer akzentuierten mit umgebenden nicht-akzentuierten Silben). Auch segmentale Einflüsse auf den F0-Verlauf lassen sich mit Hilfe dieser superpositionellen Ansätze mitmodellieren.

3.1.5 Einteilung der Intonationsmodelle

Gegeben die in den vorangegangenen Abschnitten behandelten Kenngrößen lassen sich die nun vorzustellenden Intonationsmodelle anhand eines Klassifikationsbaums wie in Abbildung 3.3 darstellen:

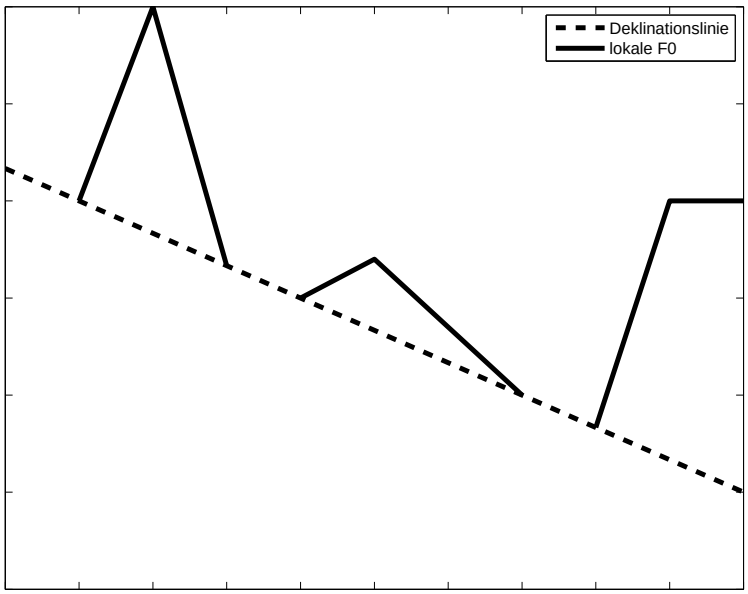

Abbildung 3.2: Superpositionale Darstellung des F0-Verlaufs als Überlagerung von globaler Deklinationslinie und lokalen F0-Bewegungen.

In der folgenden Beschreibung einzelner Modelle werden zunächst die symbolischen Intonationsbeschreibungen behandelt, gefolgt von den parametrisch-einschichtigen und abschließend den parametrisch-superpositionalen Modellen.

3.2 Tonsequenzmodell

Charakteristika: *tonbasiert, symbolisch, perzeptiv, einschichtig.*

Das Tonsequenzmodell (TSM) nach Pierrehumbert (1980) und Beckman und Pierrehumbert (1986b) fußt konzeptuell auf der level-basierten Intonationsbeschreibung des amerikanischen Strukturalismus und auf der Autosegmentalen Phonologie (Goldsmith, 1976), die Laute und suprasegmentale phonologische Phänomene auf getrennten Ebenen repräsentiert und einander über Assoziationslinien zuordnet, wodurch der angenommenen wechselseitigen Unabhängigkeit dieser Phänomene Rechnung getragen wird. Das Modell ist zudem auf Kompatibilität mit der metrischen Phonologie hin konzipiert (vgl. Abschnitt 5.2), genauer, auf die Überführung metrischer Bäume in Intonation ausgerichtet.

Nach dem hier vorgestellten Tonsequenzansatz von Pierrehumbert (1980) lässt sich eine Äußerung prosodisch segmentieren in Intonationsphrasen (*IP*), die sich nach einer Modellaktualisierung durch Beckman und Pierrehumbert (1986b) weiter in intermediäre Phrasen (*ip*) unterteilen lassen. Der F0-Verlauf innerhalb dieser Phrasen wird nun als Abfolge von Tönen beschrieben, die den akzentuierten Silben und Silben im Umfeld von Phrasengrenzen zugeordnet werden. Beide Phrasentypen bestehen aus mindestens einem

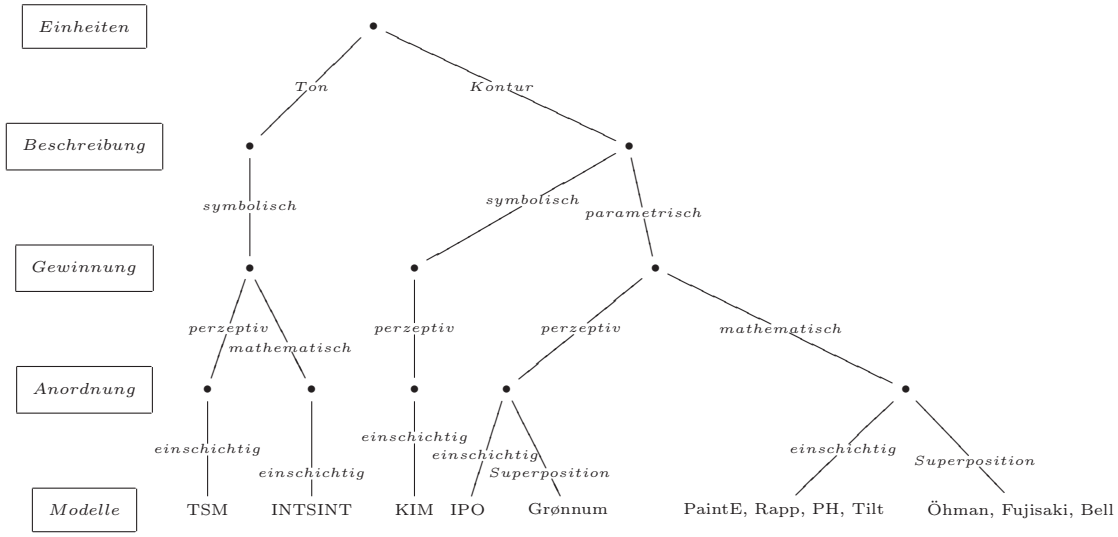

Abbildung 3.3: Unterteilung der Intonationsmodelle.

Tonhöhenakzent (*pitch accent*) und einem *Phrasenton* (*phrase accent*), der dem letzten Tonhöhenakzent in der Phrase folgt. Intonationsphrasen werden zusätzlich durch einen *Grenzton* (*boundary tone*) abgeschlossen.

Tonhöhenakzente verleihen den mit ihnen assoziierten Silben Prominenz, Phrasen- und Grenzton determinieren den Intonationsverlauf zwischen dem letzten Tonhöhenakzent und der Phrasengrenze.

Dem in der Britischen Schule hervorgehobene sogenannten *nuklearen Akzent* kommt bei Pierrehumbert (1980) keine über die Funktion von Tonhöhenakzenten hinausgehende Bedeutung zu. Während bei Palmer (1922) der nukleare Akzent (*nucleus* in dessen Terminologie) der prominentesten Silbe zugeordnet wird und das einzige obligatorische Element der Intonationskontur darstellt,[1] definiert Pierrehumbert (1980) den nuklearen Akzent lediglich über seine Position als den letzten Tonhöhenakzent einer Intonationsphrase ohne dessen Prominenz zu spezifizieren.

Das Toninventar wurde gegenüber dem vierstufigen System des amerikanischen Strukturalismus auf zwei elementare Töne reduziert (H=hoch, L=tief, jeweils in Relation zum vorangehenden Ton), die sich zu komplexen Tönen kombinieren lassen.

Das bisher Gesagte lässt sich zur regulären Intonationsgrammatik in Abbildung 3.4 zusammenfügen.

Im Labelinventar werden elementare Töne mit '+' zu komplexen verbunden, wobei '$*$' dabei den Ton mit der akzentuierten Silbe verknüpft. $H + L^*$ bedeutet also beispielsweise, dass die Tonhöhe von einem hohen Punkt aus vor der akzenttragenden Silbe in einen tiefen Stimmbereich abfällt (*früher Gipfel*). H wird hierbei als *Leitton* (*leading tone*) bezeichnet. In $H^* + L$ ist L der *Folgeton* (*trailing tone*). Grenztöne, die am Rand

[1] Nach Palmer besteht eine Intonationskontur aus einem fakultativen *head*, einem obligatorischen *nucleus* auf der prominentesten Silbe, sowie einem fakultativen vom *nucleus* determinierten *tail*.

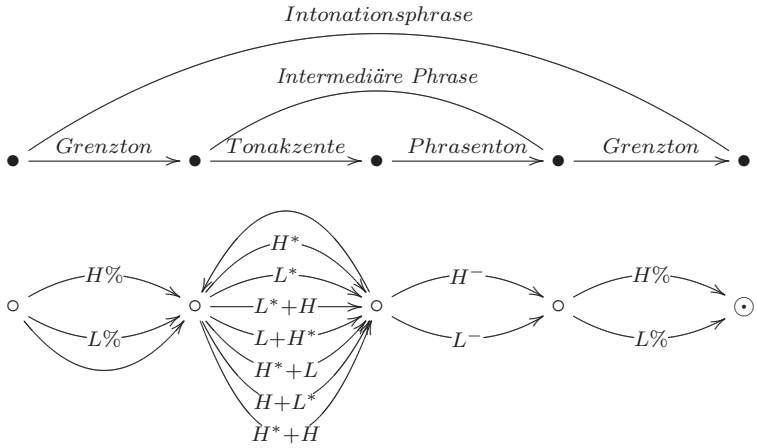

Abbildung 3.4: Finite-State-Grammatik für Intonationsphrasen von Beckman und Pierrehumbert (1986b) für das Amerikanische Englisch; nach (Ladd, 1996, S. 81).

von Intonationsphrasen auftreten, werden mit '%' markiert, Phrasentöne am Ende von intermediären Phrasen mit '−'.

Pierrehumberts Tonsequenzmodell entspricht einem endlichen Automaten, der bei jedem Übergang von einem Zustand in den nächsten einen Ton generiert. Das bedeutet, dass die Realisierung jedes Tons nur von dem zuletzt vorangehenden Ton abhängt und nicht von früheren oder folgenden.

In diesem streng lokalen Ansatz werden auch globale F0-Bewegungen als lokale Ereignisse beschrieben, so beispielsweise die in Folge der Deklination abnehmende Höhe der H-Töne als Sequenz von *Downsteps*.

Die Verbindung zwischen den Tönen erfolgt mittels (beispielsweise linearer) Interpolation. Eine besondere Art der Verknüpfung stellt das sogenannte *linking* von bitonalen Tönen dar (Gussenhoven, 1984). Hierbei wird der Folgeton des vorangehenden Tonakzents abgespalten und entweder mit dem folgenden Akzent verbunden (*partielles linking*) oder ganz entfernt (*vollständiges linking, Hutkontur*). Das folgende Beispiel aus Mayer (1997) soll dies illustrieren:

kein linking:
$[Detektivromane/L*+H]_{ip}$ $[sollen\ keine\ Literatur/L*+H\ sein]_{ip}]_{IP}$
partielles linking:
$[Detektivromane/L*\ sollen\ keine\ Literatur/+HL*+H\ sein]_{IP}$
komplettes linking:
$[Detektivromane/L*\ sollen\ keine\ Literatur/L*+H\ sein]_{IP}$

Gussenhoven spricht hier von zwei phonologischen Ebenen: der abstrakten Ebene der *Tonsegmente* und der sich aus diesen durch Operationen wie das Linking ergebenden Ebene der *phonologischen Oberflächenformen*. Linking kann wie im obigen Beispiel zur Aufhebung von Grenzen zwischen intermediären Phrasen führen.

3.3 INTSINT-Modell

Charakteristika: *tonbasiert, symbolisch, mathematisch, einschichtig.*

Das INTSINT-Modell (INternational Transcription System for INTonation, Hirst und Di Cristo, 1998) ist wie auch der im vorangegangenen Abschnitt beschriebene Tonsequenzansatz symbolisch, tonbasiert und einschichtig, im Gegensatz dazu aber weniger theoriegeleitet, da ein formuliertes Ziel bei der Entwicklung dieses Modells seine Sprachunabhängigkeit war: Eine Beschreibung beliebiger Intonationssysteme soll ohne Anpassungen wie Veränderungen im Label-Inventar möglich sein.

Die prosodische Analyse beginnt mit einer Segmentierung der F0-Kontur in intonatorische Einheiten (*intonation units*, wie beispielsweise Intonationsphrasen). Innerhalb dieser Einheiten wird die F0-Kontur als Abfolge von Zielpunkten verstanden. Die Tonhöhe jedes Zielpunkts kann erstens in Abhängigkeit des zuletzt vorangehenden Zielpunkts beschrieben werden (*higher, lower, same*), wobei auch hier detailliertere Abstufungen mittels *Upstep* und *Downstep* möglich sind. Zweitens kann der Zielton bei sehr starker F0-Auslenkung auch global bezogen auf die F0-Spannweite des Sprechers beschrieben werden als *top* oder *bottom*. Grenztöne aus dem TSM werden hier allgemeiner als *initiale* und *finale Töne* bezeichnet.

3.4 Kieler Intonationsmodell

Charakteristika: *konturbasiert, symbolisch, perzeptiv, einschichtig.*

Im Kieler Intonationsmodell (KIM), das von Kohler (1991) für das Deutsche entwickelt wurde, wird Intonation als Abfolge von Gipfel- und Talkonturen verstanden. Silbenkerne sind hierbei als Bündel distinktiver Merkmale repräsentiert, die Mikroprosodie, Wortbetonung, prosodische Struktureigenschaften, Konturtyp, F0-Alinierung und Sprechgeschwindigkeit codieren. Mittels handgefertigter kontextsensitiver Ersetzungsregeln des aus der generativen Phonologie von Chomsky und Halle (1968) übernommenen Typs $A \longrightarrow B|X\_Y$ werden diese Merkmalsbündel sukzessive in F0-Werte überführt. Wie im TSM wird Deklination lokal in Form von Downsteps modelliert.

Ein besonderes Gewicht kommt in diesem Modell der zeitlichen Alinierung von Kernen akzentuierter Silben und F0-Gipfeln zu (siehe hierzu Kapitel 5).

Zum Kieler Intonationsmodell wurde ein entsprechendes Etikettierungssystem namens PROLAB entwickelt (Kohler, 1995a).

3.5 Maximumbasierte Beschreibung nach Heuft und Portele

Charakteristika: *konturbasiert, parametrisch, mathematisch, einschichtig.*

Im parametrischen Modell von Heuft et al. (1995) wird die F0-Kontur als Abfolge von F0-Maxima verstanden, wobei jedes Maximum, wie auch in Abbildung 3.5 zu sehen, durch die folgenden Parameter charakterisiert ist:

- *Delay:* zeitlicher Abstand des F0-Maximums zum Beginn des Nukleus der akzentuierten Silbe,
- *Amplitude* des Maximums relativ zum Abstand zwischen Base- und Topline,
- *Steilheit* des Anstiegs und des Falls vor und nach dem Maximum.

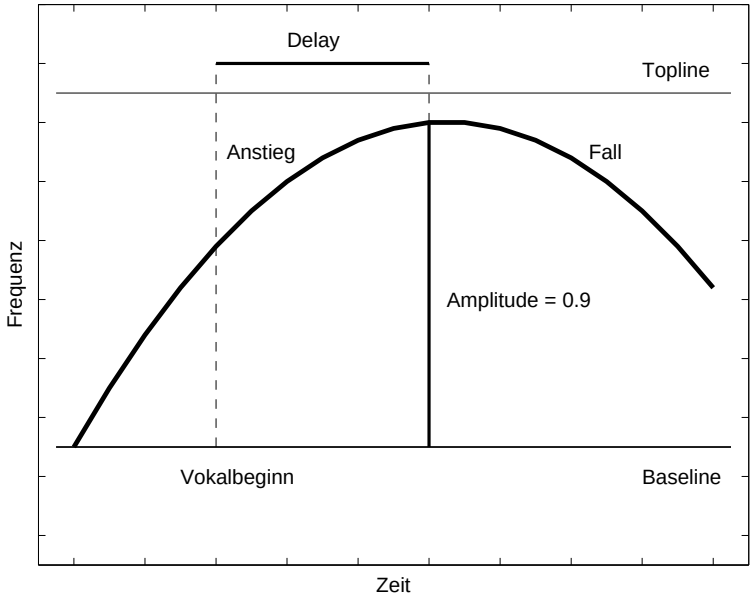

Abbildung 3.5: Maximumbasierte F0-Parametrisierung nach Portele & Heuft.

3.6 Tilt-Modell

Charakteristika: *konturbasiert, parametrisch, mathematisch, einschichtig.*

Die F0-Parametrisierung an Akzenten und Phrasengrenzen durch das Tilt-Modell (Taylor, 1995) ist in Abbildung 3.6 veranschaulicht. Dieses Modell ist eine Weiterentwicklung des RFC-Modells (*rise/fall/connection*; Taylor, 1995), in dem F0-Konturen mit folgenden vier Parametern beschrieben werden: Amplitude und Dauer des F0-Anstiegs (A_r, D_r) sowie des F0-Abfalls (A_f, D_f). Hierfür sind drei Ereigniszeitpunkte zu definieren: Ereignisstart, F0-Gipfel und Ereignisende. Im Tilt-Modell werden die vier RFC-Parameter zu den drei Parametern Amplitude A, Dauer D und Tilt zusammengefasst, wobei unter dem Tilt die Form der F0-Kontur zu verstehen ist.

$$\text{tilt} \quad = \quad \frac{|A_r| - |A_f|}{2 \cdot (|A_r| + |A_f|)} + \frac{D_r + D_f}{2 \cdot (D_r + D_f)} \tag{3.1}$$

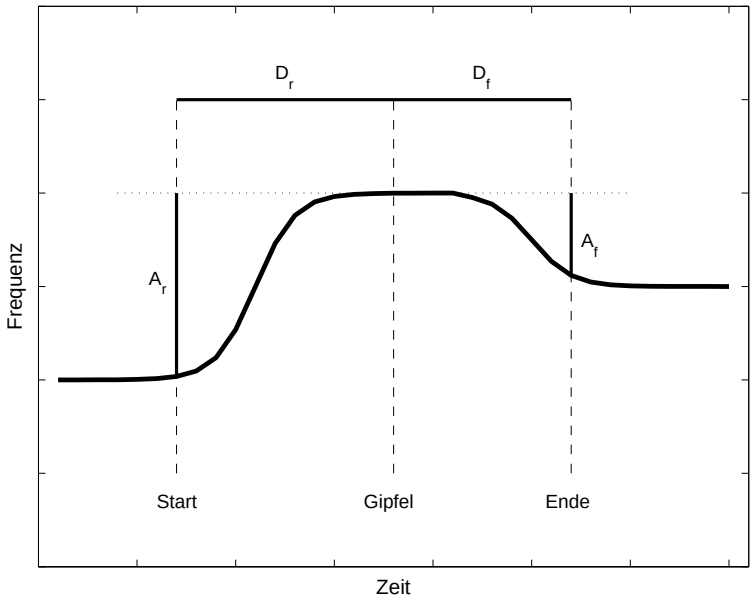

Abbildung 3.6: F0-Parametrisierung im RFC/Tilt-Modell; nach Dusterhoff und Black (1997).

$$\begin{aligned} A &= |A_r| + |A_f| \\ D &= D_r + D_f \end{aligned}$$

D_r und D_f bestimmen das Alignment der F0-Kontur auf der betroffenen Silbe. Zur F0-Stilisierung lassen sich die hierzu nötigen vier RFC-Parameter folgendermaßen rekonstruieren:

$$\begin{aligned} A_r &= \frac{A \cdot (1 + \text{tilt})}{2} \\ A_f &= \frac{A \cdot (1 - \text{tilt})}{2} \\ D_r &= \frac{D \cdot (1 + \text{tilt})}{2} \\ D_f &= \frac{D \cdot (1 - \text{tilt})}{2} \end{aligned} \tag{3.2}$$

3.7 Rapp-Modell

Charakteristika: *konturbasiert, parametrisch, mathematisch, einschichtig.*

Im parametrischen Modell von Rapp (1998b) wird der F0-Verlauf auf akzentuierten und nachfolgenden Silben, wie Abbildung 3.7 zeigt, als Addition von Tangens-hyperbolicus- und Gaußfunktion wie folgt stilisiert:

$$y(t) \quad = \quad \alpha \cdot \tanh(\beta \cdot (t - \gamma)) + \delta \cdot e^{-(\epsilon \cdot (t-\zeta))^2} + \eta \tag{3.3}$$

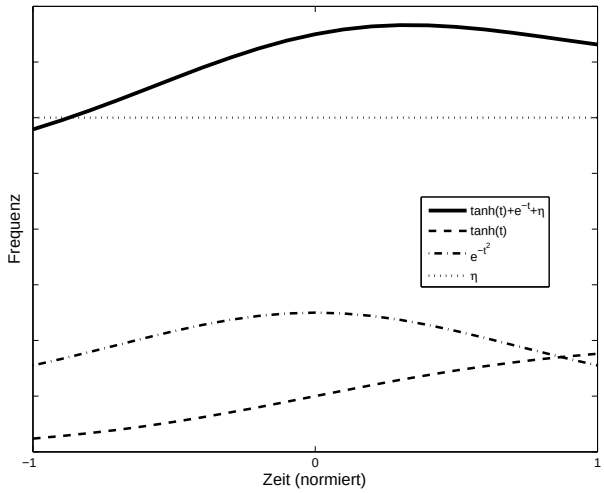

Abbildung 3.7: Rapp-Modell: Stilisierung mittels Tangens hyperbolicus und Gaußfunktion.

Mit der Tangens-hyperbolicus-Funktion lassen sich F0-Anstieg und -Abfall stilisieren. Die Gaußfunktion dient der Erfassung von kombinierten Auf- und Abbewegungen. Die Parameter tragen Folgendes zur Stilisierung bei:

- α: Tonhöhendifferenz zwischen akzentuierter und postakzentuierter Silbe,
- β: Steilheit des F0-Anstiegs oder Abfalls,
- γ: Zeitpunkt des Anstiegs oder Abfalls,
- δ: Höhe des Gipfels,
- ϵ: Steilheit des Gipfels,
- ζ: Startzeitpunkt des Gipfels,
- η: F0-Baseline.

3.8 PaintE-Modell

Charakteristika: *konturbasiert, parametrisch, mathematisch, einschichtig.*

Das parametrische PaintE-Modell (*PArametric INTonation Event*) von Möhler (1998b) beschreibt den F0-Verlauf in auf intonatorisch relevante Silben zentrierten 3-Silben-Fenstern in Form zweier überlagerter Sigmoidfunktionen, wie in Abbildung 3.8 zu sehen.

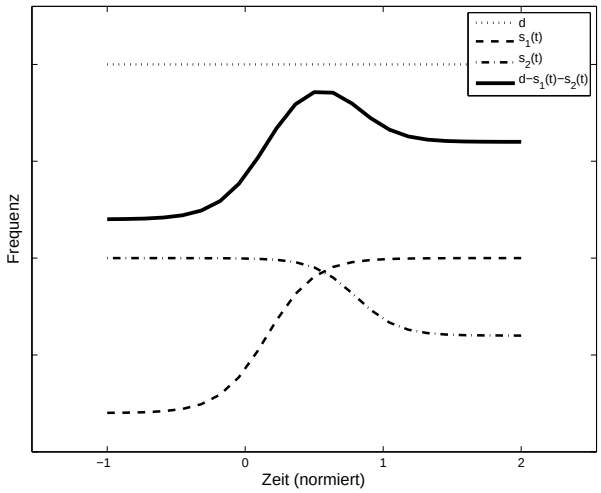

Abbildung 3.8: PaintE: Stilisierung mittels zweier Sigmoiden s_1 und s_2.

Die Stilisierungsfunktion ist folgendermaßen gegeben:

$$y(t) = d - \frac{c_1}{1 + e^{-a_1(b-t)+\gamma}} - \frac{c_2}{1 + e^{-a_2(t-b)+\gamma}}, \tag{3.4}$$

wo a_1, a_2 die Steigungen bezeichnen, b die Alinierung der Funktion auf dem Silbentriplett, c_1, c_2 die Amplituden der Sigmoiden, d den Gipfel und γ einen Alinierungsparameter. Die Zeit t ist auf das Intervall $[-1\ 2]$ normiert, wobei die Zielsilbe den Bereich zwischen 0 und 1 umspannt.

Konturklassen In einer erweiterten Fassung des Modells (Möhler und Conkie, 1998) werden die Parametervektoren mittels Vektorquantisierung geclustert. Daraus resultiert eine Überführung der parametrischen Intonationsbeschreibung in eine abstraktere symbolische Beschreibung.

3.9 IPO-Modell

Charakteristika: *konturbasiert, parametrisch, perzeptiv, einschichtig.*

Das Eindhovener IPO-Modell (t'Hart et al., 1990), ursprünglich für das Niederländische entwickelt und mittlerweile auf diverse Sprachen übertragen, so auch auf das Deutsche (Adriaens, 1991), ist konturbasiert, parametrisch und einschichtig. Die Intonationseinheiten werden anhand der perzeptiven Urteile von Versuchspersonen gewonnen.

Stilisierung Ausgangspunkt der Intonationsmodellierung ist die *Close-copy*-Stilisierung von F0-Verläufen, die zwei Bedingungen zu erfüllen hat:

- Sie soll aus der geringstmöglichen Anzahl aneinandergereihter Geradenstücke bestehen.

- Die Bedingung der *perzeptuellen Gleichheit* muss erfüllt sein, Originalkontur und Stilisierung dürfen sich perzeptiv nicht voneinander unterscheiden.

Aus den Close-copy-Stilisierungen lassen sich nun allgemein verwendbare *standardisierte Stilisierungen* gewinnen, indem einander *perzeptiv äquivalente* Stilisierungen zusammengefasst werden. Perzeptive Äquivalenz zwischen zwei F0-Konturen ist dann gegeben, wenn der eine Verlauf perzeptiv als erfolgreiche Wiederholung des anderen bewertet wird.

Diese standardisierten F0-Bewegungen werden nun im nächsten Schritt modelliert als Transitionen zwischen parallel verlaufenden Deklinationslinien. Das sich damit ergebende Modell besteht, wie in Abbildung 3.9 zu sehen, aus folgenden Komponenten:

- **Standardisierte Tonhöhenbewegungen**, die charakterisiert sind durch die vier Parameter Richtung (auf- oder abwärts), Frequenzumfang (in Halbtönen), Geschwindigkeit (in Halbtönen pro Sekunde) und Alinierung zum Silbenanfang (in Millisekunden).

- Parallel verlaufende **Deklinationslinien**, die charakterisiert sind durch Deklinationsgeschwindigkeit in Halbtönen pro Sekunde und die als Start- und Zielpunkte der Tonhöhenbewegungen fungieren.

Die Inventargröße ist sprachabhängig. So wurden für das Englische (Brown et al., 1980) und Deutsche (Adriaens, 1991) jeweils drei Deklinationslinien angesetzt, für das Niederländische zwei (t'Hart et al., 1990). Die Anzahl der standardisierten Tonhöhenbewegungen variiert von zehn (Niederländisch) über elf (Deutsch) bis zu 27 (Englisch).

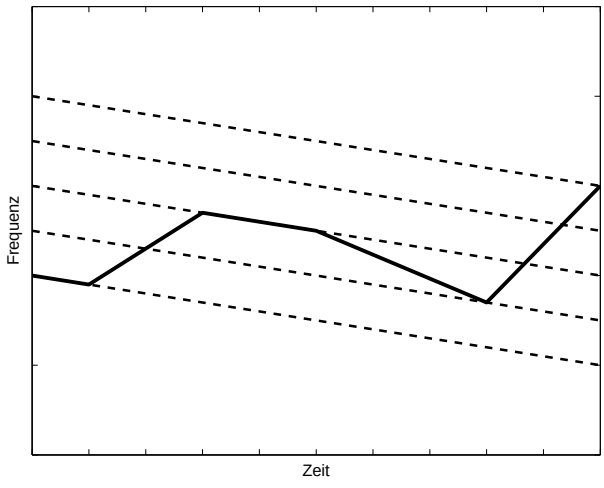

Abbildung 3.9: F0-Generierung mittels der standardisierten Intonationsbausteine.

Manuell erstellte Regeln zur Intonotaktik bestimmen erlaubte Kombinationsmöglichkeiten der gewonnenen Intonationsbausteine. Hierbei werden die Bausteine auch hierarchisch organisiert (Noteboom, 1997). Aufeinanderfolgende eng zusammengehörige standardisierte Tonhöhenbewegungen werden zu *Konfigurationen* zusammengefasst. Die Einordnung dieser Konfigurationen in Präfix-, Wurzel- und Suffix-Konfigurationen regelt die erlaubten Konfigurationsabfolgen innerhalb der nächstgrößeren Einheiten, der *Pitch-Konturen*, die nicht weiter spezifiziert eine inhaltlich zusammengehörige Wortfolge umspannen (Noteboom, 1997).

Anhand von Perzeptionsexperimenten, in denen Versuchspersonen die Aufgabe hatten, Pitch-Konturen nach eigenem Belieben in Klassen einzuteilen (Collier und t'Hart, 1972; Gussenhoven, 1983), konnten Pitch-Konturen jeweils zu Oberflächenformen zugrundeliegender *basic intonation patterns* gebündelt werden.

3.10 Bierwisch-Modell

Charakteristika: *symbolisch, perzeptiv, einschichtig.*

Bierwisch (1966) leitet regelbasiert in ihrem Modell für die Intonation des Deutschen, das in der Tradition der Generativen Grammatik steht, die prosodische Struktur einer Äußerung von ihrer syntaktischen Struktur ab und generiert daraus anhand von Ersetzungsregeln eine silbenbasierte phonetische Transkription zur Angabe relativer Tonhöhen sowie -bewegungen.

3.11 Öhman-Modell

Charakteristika: *konturbasiert, parametrisch, mathematisch, superpositional.*

Das erste superpositionelle Modell der Intonation wurde von Öhman und Lindqvist (1965) sowie Öhman (1967) für das Schwedische und für dänische Dialekte vorgestellt. Stärker noch als beim später entwickelten Fujisaki-Modell steht hier die Modellierung der Produktion der Intonation im Vordergrund. Die F0-Kontur $f_0(t)$ wird hier wie auch in Abbildung 3.10 zu sehen durch ein Larynx-Modell synthetisiert, das von den folgenden drei Komponenten angesteuert wird:

- der Stimmlippenspannung als Summe der Ausgaben $g_s(t)$ und $g_w(t)$ zweier Filter für Satz- und Wortintonation; letztere dient der Intonationsmodellierung auf Wortebene in den hier behandelten Tonakzentsprachen,
- einem akustischen Interaktionssignal für Fluktuationen des sub- und supraglottalen Luftdrucks und
- einem artikulatorischen Interaktionssignal basierend auf nicht-phonatorischen Bewegungen des hyo-thyroidschen Hebelsystems.

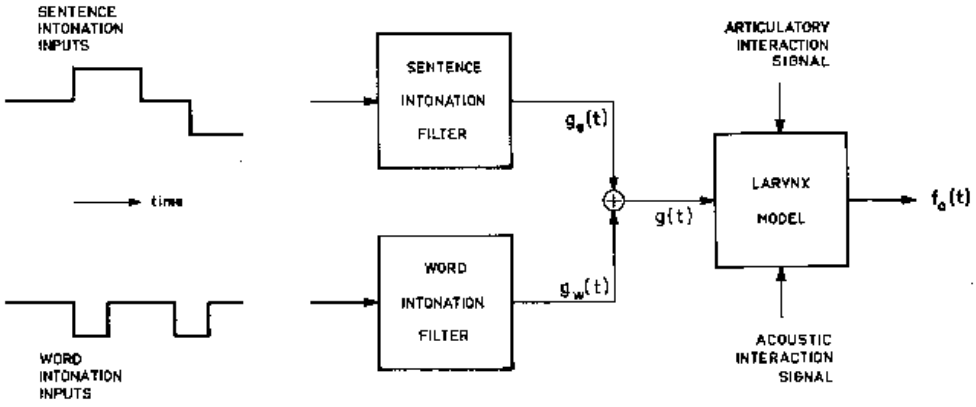

Abbildung 3.10: Das Öhman-Modell; aus (Öhman, 1967, S. 21).

Das Design der Filter für Satz- und Wortintonation richtet sich nach den dynamischen Charakteristika der mechanischen und peripher-neuronalen Komponenten des laryngalen Systems. Das Input-Signal der Filter besteht aus einem endlichen Inventar von Stufenfunktionen mit variierten Amplituden und Zeit-Onsets. Diese Stufenfunktionen repräsentieren die zentralnervösen Kommandos zur Codierung intonatorischer Ereignisse.

3.12 Fujisaki-Modell

Charakteristika: *konturbasiert, parametrisch, mathematisch, superpositional.*

Das parametrische Fujisaki-Modell (Fujisaki, 1987) kann als das weitestverbreitete und einflussreichste superpositionale Modell bezeichnet werden. Es wurde für diverse Sprachen adaptiert, so auch für das Deutsche (Möbius, 1993a; Mixdorff, 1998).

Logarithmierte F0-Konturen werden als additive Superposition von einem sprecherabhängigen F0-Grundwert, einer Phrasenkomponente und einer Akzentkomponente beschrieben. Graphik 3.11 zeigt die prinzipielle Wirkungsweise: Phrasenkommandos A_p (Impulse) und Akzentkommandos A_a (Rechteckfunktionen) regen die entsprechenden Systeme zur Phrasen- beziehungsweise Akzentsteuerung an. Diese Systeme sind kritisch gedämpft, d. h. ihre Schwingungsamplitude sinkt im zeitlichen Verlauf, und sie schwingen bei Rückkehr in die Ruhelage nicht darüber hinaus und beschränken somit den F0-Wertebereich nach unten.

Die Phrasen- und Akzentkommandos lassen sich zur Festlegung der prosodischen Struktur, also der Lokalisierung von Phrasengrenzen und Akzenten heranziehen.

Die Phrasenkomponente generiert die globale Intonationskontur innerhalb von Intonationsphrasen. Ein positives A_p markiert den *Pitch Reset* zu Beginn einer Intonationsphrase, oder progressiven F0-Verlauf sowie die Intonation für Entscheidungsfragen zum Ende einer Phrase. Ein negatives A_p signalisiert Finalität (*final lowering*).

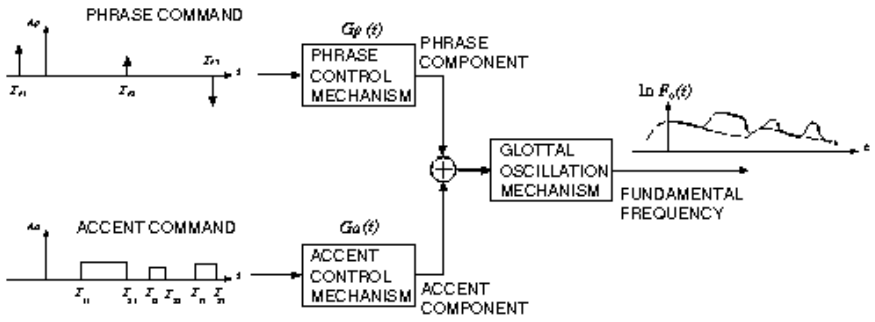

Abbildung 3.11: Komponenten des Fujisaki-Modells; aus: Fujisaki und Hirose (1984).

Mit der Akzentkomponente werden in der Regel lokale F0-Bewegungen auf akzentuierten Silben modelliert. Die im Zuge der Deklination fallende *Topline* lässt sich mit fortschreitend abnehmenden A_a-Amplituden realisieren.

Stilisierung Die Stilisierungsfunktionen sind wie folgt gegeben:

$$lnF_0(t) = lnF_{\min} + \sum_i A_{pi} C_p(t - T_{pi}) + \sum_j A_{aj} \big[C_a(t - T_{1j}) - C_a(t - T_{2j}) \big] \tag{3.5}$$

$$C_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & : \quad t \geq 0 \\ 0 & : \quad t < 0 \end{cases} \quad C_a(t) = \begin{cases} 1 - (1 + \beta t) e^{-\beta t} & : \quad t \geq 0 \\ 0 & : \quad t < 0 \end{cases}$$

und sind durch die folgenden Parameter beschrieben:

- T_p: Zeitpunkt des Phrasenkommandos,
- T_1, T_2: Start- und Endzeitpunkt des Akzentkommandos,
- A_p, A_a: Amplituden der Kommandos,
- α, β: Dämpfungsfaktoren des Phrasen- und Akzentsystems, die die Dauer der F0-Bewegungen mitbestimmen.

3.13 Bell-Labs-Modell

Charakteristika: *parametrisch, mathematisch, superpositional.*

Das Bell-Labs-Modell (van Santen et al., 1998) ist wie das Fujisaki-Modell parametrisch und superpositional, aber nicht rein konturbasiert.

Phrasenkurven werden als zweiteilige Kurven mittels nicht-linearer Interpolation zwischen drei F0-Zielpunkten generiert. Die Zielpunkte sind: Phrasenbeginn, Beginn der

letzten Akzentgruppe der Phrase und Phrasenende. Eine Akzentgruppe ist hier wie die Einheit des prosodischen Fußes in Nespor und Vogel (1986) definiert als Sequenz einer betonten Silbe und aller darauf folgenden unbetonten Silben bis hin zur nächsten Betonung beziehungsweise zum Äußerungsende.

Akzentkurven werden stets solchen Akzentgruppen zugeordnet und anstelle der Fujisaki-Akzentkommandos durch ein lineares Alinierungsmodell erzeugt. Hierbei wird im Gegensatz zum rein konturbasierten Ansatz des Fujisaki-Modells auch Gewicht auf F0-Targets gelegt.

Zusätzlich zur Phrasen- und Akzentkomponente geht eine Mikrointonationskomponente in die Superposition mit ein, die aus segmentellen F0-Perturbationen besteht.

Akzentalinierung Untersuchungen für das Amerikanische Englisch (van Santen und Hirschberg, 1994) haben ergeben, dass die zeitliche Alinierung des F0-Gipfels auf den Silbennukleus von der Länge der Akzentgruppe abhängt. Hierzu wurde die Akzentgruppe unterteilt in a) Onset und b) Reim der akzentuierten Silbe sowie c) dem Rest der Akzentgruppe. Eine Verlängerung jedes dieser Teile führte in unterschiedlichen Ausmaßen zu einer Rückverlagerung des F0-Gipfels. Diese Befunde wurden folgendermaßen modelliert:

$$T_p(a) \quad = \quad \sum_j \alpha_{sj} \cdot D_j(a) + \mu_s. \tag{3.6}$$

T_p steht für den Zeitpunkt des F0-Gipfels in Akzentgruppe a, $D_j(a)$ ist die Dauer des j-ten Teils der Akzentgruppe. Jede dieser Dauern geht mit einem spezifischen Gewicht in die Berechnung ein, das zudem abhängig ist vom Strukturtyp s der Akzentgruppe. Folgende vier Typen werden hierbei unterschieden: polysyllabische vs. monosyllabische Akzentgruppe mit sonoranter vs. stimmloser vs. mit stimmhaftem Obstruenten belegter Coda. μ_s schließlich ist der strukturtypabhängige Zeitmittelwert.

Die zeitliche Alinierung von F0-Gipfeln lässt sich auf eine Alinierung beliebiger Zielpunkte i verallgemeinern:

$$T_i(a) \quad = \quad \sum_j \alpha_{isj} \cdot D_j(a) + \mu_{is}. \tag{3.7}$$

Hierdurch wird dem hybriden Charakter dieses Modells hinsichtlich Kontur- vs. Tonbasiertheit Rechnung getragen.

3.14 Grønnum-Modell

Charakteristika: *konturbasiert, parametrisch, superpositional.*

In dem superpositionalen konturbasierten Modell von Grønnum (1995) für die dänische Intonation wird der F0-Verlauf, wie in Abbildung 3.12 zu sehen, als Überlagerung von drei Konturtypen beschrieben:

- Die globale Textkontur (*textual contour*), ein linearer F0-Abstieg von etwa einer halben Oktave.
- Die Textkontur wird überlagert von Äußerungskonturen (*utterance contours*), die um etwa 3 Halbtönen linear abfallen. Lange Äußerungskonturen werden in Sequenzen linear abfallender Phrasenkonturen (*phrase contours*) zerlegt.
- Aufgesetzt auf diese Äußerungskontur sind Akzentgruppen (*stress group patterns*), die wie prosodische Füße in Nespor und Vogel (1986) definiert sind, also aus einer betonten und aller darauffolgenden unbetonten Silben bestehen. Ihre Standardform besteht in einem kurzen Abfall gefolgt von einer Steigung über den Nukleus der betonten Silbe sowie einem Abfall bis zum Ende der Einheit.

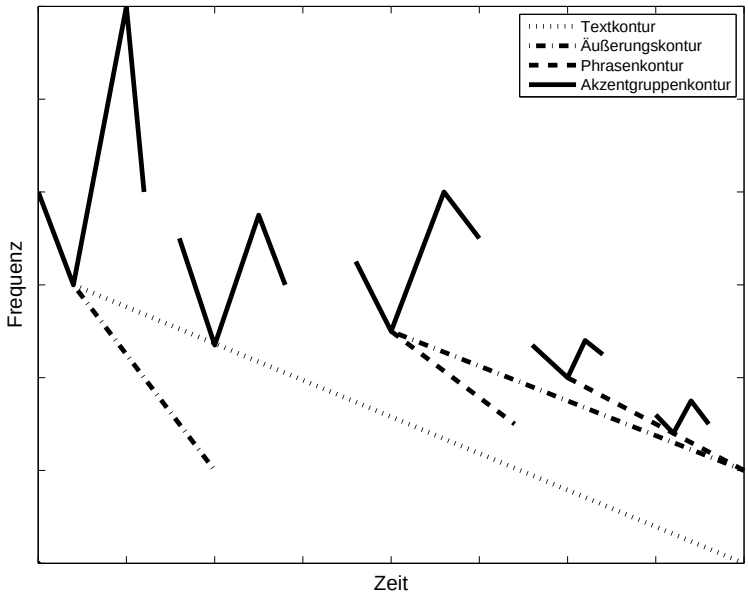

Abbildung 3.12: Grønnum-Modell: Überlagerung von Text-, Äußerungs-, Phrasen- und Akzentgruppenkonturen; nach (Grønnum, 1995, S. 128).

Die längeren Struktureinheiten beeinflussen jeweils die Ausprägung der kürzeren überlagernden Komponenten: so verringern sich beispielsweise die Onset-Amplituden der Äußerungskonturen im Laufe einer Textkontur ebenso wie die Amplituden der Akzentgruppen innerhalb einer Text- und Äußerungskontur. Grønnum macht allerdings keine Angaben zur konkreten mathematischen Zerlegung der F0-Konturen in die angenommenen Komponenten.

3.15 Einsatzmöglichkeiten der Modelle

Intonationsmodelle dienen der Vermittlung zwischen linguistischer Information, physiologischen Parametern und Wahrnehmungsapparat (im Folgenden: Hintergrund) auf der einen Seite und konkreter F0-Kontur als Oberflächenerscheinung auf der anderen Seite durch Bereitstellung einer geeigneten Intonationsrepräsentation. Abbildung 3.13 fasst zusammen, inwieweit die besprochenen Modelle in dieser Vermittlerrolle eingesetzt werden.

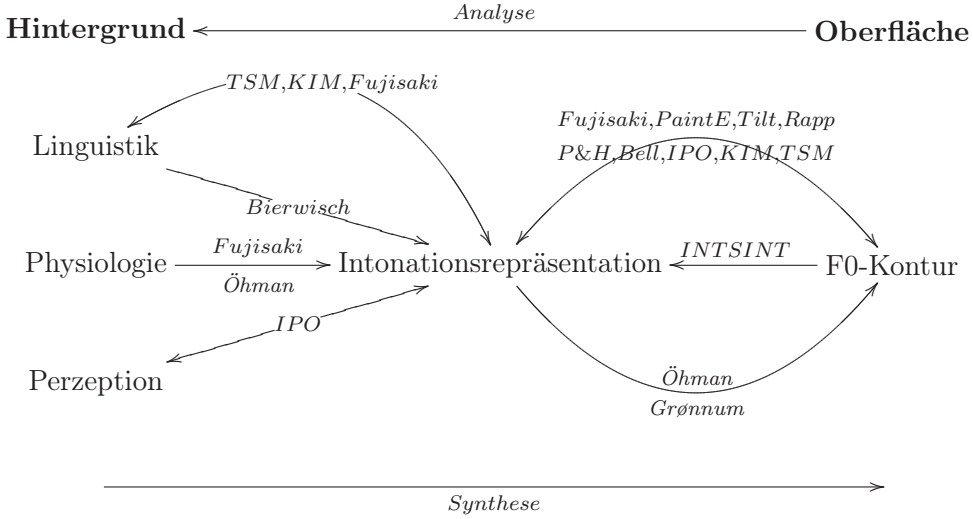

Abbildung 3.13: Einsatzbereiche der Intonationsmodelle.

Die behandelten Modelle werden zum momentanen Forschungsstand in unterschiedlichem Ausmaß zur linguistischen, physiologischen oder perzeptiven Verankerung von F0-Konturen genutzt. Auch sind sie nicht gleichermaßen zu Analyse- oder Synthesezwecken in Verwendung. Das Fujisaki-Modell beispielsweise erlaubt sowohl eine Vermittlung zwischen Signal und abstrakter Repräsentation als auch zwischen Repräsentation und linguistischem und physiologischem Hintergrund. Die Verbindungen sind reziprok, was dieses Modell sowohl zur F0-Analyse als auch zur -Synthese qualifiziert.

Dagegen wird die Transformationskette zum jetzigen Entwicklungsstand beispielsweise vom Bierwisch-, INTSINT- und Rapp-Modell nicht vollständig abgedeckt. Während sich das Bierwisch-Modell auf die syntaxgeleitete Generierung der Intonationsrepräsentation beschränkt, vermitteln Rapp und INTSINT zwischen F0-Kontur und abstrakter Repräsentation. Bierwisch- und INTSINT-Modell sind hierbei unidirektional dahingehend, dass ersteres allein für Synthese und letzteres allein für Analyse verwendet wird. Das Rapp-Modell dagegen eignet sich bidirektional sowohl zur F0-Analyse als auch zur F0-Synthese.

Einige Modelle decken nur zusammen mit entsprechenden Zusatzmodulen die Transformationskette vollständig ab. So sind symbolische Ansätze wie das TSM nur in Zusammenhang mit entsprechenden Zusatzmodulen zur F0-Synthese aus der Intonationsrepräsentation nutzbar.

Kapitel 4

Gewinnung der Intonationsrepräsentation

Nach einer kurzen Vorstellung experimentalphonetischer Ansätze zur Gewinnung abstrakter intonatorischer Einheiten soll in den folgenden Abschnitten auf die modellabhängige Extrahierung solcher Einheiten aus den F0-Konturen eingegangen werden.

Symbolisch beschriebene Intonationseinheiten (vgl. Abschnitt 3.1) werden in der Regel theoriegeleitet durch manuelle Etikettierung gewonnen. Darauf aufbauend können Klassifikatoren zur automatischen Etikettierung trainiert werden. Eine parametrische Beschreibung von Intonationseinheiten lässt sich häufig mittels Analyse durch Synthese erzielen.

4.1 Experimentalphonetische Ermittlung

Die Rückführung der konkreten F0-Realisierung, also der phonetischen Oberflächenform, in diskrete Einheiten wurde experimentalphonetisch sowohl in Perzeptions- als auch in Produktionsstudien unternommen. Einen Überblick hierüber gibt Gussenhoven (2006).

Perzeption Im IPO-Modell wird die F0-Stilisierung sowie die Gewinnung abstrakter Intonationseinheiten mittels Experimenten zur Feststellung perzeptiver Gleichheit und perzeptiver Äquivalenz betrieben.

Auch können Befunde zur kategorialen Wahrnehmung (Kohler, 1987, 1991) zu diesem Zweck herangezogen werden, indem anhand der dort ermittelten Kategoriegrenzen eine Partitionierung der phonetischen Oberflächenformen als Realisierungen verschiedener abstrakter Einheiten vorgenommen wird.

Produktion Auf Produktionsseite lassen sich Imitationsexperimente anführen, die ergaben, dass sich bei Reproduktion variierter F0-Verläufe durch Versuchspersonen Kategorien herausbilden (Pierrehumbert und Steele, 1989).

Weiter wurde festgestellt, dass intonatorische Selbstimitation eines Sprechers ausgehend von verschiedenen Startpunkten zu wenigen stabilen Mustern konvergiert (Braun et al., 2006). Hierbei wurde der Versuchsperson zunächst eine Äußerung eines anderen Sprechers zur intonatorischen Imitation präsentiert und im Anschluss daran in einer repetitiven Imitationsaufgabe die jeweils von ihr zuletzt getätigte Imitation. Die Autoren bezeichnen die sich daraus ergebenden diskreten Intonationsmuster als *Attraktoren*, worunter Werte verstanden werden, zu denen bestimmte Funktionen wie beispielsweise die Quadrierung reeller Zahlen mit Betrag kleiner 1 oder eben die hier untersuchte Imitation bei rekursiver Anwendung konvergieren.

4.2 Manuelle Etikettierung

Bei der manuellen Etikettierung vergeben Experten prosodische Labels. Üblicherweise beziehen sich Akzentlabels auf Silben oder Wörter und Grenzlabels auf Wortgrenzen. Das Labelinventar ist abhängig vom zugrundeliegenden Intonationsmodell. Insofern lässt sich dieses Vorgehen als Befragung von ausgebildeten Versuchspersonen verstehen, bei der die Antwortalternativen durch eine bestimmte Theorie vorgegeben sind.

4.2.1 Label-Inventare

Prosodische Label-Inventare können im einfachsten Fall durch Markierung von Akzenten und Phrasengrenzen zur Beschreibung der prosodischen Struktur einer Äußerung dienen. Ein im Rahmen des VERBMOBIL-Projekts entwickeltes Inventar (Reyelt und Batliner, 1994) erlaubt zusätzliche Angaben zu Stärke und Typ der Phrasengrenze, zur Prominenz des Akzents und zum phrasenfinalen F0-Verlauf: Phrasengrenzen können als schwach, stark oder irregulär – beispielsweise in Verbindung mit Häsitationen – klassifiziert werden, und Akzente als Neben-, Haupt- oder emphatischer Akzent. Phrasenfinale F0-Verläufe werden hier unterteilt in final (fallend), progredient (gleichbleibend) und fragend (steigend).

Genauere Angaben zu den tonalen Eigenschaften der Akzente erlaubt das ToBI-System (*tonal and break indices*), das auf dem Pierrehumbertschen Tonsequenzansatz basiert und ursprünglich für die Intonation des Englischen entwickelt (Silverman et al., 1992) und mittlerweile auf viele Sprachen adaptiert wurde, so auch auf die Intonation des Deutschen (Reyelt et al., 1996). Das Labelinventar besteht hier aus einzelsprachlich angepassten Tonakzent-, Phrasenton- und Grenzton-Etiketten sowie Grenzlabels, die die unterschiedlich starke Markiertheit der Grenzen zwischen klitischen Verbindungen bis hin zu Intonationsphrasen codieren. Eine Motivation der deutschen Fassung GToBI (German ToBI) war auch die Vereinheitlichung diverser bestehender Labelinventare wie das soeben beschriebene VERBMOBIL-Inventar, das zum Kieler Intonationsmodell gehörende PROLAB sowie einige bereits an ToBI angelehnte Inventare (Grice und Benzmüller, 1995; Mayer, 1995).

In Anbetracht der Tatsache, dass die prosodischen Urteile unterschiedlicher Labeller nicht immer übereinstimmen müssen, werden mitunter “harte” Labels durch relative Häu-

figkeiten ersetzt, mit der sie vergeben werden. Ein Beispiel hierfür sind die *Prominence-Scores* in Mo et al. (2009), in deren Studie für jedes Wort einer Äußerung nach Etikettierung durch mehrere Labeller die relative Häufigkeit angegeben wird, mit der es als akzentuiert etikettiert wurde. Bei einem derartigen Vorgehen werden häufig Experten durch naive Versuchspersonen ersetzt.

4.2.2 Evaluierung

Grundsätzlich ist die Frage nach der phonetischen Berechtigung eines Etikettiersystems damit verbunden, inwieweit

1. unterschiedliche Labeller dasselbe Segment gleich etikettieren und

2. ein Labeller bei wiederholter Bearbeitung eines Segments dasselbe Etikett vergibt.

Man spricht hierbei von *Inter-* und *Intra-Labeller-Konsistenz*. Ist diese Konsistenz nicht gegeben, so stellt sich die Frage nach der perzeptiven Adäquatheit der gewählten prosodischen Abstrahierung. Die Inter-Labeller-Konsistenz der ToBI-Etikettierung wurde beispielsweise für das Englische von Pitrelli et al. (1994) und für das Deutsche von Grice et al. (1996) untersucht. In beiden Fällen wurde die Inter-Labeller-Konsistenz definiert als Prozentsatz der Übereinstimmungen jeweils zweier Etikettierer auf Wortebene. Für Tonakzente wurde in Grice et al. (1996) in rund 71 % der Urteilspaare Übereinstimmung gefunden, in Pitrelli et al. (1994) lag die Übereinstimmung bei 68 %.

In Gut und Bayerl (2004) werden Intra- und Inter-Labeller-Konsistenz mittels Korrelationen in Gestalt von κ-Scores (Cohen, 1960) für Labeller-Paare angegeben. Für ToBI-Annotationen liegen auch hier leider nur die Resultate zur Inter-Labeller-Konsistenz vor: der mittlere κ-Wert über alle Labeller-Paare beträgt 0.33 und ist damit eher niedrig.

Wightman (2002) stellt in einem Überblick über mehrere Inter-Labeller-Konsistenz-Studien zu ToBI hohe Übereinstimmung in der Bestimmung der prosodischen Struktur fest, dagegen aber sehr viel niedrigere Übereinstimmung bei der Wahl konkreter Tonakzent-Labels. Er räsoniert in diesem Zusammenhang über mögliche Reduzierungen des Labelinventars, etwa den Verzicht auf den – nicht unumstrittenen – Downstep (Dainora, 2001), oder eine Reduzierung, wie sie im VERBMOBIL-Projekt vorgenommen wurde (Niemann et al., 1997).

4.3 F0-Vorverarbeitung bei automatischer Extrahierung

Um Features aus F0-Konturen zu deren Klassifizierung extrahieren zu können, müssen diese zunächst vorverarbeitet werden. Die Vorverarbeitung umfasst die Detektion und Korrektur von Messfehlern, die Glättung der Konturen, unter anderem zur Abschwächung mikrointonatorischer Effekte, sowie die Interpolation über stimmlose Signalabschnitte. Außerdem werden häufig die Hertz-Werte in perzeptiv adäquatere Halbton- oder ERB-Werte transformiert und eine Zeitnormalisierung durchgeführt.

4.3.1 Identifizierung und Korrektur von Messfehlern

Im Wesentlichen sind hier grobe F0-Diskontinuitäten zu detektieren, wie sie beispielsweise in Form von Oktavsprüngen auftreten. Zur automatischen Korrektur der Fehler eignen sich Listenkorrekturverfahren (Reddy, 1967; Specker, 1984), Interpolation oder Glättungsverfahren (s.u.).

4.3.2 Interpolation

Interpolation dient der Überbrückung von Messfehlern sowie F0-Lücken in stimmlosen Signalabschnitten (ohne Pausen), um eine stetige Funktion zu erhalten. Dies kann mit Hilfe von Polynomen geschehen, die an die zum zu überbrückenden Intervall zeitlich benachbarten F0-Werte angepasst werden. Diese stückweise Überbrückung wird auch als *Polygonzug* (bei Polynomen erster Ordnung) beziehungsweise *Spline-Interpolation* (bei höherer Ordnung) bezeichnet. Üblicherweise vermeidet man bei der Anpassung eine zu hohe Polynom-Ordnung, da diese mit einer zunehmenden Instabilität des Polynoms einhergeht. Instabilität meint, dass es zu hohen Ausschlägen des Polynoms zwischen den Interpolationspunkten kommen kann.

4.3.3 Glättung

Glättung dient der Abschwächung kurzzeitig auftretender Schwankungen in der F0-Kontur. Diese können durch Einflussgrößen wie Messfehler und mikroprosodische Effekte bedingt sein. Insofern dient die Glättung einer weiteren F0-Korrektur sowie der Abschwächung mikroprosodischer Effekte. Übliche Glättungsverfahren bestehen der Tiefpassfilterung von F0-Konturen beispielsweise mittels *Moving-Average-* oder *Savitzky-Golay-Filtern* (Savitzky und Golay, 1964).

Moving-Average Moving-Average-Filter ersetzen jeden F0-Wert $y[t]$ durch den Mittelwert in einem t-zentrierten Fenster der Länge $2n+1$. Je größer das gewählte Fenster, desto stärker ist die resultierende Glättung.

$$y[t] \quad = \quad \frac{\sum_{i=t-n}^{t+n} y[i]}{2n+1} \tag{4.1}$$

Medianglättung Die Anwendung der Moving-Average-Glättung zur Messfehlerkorrektur bringt den Nachteil mit sich, dass Oktavfehler über die Mittelwertbildung in nichtharmonische Grobfehler umgewandelt werden, die unter Umständen perzeptiv als noch störender empfunden werden.

Um diesen Effekt bei der Behandlung von Grobfehlern zu vermeiden, schlugen Rabiner et al. (1975) vor, jeden F0-Wert $y[t]$ nicht durch das arithmetische Mittel, sondern durch den Median-Wert in einem t-zentrierten Fenster der Länge $2n+1$ zu ersetzen:

$$y[t] \quad = \quad \text{median}(y[t-n \ldots t+n]) \tag{4.2}$$

Savitzky-Golay Bei der Savitzky-Golay-Filterung wie beispielsweise in Jan Van Santen et al. (2004) wird $y[t]$ ersetzt durch einen Wert, der sich durch eine Polynomanpassung maximal n-ter Ordnung im auf t zentrierten Fenster der Länge $2n+1$ ergibt. Je niedriger die Ordnung des Polynoms, desto stärker die Glättung.

$$y[t] \quad = \quad \text{polyfit}(y[t-n \ldots t+n])[n+1] \tag{4.3}$$

Allgemein erhält eine Savitzky-Golay-Filterung die relativen Maxima und Minima der Originalkontur eher als eines der oben beschriebenen Mittelwertfilter.

Zu Glättungsverfahren, die gezielt zur Beseitigung der Mikroprosodie entwickelt wurden, zählen MOMEL (Hirst und Espesser, 1993) und WAM (Reichel und Winkelmann, 2010).

MOMEL MOMEL (*MOdelisation MELodique*) findet unter anderem im Rahmen des INTSINT-Modells (siehe Abschnitt 3.3) Anwendung und dient dort neben der Glättung von F0-Konturen auch ihrer Stilisierung durch die Extrahierung von F0-Zielpunkten, die im darauffolgenden Schritt automatisch mit INTSINT-Labels versehen werden. Das MOMEL-Verfahren läuft in folgenden Schritten ab:

- In jedem Fenster einer gefensterten F0-Kontur wird iterativ eine Parabel p unter Minimierung des quadratischen Fehlers zwischen p und der Originalkontur angepasst.
- In jedem Iterationsschritt werden hierbei nach der Anpassung Original-F0-Werte ab einer gewissen Abweichung d von der Parabel entfernt.
- Sobald keine F0-Werte mit einer Abweichung größer gleich d mehr auftreten, terminiert die Iteration.
- Nach einer weiteren Reduzierung der Extremwerte der Parabeln auf Grundlage ihrer Abweichung von lokalen Mittelwerten dienen die verbleibenden Werte schließlich als Stützstellen für eine quadratische Spline-Funktion zur Glättung der F0-Kontur.

WAM Im WAM-Modell (*Weighting Against Micromelody*) wird die Beseitigung der Mikroprosodie als Gewichtung der F0-Kontur $w_c(y)$ verstanden, wobei die Gewichtungsfunktionen w_c folgendermaßen zu ermitteln sind:

- Zunächst werden vokalische Segmente hinsichtlich der folgenden mikroprosodisch relevanten Faktoren klassifiziert: Stimmhaftigkeit des vorangehenden und des folgenden Konsonanten (stimmhaft vs. stimmlos) sowie Zungenhöhe des Vokals (hoch vs. mittel vs. tief)
- Für jede der resultierenden $2 \cdot 3 \cdot 2 = 12$ Klassen c wird der zeitnormalisierte F0-Median-Vektor ermittelt.

- Zusätzlich wird für jeden Sprecher eine Referenz festgelegt als der Median aller vokalischen F0-Konturen.
- Punktweise Division dieser Referenz durch jeden Medianvektor ergibt einen Gewichtsvektor v_c für jede Vokalklasse c.
- Die klassenspezifischen Gewichtungsfunktionen w_c resultieren schließlich aus dem jeweiligen Gewichtsvektor v_c durch Approximation eines Polynoms dritter Ordnung an den Vektor.

Der Glättungsvorgang besteht dann aus zwei Schritten:

- Kategorisierung des zu glättenden vokalischen Segments anhand des Stimmtons der Umgebung sowie der formantgeleiteten Bestimmung der Zungenhöhe.
- Anwendung der entsprechenden Gewichtungsfunktion $w_c(y)$ zur Glättung der F0-Kontur y.

Während MOMEL ausschließlich auf der F0-Kontur basiert, benötigt WAM zusätzlich vorgeschaltete Module zur Stimmton- und Formantdetektion. Dafür gewährleistet WAM in stärkerem Ausmaß eine Trennung von Mikro- und Makroprosodie mit resultierender Beseitigung mikroprosodischer Einflüsse bei Beibehaltung der Makroprosodie (Reichel und Winkelmann, 2010).

4.3.4 Frequenz-Transformationen

Die in Abschnitt 2.5 besprochenen Aspekte der Intonationswahrnehmung legen eine Transformation der Hertz-Werte in perzeptiv adäquatere Skalen nahe. Außerdem sind bei der Intonationsanalyse weniger die sprecherabhängigen absoluten F0-Werte von Interesse als vielmehr F0-Verhältnisse. Die Transformation kann im einfachsten Fall in einer simplen Logarithmierung der F0-Werte y bestehen, wie sie in Implementierungen des Fujisaki-Modells unternommen wird. Häufig werden auch die Hertz- in Halbtonwerte umgewandelt:

$$y_{\mathrm{HT}} \quad = \quad 12 \cdot \log_2 \frac{y_{\mathrm{Hz}}}{b}, \tag{4.4}$$

wobei b für den Referenzwert (beispielsweise 1) steht, der auch einen Basis-F0-Wert, wie er im Fujisaki-Modell in die Kontur mit einfließt, repräsentieren kann.

Eine weitere übliche Transformation bildet Hertz-Werte auf ERB-Werte ab. Nach Hermes und van Gestel (1991) gilt folgender Zusammenhang:

$$y_{\mathrm{ERB}} \quad = \quad 16.7 \cdot \log_{10}(1 + \frac{y_{\mathrm{Hz}}}{165.4}) \tag{4.5}$$

4.3.5 Stilisierung

Zur Berücksichtigung perzeptiver Constraints ist neben der Frequenz-Transformation auch eine perzeptiv motivierte F0-Stilisierung möglich. Diese Stilisierung kann auch parametrischen Intonationsbeschreibungen vorgeschaltet werden, auch wenn ihnen an sich schon eine Stilisierung innewohnt.

Mertens und d'Allessandro (1995) stellen hierzu ein Verfahren vor, dass auf Unterschiedsschwellen bei der Wahrnehmung von Glissandos beruht, namentlich der *glissando threshold G* und der *differential glissando threshold DG* (t'Hart et al., 1990); beide wurden in Abschnitt 2.5 vorgestellt.

Der F0-Verlauf über den Silbenkernen einer Äußerung wird – für jeden Kern getrennt – geglättet, in tonale Segmente zerlegt und schließlich anhand der F0-Stützstellen an den Segmenträndern linear approximiert.

Die Glättung repräsentiert die in d'Alessandro und Castellengo (1994) gefundene Kurzzeitintegration von Vibrato-Tönen und besteht in einer F0-Mittelwertbildung über Zeitbereiche, in denen eine F0-Änderung kleiner G festgestellt wird. Die geglättete Kontur wird nun rekursiv an Punkten hinreichend großer Abweichung von einer durch den Silben-F0-Verlauf gezogenen Geraden in tonale Segmente zerlegt und dort wiederum linear repräsentiert. Im nächsten Schritt werden benachbarte tonale Segmente zusammengefasst, wenn die Differenz ihrer F0-Steigungen kleiner DG ist. Die abschließende F0-Stilisierung ergibt sich durch lineare Interpolation zwischen F0-Zielpunkten, die sich zu Beginn und Ende der tonalen Segmente befinden.

4.3.6 Zeitnormalisierung

Zur Abstrahierung der F0-Kontur von temporalen Größen wie Rhythmus und Sprechgeschwindigkeit kann eine Normalisierung der Zeit t im betrachteten Segment (zum Beispiel einer Silbe) auf ein konstantes Intervall durchgeführt werden, beispielsweise auf das Intervall [0 1] durch:

$$t_{\text{norm}} = \frac{t - \min(t)}{\max(t) - \min(t)} \tag{4.6}$$

4.4 Automatische Klassifizierung

Motiviert durch die Tatsache, dass manuelle prosodische Etikettierung von Sprachdaten sehr zeit- und personalaufwendig ist, wurden diverse Versuche unternommen, anhand des handetikettierten Materials mittels Methoden überwachten Lernens Modelle zur automatischen Lokalisierung und Klassifizierung prosodischer Ereignisse zu trainieren. Mit der Lokalisierung von Akzenten und Phrasengrenzen wird die prosodische Struktur einer Äußerung extrahiert, die Klassifizierung schließlich liefert die theorieabhängigen Kategorien von Akzenten und Phrasengrenzen. Lokalisierung und Klassifizierung können nacheinander (Wagner, 2009) oder auch in einem Schritt (Schweitzer und Möbius, 2009) vorgenommen werden.

Die automatische prosodische Etikettierung kann auf Silbenebene erfolgen oder auf Wortebene, letzteres für den Fall, dass Akzente nur auf worthauptbetonten Silben und Phrasengrenzen nur zwischen Wörtern zugelassen sind.

4.4.1 Merkmale

Signalbasierte Merkmale Intonationsrelevante Merkmale, die sich durch Signalanalyse ermitteln lassen, umfassen F0-Features wie Maximum, Spannweite und Steigung, temporale Features wie z-transformierte Silben- und Nukleus-Dauern und Energie-Features. Eine umfassende Darstellung findet sich beispielsweise in Kießling (1997).

Zusätzlich werden Features herangezogen, die indirekt aus einer F0-Parametrisierung resultieren. So verwenden Batliner et al. (1999) lineare Regressionskoeffizienten und Schweitzer und Möbius (2009) PaintE-Parameterwerte als zusätzliche Merkmale.

In der Regel werden diese Features für die zu klassifizierende Silbe sowie den umgebenden Silben extrahiert.

Textbasierte Merkmale Im Falle einer vorliegenden Signal-Text-Alinierung können auch textbasierte Features herangezogen werden wie beispielsweise Part-of-Speech-Information (POS) und Interpunktion (Vereecken et al., 1998). Da diese Merkmale auch zur Generierung von Intonationskonturen Verwendung finden, erfolgt eine genauere Ausführung an entsprechender Stelle in Kapitel 6.

4.4.2 Klassifikatoren

Zu den für die prosodische Etikettierung herangezogenen Klassifikatoren zählen neuronale Netze (Ananthakrishnan und Narayanan, 2008; Wagner, 2009), C4.5-Entscheidungsbäume und prädikatenlogische Lernverfahren (Rapp, 1998b), Klassifikations- und Regressionsbäume (Bulyko und Ostendorf, 2001) sowie instanzbasierte Lernverfahren (Schweitzer und Möbius, 2009), die auf Grund ihrer konzeptuellen Beschaffenheit auch eine exemplartheoretische Modellierung (Johnson, 1997) der Intonation ermöglichen.

Ergänzend zu den hier aufgelisteten statischen Klassifikatoren, lassen sich mit Hilfe von Hidden-Markov-Modellen auch die Transitionswahrscheinlichkeiten mitmodellieren, mit der prosodische Ereignisse aufeinander folgen (Rapp, 1998b; Brindöpke et al., 1998).

Die höchsterzielten Performanzen liegen derzeit in Abhängigkeit der Schwierigkeit der Klassifikationsaufgabe zwischen etwa 80 % (Vorhersage des ToBI-Inventars) und 95 % (dichotome Vorhersage von Phrasengrenzen).

4.5 Analyse durch Synthese

Während symbolisch beschriebene prosodische Ereignisse mittels manueller Etikettierung oder maschineller Klassifikation gewonnen werden, basiert die Extrahierung der Ereignisse bei parametrischen Modellen in deren Erzeugung. In dem dafür vorgesehenen Rahmen

der Analyse durch Synthese wird eine F0-Kontur analysiert, indem sie durch eine gewählte Stilisierungsfunktion resynthetisiert wird. Die daraus resultierenden Parameterwerte der Stilisierungsfunktion dienen dann als Beschreibung der Kontur.

Allgemein basiert die Stilisierung der F0-Kontur auf einer Anpassung der Stilisierungsparameterwerte dahingehend, dass die Distanz zwischen Original und synthetisierter Kontur minimal wird. In Abhängigkeit der gewählten Stilisierungsfunktion lässt sich die Anpassung **analytisch** oder **numerisch** vollziehen. Während mit Hilfe von analytischen Verfahren die **global** beste Lösung zur F0-Approximation gefunden werden kann, liefern numerische Verfahren unter anderem in Abhängigkeit der Parameterinitialisierung nur eine **lokal** beste Lösung, die sich bei jedem Durchlauf des Verfahrens ändern kann. Die mit dem letzteren Ansatz verbundenen Nachteile werden im nachfolgenden Abschnitt sowie in Abschnitt 7 erörtert.

Alle der in dieser Arbeit besprochenen parametrischen Modelle beruhen auf Stilisierungsfunktionen, die keine analytische Approximation ermöglichen. Auf die Verfahren zweier dieser Modelle (Fujisaki und Tilt) soll im Folgenden ein wenig genauer eingegangen werden.

Fujisaki-Modell

Die F0-Stilisierung im Rahmen des superpositionalen Fujisaki-Modells verlangt zunächst eine Aufspaltung der Kontur in eine Phrasen- und eine Akzentkomponente. Mixdorff (2002) verwendet hierzu beispielsweise eine Hochpassfilterung der F0-Kontur zur Trennung der niederfrequenten Phrasen- von den hochfrequenten Akzentanteilen und passt die Systemparameter mittels Gradientenabstiegsverfahren getrennt an die jeweiligen Komponenten an. Abschließend erfolgt eine weitere Feinanpassung der Parameter an die komplette Kontur.

Problematisch ist die fehlende Injektivität der Relation zwischen Parameterwerten und F0-Kontur, unterschiedliche Parameterbelegungen können also zur selben Kontur führen, was, wie später in Abschnitt 7 eingehender thematisiert, unter anderem die linguistische Interpretierbarkeit des Modells beeinträchtigt.

Ansätze zur Entschärfung dieses Problems bestehen beispielsweise darin, bestimmte Parameter als Konstanten zu betrachten (Mixdorff, 2002; Pfitzinger et al., 2009), oder lokale und globale Parameterwerte gleichzeitig zu schätzen (Agüero et al., 2004), um eine optimale Aufteilung der Kontur in ihre globale und lokale Komponenten zu finden. Weitere Arbeiten beziehen linguistische Constraints bei der Stilisierung mit ein (Sakurai et al., 2003), wozu auch gehört, die temporalen Parameter T_p und T_a (vergleiche Abschnitt 3.12) fest in der prosodischen Struktur zu verankern. Letzteres setzt allerdings eine vorangehende Lokalisierung von Phrasengrenzen und Akzenten voraus.

In Pfitzinger et al. (2009) werden mehrere Extraktorverfahren für Fujisaki-Parameter hinsichtlich mittlerer quadratischer Abweichung zwischen Original und Stilisierung sowie algorithmischer Komplexität miteinander verglichen. Angemerkt wird hierbei aber auch, dass diese rein mathematische Evaluierung nur bedingt aussagekräftig ist, da auf Grund der Übermächtigkeit des Fujisaki-Modells Konturen je nach Stilisierungsverfahren mit

beliebiger Genauigkeit, aber auf Kosten der linguistischen Verankerung realisiert werden können.

Tilt-Modell

Die Gewinnung der Tilt-Parameterwerte wird in Taylor (2000) genauer beschrieben und verläuft grob in folgenden Schritten:

- Lokalisierung des Intonationsereignisses,
- Ermittlung der RFC-Parameter und
- Überführung dieser Parameter in Tilt-Parameter.

Als Ereignisdetektoren fungieren Hidden-Markov-Modelle, die anhand von F0- und Energiewerten das akustische Signal in intonatorisch relevante und nicht-relevante Signalabschnitte segmentieren. Um die extrahierten Ereignisgrenzen herum werden Suchfenster aufgespannt, innerhalb derer dasjenige zeitliche Start- und Endpunkt-Paar ermittelt wird, das zu einer F0-Stilisierung mit kleinstmöglicher Abweichung zur Originalkontur führt. Die Tilt-Parameter ergeben sich schließlich aus den RFC-Parametern der Stilisierungsfunktion wie in Gleichung 3.1 angegeben.

Kapitel 5

Linguistische Interpretation

Nach einem kurzen Überblick über prinzipielle Problematiken und gängige Untersuchungsmethoden der linguistischen Analyse von Intonationsmustern wird zunächst die prosodische Strukturierung einer Äußerung linguistisch beleuchtet und im Anschluss ein Überblick über einige Interpretationsansätze symbolischer und parametrischer Intonationsbeschreibungen gegeben.

5.1 Problemstellung

Ziel ist letztlich die Ermittlung der Bedeutung von Intonationskonturen. Bei symbolischen Konturbeschreibungen wird hierzu versucht, die konkrete Realisierung auf diskrete bedeutungstragende Einheiten, gelegentlich *Intoneme* (Isačenko und Schädlich, 1964) genannt, zurückzuführen. Parametrische Intonationsbeschreibungen versuchen, Komponenten der F0-Parametrisierung mit linguistischen Einflussgrößen in Beziehung zu setzen.

Untersuchungsmethoden Erkenntnisse zur linguistischen Interpretation intonatorischer Ereignisse lassen sich über Korpusanalysen oder Perzeptionsexperimente gewinnen. Im ersten Fall liegen linguistisch (z. B. diskursanalytisch) sowie intonatorisch annotierte oder parametrisierte Sprechdaten vor, die hinsichtlich systematischer Zusammenhänge dieser beiden Beschreibungsebenen statistisch oder linguistisch-impressionistisch untersucht werden. Im zweiten Fall werden Versuchspersonen natürliche oder systematisch intonatorisch variierte Stimuli zu einer bestimmten linguistischen Beurteilung präsentiert, beispielsweise in Form von Akzeptanzurteilen, bei denen Versuchspersonen die Adäquatheit einer Intonationskontur in konkreten linguistischen Kontexten einschätzen sollen (Kohler, 1987; Kleber, 2006).

Abstraktheitsebene der Bedeutung Eine eindeutige Interpretation intonatorischer Muster wird enorm erschwert durch die oftmals beobachtete 1-zu-N-Beziehung zwischen Form und Funktion der Intonation. So stellten Ward und Hirschberg (1985) fest, dass dieselbe (TSM-repräsentierte) Kontur kontextabhängig als Zeichen von Unsicherheit, Un-

gläubigkeit, Höflichkeit oder Ironie ausgelegt werden kann. In Pierrehumbert und Hirschberg (1990) wird von fehlenden Konturunterschieden zwischen W-Fragen und Deklarativsätzen berichtet. Ein gängiger Ansatz zur Entschärfung dieses Problems besteht darin, die Bedeutung von Konturmustern hinreichend abstrakt zu formulieren und kontextabhängig zu spezifizieren (Pike, 1945; Gussenhoven, 1984). Für eine detaillierte Beschreibung dieser Problematik siehe Peters (2006), S. 101ff.

Alternative Codierungsmuster In vielen Sprachen kann die Codierung derselben linguistischen Information auch mit anderen Mitteln als der Intonation erfolgen. In erhöhtem Maße gilt dies für Sprachen mit relativ freier Wortstellung, wie dem Deutschen, in dem beispielsweise Hervorhebung sowohl intonatorisch durch Akzentuierung als auch syntaktisch beispielsweise durch Linksversetzung bewerkstelligt werden kann.

5.2 Prosodische Struktur

5.2.1 Phrasierung

Konstituentenstruktur

Phrasierung dient der Zusammenfassung inhaltlich zusammengehöriger Äußerungsteile, was sich in Bezeichnungen für prosodische Phrasen als *sense units* (Selkirk, 1984) widerspiegelt. Die Phrasierung muss aber nicht notwendigerweise aus der syntaktischen Struktur einer Äußerung ableitbar sein, wie das folgende Beispiel aus Nespor und Vogel (1986) zeigt:

> This is [the cat that caught [the rat that stole [the cheese]$_{NP}$]$_{NP}$]$_{NP}$.
> [This is the cat]$_{IP}$ [that caught the rat]$_{IP}$ [that stole the cheese]$_{IP}$.[1]

Analog zur Syntax lässt sich aber auch die prosodische Phrasierung einer Äußerung in Form einer Konstituentenstruktur angeben. Nespor und Vogel (1986) gehen von einer streng-hierarchischen Struktur basierend auf der *Strict Layer Hypothese* nach Selkirk (1984) aus mit den Kennzeichen *Exhaustivität* (jedes Segment einer Ebene wird komplett von einem Segment der nächsthöheren Ebene dominiert) und *Non-Rekursivität*. Ladd (1986) hingegen plädiert für eine rekursive prosodische Konstituentenstruktur, da diese die Darstellung von Abhängigkeiten nicht benachbarter (beispielsweise durch Einschübe getrennter) Intonationsphrasen erlaubt.

Performance Structure

Gee und Grosjean (1983) entwickelten auf Grundlage empirischer Analysen die prosodische hierarchische *Performance Structure* von Äußerungen, die sie von gemittelten und satznormalisierte Pausendauern zwischen benachbarten Wörtern ableiteten. Mittels des von den Autoren vorgestellten ϕ-Phrasen-Algorithmusses lässt sich die syntaktische

[1] NP: Nominalphrase, IP: Intonationsphrase

Struktur eines Satzes in seine prosodische *Performance Structure* überführen. ϕ-Phrasen werden hierbei im Allgemeinen durch Segmentierung des Satzes hinter jedem Inhaltswort, das den Kopf einer syntaktischen Konstituente bildet, gewonnen. Zu den Ausnahmen zählen attributive Adjektive, die keine eigene ϕ-Phrase bilden können. Hier ein Beispiel einer entsprechenden Zerlegung eines Satzes (Gee und Grosjean, 1983, S. 445, Köpfe sind fettgedruckt):

[**John**] [**asked**] [the strange young **man**] [to be **quick**] [on the **task**]

Chunk Parsing

Abney (1991) entwickelte zur prosodisch motivierten flachen syntaktischen Analyse einen *Chunk-Parser*. Allgemein sind Chunks hier definiert als Inhaltswörter, die als *major heads* fungieren können, zugehörige Funktionswörter, sowie Inhaltswörter, die sich zwischen den *major heads* und deren zugehörigen Funktionswörtern befinden. Ein Beispiel (die *major heads* sind fett gedruckt):

[**John**] [**asked**] [the strange young **man**] and [**nodded**]

Durch Eingliederung unverknüpfter Wörter (*orphan nodes*; *and* im obigen Beispiel) in den nachfolgenden Chunk lassen sich Chunks in ϕ-Phrasen überführen.

Detachment-Regel

Hirst (1993) trägt mit seiner *Detachment-Regel* der Variabilität der prosodischen Phrasierung einer Äußerung Rechnung. Gemäß dieser rekursiven Regel können optional bestimmte syntaktische Konstituenten am rechten Rand eines Syntaxbaums prosodisch vom vorangehenden Satzteil separiert werden. Diese Regel gilt für Konstituenten der Kategorien *Satz, Nominalphrase, Verbalphrase* und *Präpositionalphrase*. Der Satz *Jane gave Mary the book* lässt sich demzufolge unter anderem in folgenden Phrasierungsvarianten realisieren (Hirst, 1993, S. 785f):

[Jane gave Mary the book]
[Jane] [gave Mary the book]
[Jane] [gave] [Mary] [the book]

5.2.2 Akzente

Die Aufgabe der Akzentuierung besteht in der Markierung des Fokus einer Äußerung. Fokus lässt sich definieren als 'Informationszentrum eines Satzes, auf das das Mitteilungsinteresse des Sprechers gerichtet ist' (Bußmann, 1990), etwa die Einführung neuer Information im Diskursverlauf oder die Herausstellung von Kontrasten.

Die Ausdehnung eines Fokus kann sich von einem einzelnen Morphem oder Wort hin zu einem ganzen Satz erstrecken. Zur Unterscheidung wird hier auch von *engem* und *weitem* Fokus gesprochen (Ladd, 1980). Das im Fokusbereich akzentuierte Wort wird als

Fokusexponent bezeichnet. Der Teil der Äußerung, der nicht im Fokus steht, bildet den *Hintergrund*.

Alternativ zur Aufteilung einer Äußerung in Fokus und Hintergrund kann auch eine Zerlegung in *Topic* und *Comment* (in der Regel synonym zu: *Thema* und *Rhema*) vorgenommen werden (Halliday, 1967a). Unter *Topic* wird der Gegenstand, über den der Sprecher etwas mitteilen möchte, verstanden, und unter *Comment* das, was darüber ausgesagt wird.

Der Fokusbereich umspannt den gesamten Satz, wenn dieser als Ganzes neue Information beinhaltet, zum Beispiel als Antwort auf die Frage: 'Was gibt es Neues?' (sogenannte: *all-new*-Sätze). Der in diesen Fällen vergebene Akzent wird als *neutraler Akzent* bezeichnet.

Linguistische Ansätze zur Lokalisierung des neutralen Akzents finden sich beispielsweise in der generativen Grammatik (Chomsky und Halle, 1968; Zubizarreta, 1998) und der metrischen Phonologie (Liberman, 1975; Liberman und Prince, 1977).

Syntax

Chomsky und Halle (1968) formulierten für das Englische zwei grundlegende phonologische Regeln zur Zuweisung des neutralen Akzents, die auf der syntaktischen Oberflächenstruktur operieren: die *compound stress rule (CSR)* und die *nuclear stress rule (NSR)*.

> CSR: bei Komposita fällt der stärkste Akzent auf das erste Glied (Bsp.: **black**board)
> NSR: in einer syntaktischen Phrase fällt der stärkste Akzent auf die letzte Konstituente (Bsp.: [a black **board**]$_{NP}$)[2]

Dass diese Regeln nicht immer zu richtigen Vorhersagen führen, zeigen folgende Beispiele:

> *Mary* [[***kissed***]$_V$ [*him*]$_{NP}$]$_{VP}$
> *Er hat sich* [[*das* ***Knie***]$_{NP}$ *gestoßen*]$_{VP}$.

Hier fällt in der Verbalphrase der stärkste Akzent nicht auf die letzte Konstituente.

Cinque (1993) lokalisiert den neutralen Akzent im Rahmen seiner *null theory of phrase and compound stress* auf der syntaktisch am tiefsten eingebetteten Konstituente, womit in der obigen Verbalphrase [[*das* ***Knie***]$_{NP}$ *gestoßen*]$_{VP}$ eine korrekte Akzentzuweisung gelingt.

Phonologie

In der metrischen Phonologie werden Prominenzverhältnisse in einem Satz als binär verzweigender *metrischer Baum* dargestellt, dessen terminale Elemente die Silben sind: Jeder Knoten hat zwei Tochterknoten, die benachbarte Silben beziehungsweise syntaktische Konstituenten dominieren. Durch Labeln des einen Tochterknotens als *strong* und des

[2] Beispiele aus Winkler (1997).

anderen als *weak* werden diese Silben oder Konstituenten in Prominenzrelation zueinander gesetzt. Verfolgt man ausgehend von der Baumwurzel den durch die *strong*-Knoten vorgegebenen Pfad, so trifft man schließlich auf die prominenteste Silbe in der Äußerung: dem *designated terminal element*. Man trifft hier auf eine leicht modifizierte Version der von Chomsky und Halle formulierten Akzentuierungs-Grundregeln:

> Gegeben sei die syntaktische Konstituente $[AB]_C$ mit den Tochterknoten A und B:
>
> - NSR: Ist C eine Phrase, so ist B *strong s*.
> - CSR: Ist C ein Wort oder Teil eines Wortes, so ist B *strong*, wenn es sich weiter verzweigt, ansonsten ist A *strong* und B *weak w*.

Das Beispiel in Abbildung 5.1 soll zeigen, wie diese Regeln angewendet werden.

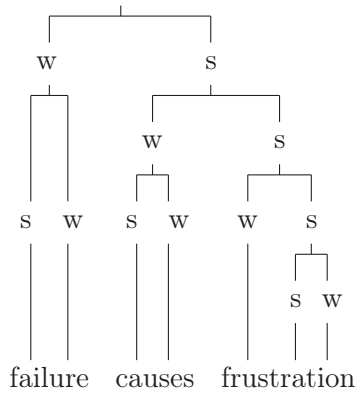

Abbildung 5.1: Metrischer Baum.

Auf Phrasenebene bestimmt dort die NSR, dass der rechte Tochterknoten prominenter ist als der linke. Auf lexikaler Ebene wirkt die CSR, derzufolge bei Zweisilbern die erste Silbe prominenter ist als die zweite und bei Dreisilbern (nach zweimaliger Anwendung der CSR) die zweite Silbe die größte Prominenz zugewiesen bekommt.

Eine zweite Möglichkeit, die Prominenzverhältnisse in einer Äußerung darzustellen, bietet das *metrische Gitter*. Hier wird die Prominenz einer Silbe als Säule von Schlägen dargestellt: je höher die Säule, desto prominenter die Silbe. Durch Regeln zur Umgestaltung dieses Gitters lassen sich auch Constraints wie die Vermeidung von *stress clashes*, also von unmittelbar aufeinanderfolgenden akzentuierten Silben, realisieren.

Semantik

Die Existenz eines neutralen Akzents, der sich anhand der syntaktischen Struktur vorhersagen ließe, wird von Bolinger (1972) bestritten. Er gibt folgendes Gegenbeispiel:

> *I have a point to* ***emphasize****.*
> *I have a* ***point*** *to make.*

Bolinger führt die unterschiedlichen Akzentuierungen bei gleicher syntaktischer Struktur darauf zurück, dass *emphasize* ein größeres semantisches Gewicht hat als *point*, während für *make* das Gegenteil gilt. Nach Bolinger besteht die Motivation der Akzentuierung eines Worts in seinem semantischen Gewicht relativ zur Textumgebung, was sich im Wesentlichen aus der Vorhersagbarkeit des Worts aus dem Kontext ergibt. Mit dem Kontext ist die Wortumgebung ebenso gemeint wie die Situation, in der sich der Sprecher äußert. Dadurch ergibt sich eine Vielzahl von Einflussfaktoren, weshalb Bolinger pessimistisch konstatiert:

> *Accent is predictable (if you're a mind reader)* (Bolinger, 1972, Aufsatztitel).

Auch in Gussenhoven (1999) bezieht sich die Fokusmarkierung nicht auf syntaktische, sondern auf semantische Konstituenten, die er in *Argumente* (Subjekt, Verb-Argumente), *Prädikate* (Verben, prädikative Adjektive, etc.) und *Modifikatoren* (Adverbien) unterteilt. Steht eine dieser semantischen Konstituenten im Fokus, muss sie gemäß seiner *Sentence Accent Assignment Rule (SAAR)* akzentuiert werden – mit Ausnahme von fokussierten Prädikaten, die sich neben ebenfalls fokussierten Argumenten befinden. Argumente werden hier also gegenüber Prädikaten als stärkere Akzentattraktoren begriffen.

Ein weiteres semantisches Motiv der Akzentuierung besteht in der Herausstellung von *Kontrasten*. Zu Analyseverfahren von Kontrastkonstruktionen sei auf Arbeiten von Prevost (1995) und van Deemter (1998) verwiesen.

Diskurs

In diskursanalytische Ansätzen (Chafe, 1976; Vallduví, 1993) wird Akzentuierung im Kontext gegebener (Hintergrund) und neuer Information (Fokus) behandelt. Die Beurteilung einer Information als gegeben oder neu beruht hierbei auf wechselseitigen Annahmen (*mutual beliefs*) der Dialogpartner über den Kenntnisstand des jeweils anderen. Als gegeben wird diejenige Information angenommen, die

- im bisherigen Diskursverlauf bereits übermittelt wurde oder
- zum von Sprecher und Hörer geteilten Weltwissen gehört oder
- aus dem situativen Kontext erschlossen werden kann.

Die Fokus-Hintergrund-Struktur einer Äußerung muss sich nicht mit ihrer Topic-Comment-Struktur decken, wie Vallduví (1993) in folgendem Beispiel verdeutlicht:

> *What about Mary? What did she do?*
> $[Mary]_{topic}$ $[[gave\ a\ shirt\ to\ \mathbf{Harry}.]_{focus}]_{comment}$
> *What about Mary? What did she give to Harry?*
> $[Mary]_{topic}$ $[gave\ [a\ \mathbf{shirt}]_{focus}\ to\ Harry.]_{comment}$

Nach Vallduví zeigt dieses Beispiel, dass die Akzentsetzung (dick hervorgehobene Wörter) der Fokus-Hintergrund-Struktur und nicht der Topic-Comment-Struktur folgt, die in beiden Sätzen trotz unterschiedlicher Akzentuierung dieselbe ist.

Feinere Unterteilungen der Konzepte “Gegeben” und “Neu” finden sich bei Prince (1981).

Einen Rahmen zur Auflösung anaphorischer Beziehungen bei der Identifizierung gegebener und neuer Information liefert die *Centering*-Theorie (Grosz et al., 1995), in der satzweise Diskursreferenten[3] geordnet in vorwärts- und rückwärtsbezogenen Zentren (*forward-* und *backward-looking centers*) gesammelt werden. Anaphernauflösung funktioniert in Form regelgeleiteter Zuordnungen von Elementen aus rückwärtsbezogenen zu Elementen aus vorwärtsbezogenen Zentren.

Ein von Grosz und Sidner (1986) entwickeltes Diskursmodell findet im Zusammenhang mit der Interpretation von prosodischer Struktur und Intonation verschiedentlich Verwendung. Die Struktur eines Diskurses gliedert sich hier in drei Komponenten, die *linguistische Struktur* (der geschriebene oder gesprochene Text), die *attentionale Struktur (attentional structure)* als Repräsentation der relativen Salienz (Hervorgehobenheit) von Diskursentitäten und die *intentionale Struktur (intentional structure)*, unter der die Intentionen des Sprechers zu verstehen sind, die zusammengenommen den Zweck des Diskurses ergeben. Hirschberg und Pierrehumbert (1986) sowie Hirschberg et al. (1987) stellten beispielsweise den positiven Zusammenhang zwischen der Salienz von Diskurseinheiten (als Teil der attentionalen Diskursstruktur) und deren Akzentuierung heraus, ebenso wie die Abhängigkeit der Akzentsetzung von der Sprecherintention – beispielsweise der Intention, dem Hörer Bekanntheit einer Diskurseinheit zu signalisieren.

5.3 Intonation

5.3.1 Interpretation symbolisch beschriebener Ereignisse

Intoneme nach Stock und Zacharias

In der Tradition von Isačenko und Schädlich (1964) und ihrem experimentell fundierten Konzept der kommunikativ motivierten Tonhöhenwechsel (siehe Abschnitt 3.1.1) nehmen Stock und Zacharias (1982) drei durch solche Tonhöhenwechsel charakterisierte phonologisch distinkte abstrakte “*Intoneme*” an:

- $I\downarrow$: Informationsintonem, charakterisiert durch Tonhöhenwechsel nach unten und zur Markierung des Abschlusses einer Informationseinheit.
- $N\uparrow$: Nonterminalitäts-Intonem mit Tonhöhenwechsel nach oben zur Markierung der Nichtabgeschlossenheit der Äußerung
- $C\uparrow$: Kontaktintonem mit Tonhöhenwechsel nach oben zur Kontaktaufnahme mit dem Hörer beispielsweise durch eine Frage.

Diese Intoneme können auch im Rahmen des Pierrehumbertschen TSM als Grenztöne modelliert werden.

[3] Unter *Diskursreferenten* sind nach Karttunen (1976) Repräsentanten von Personen oder Dingen der in der Äußerung beschriebenen Welt zu verstehen. Zu den lexikalisch-syntaktischen Mitteln zur Einführung von Diskursreferenten gehören Nominalphrasen und Pronomen.

Interpretation von Tönen im TSM

Gussenhoven (1984) Gussenhoven unterscheidet zwischen drei abstrakten semantischen Konzepten, die er den nuklearen Tonfolgen HL (Fall), HLH (Fall-Anstieg) und LH (Anstieg) zuordnet: *Addition*, der Hinzufügung neuer Information, *Selection*, des Aufgreifens einer gegebenen Information sowie *Testing*, bei dem der Sprecher sich hinsichtlich der Gegebenheit von Information nicht festlegt.

Pierrehumbert und Hirschberg (1990) Pierrehumbert und Hirschberg stellen ein kompositionelles Modell für das Englische vor, das Intonationsbausteinen jeweils auf unterschiedlicher intonatorischer Ebene eine abstrakte Bedeutung zuordnet.

Tonakzente spezifizieren den Status unter anderem von Diskursreferenten: H^*-Tonakzente codieren hierbei Neuheit, L^*–Tonakzente Bekanntheit. Mit $L+H$–Verbindungen lässt sich eine Hervorhebung erzielen, beispielsweise um Unsicherheit (L^*+H) oder Kontrastierung ($L+H^*$) auszudrücken. $H+L$–Akzente hingegen übermitteln Inferierbarkeit von Information (v.a. H^*+L), sei es aus der Diskursvorgeschichte, aus dem situativen Kontext oder aus dem geteilten Weltwissen.

Phrasentöne reflektieren die Verbindungsstärke zwischen intermediären Phrasen. Ein H^--Ton signalisiert hier die inhaltliche Zusammengehörigkeit einer Phrase mit der folgenden. Dadurch sind auch unterschiedliche semantische Interpretationen von Konjunktionen möglich. So ruft der H^--Ton in:

George ate chicken soup/**H–** and got sick

für die Konjunktion "and" eher eine kausale als eine beiordnende Funktion hervor (Pierrehumbert und Hirschberg, 1990, S. 304).

Grenztöne geben Anhaltspunkte zur Orientierung der aktuellen Intonationsphrase im Diskurs im Hinblick auf Abgeschlossenheit und Gerichtetheit. In

(1) *My car manual is almost unreadable*/**LL%**
(2) *<u>It</u>'s quite annoying*/**LH%**
(3) *I spent two hours figuring out how to use the jack*

ist Phrase (1) durch den finalen Intonationsverlauf als abgeschlossen markiert und Phrase (2) dadurch und durch ihren progredienten F0-Verlauf am Ende als vorwärtsgerichtet. *It* referiert in diesem Fall auf Phrase (3).

Dagegen dreht sich die Gerichtetheit von Phrase (2) im folgenden Beispiel um:

(1) *My car manual is almost unreadable*/**LH%**
(2) *<u>It</u>'s quite annoying*/**LL%**
(3) *I spent two hours figuring out how to use the jack.*

It bezieht sich nun auf Phrase (1) (Pierrehumbert und Hirschberg, 1990, S. 305).

Dem kompositionellen Modell nach Pierrehumbert und Hirschberg zufolge codieren also elementare Tonakzente die Verfügbarkeit der übermittelten Information, Folge- und

Leittöne die Bewertung der Information, Phrasentöne die Verbindungsstärke zwischen Informationseinheiten und schließlich Grenztöne die Relation der aktuellen Intonationsphrase mit der folgenden.

Peters (2006) Peters verzichtet in seinem Ansatz zur Beschreibung des Deutschen auf die Ebene der intermediären Phrasen und damit wie auch Féry (1993) auf Phrasentöne. An deren Stelle treten Folgetöne, die ebenso wie die L-Phrasentöne bei Pierrehumbert Abgeschlossenheit von Informationseinheiten repräsentieren können. Anstelle der verbindenden H-Phrasentöne sorgt hier das *Linking*, also der Wegfall des Folgetons mit resultierender Hutkontur für die Markierung inhaltlicher Zusammengehörigkeit, was analog zum *chicken soup*-Beispiel oben durch Uminterpretation der Konjunktion eine additive in eine elaborative Äußerung umwandeln kann.

Sein kompositionales Modell unterscheidet sich von dem nach Pierrehumbert und Hirschberg (1990) im Wesentlichen darin, dass hier die Abgeschlossenheit einer Informationseinheit nicht durch Phrasentöne, sondern durch die Anwesenheit von Folgetönen markiert wird.

Mayer (1997) Weitere Untersuchungen zur linguistischen Deutung der Töne finden sich beispielsweise bei Mayer (1997), in dessen Arbeit ihre Funktion im Deutschen bei der Disambiguierung von Satzadverben, der Anaphernresolution sowie der Diskursstrukturierung behandelt werden.

KIM

Im Kieler Intonationsmodell wird die linguistische Relevanz der Alinierung von F0-Maximum und Silbenkern hervorgehoben. Unterschieden wird zwischen *frühem, mittlerem* und *spätem Gipfel*. Bei frühen Gipfeln befindet sich das F0-Maximum zeitlich vor, bei mittleren Gipfeln auf und bei späten Gipfeln hinter dem Nukleus. Bereits in Abschnitt 2.5.3 vorgestellte Perzeptionsexperimente (Kohler, 1987) haben ergeben, dass diese Gipfel unterschiedliche Diskursfunktionen haben, nämlich die Markierung gegebener, neuer sowie überraschend neuer Information.

Entsprechungen lassen sich bei Pierrehumbert und Hirschberg (1990) finden, wo wie weiter oben bereits ausgeführt $L + H$–Akzente eine Hervorhebung und $H + L$–Akzente eine Inferierbarkeit von Äußerungsteilen vermitteln. $L^* + H$ haben ihr Äquivalent in späten Gipfeln, die für das mit "Hervorhebung" vereinbare Konzept "überraschend neue Information" stehen. Entsprechendes gilt für die mit frühem Gipfel beziehungsweise mit $H + L^*$–Akzent codierbare Inferierbarkeit von Information.

5.3.2 Interpretation parametrisch beschriebener Ereignisse

Fujisaki-Modell

Möbius (1993a) stellte im Deutschen durch Analyse der Parameterwerte bei der Intonationsstilisierung Zusammenhänge zwischen der Amplitude der Phrasenkommandos und

dem Satzmodus fest. Die Amplituden der Akzentkommandos sind abhängig von der Position der akzentuierten Silbe in der Intonationsphrase (im Zuge der Deklination kleiner werdend) und von der Wortart (bei Nomen beispielsweise größer). Die Dauer von Akzentgruppen ist unter anderem abhängig von Satzmodus und Position der Akzentgruppe in der Äußerung.

Mixdorff (1998) ließ deutsche Sprecher konstante Wortfolgen unter Variation des Satzmodus und der Fokus-Hintergrund-Struktur äußern, um nach Stilisierung der produzierten F0-Konturen die Fujisaki-Parameter im Hinblick auf die linguistischen Variablen interpretieren zu können. Darüber hinaus führte er mit synthetischen Stimuli, deren F0-Verläufe mittels des Fujisaki-Modells systematisch variiert wurden, Perzeptionsexperimente zur Beurteilung des Satzmodus durch. Für die Unterscheidung von Aussage gegenüber nicht-terminalen Intonationsverlauf konnte vor allem der Offset-Zeitpunkt T_2 des nuklearen Akzentkommandos verantwortlich gemacht werden. Frageintonation ließ sich gegenüber nicht-terminalem Verlauf durch Einsatz eines zusätzlichen Akzentkommandos am Phrasenende hervorrufen. Ein enger Fokus zeichnete sich durch eine Erhöhung der Amplitude A_a des entsprechenden Akzentkommandos aus, ein weiter Fokus dagegen mitunter durch Zusammenfall benachbarter Akzentkommandos, was im Kontext des TSM als Hutmuster erzeugendes *Linking* interpretiert werden kann.

Kapitel 6

Intonationsgenerierung

Die textbasierte Generierung der Intonation beispielsweise im Rahmen der Text-to-Speech-Synthese umfasst im Großen und Ganzen die folgenden Schritte:

1. Prosodische Strukturierung von Texten, also Lokalisierung von Akzenten und Phrasengrenzen. Im Falle symbolischer Intonationsbeschreibungen werden diese Ereignisse in Abhängigkeit der Detailliertheit des Modells noch weiter hinsichtlich ihrer tonalen Eigenschaften spezifiziert.

2. Generierung der Intonationskontur anhand der spezifizierten strukturgebenden Stützstellen.

In der konkatenativen Sprachsynthese findet auch eine signalbasierte Intonationssteuerung statt, die darin besteht, dass intonatorisch relevante akustische Merkmale in die die *Unit Selection* steuernde Kostenfunktionen mit eingehen (Bulyko und Ostendorf, 2001; Clark und King, 2006).

6.1 Textbasierte Vorhersage prosodischer Struktur

6.1.1 Phrasengrenzen

Lokalisierungen von Phrasengrenzen basieren beispielsweise auf Part-of-Speech-(POS)-Informationen wie der *Chink-Chunk*-Algorithmus (Liberman und Church, 1992). Dieses Verfahren teilt Wortarten auf in solche, die tendenziell eher phraseninitial auftreten (*Chinks*, beispielsweise Präpositionen und Artikel im Deutschen) und solche, die eher phrasenfinal zu finden sind (*Chunks*, zum Beispiel Nomen). Phrasengrenzen werden dann hinter *Chunks* gesetzt, auf die ein *Chink* folgt.

Auch statistische Ansätze (Taylor und Black, 1998) nutzen die POS-Informationen, indem sie das Etikettierproblem als Suche nach der wahrscheinlichsten binären Grenzlabel-Sequenz $\hat{G}$ gegeben eine beobachtete POS-Folge W modellieren:

$$\hat{G} = \arg\max_{G} [P(G|W)] \quad (6.1)$$

$$= \arg\max_{G} [P(W|G) \cdot P(G)] \quad (6.2)$$

Diesem Ansatz liegt allgemein das Noisy-Channel-Modell zugrunde: Beobachtbar ist als Ausgabe eines verrauschten Kanals die POS-Sequenz W, aus der die wahrscheinlichste Eingabe in den Kanal $\hat{G}$ rekonstruiert werden muss (Gleichung 6.1). Mittels der Bayes'schen Umformung lässt sich dieser Ausdruck in Gleichung 6.2 zerlegen in Transitions- $P(G)$ und Emissionswahrscheinlichkeiten $P(W|G)$, anhand derer durch den Viterbi-Algorithmus der wahrscheinlichste Pfad durch ein Hidden-Markov-Modell zur Generierung von W gefunden werden kann – und damit $\hat{G}$.

Weitere Verfahren beruhen auf in Abschnitt 5.2 vorgestellten flachen syntaktischen Analyseverfahren, beispielsweise durch den ϕ-Phrasen-Algorithmus (Gee und Grosjean, 1983; Bachenko und Fitzpatrick, 1990) oder den prosodisch motivierten Chunk-Parser (Abney, 1991).

Maschinelle Lernverfahren wie beispielsweise Entscheidungsbäume (Veilleux, 1994) werden mit Kombinationen von textbasierten Merkmalen trainiert. Für weiterführende Darstellungen textbasierter Vorhersagemethoden siehe beispielsweise Reichel (2002).

6.1.2 Akzente

Textbasierte Verfahren zur Kontrolle der Akzentvergabe stützen sich auf POS-Informationen zur Deakzentuierung von Funktionswörtern sowie auf höhere linguistische Analysen wie in Abschnitt 5.2 beschrieben. Beispiele hierfür finden sich in Hirschberg (1993), wo gegebene Information anhand von Inhaltswort-Stapeln, die an Paragraphenden geleert werden, identifiziert wird und zur Deakzentuierung der betroffenen Wörter führt. In van Deemter (1998) finden sich neben der Identifizierung neuer und gegebener Information auch Kontrastpaaranalysen. Statistische Ansätze ziehen bei der Akzentvergabe die beispielsweise mittels N-Gramm-Wahrscheinlichkeiten angegebene globale und lokale Wortvorhersagbarkeiten heran (Pan und McKeown, 1999; Pan und Hirschberg, 2000). Es werden Performanzen bis zu 95 % für Phrasengrenzen und 90 % für Akzente erreicht.

6.1.3 Tonale Spezifikationen

Die genauere Spezifikation der strukturgebenden Ereignisse in Tonsequenzansätzen erfolgt häufig erst nach Festlegung der prosodischen Struktur. Im Text-to-Speech-System MARY (Schröder und Trouvain, 2003) werden Akzenten und Phrasengrenzen regelbasiert in Abhängigkeit der Position im Satz und des Satztyps die tonalen Label zugewiesen. In Black und Campbell (1995) geschieht dies maschinell mit Hilfe von CART-Klassifikatoren (Breiman et al., 1984), die als Features (manuell gelabelte) Sprech- und Dialogakte verwenden. Die berichteten Performanzen sind hier auf Grund der erhöhten Anzahl möglicher Klassen allgemein niedriger als bei der dichotomen prosodischen Strukturierung.

6.2 Konturgenerierung

Während aus parametrischen Intonationsrepräsentationen die F0-Kontur unmittelbar abgeleitet werden kann, kommen abstrakt-symbolische Repräsentationen nicht ohne zusätzliche Module zur Umsetzung der prosodischen Etiketten in konkrete F0-Werte aus.

6.2.1 Bei parametrischer Intonationsbeschreibung

Bei parametrischen Intonationsbeschreibungen besteht die Aufgabe darin, aus dem Text heraus die Werte der Stilisierungsparameter vorherzusagen.

In Möbius (1993a, 1995) werden die Parameterwerte des Fujisaki-Modells durch handgefertigte Regeln erzeugt, die auf den in Möbius (1993a) untersuchten Zusammenhängen basieren, wie sie in Kapitel 5 teilweise vorgestellt wurden. Die Regeln sind im Synthesesystem HADIFIX (Portele et al., 1992) implementiert, und sie wurden in Perzeptionsexperimenten (Möbius und Pätzold, 1992; Möbius, 1993b) auf ihre Adäquatheit hin evaluiert.

Mixdorff (1998) behandelt in seinem regelbasierten Ansatz zur Vohersage der Parameterwerte des Fujisaki-Modells die Parameter F_b, α und β als Konstanten. Der Text wird prosodisch in Intonemsegmente nach Isačenko und Schädlich (1964) (vgl. Abschnitt 3.1.1) gegliedert und die Akzentkommando-On- und -Offsets mit diesen Segmenten aliniert. Phrasenkommandos werden zeitlich kurz vor Phrasengrenzen positioniert. Akzentkommandoamplituden werden in Abhängigkeit von Akzentstärke (drei Stufen) und Silbenposition innerhalb der Äußerung bestimmt. Die Phrasenkommandoamplituden schließlich werden durch einen Regressionsbaum unter anderem anhand der Phrasenlänge vorhergesagt.

Dusterhoff et al. (1999) prädizieren die Tilt-Parameterwerte mit Regressionsbäumen (Breiman et al., 1984) anhand leicht extrahierbarer Features unter anderem zur Position der Silbe in der aktuellen prosodischen Phrase und zu ihrer rhythmischen Einbettung in Form des Abstands zu vorangehenden und folgenden Akzenten.

6.2.2 Bei symbolischer Intonationsbeschreibung

Handgefertigte regelbasierte Verfahren finden sich beispielsweise bei Anderson et al. (1984) und Jilka et al. (1999) für das Amerikanische Englisch. F0-Zielwerte für jedes Tonsymbol werden hier zeitlich relativ zum Silbennukleus und auf der Frequenzachse relativ zu Topline und Baseline ermittelt. Die Angaben zu dieser relativen Positionierung werden von Faktoren wie Akzenttyp, metrischer Prominenz der assoziierten Silbe, Position innerhalb der Phrase, Phrasenlänge sowie vorangehenden F0-Werten abgeleitet.

In ihrem statistischen Ansatz weisen Black und Hunt (1996) den Vokalen jeder tonal markierten Silbe jeweils drei F0-Werte zu, die mittels linearer Regression berechnet werden, unter Verwendung von unter anderem folgenden Prädiktoren: Tonlabel, Grenzlabel, Wortbetonung und Position der Silbe in der Intonationsphrase. Einige Merkmale (wie Tonlabel) werden in einem 5-Silben-Fenster extrahiert. Kategoriale Prädiktoren werden binär codiert.

Kapitel 7

Diskussion

Der in diesem Teil der Arbeit gegebene Forschungsüberblick schließt mit einer Sammlung von Anforderungen an die Intonationsmodellierung und diskutiert, inwieweit die vorgestellten Intonationsmodelle diesen Anforderungen genügen.

7.1 Anforderungen an ein Intonationsmodell

Angemessene Abstrahierung vom Signal Da anzunehmen ist, dass der Sprecher nicht jeden F0-Wert einer Äußerung einzeln plant, sollte die Repräsentation in einer Datenreduktion in Form einer Abstrahierung von der konkreten F0-Kontur bestehen. Diese Abstrahierung soll

- relevanten Aspekte des Signals erfassen,
- so gestaltet sein, dass das Signal ausgehend von der abstrakten Repräsentation so originalgetreu wie nötig reproduziert werden kann,
- selbst reproduzierbar sein, wenn sie wiederholt auf demselben Signal vorgenommen wird.

Interpretierbarkeit Die abstrakte Repräsentation des F0-Verlaufs sollte so weit wie möglich linguistisch und physiologisch interpretierbar und vorhersagbar sein.

Automatisierbarkeit Eine Automatisierbarkeit der Gewinnung der Intonationsrepräsentation bringt diverse Vorteile mit sich, so zum Beispiel die folgenden:

- Das Modell lässt sich an größeren Datenmengen testen.
- Vorhandene Datenbanken lassen sich ohne großen Aufwand nach Modifizierungen des Modells aktualisieren.

- Das Modell lässt sich mit wenig Aufwand auf neue Daten, beispielsweise anderen Sprachen oder auf andere Domänen, beispielsweise Sprechgeschwindigkeitsverläufe, anwenden und testen.

In den folgenden Abschnitten werden die Ansätze der Intonationsmodellierung gemäß der unter Punkt 3.1 vorgeschlagenen Unterteilungskriterien dahingehend beleuchtet, inwieweit sie die aufgeführten Anforderungen erfüllen.

7.2 Angemessene Abstrahierung vom Signal

Datenreduktion Alle besprochenen Intonationsbeschreibungen sorgen – wenn auch in unterschiedlichem Ausmaß – für eine Abstrahierung der F0-Kontur. Zur größten Datenreduktion führen hierbei TSM-Ansätze mit ihrem endlichen geringen Inventar an Tönen sowie parametrische Ansätze wie PaintE, Rapp, Tilt und von Portele & Heuft, die mit einer geringen Anzahl von Parametern auskommen. Eine weit geringere Abstrahierung ist mit KIM zu erreichen, das F0-Verläufe teilweise sehr detailliert nachzeichnet, beispielsweise allein sieben phrasenfinale Intonationskonturen unterscheidet.

Bewahrung relevanter intonatorischer Aspekte Tonbasierte Ansätze müssen sich der Kritik stellen, relevante Eigenschaften des F0-Verlaufs zwischen den Tontargets bei der Modellierung zu eliminieren. So lassen sich beispielsweise im Neapolitanischen Fragen von Aussagen anhand der Form des F0-Verlaufs zwischen pränuklearem und nuklearem Ton unterscheiden (Petrone und D'Imperio, 2008): in Fragen ist der Verlauf konkav, in Aussagen linear. Denkbare Behandlungen solcher Phänomene im Rahmen tonbasierter Modellierung wären (a) die Einfügung eines dritten Targets zwischen pränuklearem und nuklearem Akzent zur Spezifizierung der konkaven Form, oder (b) die Erweiterung des Modells um Interpolationsregeln. Lösungsvorschlag (a) läuft aber dem Grundgedanken zuwider, dass Tontargets nur mit akzentuierten oder phrasenfinalen Silben verbunden sein sollten, und Lösung (b) wirft letztlich die Frage auf, warum bei der Intonationsbeschreibung dann nicht gleich auf konturbasierte Ansätze zurückgegriffen werden sollte.

Reproduzierbarkeit des Signals Hierin sind die parametrischen Modelle auf Grund ihrer größeren Signalnähe gegenüber den symbolischen klar im Vorteil. Während sich aus einer parametrischen Repräsentation durch Belegung der Parameter die F0-Kontur unmittelbar ergibt (Analyse durch Synthese), kommt eine symbolische Repräsentation nicht ohne zusätzliche Methoden zur Erlernung des Zusammenhangs zwischen Repräsentation und konkreter F0-Kontur aus.

Reproduzierbarkeit der Abstrahierung Die Reproduzierbarkeit der Abstrahierung lässt sich im Falle manueller symbolischer Etikettierung in Form des Intra- und Inter-Labeller-Agreements ausdrücken. Hier werden wie für TSM bereits angeführt teilweise sehr niedrige Werte erzielt, die nicht dafür sorgen, Bedenken an der Eignung dieser Ansätze zu zerstreuen.

Auch für die besprochenen parametrischen Ansätze ist Reproduzierbarkeit aus den folgenden Gründen nicht gewährleistet:

- Für die verwendeten Stilisierungsfunktionen lassen sich die Parameterwerte nur numerisch schätzen. Das bedeutet, es ist bei der Anpassung der Parameter nur das Auffinden eines lokalen Optimums garantiert, nicht aber die bestmögliche Belegung. Je nach Initialisierung können bei wiederholter Analyse derselben F0-Kontur unterschiedliche Ergebnisse herauskommen.

- Die Relation zwischen Parameterwerten und F0-Kontur ist nicht injektiv. Unterschiedliche Parameterbelegungen können also zur selben Kontur führen. Dies gilt insbesondere bei superpositionellen Modellen in Bezug auf unterschiedliche Aufteilungsmöglichkeiten der Kontur in globale und lokale Bestandteile.

Abbildung 7.1 macht deutlich, wie dieselbe F0-Kontur durch das PaintE-Modell (vgl. Abschnitt 3.8) in Abhängigkeit der Parameterinitialisierung unterschiedlich in sigmoidale und konstante Bestandteile zerlegt wird.

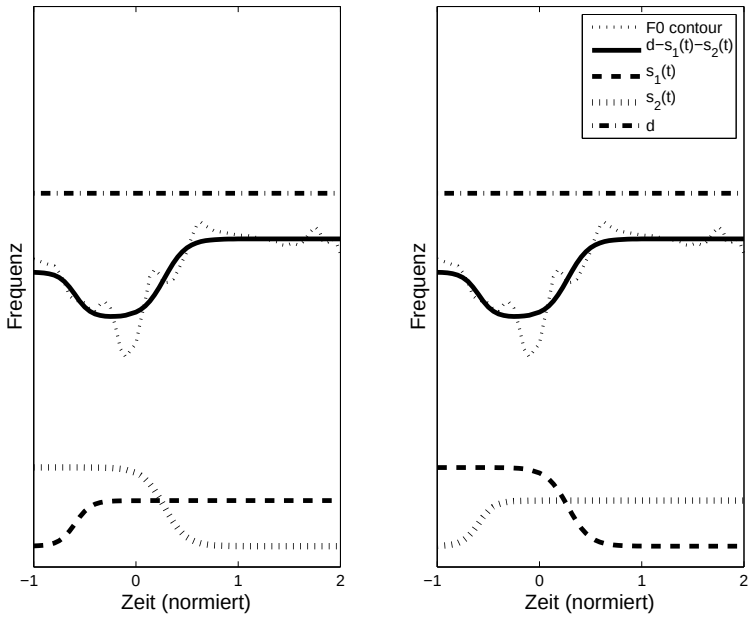

Abbildung 7.1: PaintE-Modell: fehlende Injektivität der F0-Zerlegung. Unterschiedliche Parametrisierung derselben Kontur in Abhängigkeit der Parameterinitialisierung.

Diese fehlende Reproduzierbarkeit macht insbesondere bei der linguistischen Interpretation Schwierigkeiten.

7.3 Interpretierbarkeit

Linguistische Interpretierbarkeit In der Linguistik werden bevorzugt Zusammenhänge zwischen *Symbolen* untersucht, also zwischen *kategorialen* Variablen. Hierfür werden diverse Methoden bereitgestellt, beispielsweise die Formulierung von Ersetzungsregeln $A \longrightarrow B$ zur Umwandlung der Kategorie A in die Kategorie B. In dieser methodologischen Tradition lässt sich auch eine symbolische Intonationsrepräsentation wie im TSM leichter linguistisch interpretieren als eine parametrische mit *kontinuierlichen* Variablen. Andererseits lässt sich aus dieser Tradition keineswegs ableiten, dass solche kontinuierlichen Variablen linguistisch allgemein nicht interpretierbar wären. Eine Diskussion zu dieser Thematik findet sich beispielsweise in Taylor (1995).

Experimentelle Befunde dazu, dass linguistische Phänomene neben kategorialer auch graduelle Entsprechungen in Produktion und Perzeption der Intonation aufweisen, finden sich in Baumann et al. (2006), wo die Produktion von engem und weiten Fokus sowohl kategoriale als auch graduelle Unterschiede zeigte, erstere in Form unterschiedlicher Akzenttypen (in Form von GToBI-Etiketten), letztere durch Variation von F0-Bewegung und -Alinierungen sowie Segmentdauern. Hierbei war zudem eine hohe Variabilität unter den Sprechern bei der Wahl kategorialer und gradueller Mittel zur Fokusrealisierung zu beobachten. Auf Perzeptionsseite konnten wie in Abschnitt 2.5.3 ausgeführt beispielsweise Ladd und Morton (1997) keine kategoriale Wahrnehmung von normalem gegenüber emphatischem Akzent nachweisen.

Nichtsdestotrotz konzentriert sich die Forschungsliteratur weiterhin mit großer Mehrheit auf die linguistische Verankerung symbolischer Intonationsbeschreibungen, vornehmlich zum TSM. Vergleichbare Analysen parametrischer Ansätze sind seltener (Möbius, 1993a; Mixdorff, 1998).

Stattdessen wurden oftmals Versuche unternommen, parametrische Beschreibungen wie Fujisaki, PaintE und Tilt in symbolische wie GToBI oder PROLAB zu überführen (Taylor, 2000; Mixdorff und Pfitzinger, 2009). Möhler und Conkie (1998) stellten wie bereits ausgeführt mit der Parametervektor-Quantisierung eine Methode vor, wie eine parametrische Beschreibung theoriefrei in eine symbolische übersetzt werden kann, die sich dann auf traditionell-linguistischem Wege analysieren ließe.

Neben traditionellen Gründen mag ein weiterer Grund für die mangelnde linguistische Deutung von Parametern in der fehlenden Injektivität der F0-Parametrisierung liegen. Wenn dieselbe F0-Kontur auf eine linguistische Intention, zugleich aber auf mehrere unterschiedliche Parameterbelegungen zurückgeführt werden kann, wird deren linguistische Interpretation schwierig.

Der starke Forschungsschwerpunkt auf der linguistischen Interpretation symbolischer Ansätze brachte auch mit sich, dass primär *Töne* und nicht *Konturen* Gegenstand der Analyse waren. Einen Befund gegen das Primat der Töne in der linguistischen Analyse lieferte Dainora (2002). Auf Grundlage von durch Korpusanalyse ermittelten bedingten Wahrscheinlichkeiten von Tönen stellte sie fest, dass die Wahl des Grenztons nahezu vollständig durch vorangehenden nuklearen Akzent und Phrasenton determiniert ist. Dieser Befund widerspricht der Autorin zufolge streng kompositionalen Ansätzen wie dem von

Pierrehumbert und Hirschberg (1990), in denen die Funktion der Grenztöne als unabhängig von Tonakzenten und Phrasentönen angenommen wird. Vielmehr lässt er den Schluss zu, dass anstelle einzelner Töne feste Tonverbindungen und somit letztlich Konturen Träger der linguistischen Information sind.

Phonetische Interpretierbarkeit Im Zusammenhang mit TSM ist vielerorts die streng lokale Ausrichtung der Intonationsbeschreibung kritisiert worden, die empirischen Befunden zuwiderläuft, nach denen eine gewisse Vorausplanung (*look ahead*) der Intonationskontur stattfindet, beispielsweise bei der Planung des Gefälles der Deklinationslinie in Abhängigkeit der Äußerungslänge (je kürzer die Äußerung, desto steiler; Cooper und Sorensen (1981); Thorsen (1985)). Auch perzeptive Befunde anhand von Satzvervollständigungsaufgaben, in denen wahlweise gegebene oder neue Diskursreferenten einzusetzen waren (Féry et al., 2009) zeigen, dass Hörer anhand präsentierter Intonationsabschnitte bereits Annahmen über die Form des noch ausstehenden F0-Verlaufs machen. Dies wirft Fragen auf zur Adäquatheit einer Repräsentation des TSM als endlichen Automaten, der konzeptuell keine Vorausschau erlaubt.

Auch einige parametrische Modelle laufen Gefahr, sich einer phonetischen Plausibilität zu versperren. Grund hierfür ist deren Übermächtigkeit, das heißt, neben prinzipiell möglichen können auch beliebig viele unmögliche Konturen erzeugt werden. Hinzu kommt, dass mögliche Konturen auch phonetisch unplausibel generiert werden können, so lassen sich etwa mit dem Fujisaki-Modell durch hinreichend nah aufeinanderfolgenden Akzentkommandos beliebige Konturen mit beliebiger Genauigkeit stilisieren. Ein weiteres Beispiel ist die oft gegen das Fujisaki-Modell zu Felde geführte Erzeugung einer globalen F0-Inklination innerhalb einer Intonationsphrase mit einer Kaskade von Phrasenkommandos sukzessiv wachsender Amplitude.

Physiologische Interpretierbarkeit Fast alle der hier besprochenen Modelle behandeln die Intonationskontur nicht im Kontext phonatorischer Produktionsmechanismen, mit Ausnahme zweier generisch orientierter superpositionaler Modelle von Öhman und von Fujisaki. Die in diesen Modellen postulierten Zusammenhänge zwischen Physiologie und F0-Kontur sind teilweise empirisch unmittelbar belegbar, wie der zeitliche Zusammenfall von präphonatorischer laryngaler Aktivität und Phrasenkommando bei Fujisaki (1987). Häufig aber können diese angenommenen Zusammenhänge nur indirekt über eine akzeptable Generierung der F0-Kontur sowie phonetisch-physiologische Plausibilität erschlossen werden.

7.4 Automatisierbarkeit

Grundsätzlich lässt sich feststellen: je theoriefreier das Modell, desto leichter automatisierbar. In diesem Sinne eignen sich parametrische Modelle in der Regel eher zu einer Automatisierung der F0-Analyse und -synthese als symbolische.

Alle der vorgestellten parametrischen Modelle setzen eine vorab unternommene prosodische Strukturierung der Daten voraus, mit zwei Ausnahmen: das Fujisaki-Modell und

das Tilt-Modell benötigen wie beschrieben zur F0-Stilisierung nicht unbedingt Vorwissen über die Positionen der Akzente, vor allem das Fujisaki-Modell läuft aber Gefahr, ohne dieses Vorwissen nur schlecht interpretierbare Ergebnisse zu liefern.

Im nächsten Teil wird nun das in dieser Arbeit entwickelte PKS-Intonationsmodell auch im Hinblick auf die hier diskutierten Anforderungen vorgestellt.

Teil II

Das PKS-Intonationsmodell

Überblick Inhalt dieses Teils der Arbeit die Vorstellung des hier entwickelten parametrischen konturbasierten und superpositionalen Intonationsmodells (PKS). Nach Darlegung einiger Vorüberlegungen und allgemeiner Modell-Charakteristika erfolgt eine Beschreibung der Datenvoraussetzung und -vorverarbeitungsschritte. Im Anschluss daran werden die Modellkomponenten im Detail beschrieben. Den Abschluss dieses Teils bildet die Präsentation der mathematischen und perzeptiven Modellevaluierung sowie eine Diskussion dieser Ergebnisse sowie einiger Aspekte des Modells.

Kapitel 8

Charakteristika und Architektur

8.1 Vorüberlegungen

Angesichts der in Kapitel 7 gegebenen Anforderungen

- angemessene Abstrahierung vom Signal,
- Interpretierbarkeit,
- Automatisierbarkeit

können zum Modell-Design bezüglich der in Abschnitt 3.1 vorgestellten Kriterien

- Einheiten der F0-Abstrahierung: ton- vs. konturbasiert,
- Beschreibung der Einheiten: symbolisch vs. parametrisch,
- Gewinnung der Einheiten: perzeptiv vs. objektiv-mathematisch,
- Anordnung der Einheiten: einschichtig vs. superpositional

folgende Überlegungen angestellt werden:

Automatisierbarkeit Vornehmliches Ziel dieser Arbeit ist die Entwicklung eines Intonationsmodells, das sowohl eine rein datenbasierte automatische Intonationsbeschreibung ermöglicht als auch eine automatische Generierung von F0-Konturen.

Hierfür bietet sich eine konturbasierte F0-Abstrahierung an, da sie in der Synthese keiner Zusatzregeln zur Übersetzung von Tönen in F0-Verläufe bedarf.

Ferner ist eine parametrische Beschreibung gegenüber einer symbolischen zu bevorzugen, da sie dem datenbasierten und theoriefreien Ansatz in dieser Arbeit eher entspricht.

Die Gewinnung der Einheiten sollte auf objektiv-mathematischem Wege möglich sein, damit auf manuelle Etikettierung oder Befragung von Versuchspersonen verzichtet werden kann.

Abstrahierung Die F0-Abstrahierung muss einerseits hinreichend signalnah sein, damit relevante Aspekte nicht verloren gehen. Andererseits macht es Sinn, auf Beschreibungsökonomie zu achten, um die Mitmodellierung von Rauschen ebenso wie die Übermächtigkeit eines Modells zu verhindern. Diese als Occams Messer bekannte Abwägung betrifft vor allem die Wahl der Stilisierungsfunktion zur parametrischen F0-Beschreibung. Sehr komplexe und daher mächtige F0-Stilisierungsfunktionen wie beispielsweise im Fujisaki-Modell sollen in dieser Arbeit vermieden werden, zumal diese Funktionen wie schon beschrieben darüber hinaus keine Reproduzierbarkeit der Abstrahierung garantieren.

Hinsichtlich der Anordnung von Einheiten vermag ein superpositionaler Ansatz gegenüber einem einschichtigen zur Beschreibungsökonomie beitragen dahingehend, dass er ermöglicht, globale Phänomene wie die Deklination als solche zu beschreiben, und nicht umständlicher als Aneinanderreihung mehrerer lokaler Ereignisse (wie beispielsweise Downsteps in einschichtigen Tonsequenzansätzen).

Interpretierbarkeit Auf Grund der Wahl eines weitestgehend theoriefreien Ansatzes ist die gewonnene Intonationsrepräsentation erst post hoc auf phonetische oder linguistische Interpretierbarkeit hin untersuchbar. Daher ist es wichtig, eine parametrische Beschreibungsform zu wählen, die eine solche Untersuchung ermöglicht. Es macht also Sinn, bei der Beschreibung auf Plausibilität zu achten, die sich beispielsweise in der Entscheidung für einen superpositionalen Ansatz äußern kann vor dem Hintergrund experimenteller Befunde wie von Cooper und Sorensen (1981) sowie Thorsen (1985) zur Vorausplanung von Deklinationskonturen.

Bei im obigen Abschnitt geforderten einfachen Modellen besteht die Gefahr, dass einzelne Beschreibungsparameter zu viele Aspekte der Kontur codieren, so dass eine Interpretation scheitern muss. Eine Lösung kann hier darin bestehen, eine abstrakte symbolische Zwischenebene beispielsweise durch Parameter-Clustering zu schaffen und diese dann zur weiteren Interpretation heranzuziehen.

8.2 Allgemeine Charakteristika

Das in dieser Arbeit entwickelte PKS-Intonationsmodell lässt sich nach diesen Vorüberlegungen in Bezug auf die in Abschnitt 3.1 gegebene Taxonomie folgendermaßen charakterisieren:

- Die Einheiten der F0-Abstrahierung sind **Konturen** ($P\boldsymbol{K}S$).
- Ihre Beschreibung erfolgt **parametrisch** ($\boldsymbol{P}KS$).
- Die Einheiten werden auf objektiv-mathematischem Wege gewonnen.
- Ihre Anordnung ist **superpositional** ($PK\boldsymbol{S}$).

Entsprechend zu den Kapiteln 4 und 6 soll an dieser Stelle ein kurzer Überblick gegeben werden über die Gewinnung der Intonationsrepräsentation und die Intonationsgenerierung durch das PKS-Modell.

8.2.1 Gewinnung der Intonationsrepräsentation

Die Entwicklung des Modells ist in Abbildung 8.1 skizziert.

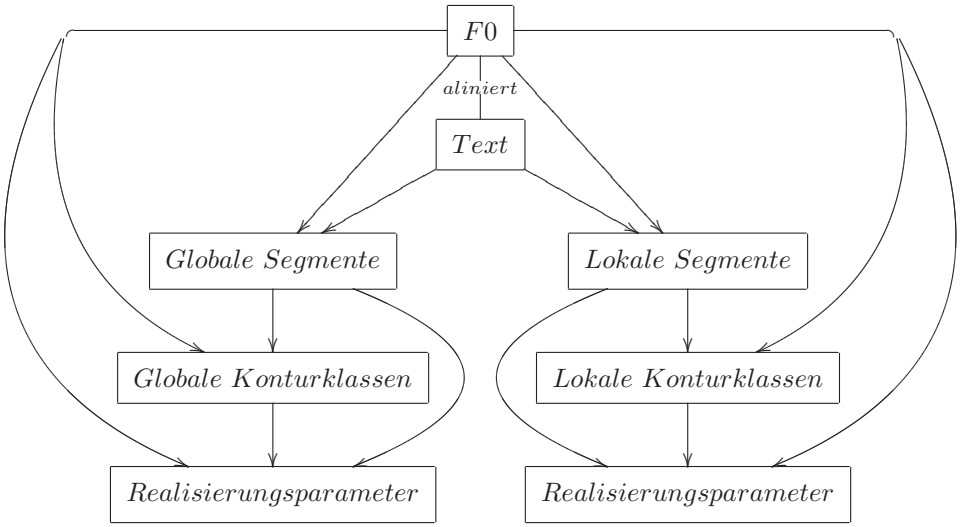

Abbildung 8.1: Entwicklung des PKS-Modells.

F0-Verläufe werden unter Zuhilfenahme von Signal und Text in globale und lokale Segmente untergliedert und in diesen Segmenten parametrisiert. Globale Segmente entsprechen hierbei Intonationsphrasen und lokale Segmente Akzentgruppen. Aus den Parametrisierungen werden im Anschluss daran phonologisch-abstrakte globale und lokale Konturklassen gewonnen. Die globalen Konturklassen repräsentieren mögliche Deklinationsverläufe, die lokalen Klassen F0-Verläufe auf akzentuierten und umliegenden nichtakzentuierten Silben. Aus dem Kontur-Vergleich zwischen den abstrakten Klassen und den in den Trainingsdaten vorliegenden F0-Verläufen ergeben sich schließlich phonetische Realisierungsparameter.

8.2.2 Intonationsgenerierung

Die Erzeugung einer Intonationskontur mittels des PKS-Modells geschieht, wie in Abbildung 8.2 skizziert, in folgenden Schritten. Auf phonologischer Ebene wird ein semantisch und diskursbezogen hinreichend analysierter Text intonatorisch in globale und lokale Segmente für Intonationsphrasen und Akzentgruppen strukturiert. Den Segmenten werden passende Konturklassen zugewiesen. Die Charakteristika der Konturklassen werden mittels phonetischer Realisierungsparameter an den Kontext angepasst, superponiert und ergeben dadurch auf akustischer Ebene den konkreten F0-Verlauf.

Es sei darauf hingewiesen, dass nicht eine vollständig ausgearbeitete textbasierte Vorhersage von prosodischer Struktur und Intonationsklassen Gegenstand dieser Arbeit ist, sondern die Schaffung einer hierfür brauchbaren Grundlage in Form der Untersuchung linguistischer Bezüge der Konturklassen (siehe Teil III).

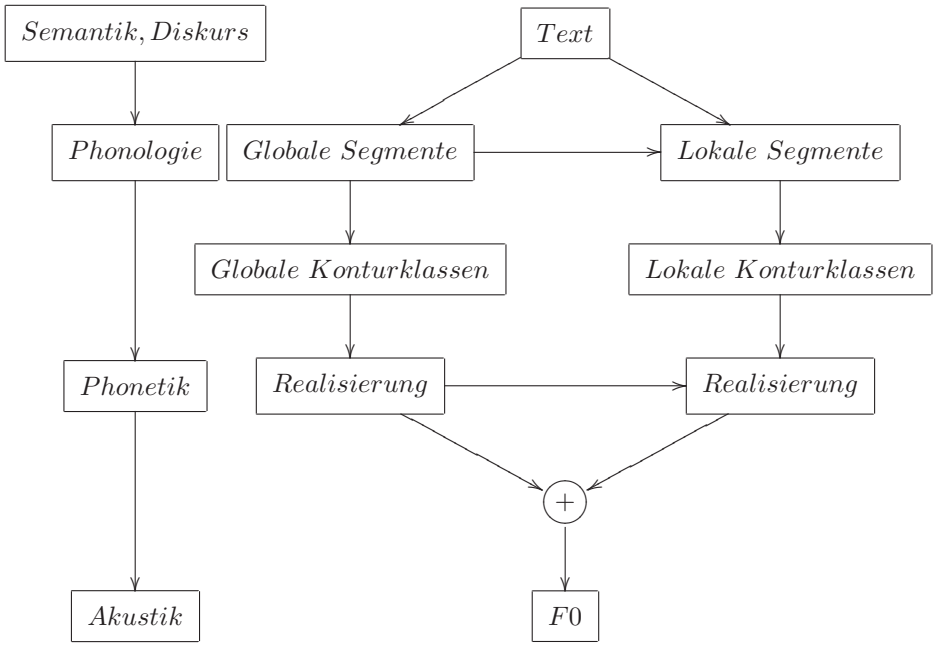

Abbildung 8.2: Architektur des PKS-Modells zur F0-Generierung.

Kapitel 9

Daten und Vorverarbeitung

9.1 Daten

Die zur Entwicklung des PKS-Modells verwendeten Daten stammen aus dem SI1000P-Korpus (Schiel et al., 1999). Das Korpus wurde 1998 am Institut für Phonetik und Sprachliche Kommunikation (heute: Institut für Phonetik und Sprachverarbeitung) in München im Auftrag der Siemens AG zu Zwecken der konkatenativen Sprachsynthese aufgenommen und besteht aus 993 Zeitungssätzen, die von zwei professionellen süddeutschen männlichen Nachrichtensprechern des Bayerischen Rundfunks vorgelesenen wurden.

Die Aufnahmen erfolgten in einem schalltoten Raum mit einem ungerichteten Sennheiser MKH20 Mikrofon, das 30 cm vom Mund positioniert wurde. Als Aufnahmegerät diente ein 4-Kanal-Datrekorder. Das Sprachsignal wurde mit 48 kHz und 16 Bit abgetastet und anschließend auf 16 kHz heruntergesampelt.

Die Annotation umfasst unter anderem die vollständige orthographische Verschriftung sowie eine automatische Segmentierung mit dem Munich *Munich Automatic Segmentation System (MAUS)* (Schiel, 1999, siehe Abschnitt 9.3.3).

Für diese Arbeit wurde nur das Material eines der beiden Sprecher (des Sprechers *AI*) herangezogen. Für eine Diskussion der Beschränkung auf nur einen Sprecher siehe Kapitel 12. Das verwendete Material umfasst insgesamt 189 Minuten 45 Sekunden und etwa 45700 realisierte Silben. Die Nachrichtensätze hängen gruppenweise thematisch zusammen, wurden aber einzeln aufgenommen.

Zur Parameteroptimierung der im folgenden Abschnitt 9.2 beschriebenen Verfahren zur Pausen- und Silbenkerndetektion diente ein 20 Sätze umfassender handsegmentierter Teil des SI1000P-Korpus, der 1011 Silben und 86 Pausen beinhaltete.

9.2 Vorverarbeitung: Überblick

Ziel Training und Anwendung des PKS-Modells stellt im Wesentlichen folgende Vorbedingungen an die Daten:

- Pausen und Silbenkerne müssen im Signal lokalisiert sein,
- der dem Signal zugrundeliegende Text muss part-of-speech-gelabelt inklusive Satzzeichenangaben vorliegen,
- Pausen und Silbenkerne müssen auf Wortebene den entsprechenden Stellen im Text zugeordnet werden können.

Ziel ist also eine Alinierung der Silbenkerne und Pausen mit Wörtern (beziehungsweise Wortzwischenräumen), worüber sich anhand der Zeitinformation eine Alinierung zwischen F0-Kontur und Text ergibt. Zur späteren Modellierung werden nur F0-Abschnitte im Bereich der extrahierten Silbenkerne herangezogen (Pfitzinger, priv. Komm.), daher kann auf eine exakte zeitliche Bestimmung von Lautsegment- oder Silbengrenzen verzichtet werden.

Zu jedem Silbenkern muss die die Wortart des ihm zugrundeliegenden Worts verfügbar sein, sowie die Angabe, ob es sich um eine wortbetonte und damit potentiell akzentuierbare Silbe handelt oder nicht.

Die Part-of-Speech-Information ist nötig zur Festlegung der prosodischen Struktur sowie zur später vorgenommenen linguistischen Interpretation der Intonation.

Vorverarbeitungsschritte Die zur Erfüllung dieser Voraussetzungen notwendigen Vorverarbeitungsschritte für Signal und Text sind in Abbildung 9.1 dargestellt und werden in den nächsten Abschnitten erläutert.

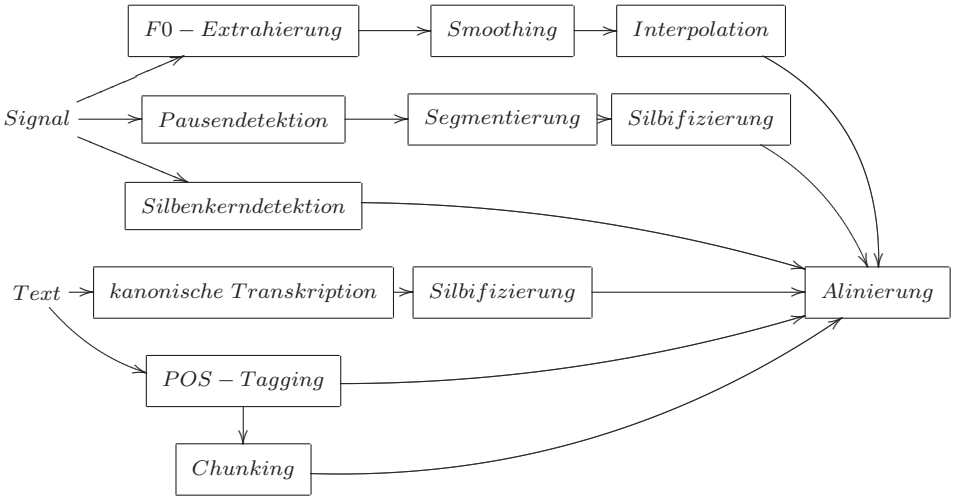

Abbildung 9.1: Flussdiagramm der Vorverarbeitungsschritte.

9.3 Signal-Vorverarbeitung

9.3.1 F0-Extrahierung und -bearbeitung

Extrahierung Die F0-Werte wurden mit einer Abtastrate von 100 Hz mittels des im über das EMU-System (Cassidy und Harrington, 1996) verfügbaren Schaefer-Vincent-Algorithmus (Schaefer-Vincent, 1983) ermittelt. Fehlerhafte Bereiche wurden automatisch anhand sprunghafter Abweichungen von der Umgebung identifiziert und für die nachfolgende Interpolation (siehe unten) auf 0 gesetzt. Es fand keine manuelle Korrektur der F0 statt.

Bearbeitung Die F0-Werte ungleich 0 wurden unter Verwendung von Gleichung 4.4 von Hertz- in Halbtonwerte überführt (mit Basis $b = 50$ Hz). Über stimmlose Signalabschnitte und detektierter Messfehler wurde linear interpoliert. Die anschließende Glättung erfolgte unter Verwendung eines in Gleichung 4.3 beschriebenen Savitzky-Golay-Filters mit Polynomordnung 3 und einer Fensterlänge von fünf Samples.

9.3.2 Pausendetektion

Zur Lokalisierung von Sprechpausen wurde ein Analysefenster w_p zusammen mit einem längeren Referenzfenster w_r mit dem selben Zeitmittelpunkt in 50 ms-Schritten über das akustische Zeitsignal geschoben. Bei hinreichend großem Energieunterschied zwischen w_p und w_r jeweils gemessen als mittlere quadratische Abweichung erfolgte eine Klassifikation des Inhalts von w_p als Pause:

$$\mathrm{RMS}(w_p) < \mathrm{RMS}(w_r) \cdot c \quad \longrightarrow \quad \text{Pause}. \tag{9.1}$$

Benachbarte Pausensegmente wurden konkateniert.

Die Parameterwerte wurden durch unbeschränkte nonlineare Optimierung mittels des Nelder-Mead-Simplex-Verfahrens (Nelder und Mead, 1965; Lagarias et al., 1998) auf dem im vorangegangenen Abschnitt 9.1 beschriebenen handsegmentierten Teilkorpus geschätzt.[1] Grob gesagt basieren solche Simplex-Verfahren darauf, im n-dimensionalen Raum (n ist hierbei die Anzahl der verwendeten Parameter) einen durch $n + 1$ Ecken aufgespannten Simplex[2] iterativ so zu modifizieren, dass die assoziierte Zielfunktion (der Fehler) lokal minimiert wird. Jede Ecke entspricht einer Parameterwertkombination. In jedem Iterationsschritt wird hierbei die Ecke mit dem höchsten Fehler nach bestimmten Verfahrensmustern durch eine neue ersetzt. Da die Nelder-Mead-Methode ohne Ableitung auskommt, zeichnet sie sich durch eine vergleichsweise hohe Robustheit bei nicht-linearen oder unstetigen Fehlerfunktionen (so wie die hier vorliegende) aus – im Gegensatz beispielsweise zu schneller konvergierenden Gradientenabstiegsverfahren.

[1] Matlab-Funktion *fminsearch*.

[2] Unter einem Simplex wird das einfachstmögliche Polytop in einem n-dimensionalen Raum verstanden.

Der hierbei über ein Entwicklungskorpus zu minimierende Gesamtfehler F wurde folgendermaßen ermittelt: das Signal wurde in n gleich lange Segmente unterteilt. Bei m fehlerhaften Segmenten, also Segmenten, in denen der Pausendetektor mindestens eine Auslassung oder einen falschen Alarm aufwies, ergab sich F als Quotient $\frac{m}{n}$. Die lokale Minimierung dieses Fehlers ergab die in Tabelle 9.1 aufgezeigte Parameterbelegung.

Länge(w_p)	0.15 s
Länge(w_r)	5 s
c	0.06

Tabelle 9.1: Nach Optimierung gewonnene Parameterwerte zur Pausendetektion. c vgl. Gleichung 9.1.

9.3.3 Lautsegmentierung

Die Segmentierung erfolgte einzeln für jeden interpausalen Signalabschnitt mit dem Hidden-Markov-Modell-basierten *Munich Automatic Segmentation System (MAUS)* (Schiel, 1999). Zur Ermittlung optimaler Werte für die Gewichtung des Phonemfolge--Wahrscheinlichkeitsmodells gegenüber dem akustischen Modell sowie für die Gewichtung von Lautelisionen wurde erneut die oben genannte nonlineare Nelder-Mead-Methode herangezogen. Zu minimieren war diesmal die Levenshtein-Distanz zwischen gegebener und vorhergesagter Phonemfolge im handsegmentierten Referenzkorpus (vgl. Abschnitt 9.1), also die minimale Anzahl nötiger Editieroperationen (Einfügung, Löschung oder Substitution), um die MAUS-Ausgabe in die Referenztranskription umzuwandeln. Die Distanz wurde mittels dynamischer Programmierung nach einem Verfahren von Wagner und Fischer (1974) berechnet.

MAUS liefert zusätzlich eine Zuordnung der Phonemfolge zu den Wörtern im Text, was die spätere Alinierung von Silbenkernen und Text stark vereinfacht.

9.3.4 Silbenkerndetektion

Zur Extrahierung des für vokalische Silbenkerne relevanten Frequenzbands wurde das Signal mit einem Butterworth-Filter zehnter Ordnung bandpassgefiltert. Dieser Filtertyp zeichnet sich dadurch aus, dass er im Durchlassbereich eine monotone Übertragungsfunktion aufweist, also keine „gewellte" Funktion, wie sie Chebychev- oder Cauer-Filter liefern, und dabei noch relativ steile Flanken an den Bandgrenzen gewährleistet – im Gegensatz beispielsweise zu ebenfalls monotonen Bessel-Filtern. Die gewählte hohe Ordnung trägt ebenfalls zur Ermöglichung steiler Flanken bei.

Im Anschluss an die Filterung erfolgte die Silbenkerndetektion: Hierzu wurden analog zur Pausendetektion ein Kurzzeitanalysefenster w_n und ein längeres Referenzfenster w_r parallel zueinander mit einer Schrittweite von 50 ms über das gefilterte Signal verschoben. Überstieg der RMS-Wert in w_n einen relativ zum gefundenen RMS-Maximum definierten

Schwellwert und war er gegenüber dem in w_r gefundenen Wert hinreichend größer, wurde w_n als Silbenkernbereich klassifiziert:

$$\mathrm{RMS}(w_n) > c_1 \cdot \max(\mathrm{RMS}) \wedge \mathrm{RMS}(w_n) > \mathrm{RMS}(w_r) \cdot c_2 \quad \longrightarrow \quad \text{Silbenkern.} \tag{9.2}$$

In Sequenzen zeitlich überlappender Silbenkernbereiche wurden nur diejenigen beibehalten, die ein lokales RMS-Maximum aufwiesen. Auf diese Weise konnte einer in der Regel fehlerhaften zu dichten Aufeinanderfolge von Silbenkernen entgegengewirkt werden. Silbenkerne wurden schließlich den absoluten Amplitudenmaxima innerhalb der verbleibenden Silbenkernbereiche zugeordnet.

Wie bei der Pausendetektion wurden die Parameterwerte mittels des Simplex-Verfahrens auf dem handsegmentierten SI1000P-Teilkorpus (vgl. Abschnitt 9.1) optimiert. Als zu minimierender Fehler wurde der mittlere zeitliche Abstand zwischen detektierten Silbenkernen und den alinierten Referenzmarken eines kleinen handsegmentierten Entwicklungskorpus (1000 Silben) herangezogen. Zur Vermeidung der Belohnung von Auslassungen beinhaltete die Alinierung die Zuordnung aller Referenzmarken zu den detektierten Kernen, also gegebenenfalls die gleichzeitige Zuordnung eines detektierten Kerns zu mehreren Referenzmarken. Die Optimierung ergab die in Tabelle 9.2 aufgeführte Parameterbelegung.

Frequenzband	245 – 3215 Hz
Länge(w_n)	90 ms
Länge(w_r)	250 ms
c_1	0.15
c_2	1.2

Tabelle 9.2: Parameterwerte zur Silbenkerndetektion. c_1, c_2 vgl. Gleichung 9.2.

Die durchschnittliche Abweichung zwischen Detektion und Referenzmarke betrug bei 1000 Referenzmarken 60 ms (45 ms ohne Berücksichtigung der Auslassungen).

9.4 Text-Vorverarbeitung

9.4.1 Part-of-Speech-Tagging

Zur Zuweisung der Wortarten kam ein in Reichel (2005a) entwickelter POS-Tagger zum Einsatz. Dieser Markov-Tagger berücksichtigt Kontext-Informationen zur statistischen POS-Disambiguierung (beispielsweise *Sucht* als Substantiv oder Verb) und zieht zur Behandlung von *Out-of-Vocabulary*-Fällen automatisch segmentierte Wortsuffixe heran, die im Deutschen häufig Informationen zur Wortart tragen. Die der Wortfolge $W = w_1 \ldots w_n$ am wahrscheinlichsten zugrundeliegende Tag-Sequenz $\hat{T}$ ergibt sich durch folgende Maximierung:

$$\hat{T} = \arg\max_{t_1 \ldots t_n} \Big[\prod_{i=1}^{n} \frac{1}{P(t_i)} \sum_j u_j P(t_i|\text{t-history}_{ij}) \sum_k v_k P(t_i|\text{w-representation}_{ik}) \Big] \quad (9.3)$$

t-history$_{ij}$ ist hierbei die POS-Vorgeschichte der Länge j zum Wort w_i und *w-representation*$_{ik}$ die k-te Repräsentation des Worts w_i in Form eines Suffix-Strings, der durch eine *Successor-Variety*-geleitete Wortsegmentierung (Nascimento und da Cunha, 1998) gewonnen wurde. Die Wahrscheinlichkeitsverteilungen wurden mittels Good-Turing (Good, 1953) geglättet und die Interpolationsgewichte u_j und v_k mit Hilfe des *Expectation-Maximisation*-Algorithmus (Dempster et al., 1977) geschätzt.

9.4.2 Chunking

Unter Chunking wird hier wie bei Abney (1991) eine prosodisch motivierte flache syntaktische Strukturierung einer Wortfolge verstanden. In Anlehnung an die von Abney gegebene Chunk-Definition sowie die ϕ-Phrasen nach Gee und Grosjean (1983) und Bachenko und Fitzpatrick (1990) (vgl. Abschnitt 5.2) sei hier ein Chunk definiert als Inhaltswort mit allen vorangehenden Funktionswörtern, wobei Chunks Satzzeichen und Sprechpausen nicht überschreiten dürfen. Sie dienen der prosodischen Gliederung der Daten in lokale Segmente (siehe Abschnitt 10.1) Die Chunk-Grammatik ist in Abbildung 9.2 zu sehen.

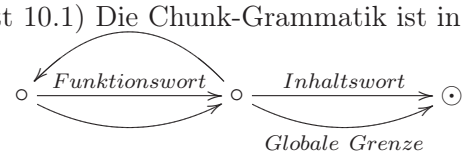

Abbildung 9.2: Finite-State-Grammatik für syntaktische Chunks.

Diese Chunk-Definition folgt zum einen den in Gee und Grosjean (1983) ermittelten prosodischen Phrasierungseinheiten und beruht zum anderen auf der Beobachtung, dass die verwendeten Daten äußerst wenig akzentuierte Funktionswörter enthalten. Letzteres gewährleistet, dass innerhalb eines Chunks in der Regel maximal ein Akzent möglich ist, was die Zuordnung des Texts zu Akzentgruppen sowie die anschließende Konturstilisierung vereinfacht. Die strikte Trennung von Inhalts- und Funktionswörtern als Träger lexikalischer gegenüber grammatikalischer Information ist genaugenommen nicht aufrechtzuhalten (Bußmann, 1990) und ist daher eher heuristischer Natur. Als Inhaltswörter wurden in dieser Arbeit Wörter mit den folgenden Wortarten festgelegt: *Substantive, Vollverben, Adjektive, Numeralia.* Alle anderen Wörter wurden als Funktionswörter klassifiziert.

9.4.3 Kanonische Transkription

Zur späteren Lokalisierung der wortbetonten Silben wurde der orthographische Text mittels maschineller Graphem-Phonem-Konvertierung (Reichel, 2005b; Reichel und Schiel,

2005) in eine kanonische Transkription mit enthaltener Wortbetonung überführt. Die Konvertierung beruht auf einem C4.5-Entscheidungsbaum (Quinlan, 1993), der anhand automatisch extrahierbarer orthographischer, morphologischer und POS-Features trainiert wurde.

9.4.4 Silbifizierung

MAUS-Segmentierung und kanonische Transkription wurden jeweils mit Hilfe eines zweistufigen automatischen Verfahrens in Silben segmentiert (Reichel, 2005b; Reichel und Pfitzinger, 2006). Hierbei werden die Silbengrenzen zunächst vor Sonoritätsminima in der Phonemkette gesetzt und anschließend anhand der in Kohler (1995b) spezifizierten und hier um spontansprachliche Phänomene erweiterten Silbenphonotaktik feinjustiert.

9.5 Alinierung

Abbildung 9.3 zeigt beispielhaft die vorzunehmenden Alinierungen zwischen den Signal- und Textebenen.

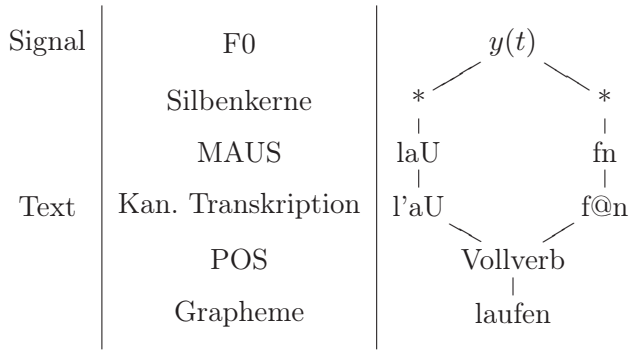

Abbildung 9.3: Alinierung der Signal- und Textebenen.

Alinierung Silbenkerne/F0–Text

Zur angestrebten Alinierung der in den Silbenkernregionen befindlichen F0-Abschnitte mit dem Text mussten die im Signal detektierten Silbenkerne nur noch mit der silbifizierten MAUS-Transkription aliniert werden, da wie beschrieben bereits eine Verknüpfung zwischen der MAUS-Transkription und dem Text auf Wortebene bestand.

Die MAUS-Silben wurden hierzu zeitlich mit den detektierten Silbenkernen abgeglichen. Falls nötig wurde hierbei ein Ausgleich zwischen mehrfachen Silbenkernen in einer Silbe mit fehlenden Silbenkernen in benachbarten Silben unternommen, wobei im Falle solcher Verschiebungen die Zeitinformation der gegenüber MAUS robusteren Silbenkerndetektion übernommen wurde.

Alinierung Silbenkerne–Wortbetonung

Die Verknüpfung der Silbenkerne mit in der silbifizierten kanonischen Transkription vorliegenden Wortbetonungsangaben fand unter Vermittlung der MAUS-Segmentierung statt. Hierzu wurden unter Verwendung der Levenshtein-Distanz und von Heuristiken wortweise die kanonische und die MAUS-Silbenfolge aufeinander abgebildet und die Wortbetonungsinformation über die bereits bestehende Maus-Silbenkern-Alinierung an die Silbenkerne durchgereicht.

9.6 Evaluierung

Zur Evaluierung der nun im Einzelnen beschriebenen Verarbeitungsschritte sei auf die an den entsprechenden Stellen angegebenen Referenzen verwiesen. Die hier vorgenommene Evaluierung beschränkte sich auf die Detektion der Pausen und Silbenkerne im Signal, da diese entscheidend ist für die Wahl der zur Stilisierung heranzuziehenden F0-Abschnitte sowie für die F0-Text-Alinierung. Hierzu fand das in Abschnitt 9.1 angeführte 1011 Silben und 86 Pausen umfassende SI1000P-Teilkorpus als Referenz Verwendung, das bereits zur Entwicklung des Pausen- und Silbenkerndetektors herangezogen wurde. Auf Grund seines geringen Umfangs musste auf eine Unterteilung in Trainings- und Testpartition verzichtet werden. Die aus diesem Grunde nur bedingt aussagekräftigen Resultate sind in Tabelle 9.3 festgehalten. Ein detektierter Silbenkern galt als Treffer, wenn die Marke sich innerhalb eines Silbenkernsegments in der Handsegmentierung befand. Eine detektierte Pause wurde als Treffer bewertet, wenn sie sich zeitlich mit einer handsegmentierten Pause überlappte. Zur Ermittlung des Fehlers wurden (echte) Einfügungen, (echte) Löschungen sowie Verschiebungen gezählt. Verschiebungen ersetzten hierbei unmittelbar benachbarte Einfügungen und Löschungen. Der Fehler ergab sich dann durch die Gesamtzahl an Einfügungen, Löschungen und Verschiebungen dividiert durch die Anzahl der Referenzmarken.

	Pausendetektion	Silbenkerndetektion
Anzahl der Referenzmarken	86	1011
Einfügungen	2	23
Löschungen	6	12
Verschiebungen	1	36
Gesamtfehler	10.47 %	7.02 %

Tabelle 9.3: Evaluierung der Pausen- und Silbenkerndetektion.

Kapitel 10

Modellentwicklung und -anwendung

10.1 Prosodische Struktur

Im Sinne der Superpositionalität des Modells wird eine Segmentierung der Daten in globale und lokale Segmente vorgenommen.

Globale Segmente

Globale Segmente, so wie sie in dieser Arbeit signal- und textbasiert extrahiert wurden, haben eine ungefähre Entsprechung zu Intonationsphrasen. Segmentgrenzen wurden gesetzt an Sprechpausen und Interpunktion. Diskontinuitäten im F0-Verlauf konnten nicht sinnvoll als Grenzsignal genutzt werden, da sich Pitch Resets (mit vorausgehender Grenze) nicht verlässlich von Grenztönen (mit folgender Grenze) und akzentuierten Silben trennen ließen.

Lokale Segmente

Lokale Segmente basieren auf dem in Abschnitt 9.4.2 vorgestellten Chunker und sind somit rein syntaktisch definiert als Sequenz von Funktionswörtern mit abschließendem Inhaltswort, beziehungsweise abschließender globaler Segmentgrenze. Diese syntaktische Definition stellt für das verwendete Datenmaterial weitgehend sicher, dass sich in einem lokalen Segment nur maximal ein Akzent befindet, es also in etwa mit einer Akzentgruppe gleichgesetzt werden kann, worauf in der Diskussion in Kapitel 12 noch genauer einzugehen ist. Zudem erleichtert die Beschränkung auf maximal ein Inhaltswort pro Segment die spätere linguistische Analyse lokaler Konturen. Das segmentfinale Inhaltswort wird im Folgenden auch als **Kernwort** bezeichnet.

Hierarchie

Die Strukturierung in globale und lokale Segmente gehorcht der Exhaustivitätsforderung der *Strict-Layer-Hypothese*, was bedeutet, dass jedes lokale Segment komplett von einem globalen Segment dominiert wird. Hier ein illustratives Beispiel:

$$\Big[[\text{viele Grüße}]_l\ [\text{aus dem winterlichen}]_l\ [\text{Alpenvorland}]_l\ \Big]_g,$$
$$\Big[\ [\text{wünschen}]_l\ [\text{meine Kollegen}]_l\ [\text{und ich}]_l\Big]_g.$$

Abbildung 10.1: Prosodische Strukturierung: Segmentierung in globale $[\ldots]_g$ und lokale $[\ldots]_l$ Segmente.

Akzentuierung

Da lokale Segmente wie eben beschrieben rein syntaktisch und nicht mittels vorausgehender Detektion akzentuierter Silben gewonnen werden, beschränkt sich die prosodische Strukturierung auf die Lokalisierung von Phrasengrenzen, was die Datenvorverarbeitung entscheidend vereinfacht. Die Akzentlokalisierung geschieht erst indirekt bei der Gewinnung der Konturklassen, wie in Abschnitt 10.3 noch beschrieben wird.

10.2 Parametrisierung

10.2.1 Vorüberlegungen

Stilisierungsfunktion In Abschnitt 7.2 wurde die Problematik der fehlenden eineindeutigen Beziehung zwischen Parametrisierung und zugrundeliegender Kontur diskutiert. Zur Umgehung dieses Problems wurden in dieser Arbeit anstelle von Funktionen hoher Komplexität einfach Polynome zur Stilisierung herangezogen. Polynome unterschiedlicher Ordnung bilden eine *Basis* B im Sinne der linearen Algebra, also eine linear unabhängige Teilmenge eines Vektorraums V, die unter anderem die folgende nützliche Eigenschaft aufweist:

> Jedes Element von V lässt sich als Linearkombination von Vektoren aus B *eindeutig* darstellen.

Die Darstellung von F0-Konturen in Form von Polynomen n-ter Ordnung, die als Linearkombination der Polynome der Ordnung 1 bis $n-1$ aufzufassen sind, gewährleistet also eine stabile Abstrahierung vom Signal, eine Eigenschaft, die keine der im Theorieteil aufgeführten parametrischen Intonationsmodellen aufweist. Entsprechend ist die polynomiale Kontur-Approximation analytisch und nicht wie bei den gegebenen Modellen nur numerisch zu erreichen, es wird hier also die global beste Anpassung erzielt.

Alternativ zur polynomialen Stilisierung kämen auch andere Parametrisierungen in Betracht, die dieselben Vorzüge aufweisen. Hierzu zählen beispielsweise die digitale Cosinustransformation, da auch die Cosinusschwingungen unterschiedlicher Frequenz eine Basis bilden, oder die Zerlegung mittels Legendre-Polynomen. Da aber die zugehörigen Basisfunktionen in ihrer Form den Polynomen entsprechender Ordnung sehr stark ähneln, ist keine bedeutende Änderung der Anpassungsgüte zu erwarten. Hierzu passt auch die Beobachtung einer hohen Korrelation zwischen den Werten von Polynom- und DCT-Koeffizienten bei der Stilisierung derselben Daten (Harrington, priv. Komm.).

Die polynomiale Approximation der Ordnung n einer F0-Kontur $y(t)$ ist folgendermaßen gegeben:

$$y(t) \quad = \quad \sum_{i=0}^{n} s_i \cdot t^i \tag{10.1}$$

Die Koeffizienten s_0 bis s_n werden mittels der Methode der kleinsten Quadrate analytisch ermittelt.

Bestimmung der Ordnung Zur Festlegung der Polynomordnung muss zum einen eine hinreichende Präzision der Stilisierung berücksichtigt werden, und zum anderen eine hinreichende Robustheit gegenüber Rauschen. Die Abwägung zwischen hoher Ordnung für die Präzision und niedriger Ordnung für die Robustheit ist bei der Stilisierung von Deklinationsgrundlinien in globalen Segmenten mit der Wahl der ersten Ordnung (also einer linearen Stilisierung) vergleichsweise leicht zu treffen, bedarf aber gewisser Überlegungen im Kontext lokaler Segmente, wie in den nachfolgenden Abschnitten dargelegt wird.

10.2.2 Globale Segmente

Ziel der Parametrisierung globaler F0-Segmente ist die Extrahierung und lineare Stilisierung des Deklinationsverlaufs. Hierzu wird der Verlauf aus der Kontur zunächst extrahiert, dann stilisiert und schließlich von der F0-Kontur abgezogen.

Stilisierung

Zur Stilisierung der Deklinationsgrundlinie wird für jede im globalen Segment enthaltener Silbe ein F0-Basiswert ermittelt. Dafür wird die F0-Kontur in einem Zeitfenster von 110 ms Länge um den Silbenkern herangezogen und daraus der Median der F0-Werte kleiner gleich dem zehnten Perzentil berechnet. Die Verwendung von Medianen schmälert die Anfälligkeit gegenüber fehlerbedingten F0-Ausreißern. Die Stilisierung der Baseline erfolgt wie in Abbildung 10.2 zu sehen durch Anpassung einer Geraden als flachstmögliche untere Tangente der Medianwertfolge m, die genau zwei Punkte aus m berührt. Die Tangente wird aus der Menge aller möglichen linearen Verbindungen von Paaren lokaler m-Minima bestimmt. Scheiternde Suchen nach einer unteren Tangente in dieser Menge linearer Verbindungen werden abgefangen, indem eine Regressionsgerade durch die lokalen m-Minima gelegt und solange parallel nach unten verschoben wird, bis sie m nicht mehr schneidet.

Zur Beseitigung des Einflusses der Segmentlänge in Form der Silbenzahl wird vor der Bestimmung der Tangentensteigung eine Zeitnormalisierung auf das Intervall [0 1] durchgeführt.

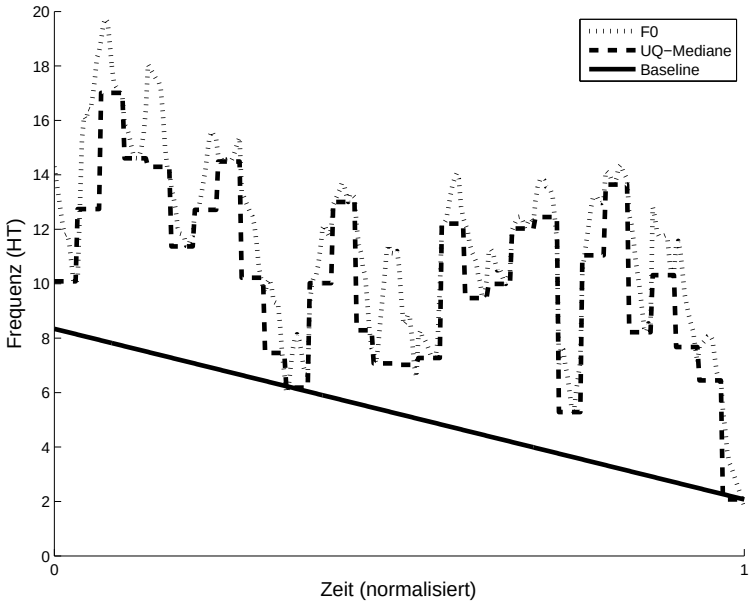

Abbildung 10.2: Stilisierung der Deklinationsgrundlinie als untere Tangente durch eine Medianfolge von silbenbezogenen unteren F0-Quartilen(UQ).

Residuum-Bildung

Zur weiteren Analyse der die Deklinationslinie überlagernden lokalen F0-Bewegungen wird durch Subtraktion der stilisierten Grundlinie $\mathrm{bl}(t)$ vom F0-Verlauf ein F0-Residuum $r(t)$ gebildet.

$$r(t) \quad = \quad y(t) - \mathrm{bl}(t) \tag{10.2}$$

10.2.3 Lokale Segmente

Grundlage für die Stilisierung der F0 über ein lokales Segment ist wie in Abbildung 10.3 zu sehen das im vorangegangenen Schritt gewonnene F0-Residuum $r(t)$ in 110 ms Fenstern um die detektierten Silbenkerne. Die Vorteile dieser Fensterung bestehen darin, dass

- auf eine Detektion der Silbengrenzen verzichtet werden kann, und
- nur die entscheidenden F0-Abschnitte im Bereich der Silbenkerne in die F0-Stilisierung mit eingehen. Dies erspart eine vorangehende Kontur-Gewichtung, wie sie anderswo mehr oder weniger arbiträr beispielsweise in Abhängigkeit der Intensität (Hermes, 1998) durchgeführt wird.

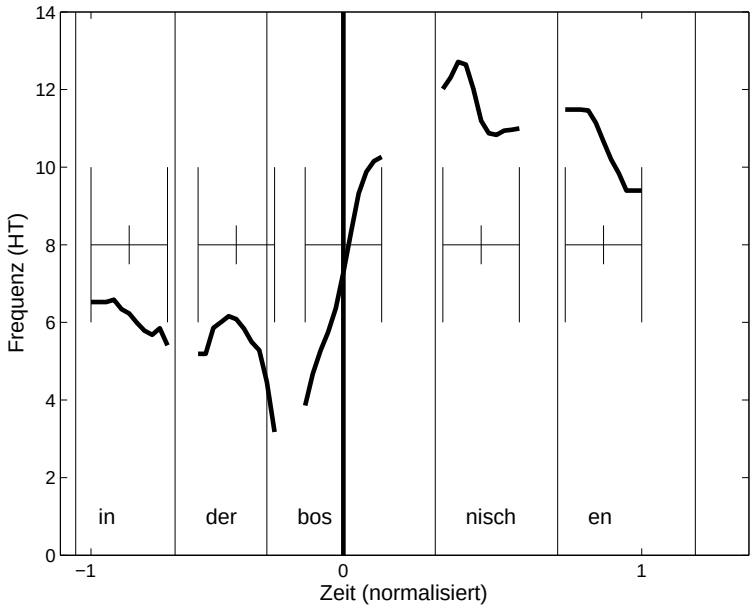

Abbildung 10.3: Zur Stilisierung herangezogene F0-Kontur-Abschnitte in Fenstern um die Silbenkerne in einem lokalen Segment. Zeitnormalisierung auf das Intervall [−1 1] mit Platzierung der 0 auf dem Kern der potentiell akzenttragenden Silbe.

Zeitnormalisierung

Analog zur Stilisierung globaler F0-Konturen wurde auch innerhalb lokaler Segmente eine Zeitnormalisierung vorgenommen (vgl. Abbildung 10.3). Zeitwerte werden auf das Intervall [−1 1] abgebildet, wobei das Intervall durch den Beginn des ersten und das Ende des letzten Silbenkernfensters begrenzt ist und der Ursprung 0 dem Kern der wortbetonten Silbe des Kernworts zugeordnet wird. Die zeitliche Normierung auf ein festes Intervall dient zum einen zur Vergleichbarmachung der Segmente durch Abstrahierung von Segmentlängenunterschieden und durch konstante Positionierung des Akzents auf den Ursprung. Dadurch wird der Unabhängigkeit von Intonationsmustern von konkreten Silbenanzahlen Rechnung getragen. Weiter ermöglicht die Normalisierung, dass instabiles Verhalten der Stilisierungspolynome außerhalb des einmal gewählten Definitionsbereichs ignoriert werden kann.

Stilisierung

Funktion Zur Stilisierung der F0-Residualkontur im lokalen Segment wurde ein Polynom dritter Ordnung herangezogen. Die hier getroffene Annahme, dass diese Ordnung ausreichend hoch ist, erscheint plausibel, da mit ihr wie in Abbildung 10.4 zu sehen ist, ein Tal und ein Gipfel modelliert werden kann, womit lokale Kontursegmente mit maximal einem enthaltenen Akzent hinreichend gut angenähert werden können.

Polynome zweiter Ordnung erwiesen sich dagegen zur Modellierung der beobachteten

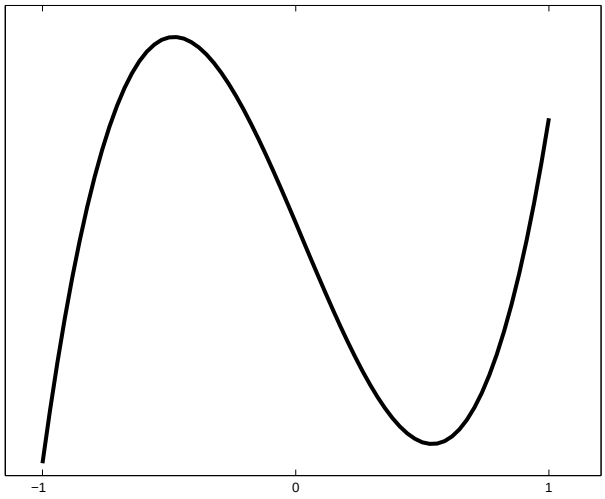

Abbildung 10.4: Beispiel eines Polynoms dritter Ordnung.

Konturen als nicht mächtig genug. Gegen höhere Ordnungen, wie sie zur Stilisierung mehrgipfliger Konturen nötig wären, spricht neben der überflüssigen Komplexität auch ihre sich mit steigender Ordnung zunehmend verschlechternde Konditionierung, das heißt ihre wachsende Anfälligkeit gegenüber unsystematischen Störungen.

Beitrag der Funktionsparameter Abbildung 10.5 zeigt den Beitrag der Koeffizienten s_j auf die Gestalt der durch das Polynom $\sum_{j=0}^{3} s_j t^j$ gegebenen Konturen. s_0 bestimmt die Höhe der F0 zum Zeitpunkt 0, also auf der akzentuierten Silbe. s_1 bestimmt die allgemeine Steigung, positive Werte bewirken einen Anstieg, negative einen Abfall der Kontur. Mit s_2 wird die Modellierung von F0-Gipfeln (negative Werte) und -tälern (positive Werte) gesteuert. Je größer $|s_2|$, desto ausgeprägter sind Gipfel oder Tal. s_3 steuert die Steigung im vorderen und hinteren Bereich der Funktion, also grob über den prä- und postakzentuierten Silben. Positive Werte führen hier zu einem F0-Anstieg, negative zu einem Abfall. Je größer $|s_3|$, desto ausgeprägter sind die Steigungen. In einer ersten Annäherung ließe sich der Koeffizient s_2 primär mit **Prominenz** in Verbindung bringen, da er die Ausgeprägtheit des Gipfels steuert, Koeffizient s_3 mit **Progredienz vs. Finalität**, da er für den F0-Verlauf im postakzentuierten Bereich zuständig ist, und Koeffizienten s_0 und s_1 mit beidem. s_0 kann über das allgemeine F0-Niveau sowohl Prominenz als auch die Wahl eines hohen oder tiefen Grenztons steuern, s_1 anhand des Vorzeichens einen progredienten gegenüber finalen F0-Verlauf sowie über das Ausmaß der Steigung die Prominenz.

Ein konkretes Stilisierungsbeispiel eines lokalen F0-Segments mittels der gewählten polynomialen Stilisierungsfunktion findet sich in Abbildung 10.6.

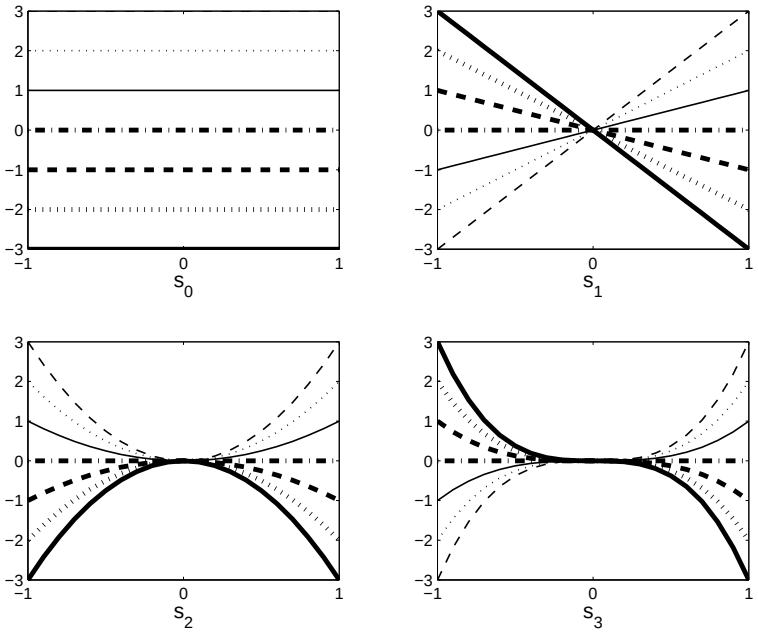

Abbildung 10.5: Auswirkungen der getrennten Variation (von -3 bis $+3$) der Polynomkoeffizienten s_0 bis s_3 in $\sum_{j=0}^{3} s_j t^j$ bei Nullsetzung der anderen Koeffizienten.

10.3 Klassifizierung der Konturen

Zur Gewinnung einer phonologischen Repräsentation der F0-Konturen werden die im vorangegangenen Stilisierungsschritt erhaltenen Polynomkoeffizientenvektoren für globale und lokale Segmente anhand ihrer Ähnlichkeit jeweils in diskrete Klassen eingeteilt.

Um zu verhindern, dass die Koeffizienten auf Grund verschiedener Wertebereiche mit unterschiedlichem Gewicht in die Ähnlichkeitsberechnung eingehen, wurden sie jeweils auf das Intervall [0 1] normiert.

10.3.1 Initiale Ermittlung der Clusterzentren

Beim Clustern ist in der Regel die optimale Anzahl von Klassen nicht per se gegeben, so auch im vorliegenden Fall. Eine Möglichkeit der Ermittlung einer geeigneten Anzahl von Klassen bietet das Subtraktive Clustern (Chiu, 1994).

Methode Beim Subtraktiven Clustern werden Clusterzentren anhand der sogenannten *Nachbarndichte* eines Punkts gewonnen. Die Nachbarndichte D_i für Punkt x_i hängt von den Abständen der benachbarten Punkte x_j innerhalb eines Umkreises mit Radius r_a ab:

$$D_i = \sum_j e^{-\frac{||x_i - x_j||}{(r_a/2)^2}} \tag{10.3}$$

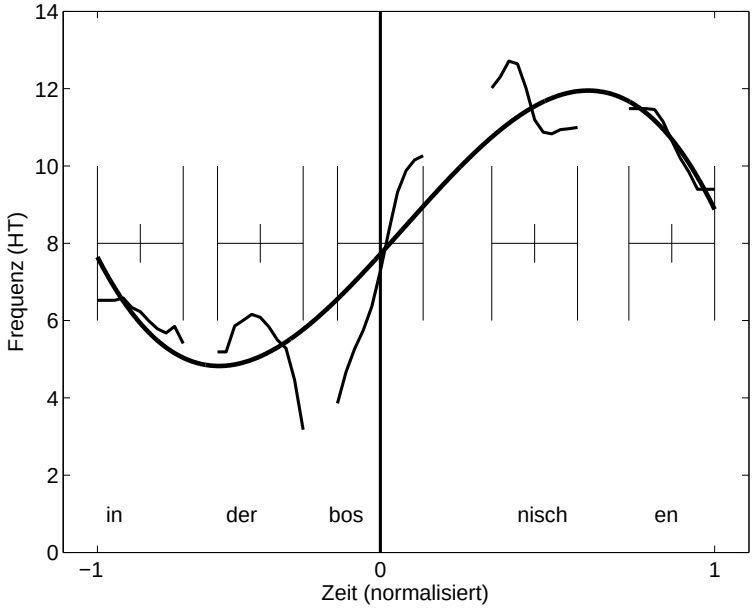

Abbildung 10.6: Polynomiale Approximation 3. Ordnung der in Abbildung 10.3 gezeigten Kontur.

D_i nimmt umso höhere Werte an, je größer die Anzahl der x_j ist, die sich innerhalb des durch r_a festgelegten Umkreises befinden, und je kleiner die Distanzen zwischen x_i und diesen x_j sind. Der Punkt x_{cz} mit der höchsten Dichte D_{cz} wird als Clusterzentrum gewählt und entfernt. Die D_k aller in einem Umkreis mit Radius r_b verbleibenden Punkte x_k werden folgendermaßen neu berechnet:

$$D_k \quad = \quad D_k - D_{cz} \cdot e^{-\frac{||x_k - x_{cz}||}{(r_b/2)^2}} \tag{10.4}$$

Dieses Update führt dazu, dass die Dichte eines Punktes x_k umso mehr reduziert wird, je näher er am gerade ermittelten Clusterzentrum liegt. Dadurch wird verhindert, dass Zentren zu nah beieinander liegen. Durch iteratives Anwenden der Gleichungen 10.3 und 10.4 werden solange Clusterzentren erzeugt, bis ein Abbruchkriterium erfüllt ist, in diesem Falle: $D^j_{cz}/D^1_{cz} < c$, d. h. das im Iterationsschritt j gefundene Dichtemaximum ist gegenüber dem zu Beginn gefundenen klein.

Parameteroptimierung Die Parameterwerte zum Subtraktiven Clustern wurden mittels des in Abschnitt 9.3.2 skizzierten Simplex-Verfahrens optimiert. Hierzu wurde ein 20 % der Daten umfassendes Teilkorpus herangezogen. Der zu minimierende Fehler e wurde aus der Silhouette abgeleitet, die aus dem Clustern der Daten-Stichprobe nach Festlegung der Clusterzentren durch obiges Verfahren resultiert. Unter der Silhouette $S(i)$ ist eine Funktion zu verstehen, die für jeden Punkt i in Cluster A misst, wie ähnlich er den Punkten in A ist verglichen zu seiner Ähnlichkeit mit den Punkten der restlichen Cluster:

$$S(i) \quad = \quad \frac{d_B(i) - d_A(i)}{\max(d_A(i), d_B(i))} \tag{10.5}$$

$d_A(i)$ steht für die mittlere Distanz zwischen i und allen Punkten desselben Clusters A, $d_B(i)$ für die mittlere Distanz zwischen i und den Punkten des i-ähnlichsten Clusters $B \neq A$. Als Distanz d wurde hier der quadrierte Euklidische Abstand herangezogen:

$$d(v, w) \quad = \quad \sum_j (v(j) - w(j))^2 \tag{10.6}$$

Die Silhouette eines Punkts i nimmt Werte zwischen -1 und 1 an. Liegt sie nahe 1, bedeutet das eine gute Zuordnung von i zu seinem Cluster ($d_A \ll d_B$), ein Wert nahe 0 zeigt, dass i nicht eindeutig einem der gegebenen Cluster zuzuordnen ist ($d_A \approx d_B$). Liegt der Silhouettenwert in der Nähe von -1, so ist i sehr wahrscheinlich dem falschen Cluster zugeordnet worden ($d_A \gg d_B$).

Der zu minimierende Fehler e ist nun wie folgt definiert:

$$e \quad = \quad 1 - \frac{\text{mean}(S) - 1}{2}, \tag{10.7}$$

ist also gleich 1 minus dem Mittelwert der auf den Bereich [0 1] abgebildeten Silhouette und nimmt Werte zwischen 0 und 1 an. Tabelle 10.1 zeigt auf diese Weise optimierten Parameterwerte.

Parameter	Wert
$r_a = r_b$	0.375
c	0.150
e	0.301

Tabelle 10.1: Optimierte Parameterwerte für die initiale Clusterzentrenermittlung. Die Radien beziehen sich auf in jeder Dimension auf [0 1] normierte Werte.

10.3.2 Konturklassen

Vorverarbeitung Vor dem Clustern wurden Ausreißer aus den Daten entfernt und die verbleibenden Vektoren in jeder Dimension bezogen auf gefundenes Minimum und Maximum auf das Intervall [0 1] normiert. Die Ausreißer, gekennzeichnet durch die Abweichung um den doppelten Interquartilsabstand vom 25. und 75. Perzentil nach unten beziehungsweise oben in mindestens einer Dimension, wurden im Anschluss an das Clustern der jeweils ähnlichsten Klasse zugewiesen.

Methode Zum Clustern wurde die *Kmeans*-Methode herangezogen, ein iteratives Verfahren für hartes und flaches Clustern: es weist jedes zu clusternde Objekt genau einer Klasse zu (hart) und strukturiert die Klassen nicht hierarchisch (flach). Nach der initialen Festlegung der k Clusterzentren nach dem oben beschriebenen Verfahren schreibt *Kmeans* im ersten Iterationsschritt sukzessive jeden Vektor der ähnlichsten Klasse zu. Zur Ermittlung der Ähnlichkeit bedarf es a) eines Abstandsmaßes (siehe unten) und b) einer Clusterrepräsentation, die bei *Kmeans* in Form des Zentroids (des Mittelwert-Vektors aller zur selben Klasse gehörigen Vektoren) gegeben ist. Nach Zuweisung eines Vektors zur ähnlichsten Klasse wird der Klassen-Zentroid entsprechend aktualisiert.

In den folgenden Iterationsschritten wird für jedes Objekt geprüft, ob es noch der ihm ähnlichsten Klasse angehört, was im Zuge der fortlaufenden Clusteraktualisierung nicht mehr zwingend der Fall sein muss. Falls nicht, wird es realloziert, also aus der aktuellen in die ähnlichste Klasse überführt. Die Iteration endet, sobald alle Cluster stabil sind, also kein Vektor mehr einem neuen Cluster zugeordnet werden muss.

Abstandsmaß Als Abstandsmaß wurde die quadrierte Euklidische Distanz zwischen den Polynomkoeffizientenvektoren herangezogen. Nachteilig ist hier festzuhalten, dass diese Distanz nicht perzeptiv motiviert, sondern rein mathematischer Natur ist. Vorhandene Ansätze zur Entwicklung objektiver Maße zur Messung der perzeptiven Distanz zwischen Intonationsverläufen, wie sie in Abschnitt 2.5.3 vorgestellt wurden, erwiesen sich als für diese Aufgabenstellung unzureichend. So bezogen sich beispielsweise bei Reichel et al. (2009) die der Modellentwicklung zugrundeliegenden Ähnlichkeitsurteile nur auf Einzelsilben und nicht auf Segmente variabler Länge.

Ferner wird die Verwendung der Euklidischen Distanz dem Bottom-Up-Charakter des primär datengetriebenen PKS-Modells eher gerecht als der Einbezug phonetischen Vorwissens und erleichtert die Anwendung des Modells auf neue Daten. Aus den Ergebnissen des Perzeptionstests in Reichel et al. (2009) ist zudem eine Abhängigkeit der Ähnlichkeitsbeurteilung von der Muttersprache ableitbar, was bedeutet, dass ein perzeptiv fundiertes Ähnlichkeitsmaß für neue Sprachen neu entwickelt werden müsste.

Klassen Die durch Clustern erhaltenen globalen und lokalen Konturklassen sind in den Abbildungen 10.7 und 10.8 zu finden. In Tabellen 10.2 und 10.3 finden sich Angaben zu ihren Häufigkeiten und durchschnittlichen Längen gemessen in der Anzahl enthaltener Silben.

Klasse	Steigung	relative Häufigkeit	Durchschnittslänge
1	−4.2176	0.46	17
2	−9.2595	0.29	15
3	1.0208	0.25	14

Tabelle 10.2: Steigungskoeffizient b_1, relative Häufigkeiten und Durchschnittslängen (Silbenzahl) globaler Konturklassen.

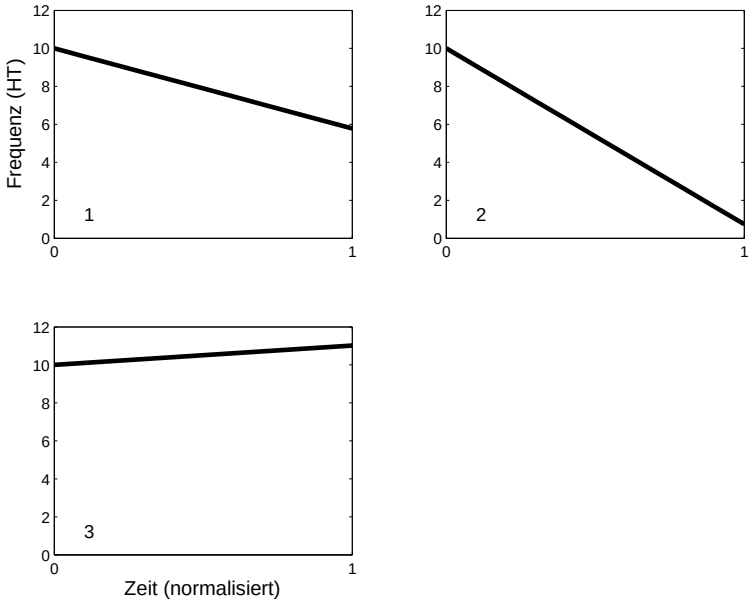

Abbildung 10.7: Globale Konturklassen. Der nicht in der Klassencharakteristik enthaltene Y-Offset ist hier konstant auf 10 HT gesetzt.

Die globalen Konturklassen zeichnen sich durch unterschiedliche Steigungen/Gefälle der Deklinationsgrundlinien aus. Erwartungsgemäß ist ein Übergewicht fallender Grundlinien zu beobachten.

Lokale Klassen unterscheiden sich durch Frequenzmaximum und -spannweite sowie progredienten gegenüber finalem Verlauf. Im Zuge der linguistischen Interpretation in Teil III werden diese Variationen im Hinblick auf semantisches Gewicht, informative Neuheit und Diskursverlauf untersucht.

Festzuhalten ist an dieser Stelle, dass die Akzentlokalisierung in diesem Modell nicht im Zuge der prosodischen Strukturierung geschieht, sondern erst post hoc nach der Klassifizierung der F0-Konturen durch Gewinnung von im unterschiedlichen Maße prominenzverleihender Konturklassen.

10.4 Phonetische Realisierungsparameter

10.4.1 Kontur-Realisierung

Die Übersetzung der abstrakten phonologischen Konturklassen, die in Form von Zentroidvektoren vorliegen, in konkrete phonetische Realisierungen wird mit linearen Regressionsmodellen bewerkstelligt. Hierbei wird bei globalen Konturen für den Steigungskoeffizienten und bei lokalen Konturen ein Modell getrennt für jeden der Polynomkoeffizienten erstellt. Die phonetischen Regressionsmodelle bewirken allgemein eine kontextabhängige Variation der abstrakten Konturzentroiden.

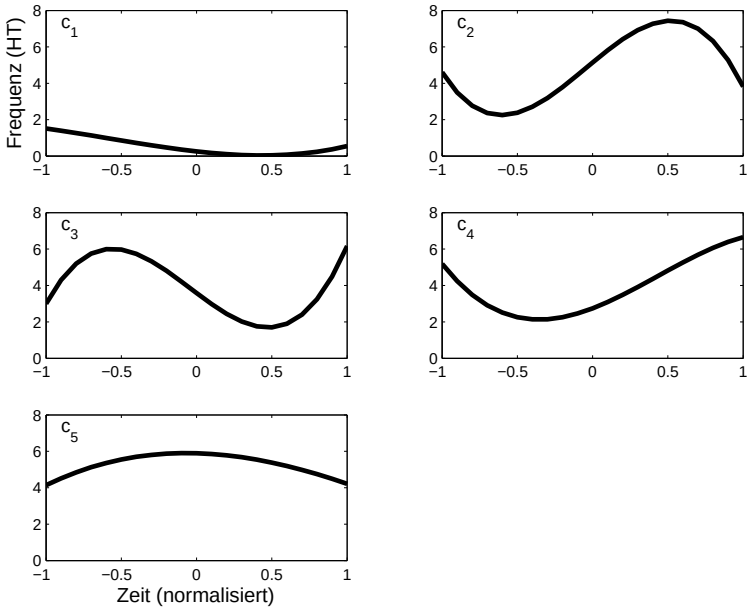

Abbildung 10.8: Lokale Konturklassen. Der Nukleus der akzentuierten Silbe befindet sich beim Zeitpunkt 0.

Anpassung globaler Konturen Es seien i und j die Indizes für globale und lokale Segmente und k die Ordnung des Polynomkoeffizienten, dann ist das Modell zur phonetischen Anpassung $r(*)$ des Zentroid-Steigungskoeffizienten $b_{1,i}$ für globale Konturen folgendermaßen gegeben:

$$r(b_{1,i}) \quad = \quad w_0 + w_1 \cdot b_{1,i} + w_2 \cdot r(b_{1,i-1}) + w_3 \cdot l_i. \tag{10.8}$$

Die Benennung der Prädiktoren findet sich in Tabelle 10.4. Mittels Hauptkomponentenanalyse wurden die Prädiktoren vor Schätzung der Gewichte orthogonalisiert.

Anpassung lokaler Konturen Die Anpassung der lokalen Konturen an der Stelle j erfolgt für jeden Zentroid-Koeffizienten s_k getrennt wie folgt:

$$r(s_{k,j}) \quad = \quad w_{0,k} + w_{1,k} \cdot s_{k,j} + \sum_{n=2}^{5} w_{n,k} \cdot r(s_{n-2,j-1}) + w_{6,k} \cdot r(b_{1,i}) + w_{7,k} \cdot p_j. \tag{10.9}$$

Tabelle 10.5 erläutert die verwendeten Prädiktoren, die vor Schätzung der Gewichte durch eine Hauptkomponentenanalyse orthogonalisiert wurden.

Aus der Wahl der Prädiktoren wird ersichtlich, dass die phonetische Realisierung letztlich als eine Verankerung der Konturklassen in den konkreten intonatorischen Kontext zu verstehen ist. Dies wird durch Überführung der Zentroid-Koeffizienten $b_{1,i}$ und

Klasse	s_0	s_1	s_2	s_3	rel. Häufigkeit	Durchschn. Länge
1	0.2537	−0.9297	0.7758	0.4436	0.22	5
2	5.1403	6.8721	−0.9293	−7.2646	0.18	4
3	3.5853	−6.2229	0.9980	7.8085	0.17	6
4	2.7439	3.1747	3.1763	−2.4418	0.20	5
5	5.8955	−0.2384	−1.7163	0.2772	0.23	5

Tabelle 10.3: Polynomkoeffizienten s_k, relative Häufigkeiten und Durchschnittslängen (Silbenzahl) lokaler Konturklassen.

$b_{1,i}$	Zentroidsteigung
$r(b_{1,i-1})$	realisierte Steigung der vorangehenden globalen Kontur
l_i	Länge der aktuellen Kontur (in Silben)

Tabelle 10.4: Prädiktoren im linearen Regressionsmodell zur phonetischen Realisierung globaler Konturklassen.

$s_{k,j}$ in kontextabhängige phonetische Realisierungen $r(b_{1,i})$ und $r(s_{k,j})$ erreicht. Als Kontext werden hierbei die globale Kontur, die Position des lokalen Segments im globalen Segment, sowie die vorangehende lokale Kontur herangezogen.

10.4.2 Pitch Reset

Unter Pitch Reset $\mathrm{pr}_{j-1:j}$ ist der Frequenzunterschied zwischen dem Ende der vorangehenden globalen Kontur $j-1$ und dem Beginn der aktuellen globalen Kontur j zu verstehen. Die Modellierung von $\mathrm{pr}_{j-1:j}$ erfolgt erneut mit einem linearen Regressionsmodell der Form

$$\mathrm{pr}_{j-1:j} = w_0 + w_1 \cdot r(b_{1,j-1}) + w_2 \cdot r(b_{1,j}) + w_3 \cdot \mathrm{pl}_{j-1:j} + w_4 \cdot \mathrm{bl}_{j-1} \tag{10.10}$$

mit in Tabelle 10.6 spezifizierten Prädiktoren. Auch hier wurden die Prädiktoren vor der Regressionsanalyse durch eine Hauptkomponentenanalyse orthogonalisiert.

In Anhang A sind die Gewichte der Hauptkomponenten für die drei phonetischen Regressionsmodelle angegeben.

10.5 F0-Generierung

Die Anwendung des Modells beispielsweise in Kontext der Text-to-Speech-Synthese ist in allgemeiner Form im in Abbildung 8.2 gezeigten Diagramm skizziert. In den folgenden Abschnitten soll auf einzelne Schritte etwas detaillierter eingegangen werden. Wie in Abschnitt 8.2.2 bereits dargelegt, ist die Entwicklung einer elaborierten textbasierten Vorhersage der Konturklassen nicht Gegenstand dieser Arbeit, da hier zunächst

$s_{k,j}$	Zentroidkoeffizient k-ter Ordnung
$r(s_{k,j-1})$	realisierter Koeffizient k-ter Ordnung im vorangehenden Segment
$r(b_{1,i})$	realisierte Steigung der aktuellen globalen Kontur
p_j	relative Position im globalen Segment

Tabelle 10.5: Prädiktoren im linearen Regressionsmodell zur phonetischen Realisierung lokaler Konturklassen.

$r(b_{1,j-1})$	realisierte Steigung der vorangehenden globalen Kontur
$r(b_{1,j})$	realisierte Steigung der aktuellen globalen Kontur
$\mathrm{pl}_{j-1:j}$	Länge der dazwischenliegenden Pause (ggf. 0)
bl_{j-1}	Baseline-Wert der letzten Silbe der vorangehenden Kontur

Tabelle 10.6: Prädiktoren im linearen Regressionsmodell zur Vorhersage des Pitch Resets $\mathrm{pr}_{j-1,j}$.

der Grundstein solcher Vorhersagen in Form linguistischer Interpretationsversuche gelegt wird (siehe Teil III).

Prosodische Struktur Zu Beginn der F0-Generierung steht die Segmentierung des mit einer Intonationskontur zu versehenden Sprachsignals in globale und lokale Segmente anhand von POS-Sequenz, Satzzeichen und Signalpausen, sowie die Detektion von Silbenkernen.

Globale Konturen Für jedes globale Segment ist eine passende globale F0-Konturklasse zu wählen, also eine Deklinationsgrundlinie, deren Steigung mittels des linearen Regressionsmodells an den vorangehenden globalen Intonationskontext anzupassen ist. Der F0-Startpunkt wird durch das Pitch-Reset-Modell in Abhängigkeit des F0-Endpunkts der vorangehenden Grundlinie ermittelt. Die Deklinationsgrundlinie liefert für jede enthaltene Silbe ein F0-Niveau.

Lokale Konturklassen Innerhalb eines globalen Segments sind für alle lokalen Segmente passende lokale Konturklassen zu wählen und mit den entsprechenden phonetischen Anpassungsmodellen für jeden Polynomkoeffizienten kontextabhängig umzuformen. Zur zeitlichen Alinierung der lokalen Kontur wird dem Silbenkern der wortbetonten Silbe des Kernworts im lokalen Segment die 0 im normalisierten Intervall $[-1\ 1]$ zugeordnet. Die Anpassung der Kontur an die konkreten Zeitverhältnisse erfolgt durch separate Denormalisierung ihres zeitlichen Präakzent- ($[-1\ 0[$) und Postakzent-Verlaufs ($]0\ 1]$):

$$t = \frac{(t_n - \min(t_n))(\max(t) - \min(t))}{\max(t_n) - \min(t_n)} + \min(t) \tag{10.11}$$

t_n ist hierbei der normalisierte und t der konkrete Zeitwert.

Superposition Die F0-Werte der lokalen Konturen werden wie in Abbildung 10.9 gezeigt auf die durch die globale Kontur gegebenen silbenabhängigen F0-Niveaus auf der Halbtonskala addiert und schließlich durch Umkehrung von Gleichung 4.4 in Hertz-Werte transformiert:

$$y_{\mathrm{Hz}} \quad = \quad 2^{\frac{y_{\mathrm{HT}}}{12}} \cdot b \tag{10.12}$$

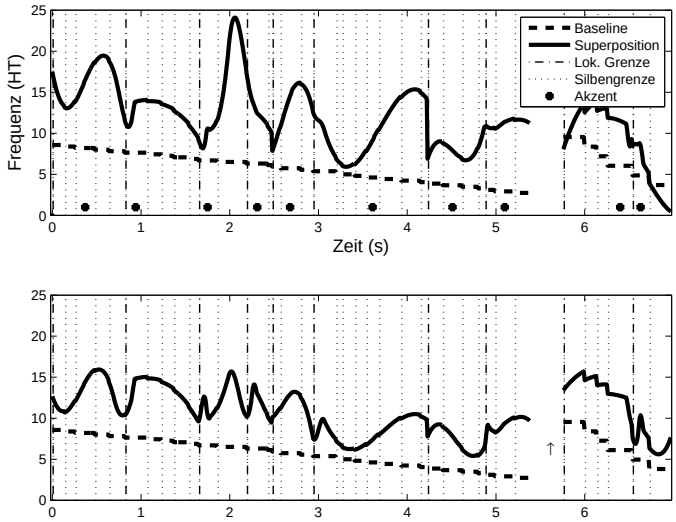

Abbildung 10.9: F0-Generierung eines globalen Segments, in dem sich 10 lokale Segmente befinden. Superposition von Deklinationsgrundlinie (gestrichelt) und den lokalen F0-Bewegungen. **Oben:** Superposition phonologischer Klassen (•: potentiell akzentuierte Silbe). **Unten:** Superposition phonetischer Realisierungen (↑: Pitch Reset). Zugrundeliegender segmentierter Text: [[*In der bosnischen*] [*Moslemenklave*] [*Bihac*] [*gingen*] [*die Kämpfe*] [*zwischen den Regierungstruppen*] [*und serbischen*] [*Verbänden*]] <*P*> [[*auch heute früh*] [*weiter*]]

Kapitel 11

Evaluierung

Die Güte des PKS-Modells wurde sowohl objektiv-mathematisch als auch anhand zweier Perzeptionsexperimente ermittelt. Die zusätzliche perzeptive Evaluierung ist nötig, da von objektiven Abstands- oder Formähnlichkeitsmaßen zwischen Intonationskonturen nicht ohne Weiteres auf deren perzipierte Ähnlichkeit oder gar auf perzipierte Natürlichkeit oder funktionale Äquivalenz geschlossen werden kann (vergleiche Abschnitt 2.5.3).

Zur mathematischen Evaluierung wurden zwei Varianten des Modells mit unterschiedlicher Anzahl lokaler Konturklassen herangezogen, zur perzeptiven Evaluierung nur die Variante mit dem größeren Potential linguistischer Interpretierbarkeit, im Hinblick auf später in dieser Arbeit vorgestellte Untersuchungen.

11.1 Mathematische Evaluierung

11.1.1 Methode

Zur objektiv-mathematischen Evaluierung wurden die in Prosodiestudien üblichen Maße herangezogen:

- der mittlere quadratische Fehler (*RMSE*) zwischen Original- und modellierter F0-Kontur und
- die Korrelation zwischen Original- und modellierter F0-Kontur.

Während der RMSE den Abstand zwischen den F0-Konturen angibt, erlaubt die Korrelation Aussagen über die Formähnlichkeit der Konturen.

Es wurde eine zehnfache Kreuzvalidierung vorgenommen, wobei jeweils 90 Prozent der Daten zur Modellentwicklung herangezogen wurden, also zur Extrahierung der Konturklassen und zur Gewinnung der phonetischen Regressionsmodelle. Auf den verbleibenden zehn Prozent wurde eine F0-Resynthese durchgeführt. Die Partitionierung wurde so vollzogen, dass die Testdaten aus zusammenhängenden Korpusteilen gebildet wurden. Zur F0-Resynthese wurden die segmentierten F0-Konturen nach Zerlegung in globale und lokale Komponenten durch die jeweils ähnlichsten globalen und lokalen Konturklassen

ersetzt und mittels der im Training gewonnenen phonetischen Realisierungsmodelle dem Kontext angepasst.[1] Die Ermittlung der ähnlichsten Konturklasse erfolgte hierbei durch Stilisierung der F0-Komponente wie in Abschnitt 10.2 beschrieben und Berechnung der quadrierten Euklidischen Distanz zwischen dem Koeffizientenvektor der Stilisierung und den Klassenzentroiden.

Die Evaluierung wurde für zwei Versionen des PKS-Modells vorgenommen, die sich im Hinblick auf die Gewinnung der lokalen Konturklassen unterscheiden.

- PKS-5: hier wurden die lokalen Konturklassen so ermittelt wie in Abschnitt 10.3 beschrieben: durch Initialisierung der Clusterzentren mittels optimiertem subtraktiven Clustern. Dieser Ansatz liefert eine relativ geringe Anzahl an lokalen Klassen, auf das gesamte Korpus angewendet sind es fünf, die in Abbildung 10.8 zu sehen sind.

- PKS-16: hier wurde unter Beibehaltung der globalen Konturklassen aus PKS-5 wie bei Möhler und Conkie (1998) die Anzahl der lokalen Cluster auf 16 festgelegt, das heißt das Kmeans-Verfahren mit 16 zufällig im Merkmalsraum lokalisierten Zentren initialisiert. Die entstandenen Klassen finden sich in Abbildung 11.1.

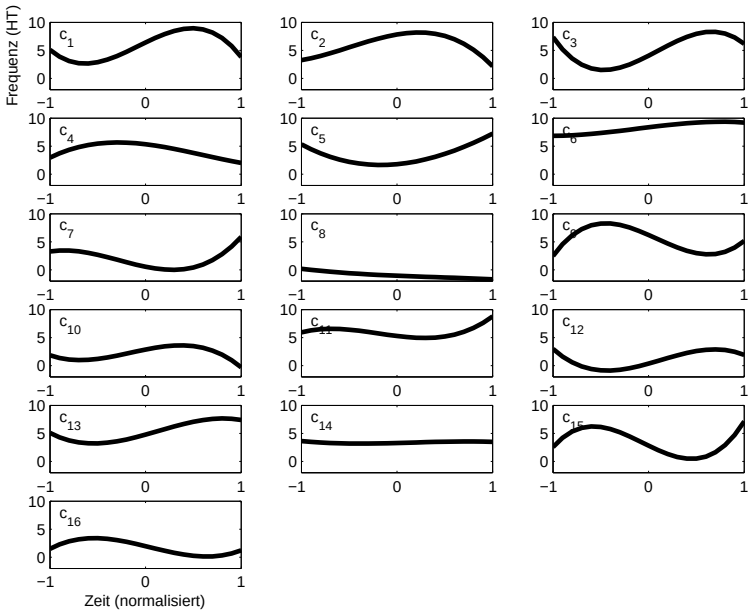

Abbildung 11.1: Lokale Konturklassen der PKS-16-Variante.

PKS-5 bringt wenige und auf Grund des gewählten Optimierungskriteriums distinkte Konturklassen hervor und wurde aus diesem Grund für die in Teil III präsentierten korpus- und perzeptiv basierten linguistischen Interpretation des Modells herangezogen.

[1] Der erste Deklinationsonset wurde auf den Original-F0-Wert gesetzt.

Dafür ist es auf Grund der geringen Clusterzahl weniger geeignet, eine mathematisch gute Anpassung an die Originalkonturen zu liefern. Ob das PKS-Modell zu einer solchen Anpassung grundsätzlich in der Lage ist, wurde mit seiner PKS-16-Variante untersucht.

11.1.2 Ergebnisse

F0-Konturen Abbildung 11.2 zeigt für PKS-5 und PKS-16 die Boxplots der RMSE-Werte und Korrelationen für Trainings- und Testdaten nach zehnfacher Kreuzvalidierung. In Tabelle 11.1 finden sich die zugehörigen Validierungsmittelwerte und Standardabweichungen für Training und Test.

		Korrelation		RMSE (Hz)	
		arithm. Mittel	std	arithm. Mittel	std
PKS-5	Training	0.46	0.03	21.26	0.42
	Test	0.47	0.08	21.14	2.79
	p	0.19		0.43	
PKS-16	Training	0.64	0.01	17.56	0.39
	Test	0.64	0.07	17.57	2.33
	p	0.68		0.79	

Tabelle 11.1: Mathematische Evaluierung der Modellvarianten PKS-5 und PKS-16. Mittelwerte und Standardabweichungen (std) der Korrelationen und mittleren quadratischen Distanzen (RMSE) zwischen Original- und modellierten F0-Konturen nach zehnfacher Kreuzvalidierung für Trainings- und Testdaten. p: empirisches Signifikanzniveau.

Es ergaben sich folgende Befunde:

- PKS-16 approximiert gemessen in Korrelation und RMSE die F0-Konturen signifikant besser als PKS-5 auf Trainings- wie Testdaten (Kruskal-Wallis-Test, für Korrelationen $\chi^2_3 = 28.46$, $p < 0.001$, für RMSE $\chi^2_3 = 22.09$, $p < 0.001$. Post-hoc-Vergleich nach Dunnett, $p < 0.01$).

- Weder in PKS-5 noch in PKS-16 kommt es zu signifikanten Unterschieden von Korrelationen und mittleren quadratischen Fehlern zwischen Trainings- und Testdaten (paarweise Mann-Whitney-Tests, PKS-16: $z < 0.27$, $p \geq 0.68$; PKS-5: $z < 0.8$, $p \geq 0.19$, wobei hier in den Testdaten sogar etwas bessere Werte erzielt werden), was für eine grundsätzliche Robustheit und Generalisierbarkeit beider Modellvarianten spricht.[2]

- Die von PKS-16 erzielten Korrelations- und Distanzwerte zeugen davon, dass das PKS-Modell grundsätzlich in der Lage ist, die Original-F0-Verläufe sowohl in der Form als auch in den absoluten Werten zu approximieren. In Tabelle 12.2 findet sich eine Gegenüberstellung zu anderen Modellen.

[2] Auch eine Arkussinus-Transformation der Korrelationen sowie die Anwendung von – hier nicht zulässigen – t-Tests führt nicht zu signifikanten Unterschieden.

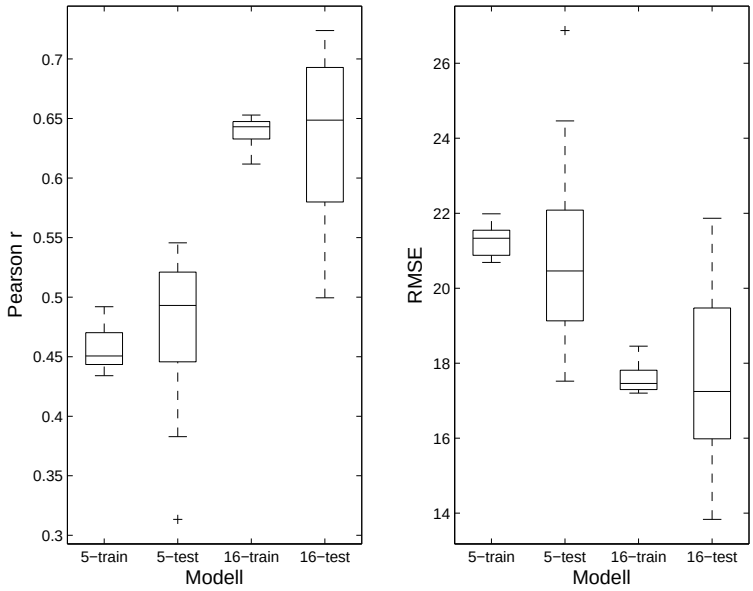

Abbildung 11.2: Evaluierung von PKS-5 und PKS-16. Pearson-Korrelationen und mittleren quadratischen Distanzen (RMSE) in Hertz zwischen Original- und modellierten F0-Konturen nach zehnfacher Kreuzvalidierung für Trainings- und Testdaten.

Phonetische Realisierungsparameter In Tabellen 11.2 und 11.3 finden sich mittlere Korrelationen und RMSE-Werte nach der Kreuzvalidierung für die phonetischen Realisierungsparameter für Trainings- und Testdaten. Folgende Resultate sind zu vermerken:

- Auch hier erzielt PKS-16 signifikant bessere Annäherungen als PKS-5 auf Trainings- und Testdaten (Kruskal-Wallis-Tests, $\chi^2_3 > 28$, $p < 0.001$).
- Vergleicht man für jedes der Modelle die Unterschiede in der Anpassungsgüte zwischen Trainings- und Testdaten ergeben sich mit Ausnahme der Steigung globaler Konturen für alle Parameter signifikante Unterschiede, wobei zu etwa gleichen Anteilen mal auf den Trainings- mal auf den Testdaten bessere Ergebnisse erzielt werden (Mann-Whitney-Tests, $p \leq 0.04$). Details finden sich in den Tabellen 11.2 und 11.3.
- Insgesamt sind die Korrelationen zwischen Modellvorhersagen und Zielwerten auch in den Testdaten hoch ($r \geq 0.78$ für PKS-16, $r \geq 0.77$ für PKS-5, mit einer Ausnahme: $r = 0.48$). Im Großen und Ganzen bestätigt sich somit auch in den phonetischen Regressionsmodellen die beim Vergleich der F0-Konturen vorgefundene Robustheit und Generalisierbarkeit auf ungesehene Testdaten.

			Korrelation		RMSE	
			arithm. Mittel	std	arithm. Mittel	std
PKS-5	Pitch Reset	Training	0.71	0.01	2.13 (HT)	0.02
		Test	0.77	0.03	2.29 (HT)	0.21
		p	0 (*)		0.004 (*)	
	Steigung	Training	0.88	0.00	2.20	0.02
		Test	0.88	0.01	2.19	0.23
		p	0.47		0.91	
PKS-16	Pitch Reset	Training	0.71	0.01	2.13 (HT)	0.02
		Test	0.78	0.03	2.28 (HT)	0.20
		p	0 (*)		0.003 (*)	
	Steigung	Training	0.99	0.00	0.63	0.02
		Test	0.99	0.00	0.62	0.16
		p	0.47		0.57	

Tabelle 11.2: Mittelwerte und Standardabweichungen (std) der Korrelationen und mittleren quadratischen Distanzen zwischen Original- und modellierten Werten für Pitch Reset und Steigungskoeffizienten der globalen Konturen nach zehnfacher Kreuzvalidierung für Trainings- und Testdaten. p: empirisches Signifikanzniveau.

11.2 Perzeptive Evaluierung

Die perzeptive Adäquatheit der Intonationsmodellierung wurde anhand zweier Perzeptionsexperimente untersucht, eines, in dem die Natürlichkeit des modellierten F0-Verlaufs zu beurteilen war, und eines, um zu prüfen ob der neue F0-Verlauf mit einem (unerwünschten) Wandel in der wahrgenommenen Sprecherintention einhergeht.

Diese Experimente bildeten die Teilexperimente 4 und 5 in einer größer angelegten Perzeptionsstudie im Rahmen dieser Arbeit (siehe Abschnitt 13.4). Teilexperimente 1–3 zur linguistischen Beurteilung der lokalen Konturklassen werden in Teil III eingehend vorgestellt.

Zur perzeptiven Evaluierung wurde nur die in der objektiven Evaluierung schlechter abschneidende PKS-5-Variante auf Grund ihres größeren Potentials einer späteren linguistischen Interpretation herangezogen und auf dem kompletten Korpus trainiert. Eine Aufteilung in Training- und Testdaten wurde hier aus folgenden Gründen verworfen: Wegen zufälliger Schwankungen der Partitionenähnlichkeiten ist eine einmalige Aufteilung wenig aussagekräftig. Würde man stattdessen die Stimuli durch n-fache Kreuzvalidierung aus mehreren Testkorpora zusammenstellen, hieße das auch, dass sie von n verschiedenen Modellen erzeugt worden wären, weshalb die beabsichtigte linguistische Interpretation auf n Modelle anstelle von einem ausgeweitet werden müsste – was hier wegen unnötig hoher Komplexität nicht beabsichtigt ist. Es sei noch einmal darauf hingewiesen, dass bereits anhand der mathematischen Evaluierung eine gute Generalisierbarkeit des PKS-Modells festgestellt worden ist.

			Korrelation		RMSE	
			arithm. Mittel	std	arithm. Mittel	std
PKS-5	s_0	Training	0.79	0.01	1.88	0.03
		Test	0.78	0.03	1.93	0.10
		p	0 (*)		0.02 (*)	
	s_1	Training	0.82	0.02	2.54	0.13
		Test	0.84	0.01	2.41	0.15
		p	0 (*)		0.04 (*)	
	s_2	Training	0.57	0.04	2.08	0.08
		Test	0.48	0.07	2.22	0.12
		p	0 (*)		0.003 (*)	
	s_3	Training	0.84	0.03	2.66	0.19
		Test	0.86	0.02	2.50	0.21
		p	0.002 (*)		0.002 (*)	
PKS-16	s_0	Training	0.89	0.01	1.41	0.03
		Test	0.87	0.01	1.50	0.08
		p	0 (*)		0.006 (*)	
	s_1	Training	0.90	0.00	1.90	0.03
		Test	0.92	0.00	1.71	0.07
		p	0 (*)		0 (*)	
	s_2	Training	0.87	0.01	1.27	0.04
		Test	0.80	0.01	1.51	0.04
		p	0 (*)		0 (*)	
	s_3	Training	0.92	0.00	1.95	0.03
		Test	0.94	0.01	1.73	0.07
		p	0 (*)		0 (*)	

Tabelle 11.3: Mittelwerte und Standardabweichungen (std) der Korrelationen und mittleren quadratischen Distanzen zwischen Original- und modellierten Werten für die Polynomkoeffizienten s_n n-ter Ordnung der lokalen Konturen. p: empirisches Signifikanzniveau.

11.2.1 Natürlichkeit

Versuchspersonen

24 Versuchspersonen im Alter zwischen 22 und 47 Jahren nahmen am Experiment teil. Es handelt sich hierbei um Studierende der Phonetik oder Mitarbeiter des Instituts für Phonetik und Sprachverarbeitung in München, und bis auf eine Ausnahme um deutsche Muttersprachler. Die Entscheidung für eine relativ homogene Versuchspersonengruppe phonetisch vorgebildeter Hörer erfolgte in Anbetracht der Gefahr, dass Laien die gestellten Fragen als zu fremd und damit als nicht beantwortbar erscheinen könnten.

Anhand der von ihnen gemachten Angaben lässt sich weiterhin Folgendes über ihre Zusammensetzung sagen:

- Geschlecht: 19 weiblich und 5 männlich
- Herkunft (Ort der Einschulung): 17 aus Süd-, 4 aus Mittel- und 2 aus Norddeutschland. Die einzige nichtdeutsche Versuchsperson kam aus Ungarn und lebt mittlerweile seit über 10 Jahren in Deutschland.

- Musikalische Vorbildung: 18 mit Vorbildung, 6 ohne.

Keine der Versuchspersonen berichtete von Hörschädigungen, die sie bei der Durchführung des Experiments beeinträchtigt hätten. Der Autor dieser Arbeit nahm nicht am Experiment teil.

Methode

Den Versuchspersonen wurden über Kopfhörer zufällig ausgewählte interpausale Äußerungssegmente aus dem SI1000P-Korpus in randomisierter Reihenfolge präsentiert, jeweils zwanzig mit Original-F0-Kontur und mit modellierter Kontur.

Die Aufgabe der Versuchspersonen bestand darin, auf einer fünfstufigen Likert-Skala mit den Endpunkten *natürlich* und *sehr unnatürlich* die Natürlichkeit der Segmente zu beurteilen. In Anhang E findet sich ein Screenshot der mit Perl-Tk erstellten Oberfläche.

Durch vorangehende Experimente zur perzeptiven linguistischen Beurteilung (siehe Teil III) sollte eine Gewöhnung der Versuchspersonen an Resynthese-Artefakte erzielt werden, um einen Einfluss dieser Artefakte auf die Natürlichkeitsbeurteilung der Intonation weitestmöglich zu reduzieren.

Stimuli

Als Stimuli wurden interpausale Segmente aus den Modellentwicklungsdaten mit einer Mindestlänge von sechs Silben herangezogen. Die TD-PSOLA-Resynthese (Charpentier und Moulines, 1989) zur F0-Modifikation erfolgte mit *Praat 5.0.29*. Um systematische Unterschiede hinsichtlich etwaiger Resyntheseartefakte zu vermeiden, wurden auch die Stimuli mit der Original-F0-Kontur auf diese Weise resynthetisiert. Zur Reduzierung eines etwaigen maschinellen Klangs auf Grund fehlender F0-Mikroperturbationen wurde auf die Modell-Konturen mit einem Verfahren von Klatt und Klatt (1990) Jitter j in Form dreier Sinusschwingungen unterschiedlicher Frequenz addiert:

$$j = \frac{fl}{50} \cdot \frac{f_0}{100} \cdot (\sin(2\pi \cdot 12.7t) + \sin(2\pi \cdot 7.1t) + \sin(2\pi \cdot 4.7t)) \qquad (11.1)$$

t bezeichnet die Zeit und fl die sogenannte *Fluttering*-Rate, die wie von Klatt und Klatt empfohlen auf 25 % gesetzt wurde.

Ergebnisse

Tabelle 11.4 zeigt die Mediane und arithmetischen Mittelwerte (*Mean Opinion Scores MOS*) der Natürlichkeitsurteile, Abbildung 11.3 die zugehörigen Boxplots und die relativen Häufigkeiten der Urteilsstufen.

Die modellierten F0-Konturen werden gegenüber den Originalkonturen als signifikant weniger natürlich beurteilt (Mann-Whitney-Test, $z = 10.66$, $p < 0.001$).

Immerhin liegt aber auch das mittlere Natürlichkeitsurteil modellierter Konturen noch signifikant über dem allgemeinen Mittelwert 3 (einseitiger Vorzeichentest für eine Stichprobe zum Medianvergleich, $z = 1.77$, $p < 0.05$).

	Urteils-Median	arithmetisches Urteilsmittel
Original	4	4.07
modelliert	3	3.12

Tabelle 11.4: Mean-Opinion-Scores für die empfundene Natürlichkeit von Original- und modellierten Konturen durch PKS-5.

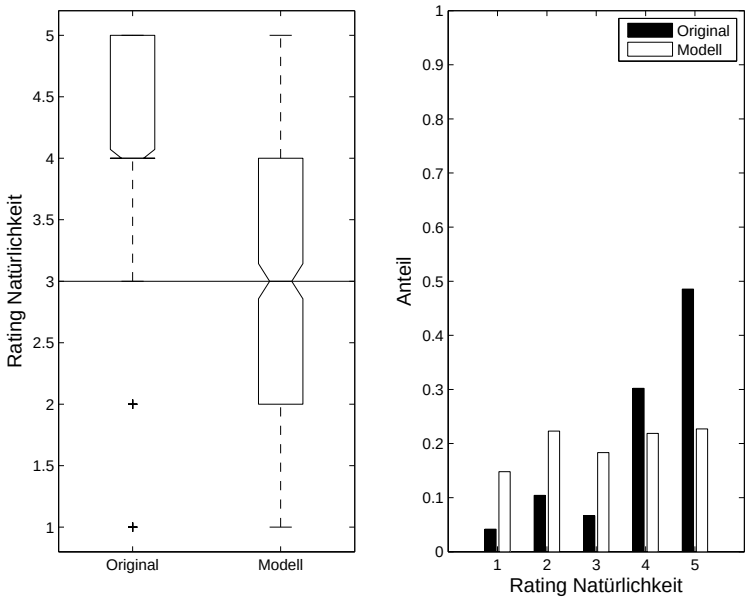

Abbildung 11.3: **Links:** Boxplots zur Beurteilung der Natürlichkeit von Original- und modellierter F0-Kontur. **Rechts:** Relative Häufigkeiten der jeweiligen Urteile.

11.2.2 Sprecherintention

Ziel dieses Teilexperiments war herauszufinden, inwieweit Unterschiede zwischen Original- und modellierter F0-Kontur auch zu Unterschieden in der perzipierten Sprecherintention führen. Untersuchte Konzepte waren:

- Informative Neuheit,
- Bedeutsamkeit (semantisches Gewicht),
- Äußerungsfinalität.

Versuchspersonen

An diesem Experiment nahmen 20 Versuchspersonen teil, die alle bereits bei der vorausgegangenen Natürlichkeitsbeurteilung sowie den später vorzustellenden Teilexperimenten 1–3 teilgenommen hatten.

Methode

Durch die Teilnahme an den vorangehenden Teilexperimenten waren die Versuchspersonen bereits für die hier gegebenen Fragestellungen sensibilisiert. Das Teilexperiment bestand aus drei Blöcken. In jedem dieser Blöcke wurden den Versuchspersonen über Kopfhörer zwölf interpausale lokale Segmente mit einer Mindestlänge von sechs Silben präsentiert, dieselben zwölf Segmente in allen drei Blöcken. Die Segmente wurden zufällig aus dem SI1000P-Korpus ausgewählt mit dem neben der Mindestlänge zusätzlichen Constraint der Ambiguität der Wortfolgen hinsichtlich der Äußerungsfinalität.[3] Im Verlauf der randomisierten Darbietung in jedem Block wurde jedes der Segmente sowohl mit Original-F0-Kontur als auch mit modellierter Kontur dargeboten.

Im ersten Block bestand die Aufgabe darin, die Stimuli auf einer fünfstufigen Likert-Skala mit den Endpunkten *bekannt* und *neu* dahingehend einzuordnen, welchen Neuheitsgrad die Intonation hinsichtlich der übermittelten Information codiert.

Im zweiten Block sollte anhand der Intonation die Bedeutsamkeit der Information auf einer fünfstufigen Skala mit den Endpunkten *belanglos* und *bedeutsam* einordnen.

Im dritten Block ging es darum, anhand der Intonation auf einer durch die Endpunkte *Fortführung* und *Abschluss* aufgespannten fünfstufigen Skala einzuordnen, wie äußerungsfinal der präsentierte Äußerungsabschnitt klingt.

Screenshots zu den verwendeten Oberflächen finden sich in Anhang E.

Ergebnisse

In Tabelle 11.5 sind die Mittelwerte der Urteile zu den behandelten Sprecherintentionen aufgeführt. Abbildungen 11.4, 11.5 und 11.6 zeigen die zugehörigen Boxplots und die Anteile der jeweiligen Urteile.

	Bedeutsamkeit	Neuheit	Finalität
Original	4	3.5	2
modelliert	3	3	2

Tabelle 11.5: Median-Werte der Urteile hinsichtlich Bedeutsamkeit, Neuheit und Finalität für dieselben lokalen Segmente mit Original- und modellierten F0-Konturen.

Hinsichtlich Neuheit und Bedeutsamkeit sind signifikante Unterschiede in der Beurteilung festzustellen (zweiseitiger Wilcoxon-Vorzeichenrangtest für abhängige Stichproben; Bedeutsamkeit: $z = -3.01$, $p < 0.005$; Neuheit: $z = -5.92$, $p < 0.001$). Die modellierten Konturen werden hierbei weniger stark in Zusammenhang mit neuer Information und Bedeutsamkeit gebracht als die Originalkonturen.

In der perzipierten Finalität gab es keine signifikanten Unterschiede zwischen Original und Modell ($z = -0.40$, $p = 0.69$).

[3] Diese Ambiguität wurde anhand der Part-of-Speech-Folge festgestellt. So kamen beispielsweise eindeutig äußerungsmediale Sequenzen wie *Konjunktion-Substantiv*-Folgen nicht in Betracht.

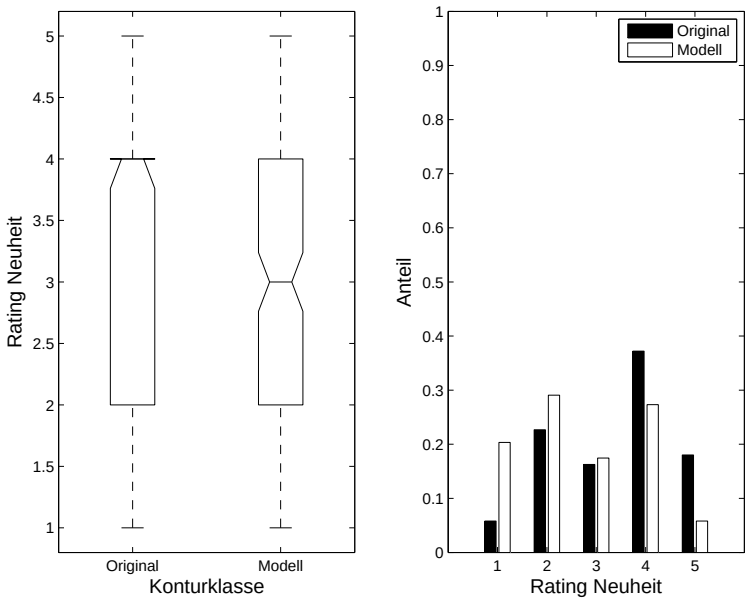

Abbildung 11.4: **Links:** Boxplots zur Beurteilung der Neuheit für Original- und modellierte F0-Kontur. **Rechts:** Relative Häufigkeiten der jeweiligen Urteile.

11.3 Zusammenfassung

Zusammenfassend lassen sich zur Evaluierung des PKS-Modells folgende Punkte festhalten:

Mathematische Evaluierung

Die Verwendung einer erhöhten Anzahl von Konturklassen bewirkt eine verbesserte Anpassung an den Original-F0-Verlauf, sowohl im Hinblick auf den mit RMSE gemessenen Abstand, als auch auf die durch Korrelation ermittelte Formähnlichkeit. Die PKS-16-Variante liefert für beide Maße bessere Ergebnisse als die PKS-5-Variante.

Beide Varianten weisen eine zufriedenstellende Generalisierbarkeit auf, da sich ihre Performanzen auf ungesehenen Testdaten nicht verschlechtern.

Perzeptive Evaluierung

Die perzeptive Evaluierung der PKS-5-Variante zeigte, dass die Modellierung im Vergleich mit dem Originalverlauf als weniger natürlich empfunden wird, wobei der Urteilsmittelwert immer noch oberhalb der mittleren Stufe liegt. Weiter codieren die modellierten Konturen in weniger starker Ausprägung die linguistischen Konzepte Bedeutsamkeit und Neuheit, zeigen aber gegenüber dem Original keinen Unterschied in der Finalitätscodierung.

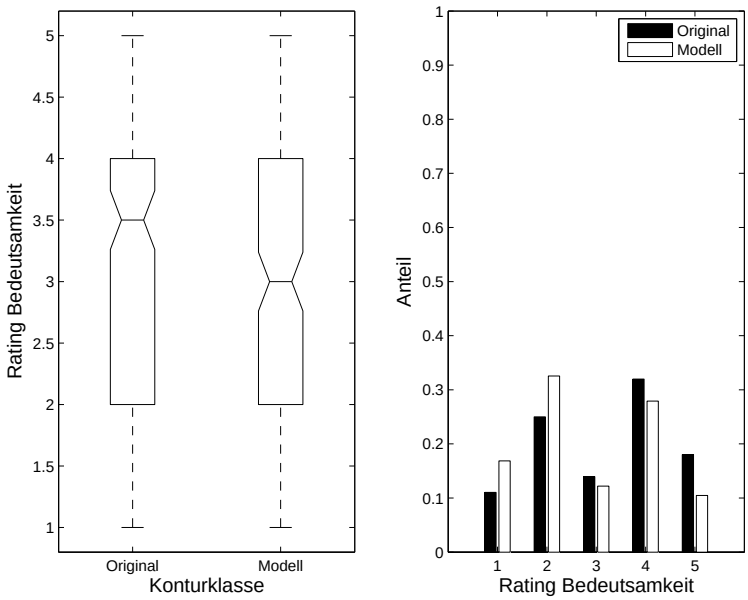

Abbildung 11.5: **Links:** Boxplots zur Beurteilung der Bedeutsamkeit für Original- und modellierte F0-Kontur. **Rechts:** Relative Häufigkeiten der jeweiligen Urteile.

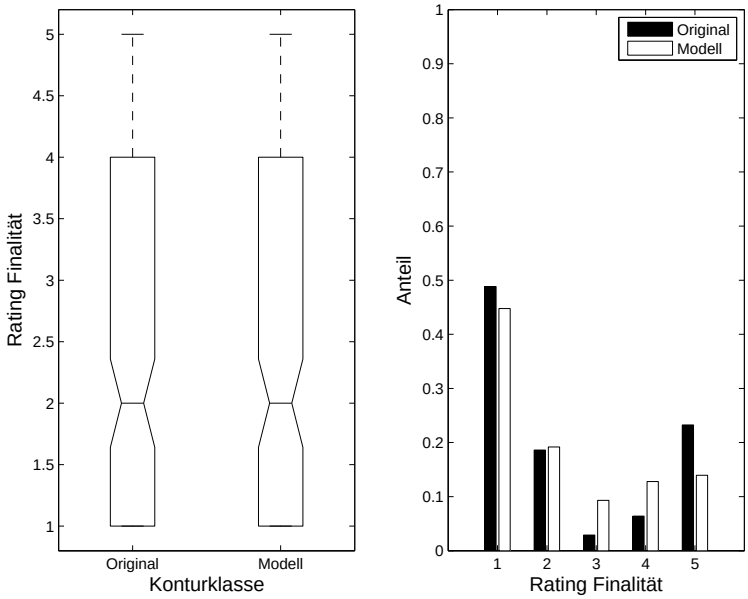

Abbildung 11.6: **Links:** Boxplots zur Beurteilung der Finalität für Original- und modellierte F0-Kontur. **Rechts:** Relative Häufigkeiten der jeweiligen Urteile.

Kapitel 12

Diskussion und Zusammenfassung des Teils II

Dieses Kapitel beinhaltet eine Diskussion der in diesem Teil der Arbeit dargelegten Modellentwicklung und -evaluierung.

12.1 Daten und Allgemeingültigkeit

Zur Entwicklung des PKS-Modells kam Material nur eines Sprechers zum Einsatz, was die Frage der Allgemeingültigkeit der extrahierten Konturklassen aufwirft. Zu rechtfertigen ist die Beschränkung auf einen Sprecher mit der Vermeidung der zwischen Sprechern vorzufindenen hohen Variabilität, wie sie beispielsweise Baumann et al. (2006) vorgefunden haben: bei der Entwicklung eines datenbasierten Modells, das wie dieses auch zur Intonationsgenerierung heranziehbar sein soll, ist es nicht zielführend, Strategien unterschiedlicher Sprecher unkontrolliert zu mischen. Stattdessen sollte eine konsistente, kommunikativ funktionierende und allgemein akzeptierte Intonationsstrategie abgebildet werden. Die Voraussetzung für Konsistenz wird durch die Beschränkung auf einen Sprecher erzielt. Kommunikatives Funktionieren und allgemeine Akzeptanz dürfte durch die Wahl eines ausgebildeten und damit hierfür qualifizierten Sprechers gewährleistet sein. Kommunikatives Funktionieren ist in dieser Studie anhand der linguistischen Interpretierbarkeit nachgewiesen worden, allgemeine Akzeptanz lässt sich aus den hohen Natürlichkeitsbewertungen der Original-F0-Konturen durch die Versuchspersonen ableiten. Aus den genannten Gründen ist es in der datenbasierten Intonationsmodellierung also durchaus nicht unüblich, sich auf nur einen Sprecher zu beschränken (Rapp, 1998b; Möhler, 1998c).

Kritisch angemerkt werden muss allerdings, dass das verwendete Korpus keine direkten Fragen enthält. Die drei auftretenden Fragen liegen als Zitate vor und sind obendrein fast alle rhetorischer Natur, was sie für eine valide Modellierung der Frageintonation unbrauchbar macht. Das Modell ist somit in seinem aktuellen Stand auf Deklarativsätze beschränkt.

12.2 Modellcharakteristika

12.2.1 Prosodische Strukturierung

Globale Segmente

Die bislang vorgenommene Segmentierung in globale Segmente ist vergleichsweise grob, da sie nur solche Segmentgrenzen identifiziert, die durch Pausen oder orthographisch durch Satzzeichen gekennzeichnet sind. Die Entwicklung eines geeigneten Verfahrens zur gezielten Detektion von F0-Diskontinuitäten in der Funktion von Grenzsignalen steht noch aus. Es ist davon auszugehen, dass das Auffinden dieser weniger stark markierten Grenzen die Anpassung und Natürlichkeit der stilisierten Kontur verbessert.

Lokale Segmente

Chunking Die lokale Segmentierung des Signals ist syntaktisch geleitet. Die Festlegung der Segmente auf eine Funktionswortsequenz mit abschließendem Inhaltswort stellt in der überwiegenden Mehrheit der Fälle sicher, dass maximal ein Akzent im Segment auftritt, was für die anschließende F0-Stilisierung entscheidend ist. Zudem vereinfacht die Beschränkung auf ein Inhaltswort die spätere linguistische Analyse (siehe Teil III dieser Arbeit).

Das hierzu vorgenommene Chunking ist angelehnt an die Arbeiten von Gee und Grosjean (1983) und Abney (1991) zur Ermittlung prosodischer Phrasierungseinheiten (vgl. Abschnitt 5.2). Eine wesentliche Abweichung zu diesen Arbeiten liegt in der hier vorgenommenen Setzung prosodischer Grenzen zwischen attributiven Adjektiven und den zugehörigen Substantiven, was beispielsweise bei Abney verworfen wird unter der Annahme, dass es sich bei attributiven Adjektiven nicht um chunkrelevante *major heads* handelt. Begründen lässt sich diese Grenze mit der in den vorliegenden Nachrichtensprecherdaten beobachteten Tendenz kurzer prosodischer Phrasen zur Verbesserung der Verständlichkeit. Aus der Kürze der Phrasen ergibt sich eine erhöhte Akzentdichte, so auch in Adjektiv-Substantiv-Sequenzen, die aus diesem Grund zur Gewährleistung der Ein-Akzent-Obergrenze in zwei lokale Segmente aufzuspalten sind.

Akzentgruppe Auf Grund der Beschränkung auf einen Akzent lässt sich das lokale Segment mit einer Akzentgruppe gleichsetzen, sofern man der allgemeinen Definition einer Akzentgruppe gemäß Stock und Zacharias (1982) folgt, derzufolge sich diese zusammensetzt aus einer akzentuierten Silbe und allen mit ihr prosodisch verbundenen unakzentuierten Silben. In einer engeren Definition wie beispielsweise bei van Santen et al. (1998) entspricht die Akzentgruppe einem prosodischen Fuß, wird also von einer akzentuierten Silbe eingeleitet gefolgt von allen unakzentuierten Silben bis hin zum nächsten Akzent. Diese engere Definition begründen van Santen et al. (1998) mit dem Befund, dass auf phonetischer Ebene die Alinierung des Akzents auf der betonten Silbe nur von der Länge der *nachfolgenden* Silbenkette abhängt, je länger, desto weiter verschiebt sich der Akzent nach hinten. Auf Ebene von Diskurs und Pragmatik jedoch zeigen Untersuchungen zum *frühen Gipfel* (Kohler, 1987), dass das F0-Maximum auch

auf der dem Akzent *vorangehenden* Silbe liegen kann. Dieser Zusammenhang lässt sich nicht herstellen, wenn Akzentgruppen nicht auch präakzentuierte Silben umfassen können. Die allgemeinere Definition der Akzentgruppe nach Stock und Zacharias (1982) ist also weiterhin vertretbar.

Resegmentierung Erste Untersuchungen zu einer Resegmentierung der Konturklassen-Sequenz mittels einer Adaption des Sequitur-Algorithmus (Nevill-Manning und Witten, 1997) ergaben einen Zusammenhang zwischen Intonation und größeren syntaktischen Einheiten, namentlich Nominalphrasen. Der Sequitur-Algorithmus, der in der Tradition von MDL-basierten Verfahren[1] eine Strukturierung von Daten als ein Komprimierungsproblem versteht, sorgt für eine Komprimierung der Konturklassen-Sequenz S, indem er aus ihr eine hierarchische Struktur in Form einer kontextfreien Grammatik ableitet, die S komplett beschreibt. Dies geschah in dieser Arbeit durch rekursives Ersetzen sich überzufällig häufig wiederholender Konturklassenpaarungen durch non-terminale Symbole. Es zeigte sich, dass auf diese Weise überwiegend solche benachbarten lokalen Segmente zusammengefasst wurden, deren Kernwörter zusammen Nominalphrasen bildeten, also beispielsweise attributive Adjektive und Nomen, oder Kardinalzahlen und Nomen.

Aus diesen Ergebnissen ließe sich also eine lokale Segmentierung in größere syntaktische Einheiten, als wie sie hier vorgenommen wurde, ableiten. Im Sinne des Constraints, maximal einen Akzent pro Segment zu erlauben, wurde allerdings für die derzeitige Version des PKS-Modells davon Abstand genommen.

Akzentsetzung

Die PKS-Modellierung verzichtet auf eine explizite Akzentsetzung. An dessen Stelle treten lokale Intonationskonturklassen, die unterschiedlichen F0-Niveaus über der Baseline und F0-Spannweiten unterschiedlich hohe Prominenz verleihen. Akzente sind also im Gegensatz zu den in Teil I besprochenen symbolischen und parametrischen Modellen keine Grundlage für die F0-Modellierung, sondern ergeben sich erst post hoc durch unterschiedlich prominente F0-Verläufe. Ein solcher Ansatz erspart wie schon dargelegt eine prosodische Etikettierung. Darüber hinaus ermöglicht er im Falle einer Einbettung in Sprachsynthesesysteme den Verzicht auf ein vorgeschaltetes Modul zur Akzentsetzung und reduziert damit die Wahrscheinlichkeit von Folgefehlern bei fehlerhaftem Output des Akzentmoduls.

12.2.2 Intonatorische Modellierung

Parametrisch, konturbasiert und superpositional

Die theoretischen und praktischen Beweggründe für die Wahl eines *parametrischen, konturbasierten* und *superpositionalen* Ansatzes der Intonationsmodellierung wie sie bereits in Kapitel 8 angesprochen wurden, seien hier noch einmal kurz zusammengefasst:

[1] MDL steht für *Minimum Description Length* (Rissanen, 1978; Grünwald, 2007) und misst den Umfang der komprimierten Daten und die Größe des zur Komprimierung herangezogenen Modells.

Die parametrische F0-Beschreibung ermöglicht eine datenbasierte und damit vergleichsweise theoriefreie Modellierung, die durch ihre Signalnähe sowohl für die F0-Analyse als auch -synthese automatisierbar ist.

Gegen den hier vertretenen konturbasierten Ansatz lässt sich seitens der Perzeptionsforschung die *Tonal Movement Coding*-Hypothese (House, 1990, vgl. Abschnitt 2.5.3) ins Feld führen, derzufolge die Wahrnehmung komplexer F0-Konturen auf einzelnen Ton-Targets beruht. Demgegenüber sprechen für eine Verwendung von Konturen anstelle von Tönen Befunde von Petrone und D'Imperio (2008) zur entscheidenden Rolle der F0-Kontur und nicht der Ton-Targets bei der Satzmodus-Codierung (vgl. Abschnitt 7.2). Hinzu kommen Befunde von Dainora (2002) zur nahezu vollständigen Determiniertheit einzelner Tönen durch vorangehende Töne, woraus sich folgern lässt, dass nicht Töne sondern eher Tonsequenzen oder Konturen als relevante intonatorische Einheiten aufzufassen sind (vgl. Abschnitt 7.3).

Die Superpositionalität des PKS-Modells ist vor dem Hintergrund von Befunden zu *Look Ahead*-Mechanismen in Intonationsproduktion und -perzeption (Cooper und Sorensen, 1981; Thorsen, 1985; Féry et al., 2009, vgl. Abschnitt 7.3) phonetisch motiviert und ermöglicht eine nicht lokal begrenzte Modellierung globaler F0-Komponenten und damit einen erhöhten Grad an Beschreibungsökonomie.

Gewinnung der Konturklassen

Zur Ermittlung der Ähnlichkeit von F0-Verläufen im Zuge der Gewinnung von Konturklassen wurde ein mathematisch-objektives Maß herangezogen. Wünschenswert wäre hier ein perzeptiv motiviertes Ähnlichkeitsmaß. Allerdings ergeben die im Abschnitt 2.5 referierten Forschungsergebnisse zur Wahrnehmung von Intonation ein derart komplexes Bild, dass gegenwärtig die Entwicklung eines Ähnlichkeitsmaßes, das neben der F0--Bewegung auch die Einflüsse der Segmentebene, non-lineare Effekte kategorialer Wahrnehmung und Muttersprache des Hörers berücksichtigt, noch nicht durchführbar ist. Hinzu kommt, dass die notwendige Datenaufbereitung für den Einsatz eines solchen Maßes über die minimalen Anforderungen des PKS-Modells weit hinausgeht.

Globale Konturen

Baseline Mit der Hinzunahme einer globalen Komponente in Form einer Baseline in die Intonationsbeschreibung unterscheidet sich das PKS-Modell von den streng lokal operierenden symbolischen Tonsequenzansätzen sowie von anderen parametrischen Modellen wie dem Rapp-Modell oder PaintE. Die gefundene Abhängigkeit der Deklination von der Länge der Intonationsphrase lässt sich nicht an der abstrakten Konturklasse selbst ablesen, da diese durch Normalisierung zeitunabhängig ist. Vielmehr fungiert die Phrasenlänge auf phonetischer Ebene als Prädiktor im Regressionsmodell zur Adjustierung des Steigungskoeffizienten.

Topline Auf eine explizite Modellierung einer Topline zusätzlich zur Baseline wurde im PKS-Modell verzichtet. Der Verringerung der F0-Spannweiten auf akzentuierten Silben

im Verlauf einer Intonationsphrase wird dafür im phonetischen Regressionsmodell zur Anpassung des Polynomkoeffizienten s_0 Rechnung getragen (siehe Tabelle 10.5), indem dort die relative Position im globalen Segment als Prädiktor mit eingeht.

Lokale Konturen

Konturklassen Einige der Konturklassen, die in Abbildung 10.8 zu sehen sind, lassen sich anhand ihrer Gestalt Intonationsereignissen anderer Modelle zuordnen. Denkbare Zuordnungen sind in Tabelle 12.1 zusammengefasst.

PKS	TSM	KIM
c_1	L*	mittleres Tal
c_2	L*+H	später Gipfel
c_3	H*+L	früher Gipfel
c_5	H*	mittlerer Gipfel

Tabelle 12.1: Denkbare Entsprechungen der lokalen Konturklassen zu Einheiten anderer Modelle (Tonsequenzmodell TSM mit GToBI-Etiketten, Kieler Intonationsmodell KIM).

Realisierung Die phonetische Realisierung der lokalen Konturen ist im PKS-Modell abhängig unter anderem von der Position in der Intonationsphrase und der vorangegangenen lokalen Kontur. Ersteres reflektiert wie oben beschrieben die implizite Miteinbeziehung einer Topline. Letzteres verankert die Kontur im aktuellen Kontext.

Hierbei lässt sich nur der linke Kontext berücksichtigen, da es sonst zu einer zirkulären Anpassung benachbarter Konturen kommen würde. Durch die nötige Vermeidung dieser Zirkularität wird das Modell allerdings Befunden zu einer Vorausplanung der Intonation auf lokaler Ebene nicht gerecht. Zu nennen sind hier:

- die Vermeidung von *stress clashes* durch Vorverlagerung des ersten Akzents (beispielsweise in *thirteen men*),
- die Beobachtung von de Pijper (1983), dass die F0-Kontur über mehrere unakzentuierte Silben zum Akzent hin ansteigt, sowie
- das durch zwei Akzente geformte Hutmuster, das impliziert, dass der Sprecher bei der Realisierung des ersten Akzents durch Verzicht auf die fallende Tonbewegung bereits den folgenden Akzent mitberücksichtigt (Levelt, 1989).

12.3 Evaluierungsergebnisse

12.3.1 Mathematische Evaluierung

Der Vergleich der PKS-basiert resynthetisierten F0-Verläufe mit den Originalkonturen in Form von Pearsons r und mittlerem quadratischem Fehler (RMSE) zeigte eine Abhän-

gigkeit der Anpassungsgüte von der gewählten Anzahl der Konturklassen: trivialerweise steigt mit höherer Anzahl der Klassen die F0-Approximationsfähigkeit des Modells.

Unabhängig von der Klassenanzahl sind die Ergebnisse auf den Testdaten, wenn überhaupt, nur geringfügig schlechter als auf den Trainingsdaten. Dies zeugt von einer hohen Generalisierbarkeit des PKS-Modells, das offensichtlich nicht auf die Trainingsdaten überadaptiert.

Vergleichbares gilt für die Modellkomponenten zur phonetischen Realisierung, mit denen auch auf den Testdaten hohe Korrelationen und geringe Abweichungen zwischen Original- und vorhergesagten Werten erzielt werden konnten.

Andere Studien Zur Orientierung seien in Tabelle 12.2 einige Evaluierungswerte anderer Modelle angeführt, ohne dass hier ein wirklicher Anspruch auf Vergleichbarkeit erhoben werden kann, da sich sowohl Korpora als auch Evaluierungsmethoden voneinander unterscheiden. In einigen der Studien wurde beispielsweise kein unabhängiges Testkorpus verwendet, und in Möhler (2001) erfolgte die Evaluierung nur auf akzentuierten und phrasenfinalen Silben.

Auch die nur einmalig vorgenommene Aufteilung in Trainings- und Testdaten bei Möhler (2001) und Agüero und Bonafonte (2005) kann keine verlässlichen Vergleichsergebnisse liefern, da die Streuung der Performanzen auf den Testdaten bei mehrfacher Kreuzvalidierung, wie Abbildung 11.2 zeigt, groß ist.

Im Gegensatz zum PKS-Modell basieren die Performanzen der anderen Modellierungen auf einer vorab gegebenen prosodischen Annotation.

Modell, Studie	RMSE (Hz)	Korrelation	Annotation	Testkorpus
PKS-5	21.1	0.43	nein	ja
PKS-16	17.6	0.64	nein	ja
PaintE, Möhler (2001)	14.0	0.69	ja	ja
Fujisaki, Agüero et al. (2005)	21.2	0.76	ja	ja
Tilt, Taylor (2000)	7.5	0.83	ja	nein
Maximumbasiert, Heuft et al. (1996)	–	0.85	ja	nein

Tabelle 12.2: Gegenüberstellung der mathematischen Evaluierung unterschiedlicher Modelle.

12.3.2 Perzeptive Evaluierung

Die perzeptive Gütemessung erfolgte für die PKS-5-Variante – trotz deren gegenüber PKS-16 schlechteren Abschneidens in der mathematischen Evaluierung. Grund für die Wahl des PKS-5-Modells ist die geringe Anzahl und die durch die gewählte Clustering-

Optimierung hohe Distinktivität seiner Konturklassen, was es für nachfolgende Untersuchungen zur linguistischen Interpretation interessanter erscheinen lässt.

Natürlichkeit Zwar erreichen die modellierten Konturen *mean opinion scores* knapp oberhalb der mittleren Urteilsstufe, jedoch bewegen sie sich deutlich unterhalb der Urteile zu den Originalkonturen. Unterschiede in den Resynthesebedingungen können dafür nicht verantwortlich gemacht werden, da alle Stimuli gleichermaßen PSOLA-resynthetisiert wurden.

Die geringere Natürlichkeit mag zum Teil dadurch bedingt sein, dass die modellierten Konturen allgemein flacher verlaufen als die Originalkonturen. Gemittelt über alle lokalen Segmente beträgt die F0-Spannweite der Originalkonturen rund 54 Hz, während sie bei den modellierten Konturen mit 22 Hz weit darunter liegt. Dies liegt an der Glättung in der Vorverarbeitung, an der polynomialen Stilisierung sowie in der Verwendung von Zentroiden als Klassenprototypen. Die phonetischen Regressionsmodelle in ihrer aktuellen Form sorgen offensichtlich nicht für eine ausreichend hohe Varianz in der Abweichung gegenüber den prototypischen Konturen.

Es ist denkbar, dass die PKS-16-Variante auf Grund ihrer stärker ausgeprägten Fähigkeit der F0-Anpassung höhere MOS-Werte erreicht hätte, jedoch wurde dies hier auf Grund der ohnehin schon umfangreichen Experimentreihe nicht untersucht.

In jedem Fall scheint die hier vorgenommene perzeptive Evaluierung modellierter Intonation angebracht angesichts des bekannten Sachverhalts, dass perzeptive Urteile nicht zufriedenstellend aus verwendeten mathematischen Standardmaßen vorhergesagt werden können (Hermes, 1998; Clark und Dusterhoff, 1999; Reichel et al., 2009).

Sprecherintention Bezüglich der Sprecherintention wurden zwischen Original und Modellausgabe keine Unterschiede bei der Finalität, aber dafür bei der Codierung von Neuheitswert und Bedeutsamkeit festgestellt. In beiden Fällen erreicht das Modell niedrigere Urteilswerte als das Original.

Da Neuheit und Bedeutsamkeit durch eine Erhöhung der Prominenz kenntlich gemacht werden, die ihrerseits unter anderem auf deutliche F0-Bewegungen zurückzuführen ist, lässt sich auch für diese Abweichung wieder der flachere F0-Verlauf des Modell-Outputs gegenüber den Originalkonturen verantwortlich machen – trotz signifikanter Unterschiede von F0-Maxima und -Spannweiten auch in den modellierten Konturen (siehe Abschnitt 15.2.1).

Dieser Sachverhalt ist vermutlich nicht allzu negativ zu bewerten, da es plausibel erscheint, dass eine weniger starke Markierung informativ neuer und bedeutsamer Passagen in der Synthese als weniger störend empfunden wird als der umgekehrte Fall der unangemessen starken Markierung von gegebener oder unwichtiger Information.

Andere Studien Auf Grund des größeren Aufwands werden perzeptive Evaluierungen seltener durchgeführt als mathematische. Auf eine vergleichende Evaluierungsstudie von Syrdal et al. (1998) zum Amerikanischen Englisch sei an dieser Stelle exemplarisch

verwiesen, in der auf einer fünfstufigen Skala *Mean Opinion Scores* zur Natürlichkeitsempfindung für mehrere Varianten des PaintE-Modells mit Vektorquantisierung, einem regelbasierten Modell zur Überführung von ToBI-Etiketten in F0-Werte sowie für das Tilt-Modell ermittelt wurden. Knapp zusammengefasst erreichten die Modelle MOS-Werte zwischen 3.1 und 3.5, wobei das PaintE-Modell mit 16 Konturklassen am besten abschnitt und das Tilt-Modell am schlechtesten.

12.4 Mögliche Erweiterungen

Sprecherabhängiger F0-Grundwert

Da die verwendeten Trainingsdaten nur von einem Sprecher stammen, wurde in dieser Arbeit auf die Verwendung eines sprecherabhängigen Basis-F0-Werts, so wie er beispielsweise im Fujisaki-Modell gegeben ist, verzichtet. Grundsätzlich wäre das PKS-Modell aber problemlos um die Mitmodellierung eines solchen Werts, beispielsweise in Form des gefundenen F0-Minimums, erweiterbar.

Einbezug der Lautsegment-Ebene

Bislang unberücksichtigt ist auch der Einfluss der segmentalen Ebene auf den F0-Verlauf. Grund hierfür ist der Verzicht auf eine manuelle Lautsegmentierung und phonetische Transkription in den Trainingsdaten im Sinne einer weitestmöglichen Reduzierung der Korpusvoraussetzungen. Im Falle des Vorliegens einer exakten Lautsegmentierung ließe sich testen, ob beispielsweise eine mikroprosodisch geleitete zeitvariable Gewichtung von F0-Verläufen (Reichel und Winkelmann, 2010), wie sie in Abschnitt 4.3.3 ausgeführt wurde, die gewünschten Perturbationen herbeiführen kann.

12.5 Zusammenfassung des Teils II

Das hier vorgestellte PKS-Intonationsmodell dient der parametrischen, konturbasierten und superpositionalen Intonationsbeschreibung. F0-Konturen sind repräsentiert als Überlagerung diskreter globaler und lokaler Konturklassen, die mittels phonetischer Realisierungsmodelle an den aktuellen Kontext angepasst werden.

Voraussetzung für die Gewinnung der F0-Repräsentation ist eine Alinierung zwischen F0-Kontur, Silbenkernen und POS-gelabeltem Text, die automatisch herzustellen ist. Eine prosodische Etikettierung ist nicht nötig.

Die mathematische Evaluierung ergab eine positive Abhängigkeit der Anpassungsgüte modellierter F0-Verläufe von der Anzahl der lokalen Konturklassen. Die perzeptive Evaluierung zeigte, dass die modellierten Konturen im Großen und Ganzen als akzeptabel beurteilt werden, allerdings weniger natürlich als die Originalkonturen. Sprecherintentionen hinsichtlich Bedeutsamkeit und Neuheit sind in den gegenüber dem Original flacher verlaufenden modellierten Konturen weniger stark ausgeprägt.

Nach dieser Vorstellung des PKS-Modells behandelt der folgende Teil nun Ansätze zu dessen linguistischer Verankerung.

Teil III

Linguistische Interpretation

Überblick Inhalt dieses Teils der Studie ist die Untersuchung, ob die rein datenbasiert gewonnenen lokalen Intonationskonturklassen im Anschluss linguistisch interpretiert werden können. Hierfür werden linguistische Korpusanalysen zur Hypothesengenerierung und Perzeptionsexperimente zu deren Überprüfung herangezogen. Kapitel 13 beschreibt die untersuchten linguistischen Konzepte sowie das allgemeine Vorgehen zu deren Verknüpfung mit der Intonation. In den Kapiteln 14, 15 und 16 werden Korpusanalyse, abgeleitete Hypothesen sowie deren perzeptive Überprüfung für die untersuchten linguistischen Konzepte einzeln vorgestellt. In Kapitel 17 wird die Nutzung der gewonnenen Erkenntnisse zur Entwicklung des PKS-EB-Modells, eines linguistischen Vorhersagemodells der Intonation in Form eines Entscheidungsbaums (EB) beschrieben sowie die perzeptive Evaluierung dieses Modells.

Kapitel 13

Allgemeines Vorgehen

13.1 Intonatorische und linguistische Untersuchungsobjekte

Die hier vorgenommene linguistische Untersuchung bezieht sich auf die Stilisierungsparameter sowie die fünf lokalen Konturklassen des Modells PKS-5 (siehe Abbildung 10.8) und umfasst folgende linguistische Kenngrößen auf semantischer und Diskursebene:

- Semantisches Gewicht (Bedeutsamkeit),
- Informative Neuheit,
- Äußerungsfinalität.

13.2 Arbeitsschritte

Zur Verknüpfung von Intonation mit linguistischen Funktionen wurde:

- auf **Parameterebene** der Zusammenhang zwischen den **Koeffizienten der Stilisierungsfunktion** und den linguistischen Konzepten beleuchtet, und
- auf **Symbolebene** eine linguistische Interpretation der **Konturklassen** versucht.

Das Vorgehen zur Interpretation der Konturklassen bestand aus den folgenden Schritten, die in den nächsten Abschnitten etwas genauer vorgestellt werden:

1. Linguistische Analyse der Textdaten,
2. Gewinnung statistischer Zusammenhänge zwischen Linguistik und den signalbasiert gewonnenen Intonationsklassen,
3. Ableitung von Hypothesen hinsichtlich der linguistischen Funktion der Intonationsklassen,

4. Überprüfung der Hypothesen mittels Perzeptionsexperimenten.

Die gewonnenen Erkenntnisse wurden anschließend zu einem linguistischen Intonationsvorhersagemodell integriert und dessen Qualität wiederum getestet.

13.3 Korpusanalyse und Hypothesengenerierung

Das Textkorpus wurde mittels automatisierter Verfahren, die in den folgenden Kapiteln beschrieben werden, linguistisch analysiert. Die mit automatischen Verfahren einhergehenden hohen Fehlerquoten können hierbei in Kauf genommen werden, da nicht eine fehlerfreie linguistische Korpusaufbereitung das Ziel ist, sondern lediglich die Schaffung einer hinreichenden Grundlage, um das gemeinsame Auftreten von Intonationskonturklassen und linguistischen Ereignissen statistisch untersuchen zu können. Ergeben sich aus dieser Untersuchung signifikante Zusammenhänge, lassen sich Hypothesen hinsichtlich der Funktion der Konturklassen formulieren – beispielsweise *Konturklasse c_i markiert die Einführung neuer Information*. Diese werden im Anschluss durch Perzeptionsexperimente überprüft.

13.4 Allgemeines Design der Perzeptionsexperimente

13.4.1 Teilexperimente

Die Perzeptionsexperimente zum oben genannten Arbeitsschritt (4) wurden als Teile einer einzelnen Sitzung durchgeführt, die insgesamt diese fünf Teilexperimente umfasste:

- **Experiment 1:** Beurteilung der intonatorischen Markierung informativer Gegebenheit/Neuheit.
- **Experiment 2:** Beurteilung der intonatorischen Markierung der Bedeutsamkeit.
- **Experiment 3:** Beurteilung der intonatorischen Markierung der Äußerungsfinalität.
- **Experiment 4:** Bewertung der Natürlichkeit der modellierten Konturen.
- **Experiment 5:** Bewertung etwaiger mit der Intonationsmodellierung einhergehender Änderungen der perzipierten Sprecherintention.

Methode und Ergebnisse der Experimente 4 und 5 wurden bereits in Kapitel 11 vorgestellt.

Reihenfolge

Die gewählte Reihenfolge der Experimente hat folgende Gründe: Die Untersuchung informativer Neuheit in Experiment 1 wurde an den Beginn gestellt, um sicherzustellen, dass

keine Wiederholung von Zielwörtern Einfluss auf die Neuigkeitsbeurteilung der Versuchspersonen nehmen konnte. Die Natürlichkeitsbewertung der modellierten F0-Konturen in Experiment 4 erfolgte erst an vierter Stelle, um die Versuchspersonen allgemein mit resynthetisierten Sprachstimuli vertraut zu machen, so dass sie bei ihren Natürlichkeitsurteilen etwaige Artefakte auf segmentaler Ebene besser zu ignorieren in der Lage waren. Experiment 5 zur Beurteilung der Sprecherintention konnte ebenfalls erst nach Experiment 1 bis 3 ausgeführt werden, um die Versuchspersonen zunächst für linguistische Funktionen der Intonation zu sensibilisieren.

13.4.2 Stimuli

Bei den Stimuli handelte es sich in den Experimenten 1 bis 3 um lokale Segmente gemäß der Definition des PKS-Modells, also um eine Sequenz von Funktionswörtern mit abschließendem Kernwort, über der variierte lokale Intonationskonturen realisiert wurden. Die Funktionswortsequenz bildet hierbei den Trägersatz, und das Kernwort entsprach dem Zielwort des Stimulus.

Auswahlkriterien der Zielwörter

Zur weitestmöglichen Konstanthaltung von Einflussfaktoren auf phonologischer und lexikalischer Ebene (gemeint sind unter anderem Rhythmus, Konturdiskontinuitäten über stimmlosen Abschnitten, morphologische Komplexität, Wortsemantik und Worthäufigkeit) wurden die Zielwörter aus dem Deutsch-Teilkorpus des Celex (Baayen et al., 1995) und deutschsprachigen Zeitungstexten des *European Corpus Initiative Multilingual Corpus I (ECI/MCI)* (Elsnet, 2008) nach folgenden Kriterien ausgewählt:

- Sie setzen sich ausschließlich aus stimmhaften Segmenten zusammen, um diskontinuierliche Intonationskonturen zu vermeiden.
- Sie besitzen dieselben rhythmischen Eigenschaften: zweisilbig mit Betonung auf der ersten Silbe.
- Die akzentuierte Silbe ist vokalauslautend und hat als Silbenkern einen Langvokal oder Diphthong.
- Es handelt sich durchweg um Substantive, um einen etwaigen Einfluss der Wortart auf die Antworten auszuschließen. Zudem eignen sich andere Wortarten weniger gut beispielsweise bei der Beurteilung informativer Neuheit.
- Es handelt sich um semantische Konkreta.
- Es handelt sich morphologisch um Simplex-Formen, die
- alle eine festgelegte Vorkommenshäufigkeit überschreiten, um den Einflüssen morphologischer Komplexität und der Wortfrequenz auf die Beurteilung des semantischen Gewichts entgegenzuwirken. Die Häufigkeitswerte wurden alternativ aus dem Celex oder dem ECI/MCI gelesen, als unterer Schwellwert wurde 10 festgelegt.

- Sie besitzen im Allgemeinen keine Konnotationen, die starke Emotionen auslösen.
- Zur Vereinheitlichung des Trägersatzes tragen die Zielwörter weibliches Genus.

In Anhang C.1 findet sich die Liste der nach diesen Kriterien gewählten 60 Zielwörter.

Generierung

Die Stimuli für die Teilexperimente 1 bis 3 wurden mittels Mbrola (Dutoit et al., 1996), das in der Distribution des Festival-Sprachsynthesesystems (Black und Taylor, 1997) verfügbar ist, generiert. Hierfür wurde die männliche Stimme *de4* gewählt, da für diese vergleichsweise viele Units zur Verfügung standen, was sich positiv in der Qualität der konkatenativen Synthese niederschlägt. Input für die Mbrola-Generierung waren eine Transkription, die manuell angefertigt wurde, sowie für jedes Phon die Spezifizierung der Dauer und des F0-Verlaufs.

Dauermodellierung

Modell Den Lautdauern der synthetischen Stimuli liegt folgendes Modell zu Grunde:

$$\hat{d}_x \quad = \quad \bar{d}_x \cdot f. \tag{13.1}$$

$\hat{d}_x$ ist hierbei die prädizierte Dauer des Lauts x, $\bar{d}_x$ ist in Anlehnung an Klatt (1979) sowie Brinckmann und Trouvain (2003) die *intrinsische* Dauer von x. Während diese bei Brinckmann und Trouvain (2003) dem Mittelwert aller x-Realisierungen entspricht, wurde hier auf Grund des relativ geringen Umfangs an Trainingsdaten zunächst eine Gruppierung von Phonemen mit erwartbar homogenen Dauerwerten vorgenommen und die intrinsische Dauer jedes Phonems gleich dem Dauermittelwert der entsprechenden Gruppe gesetzt. Angaben zu Klassifizierung und intrinsischen Dauern der Phoneme sind in Anhang B zu finden.

Die Werte des Faktors f zur Anpassung der intrinsischen Dauer an die aktuellen Erfordernisse werden in dieser Arbeit mittels eines Regressionsbaums (Breiman et al., 1984) vorhergesagt, der dafür die Attribute *Akzentuierung, Phrasenfinalität* und *Lautklasse* nutzt. Auf Grund der geringen Menge an Trainingsdaten konnten im Sinne der Modellrobustheit nicht alle bekannten Einflussfaktoren (beispielsweise die Lautklasse des folgenden Lauts) auf die Dauer berücksichtigt werden. Regressionsbaum sowie die Werte der verwendeten Attribute sind ebenfalls in Anhang B aufgeführt.

Daten und Evaluierung Dem Training des Modells liegt ein handsegmentierter und prosodisch etikettierter Teil des SI1000P-Korpus zugrunde, der 2680 Segmente umfasst. Die Etiketten zur prosodischen Struktur umfassen Labels für Haupt- und Nebenakzente sowie für starke und schwache Phrasengrenzen. Schwache und starke Akzente sowie Phrasengrenzen wurden jeweils zu einer Kategorie zusammengefasst. Auf Grund der geringen Menge an Trainingsdaten wurde keine Trennung in Training und Testkorpus

vorgenommen, das Modell also an alle verfügbaren Daten angepasst. Insofern ist hier keine aussagekräftige Evaluierung des Modells möglich und auch nicht erklärtes Ziel. Die mittlere absolute Distanz zwischen Original- und vorhergesagten Dauern beträgt in den Trainingsdaten 17 ms.

In einem informellen Vorexperiment wurden die vorhergesagten Lautdauern in den synthetisierten Stimuli von drei phonetischen Experten als natürlich beurteilt.

Grundfrequenzmodellierung

Zu jedem der zu generierenden lokalen Segmente wurden fünf intonatorische Varianten erzeugt, eine für jede lokale Konturklasse. Hierzu wurde die Baseline konstant auf 80 Hz mit Deklinationssteigung gleich 0 gesetzt, so dass die F0-Bewegung ausschließlich durch die lokale Konturklasse bestimmt war. Die F0-Kontur ergab sich somit für jede Klasse durch Einsetzen des entsprechenden Polynoms in das zeitnormalisierte Segment, wobei der zeitliche Nullpunkt wie im PKS-Modell vorgesehen auf die Mitte des Silbenkerns der akzentuierten Silbe im Zielwort gelegt wurde. Zusätzlich zu den fünf Konturklassen wurden fünf Distraktor-Konturen in Form von Mittelwertkonturen von jeweils drei Konturklassen generiert. Diese Distraktoren sollten zum einen die Anzahl der hinsichtlich der Beurteilung ambigen Fälle erhöhen und zum anderen einer mit dem Erlernen der fünf Konturklassen einhergehenden Strategiebildung der Versuchspersonen entgegenwirken.

13.4.3 Methode

Versuchspersonen Es nahmen dieselben 24 Versuchspersonen teil, die bei der bereits in Abschnitt 11.2 beschriebenen perzeptiven PKS-Evaluierung mitgewirkt hatten.

Instruktionen Den Versuchspersonen wurde zu Beginn der Experimentreihe eine vierseitige Anleitung mit Instruktionen zu allen Teilexperimenten ausgehändigt, mit der Anweisung, vor jedem der Teilexperimente den entsprechenden Abschnitt der Anleitung durchzulesen. Insbesondere wurden die Versuchspersonen vorab instruiert, sich bei ihren Urteilen möglichst nur auf die Sprechmelodie zu konzentrieren und Resynthesearteefakte auf Lautebene nach Möglichkeit zu ignorieren. Die Instruktionen sind in Anhang D abgedruckt.

Präsentation Das Experiment selbst fand am Computer statt. Die Präsentation der Stimuli erfolgte über geschlossene Kopfhörer und mittels Perl-Tk erstellten Oberflächen, deren Screenshots in Anhang E zu sehen sind. Zur Beurteilung der Stimuli standen den Versuchspersonen fünfstufige anklickbare Likert-Skalen zur Verfügung, wie sie häufig in Perzeptionstests zur Intonation Verwendung finden (Birch und Clifton, 1995; Welby, 2003). Die Endpunkte der Skalen waren jeweils mit gegensätzlichen Antwortalternativen versehen. Auf ein *Counter-Balancing* der Endpunkte, wurde verzichtet, nachdem eine informelle Vorstudie ergab, dass Versuchspersonen dadurch zu stark von ihrer eigentlichen Aufgabe abgelenkt wurden. Jede der beiden Antwortalternativen war also über ein Teilexperiment hinweg konstant an einem Skalenende positioniert. Die Versuchspersonen

wurden angewiesen, im Falle relativer Sicherheit die entsprechenden Endpunkt auszuwählen, im Falle des Tendierens zu einem der Antwortalternativen die Knöpfe halb-links, beziehungsweise -rechts, und im Falle der Unentschiedenheit den mittleren Knopf. Es bestand die Möglichkeit, vor Fällen des Urteils die Stimuli beliebig oft anzuhören. Vor jedem Teilexperiment fand eine kurze Trainingsphase statt, in der den Versuchspersonen Stimuli derart präsentiert wurden, dass jede der Intonationsklassen einmal vorkam.

Randomisierung In den Teilexperimenten 1 bis 3 wurden Versuchspersonen lokale Segmente mit den 60 Zielwörtern in zufälliger Reihenfolge und ohne Zielwortwiederholung präsentiert. Die randomisierte Zuordnung der lokalen Konturklassen und der Distraktoren zu den Stimuli gehorchte den folgenden Constraints:

- Jede der fünf Konturklassen wurde insgesamt neunmal präsentiert.
- Jede der fünf Distraktorklassen wurde insgesamt zweimal präsentiert
- Aufeinanderfolgende Stimuli durften nicht dieselbe Intonationskontur tragen, um Abhängigkeiten bei aufeinanderfolgenden Antworten zu vermeiden.

Daraus ergab sich für die Versuchspersonen ein Umfang von 55 Trials pro Teilexperiment. Die verbleibenden 5 Zielwörter bildeten die Trainings-Items.

Weitere Angaben Das Experiment dauerte insgesamt etwa 40 Minuten, wobei es den Versuchspersonen jederzeit freistand, eine Pause zu machen. Die Teilnahme wurde mit 10 Euro für Studenten und kleinen Sachgeschenken für Mitarbeiter vergütet.

In den folgenden Kapiteln werden nun Korpusanalysen, Hypothesen sowie deren perzeptive Überprüfung für die Konzepte Bedeutsamkeit, informative Neuheit und Finalität vorgestellt.

Kapitel 14

Semantisches Gewicht

Wie in Kapitel 13 angekündigt wird zunächst die hier vollzogene Modellierung des semantischen Gewichts beschrieben, gefolgt von korpusstatistischen Befunden über dessen Zusammenhang mit den Intonationsklassen und den Hypothesen, die sich daraus ableiten lassen. Im Anschluss daran erfolgt die Beschreibung des Perzeptionsexperiments zur Überprüfung der Hypothesen.

14.1 Modellierung

14.1.1 Vorhersagbarkeit

Das semantische Gewicht eines Worts wird in dieser Arbeit im Sinne von Bolinger (1972) in Abhängigkeit seiner Vorhersagbarkeit aus dem Kontext ausgedrückt: je weniger vorhersagbar, desto höher sein Gewicht. Diese Sichtweise bietet den Vorteil einer probabilistischen Modellierung des Gewichts, wie sie auch im Rahmen der textbasierten Akzentlokalisierung beispielsweise von Pan und McKeown (1999) sowie Pan und Hirschberg (2000) zum Einsatz kam. Die Modellierung der globalen (kontextunabhängigen) sowie der lokalen (kontextabhängigen) Vorhersagbarkeit erfolgt anhand eines linear interpolierten Wahrscheinlichkeitsmodells, mit dem die Wahrscheinlichkeit $Pr(w)$ einer Wortfolge w allgemein folgendermaßen gegeben ist:

$$Pr(w) \quad = \quad \prod_i \sum_j \lambda_j \cdot P_j(w_i), \tag{14.1}$$

Die Wahrscheinlichkeit für Wort w_i speist sich hier aus mehreren Quellen P_j, die mit λ_j gewichtet und aufsummiert werden.

In dieser Arbeit kommt ein linear interpoliertes Trigramm-Modell Pr zum Einsatz, womit die Vorhersagbarkeit eines Wortes an der Textstelle i folgendermaßen gegeben ist:

$$Pr(w_i) \quad = \quad \lambda_1 \cdot P(w_i) + \lambda_2 \cdot P(w_i|w_{i-1}) + \lambda_3 \cdot P(w_i|w_{i-2}, w_{i-1}). \tag{14.2}$$

Die Unigramm-Wahrscheinlichkeit $P(w_i)$ für Wort w_i repräsentiert die kontextunabhängige globale Komponente der Vorhersagbarkeit von w_i. Seine kontextabhängige lokale Komponente liegt in der Bigramm- und Trigrammwahrscheinlichkeit ($P(w_i|w_{i-1})$, $P(w_i|w_{i-2}, w_{i-1})$) gegeben die Wortvorgeschichte w_{i-2}, w_{i-1}.

14.1.2 Gewinnung des Wahrscheinlichkeitsmodells

Korpus

Der Entwicklung der Wahrscheinlichkeitsmodelle lagen das SI1000P-Korpus sowie deutschsprachige Teile des ECI/MCI zugrunde. Die Daten umfassten rund 328000 Wort-Tokens und 44700 Wort-Types.

Smoothing

Um eine Überadaption des Wahrscheinlichkeitsmodells an die vorliegenden Daten zu verhindern und es somit realistischer und robuster gegenüber ungesehenen Daten zu machen, wurden die Häufigkeitswerte zur Reservierung von Wahrscheinlichkeitsmasse für ungesehene Ereignisse mittels Good-Turing-Smoothing (Good, 1953) gemäß Gleichung 14.3 angepasst.

$$\begin{aligned} c > k: \quad c^* &= c \\ \text{else}: \quad c^* &= \frac{(c+1)\frac{N_{c+1}}{N_c} - c\frac{(k+1)N_{k+1}}{N_1}}{1 - \frac{(k+1)N_{k+1}}{N_1}} \end{aligned} \tag{14.3}$$

N_c steht für die Anzahl der N-Gramm-Types, die c Mal auftreten Die Obergrenze anzupassender Häufigkeitswerte k wurde auf 5 gesetzt. Aus den Anpassungen ergibt sich folgende geschätzte Häufigkeit für ungesehene Ereignisse:[1]

$$0^* = \frac{1^*}{N} \tag{14.4}$$

Tabelle 14.1 zeigt die neugeschätzten Häufigkeitswerte.

Lineare Interpolation

Zur Berücksichtigung lokaler und globaler Vorhersagbarkeit wurde die Trigrammwahrscheinlichkeit mittels linearer Interpolation wie in Gleichung 14.2 in Uni-, Bi- und Trigrammkomponente unterteilt:

Das Interpolationsgewicht λ_j ist hierbei die erwartete relative Häufigkeit $E(M_j|w_{1,n})$, mit der Wahrscheinlichkeitsmodell M_j für die beobachtete Wortfolge $w_{1,n}$ in einem Entwicklungskorpus zu wählen ist. Die dieser zu erwartenden Häufigkeit zugrundeliegenden

[1] Im Falle fehlender N_c-Werte können diese nach Gale und Sampson (1995) mittels Interpolation durch linearer Regression folgendermaßen approximiert werden: $\log(N_c) = a + b \cdot \log(c)$, wobei die Koeffizienten a und b anhand der beobachteten c-N_c-Paare zu schätzen sind.

c	c^* (Unigramme)	c^* (Bigramme)	c^* (Trigramme)
0	$8.1 \cdot 10^{-6}$	$1.26 \cdot 10^{-5}$	$1.44 \cdot 10^{-5}$
1	0.62	0.43	0.22
2	1.56	1.25	0.84
3	2.53	2.16	1.63
4	3.51	3.11	2.49
5	4.50	4.08	3.40

Tabelle 14.1: Good-Turing-Neuschätzung c^* der Unigramm- und Bigramm-Häufigkeitswerte c.

Wahrscheinlichkeiten $P_j(w_i)$ (Wahrscheinlichkeit des Worts w_i im Modell M_j) werden hierzu vorab in einem Trainingskorpus geschätzt.

Zur Gewinnung der Interpolationsgewichte λ_j wurde das Textkorpus wie bei der Kreuzvalidierung in vier gleich große Partitionen unterteilt. In vier Schritten wurden jeweils drei Partitionen zu einem Trainingskorpus zur Schätzung der N-Gramm-Wahrscheinlichkeiten zusammengefasst, während die verbleibende Partition als Entwicklungskorpus dafür herangezogen wurde, auf Grundlage der ermittelten Wahrscheinlichkeitsmodelle mittels des *Expectation-Maximisation*-Algorithmus (Dempster et al., 1977) iterativ die zugehörigen Interpolationsgewichte zu berechnen. Schließlich ergaben sich die in Tabelle 14.2 gezeigten endgültigen Interpolationsgewichte λ_j als arithmetische Mittelwerte der so erzeugten vier Gewichts-Tripel.

Gewicht	Wert
λ_1	0.54
λ_2	0.44
λ_3	0.02

Tabelle 14.2: Interpolationsgewichte.

Evaluierung

Die Kreuzentropierate des so entwickelten Wahrscheinlichkeitsmodells beträgt auf den Trainingsdaten 6.55. Auf eine Evaluierung des Modells auf ungesehenen Testdaten wurde verzichtet, da nicht geplant war, hieraus Konsequenzen im Hinblick auf die zu stellenden Hypothesen zu ziehen.

14.2 Korpusstatistik und Hypothesen

14.2.1 Befunde

Interpretation der Stilisierungsparameter

Die Korrelationen zwischen Polynomkoeffizienten und Vorhersagbarkeit der Kernwörter sind zwar signifikant ($p < 0.001$), jedoch zu gering ($|r| \leq 0.3$) um daraus tragfähige Schlüsse ableiten zu können. Dasselbe gilt für F0-Spannweite und -maximum sowie für Messung der Korrelationen nach Arkussinus- oder Log-Transformation der Wahrscheinlichkeitswerte.

Semantisches Gewicht und Konturklassen

Abbildung 14.1 zeigt die Kernwort-Trigrammwahrscheinlichkeiten der einzelnen Konturklassen in Form von Boxplots.

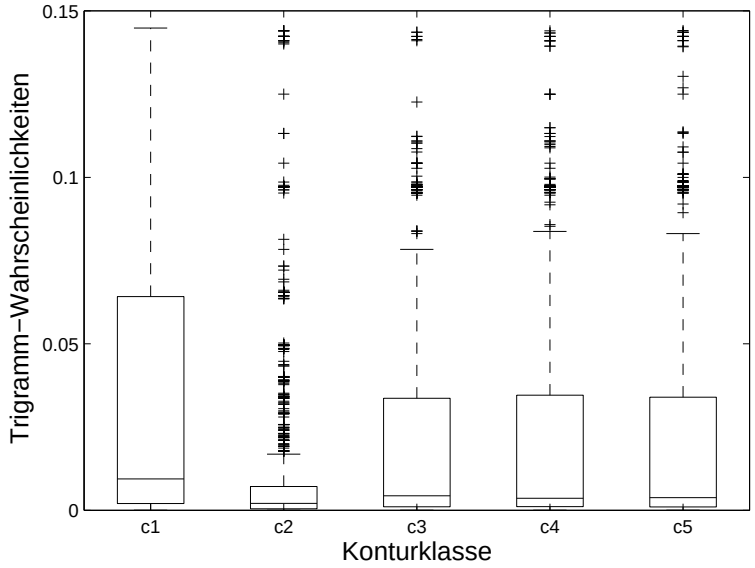

Abbildung 14.1: Boxplots der klassenabhängigen Vorhersagbarkeit der segmentfinalen Kernwörter in Form von linear interpolierten Trigrammwahrscheinlichkeiten.

Eine einfaktorielle Varianzanalyse (ANOVA) ergab signifikante Unterschiede der Wahrscheinlichkeitsmittelwerte ($F[4, 9214] = 31.7$, $p < 0.001$). Post hoc konnte die Klasse c_2 mit im Vergleich zu allen anderen Klassen signifikant niedrigerer Trigrammwahrscheinlichkeit identifiziert werden (Tukey-Kramer-Post-hoc-Test, $\alpha = 0.001$), und die Klasse c_1 als die Klasse mit signifikant höherer Wahrscheinlichkeit ($\alpha = 0.005$). Für die klassenabhängige Vorhersagbarkeit des Kernworts $\text{pred}(c_n)$ ergibt sich somit folgende Rangordnung:

$$\text{pred}(c_2) < \text{pred}(c_3), \text{pred}(c_4), \text{pred}(c_5) < \text{pred}(c_1)$$

Eine Arkussinus-Transformation der Wahrscheinlichkeiten bewirkte deutlichere Mittelwertunterschiede ($F[4, 9214] = 54.3$, $p < 0.001$), aber keine relevanten Änderungen im Post-hoc-Ergebnis.

14.2.2 Hypothesen

Gegeben der umgekehrte Zusammenhang zwischen Vorhersagbarkeit und semantischem Gewicht lassen sich die folgenden Hypothesen formulieren:

H1 Klasse c1 codiert geringe Bedeutsamkeit.

H2 Klasse c2 codiert hohe Bedeutsamkeit.

14.3 Perzeptive Validierung

14.3.1 Methode

Den Versuchspersonen wurden Aussagen der Form *Das ist eine X* mit variierten Zielwörtern und Intonationsklassen wie in den Abschnitten 13.4.2 und 13.4.3 beschrieben und in Abbildung 14.2 gezeigt präsentiert. Die Aufgabe bestand darin, die Stimuli auf einer fünfstufigen Likert-Skala mit den Endpunkten *belanglos* und *bedeutsam* hinsichtlich der vom Sprecher vermeintlich beigemessenen Relevanz der Aussage zu beurteilen.

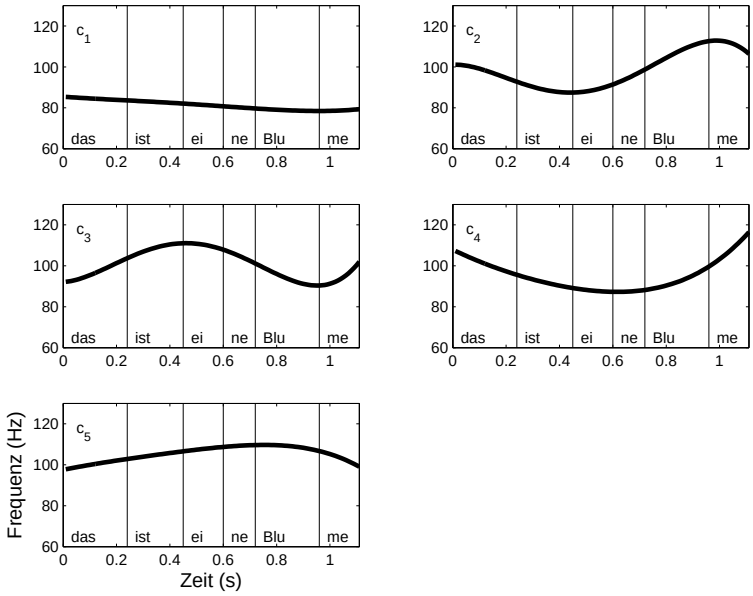

Abbildung 14.2: Stimulusbeispiel für jede Konturklasse zur Untersuchung der perzipierten Bedeutsamkeit.

14.3.2 Ergebnisse

Klasse	Median	arithm. Mittel	Interquartilsabstand	Standardabweichung
c_1	1	1.58	1	0.83
c_2	4	4.36	1	0.73
c_3	4	3.65	1	0.93
c_4	4	3.52	1	0.93
c_5	4	3.52	1	0.93

Tabelle 14.3: Mittelwerte und Streuungsmaße der Beurteilung der Konturklassen hinsichtlich Bedeutsamkeit.

Klassenabhängige Bedeutsamkeit

Abbildung 14.3 zeigt die Bedeutsamkeitsurteile in Abhängigkeit der Konturklassen in Form von relativen Häufigkeiten und Boxplots. Es bestehen hochsignifikante klassenabhängige Unterschiede (Kruskal-Wallis-Test, $\chi^2_4 = 515.36$, $p < 0.001$). Der Dunnett-Post-hoc-Test lokalisiert diese Unterschiede hinsichtlich c_1 mit niedrigerer perzipierter Bedeutsamkeit gegenüber den anderen Klassen und c_2 mit höherer Bedeutsamkeit ($\alpha = 0.01$), was folgende signifikante Abstufung hinsichtlich der Bedeutsamkeitsurteile $\mathrm{wgt}(c_n)$ ergibt:

$$\mathrm{wgt}(c_1) < \mathrm{wgt}(c_3), \mathrm{wgt}(c_4), \mathrm{wgt}(c_5) < \mathrm{wgt}(c_2)$$

Für c_1 sind signifikant niedrigere Werte und für c_2 signifikant höhere Werte als 3 (*unentschieden*) festzustellen (einseitige Vorzeichentests für eine Stichprobe zum Medianvergleich, c_1: $z = -13.26$, $p < 0.001$; c_2: $z = 13.37$, $p < 0.001$). Zudem erreicht die perzipierte Bedeutsamkeit für die verbleibenden Klassen c_3, c_4 und c_5 ebenfalls signifikant höhere Werte als 3, weshalb diese sich ebenfalls mit Bedeutsamkeit assoziieren lassen (einseitige Vorzeichentests für eine Stichprobe zum Medianvergleich, Signifikanzniveau $\alpha = 0.05$ Bonferroni-korrigiert, $z > 7.27$, $p < 0.001$).

Urteilskonsistenz

Vergleich mit Zufallsniveau Zur Untersuchung der Beurteilungskonsistenz wurde der Interquartilsabstand (die *inter quartile range IQR*) als Streuungsmaß für ordinalskalierte Daten zur Messung der Inkonsistenz herangezogen. Hierzu wurden für jede Konturklasse getrennt die IQRs aller Versuchspersonenurteile gesammelt und die Mittelwerte dieser Stichproben mit der bei Zufallsantworten zu erwartenden IQR verglichen.

Als zufällige IQR wurde unter Annahme einer Gleichverteilung der Wahrscheinlichkeiten für die Antwortalternativen 1–5 ein Wert von 2.4 ermittelt. Dieser Wert ergibt sich als Mittelwert der IQRs der fünf möglichen Antwort-Kombinationen, in denen die fünf

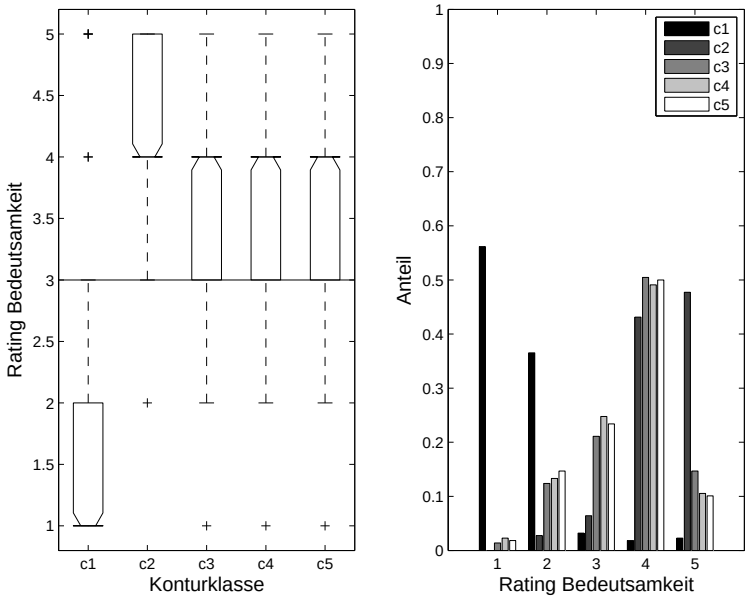

Abbildung 14.3: **Links:** Boxplots zur Beurteilung der Bedeutsamkeit in Abhängigkeit der lokalen Konturklasse. **Rechts:** relative Häufigkeiten der Urteile 1 – 5 für jede Konturklasse.

Antwort-Alternativen möglichst ausgeglichen auf neun Positionen (also die Anzahl der Präsentationen je Klasse) verteilt werden. Er entspricht der Inkonsistenz einer ratenden Versuchsperson ohne Antwort-Bias.

Ein Vergleich der klassenabhängigen IQRs mit diesem Zufalls-Inkonsistenzwert ergab, dass alle Klassen mit signifikant niedrigerer IQRs, also höherer Konsistenz, beurteilt werden konnten (einseitige Vorzeichentests für eine Stichprobe zum Medianvergleich, $z < -4.69$, $\alpha = 0.05$ Bonferroni-korrigiert, $p < 0.001$.[2] Siehe Abbildung 14.4).

Zusammen mit dem signifikanten Unterschied der klassenabhängigen Urteilsmittelwerte der mittleren Bewertungsstufe zeugt dieser Befund von einer robusten Beurteilbarkeit der Konturklassen hinsichtlich der perzipierten Bedeutsamkeit.

Paarweiser Vergleich Vergleicht man die Inkonsistenzen der Klassen untereinander (die IQRs sind in Tabelle 14.3 zu finden), so ist festzustellen, dass c_2 mit gegenüber den anderen Klassen signifikant geringerer IQR beurteilt wurde (paarweiser Levene-Test, α Dunn-Sidak-korrigiert, $p < 0.001$). Darüber hinaus sind hier keine signifikanten Unterschiede zwischen den Klassen zu finden.

Schlussfolgerungen

Festhalten lässt sich an dieser Stelle:

[2]Im Falle mehrfach durchgeführter Vorzeichentests (einer je Klasse) ist nur der am nächsten zu 0 befindliche z-Wert angegeben.)

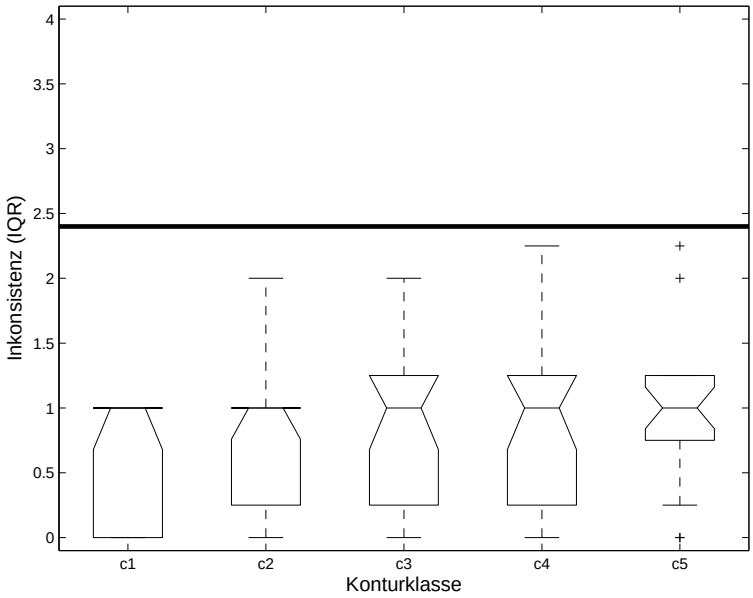

Abbildung 14.4: Klassenabhängige Urteilsinkonsistenz in Form von Interquartilsabständen (IQR) bei der Beurteilung der Bedeutsamkeit. Horizontale Linie: Inkonsistenz bei Zufallsantworten.

- Auf Parameterebene ließ sich kein interpretierbarer Zusammenhang zwischen Polynomkoeffizienten und semantischem Gewicht feststellen.
- Auf Symbolebene der Konturklassen konnte Hypothese **H1** bestätigt werden: c_1 codiert geringe Bedeutsamkeit.
- Hypothese **H2** wurde ebenfalls bestätigt: c_2 codiert hohe Bedeutsamkeit.
- Auch die restlichen Konturklassen konnten hinsichtlich der von ihnen codierten Bedeutsamkeit eingeordnet werden.
- Die Urteile fielen für alle Konturklassen und vor allem für c_2 konsistent aus.

Kapitel 15

Informative Neuheit

Wie das vorangehende Kapitel strukturiert sich auch dieses in Beschreibung der Modellierung des Informationsstatus, korpusstatistische Befunde über dessen Auswirkung auf die Intonation, daraus abgeleitete Hypothesen sowie deren experimentalphonetische Überprüfung.

15.1 Modellierung

In Abschnitt 5.2.2 bei der Vorstellung diskursbasierter Ansätze zur Akzentlokalisierung wurden drei Arten der Gegebenheit von Information unterschieden: (1) im Diskursverlauf bereits übermittelt, (2) zum geteilten Weltwissen gehörig und (3) aus dem situativen Kontext erschließbar. Mangels Möglichkeit der Modellierung von Weltwissen und in Anbetracht des konstanten situativen Kontexts *Verlesen von politischen Zeitungstexten* konzentrierte sich die Korpusanalyse dieser Arbeit auf die Identifizierung von im Diskursverlauf bereits übermittelter Information, also der ersten Art der Gegebenheit. Hierzu wurde eine Segmentierung des Nachrichtenkorpus in thematische Einheiten mit anschließender Koreferenzresolution innerhalb dieser Einheiten vorgenommen. In einem Text werden zwei Wörter als *koreferent* bezeichnet, wenn sie sich auf dasselbe außersprachliche Objekt (denselben Referenten) beziehen.

15.1.1 Allgemeines Verfahren

Zu Beginn jedes Themenblocks wird wie in Abbildung 15.1 gezeigt eine leere Diskursreferenten-Menge $\mathcal{R}$ erzeugt, die dann beim Durchlaufen des Blocks inkrementell aufgefüllt wird. Der Reihe nach wird hierbei für jedes im Block auftretende Nomen n nach seiner Normalisierung (siehe unten) geprüft, ob ein Eintrag in $\mathcal{R}$ enthalten ist, auf den n koreferiert. Falls ja, wird es als *informativ gegeben* markiert, falls nein, als *neu*. Nach dieser Überprüfung wird n in $\mathcal{R}$ aufgenommen. $\mathcal{R}$ lässt sich hierbei als vereinfachte, da ungeordnete Form eines *vorwärtsgerichteten Zentrums* im Sinne der *Centering*-Theorie verstehen (siehe Abschnitt 5.2.2).

Start eines neuen Themenblocks T:
Diskursreferentenmenge $\mathcal{R} = \{\ \}$

foreach Nomen n in T
    $n \leftarrow$ **normalisiere**(n)
    foreach $r \in \mathcal{R}$
        if coref_of(n, r): markiere n als **gegeben**
    end
    if $\neg(n$ **gegeben**$)$: markiere n als **neu**
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{n\}$
end

Abbildung 15.1: Verfahren zur Markierung neuer und gegebener Information. coref_of(n, r) bedeutet: "n ist koreferent zu r".

Problem Die in Abschnitt 9.1 dargelegte Aufnahmesituation des Korpus bringt folgendes Problem mit sich: da die Sätze einzeln aufgenommen wurden, ist nicht mehr nachzuvollziehen, inwieweit der Sprecher die Sätze als einzelne Themenblöcke realisiert oder zu Themenblöcken zusammenfasst, um darin neue von gegebener Information zu unterscheiden. Es ist nicht zu klären, ob er – in der Terminologie des oben beschriebenen Verfahrens – die Diskursreferentenmenge $\mathcal{R}$ vor jedem Satz oder erst zu Beginn des nächsten Themenblocks leert. Daher besteht Ungewissheit darüber, ob der Sprecher innerhalb desselben Themenblocks über Satzgrenzen hinweg Koreferenzen entsprechend intonatorisch kennzeichnet.

Aus diesem Grund wurden in dieser Arbeit zwei separate Textsegmentierungen mit entsprechend unterschiedlichen resultierenden Koreferenzbeziehungen vorgenommen: eine satzweise Segmentierung und eine thematische Segmentierung. Zunächst wird nun Textvorverarbeitung und thematische Segmentierung mittels automatisierter diskursanalytischer Verfahren beschrieben und im Anschluss daran die Koreferenzresolution innerhalb der ermittelten Textsegmente.

15.1.2 Vorverarbeitung: Wortnormalisierung

Die Wörter im Text wurden auf Kleinbuchstaben normalisiert und in Abhängigkeit des Verwendungszwecks auf zweierlei Arten ihrer Flektionsendungen und Suffixe entledigt:

- Zur Diskurssegmentierung erfolgte ein string-basiertes Stemming, das darin besteht, wortfinal die längstmögliche Buchstabenfolge aus der für Endungen im Deutschen üblichen Buchstabenmenge {*e, n, r, s, t*} vom Wortende zu entfernen, sofern der verbleibende Wortstring mindestens einen Vokal und eine Mindestlänge von vier Buchstaben aufweist. Beispiel: *kratzten* $\longrightarrow$ *kratz*.

- Zur Koreferenzresolution beruhte das Stemming auf einer automatisierten morphologischen Analyse nach Reichel und Weilhammer (2004) und Reichel (2005b) mit

anschließender Entfernung der als Suffix oder Flektionsendung identifizierten Wortteile. Beispiel: *Vorhaltungen* $\longrightarrow$ *vor/PRFX + halt/V + ung/SFX + en/INFL* $\longrightarrow$ *vorhalt*.

15.1.3 Diskurssegmentierung

Zur Segmentierung des Texts wurde auf den von Hearst (1997) entwickelten *TextTiling*-Algorithmus zurückgegriffen. Dieses Verfahren besteht aus drei Komponenten:

- dem *Cohesion-Scorer*,
- dem *Depth-Scorer* und
- dem *Boundary-Selector*.

Cohesion-Scorer Diese Komponente misst das Ausmaß der Themenkontinuität anhand der Textähnlichkeit benachbarter, durch eine Satzgrenze getrennter Textfenster. Die Länge der Fenster wurde in dieser Arbeit auf 35 gesetzt. Zur Ermittlung der Ähnlichkeit wurden die Textfenster F_x und F_y als binäre Term-Vektoren repräsentiert, gewichtet mit dem Informationsgehalt der Terme:

$$\begin{aligned} F_x &= [I(w_1) \cdot E_x(w_1), I(w_2) \cdot E_x(w_2), \ldots I(w_n) \cdot E_x(w_n)] \\ F_y &= [I(w_1) \cdot E_y(w_1), I(w_2) \cdot E_y(w_2), \ldots I(w_n) \cdot E_y(w_n)], \end{aligned}$$

wobei $I(w_i)$ den Informationsgehalt des Terms $w_i \in V$ bezeichnet, und $E_x(w_i)$ den Wert 1 annimmt, wenn sich Term w_i im Fenster f_x befindet, ansonsten 0. Das Lexikon V umfasst hierbei alle im SI1000P-Korpus auftretenden Substantive und Eigennamen.

Der Informationsgehalt des Terms w_i bezeichnet die nötige Anzahl an Bits zur Codierung von w_i:

$$I(w_i) = -\log_2 P(w_i) \text{ [Bit]} \tag{15.1}$$

Er ist umso höher, je niedriger die Auftretenswahrscheinlichkeit von w_i. Seine Verwendung ist dadurch motiviert, das sich selten auftretende Wörter allgemein eher zur Themenidentifizierung und damit zu deren Abgrenzung gegeneinander eignen als häufige Wörter.

Als Metrik zum Vergleich von F_x und F_y wurde die Cosinus-Ähnlichkeit gewählt:

$$s = \frac{F_x \cdot F_y}{||F_x|| \cdot ||F_y||} \tag{15.2}$$

Der Cohesion Scorer liefert auf diese Weise für jedes Paar aufeinanderfolgender Textsegmente einen Kohäsionswert. Die Sequenz dieser Kohäsionswerte wurde wie in Manning und Schütze (2001) vorgeschlagen zur gewünschten Vernachlässigung gering ausgeprägter Text-Diskontinuitäten mit einem Moving-Average-Filter der Fensterlänge 3 geglättet.

Depth-Scorer Der Depth Scorer ermittelt die Tiefen der lokalen Minima in dieser Kohäsionswert-Sequenz.

Boundary-Selector Der Boundary-Selector entscheidet anhand der Ausgabe des Depth-Scorers, welche benachbarten Segmente ausreichend unähnlich sind, um dazwischen einen Themenwechsel anzunehmen. Der zu überschreitende Schwellwert s ist abhängig von Mittelwert μ und Standardabweichung σ aller ermittelten *Depth-Scores*: $s = \mu - c \cdot \sigma$, wobei c auf 0.5 gesetzt wurde.

Heuristiken Das TextTiling-Verfahren wurde in dieser Arbeit durch folgende Heuristik ergänzt: Themenfortsetzung ist gekennzeichnet durch satzinitiale Konjunktionen, vor dem ersten Nomen auftretende Pronomen und Pronominaladverbien.

Performanz In einer Vorstudie auf einem anderen Nachrichtenkorpus, dem *IMS Radio News Corpus* (Rapp, 1998b) klassifizierte das Verfahren 90 % von 103 Satzpaaren korrekt als Themenwechsel beziehungsweise -fortführung.

15.1.4 Koreferenzresolution

Innerhalb der extrahierten Texteinheiten (Sätze beziehungsweise Themenblöcke) wurden im nächsten Schritt Koreferenzrelationen zwischen Substantiven identifiziert in Form von Hyperonym-Hyponym-Paaren – mit dem Hyperonym als Koreferent des Hyponyms. In diesem Sinne bildet die Koreferenzialität K eine antisymmetrische und transitive Relation auf das Vokabular V.

Folgende Verfahren zur Hyperonym- und damit zur Koreferenzdetektion kamen in dieser Arbeit zum Einsatz: Kompositumanalyse und Textmusterverwertung. Auf Grund der transitiven Eigenschaft der Koreferenzrelation lassen sich die so extrahierten Hyperonym-Hyponym-Paare über ihre reflexiv-transitive Hülle miteinander verknüpfen.

Kompositumanalyse

Nach der Kompositazerlegung gemäß Reichel (2005a) werden weniger spezifische (morphologisch weniger komplexe) Teile als Hyperonyme und damit als Koreferenz (*coref_ of*) zu den spezifischeren gesetzt. Beispiel *Bundesinnenminister*

coref_of(*Minister, Innenminister*)
coref_of(*Minister, Bundesinnenminister*)
coref_of(*Innenminister, Bundesinnenminister*)

Dieses Vorgehen ist genaugenommen nur für endozentrische Determinativzusammensetzungen wie im obigen Beispiel adäquat, nicht aber für Kopulativ- (*Hosenrock*) oder exozentrische Komposita (*Rotkehlchen*), die im Ganzen keine Hyponyme ihrer finalen Komponenten darstellen. Auf Grund der stark eingeschränkten Produktivität dieser beiden Kompositionstypen (Fabricius-Hansen et al., 2009) können damit verbundene Fehler in der semantischen Analyse jedoch vernachlässigt werden.

Textmuster

Gemäß eines von Hearst (1992) vorgeschlagenen Verfahrens wurden Textmuster gesucht, in denen sich Koreferenzbeziehungen wiederfinden. Im verwendeten Text ließ sich nur ein solches Muster identifizieren:

***minister|professor|chef** *NAME** **NAME**

So führt beispielsweise die Wortfolge *Bundesinnenminister Kanther* zur folgenden Koreferenzrelation:

coref_of(*Bundesinnenminister, Kanther*)

Reflexiv-transitive Hülle

Mittels der reflexiv-transitiven Hülle über die Koreferenzrelation:

$$\forall a \in V \quad : \quad \text{coref\_of}(a, a) \tag{15.3}$$

$$\forall a, b, c \in V \quad : \quad (\text{coref\_of}(a, b) \wedge \text{coref\_of}(b, c)) \longrightarrow \text{coref\_of}(a, c) \tag{15.4}$$

lassen sich nun die mit den vorangehenden Verfahren ermittelten Koreferenzen folgendermaßen weiter verknüpfen:

coref_of(*Minister, Minister*)
coref_of(*Minister, Innenminister*)
coref_of(*Minister, Bundesinnenminister*)
coref_of(*Minister, Kanther*)
coref_of(*Innenminister, Innenminister*)
coref_of(*Innenminister, Bundesinnenminister*)
coref_of(*Innenminister, Kanther*)
coref_of(*Bundesinnenminister, Bundesinnenminister*)
coref_of(*Bundesinnenminister, Kanther*)
coref_of(*Kanther, Kanther*)

15.2 Korpusstatistik und Hypothesen

15.2.1 Befunde

Interpretation der Stilisierungsparameter

Die zugrundeliegende Einteilung der lokalen Segmente in informativ neu und gegeben erfolgte für die nachfolgend angeführten Untersuchungen anhand der TextTiling-, also nicht der satzweisen, Segmentierung. Abbildung 15.2 zeigt die Mittelwerte und Streuungen der Polynomkoeffizienten in Abhängigkeit des Informationsstatus. Nur für die allgemeine F0-Anhebung s_0, die bei neuer Information höher ist als bei gegebener, sind die Unterschiede signifikant (Welch-Test, $\alpha = 0.05$; für s_0: $t_{245} = 7.10$, $p < 0.005$).

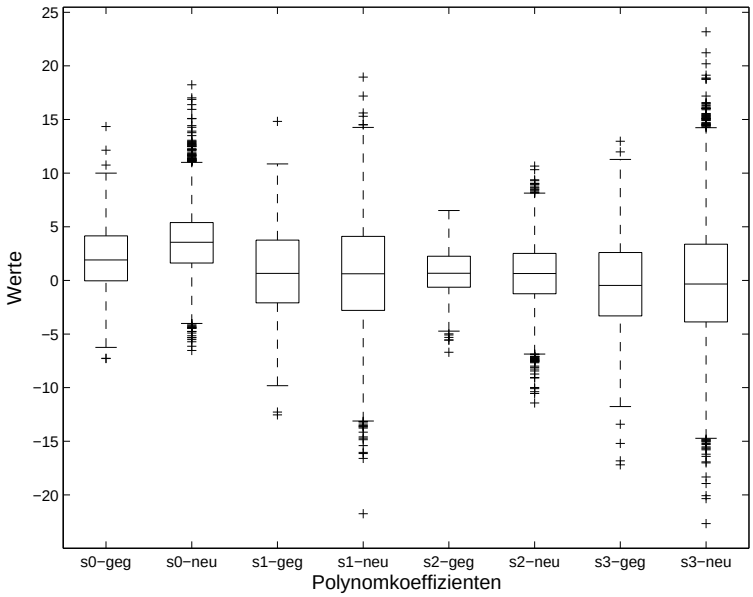

Abbildung 15.2: Polynomkoeffizienten s_j in Abhängigkeit des Informationsstatus; **geg**: gegebene Information, **neu**: neue Information.

Dagegen lässt sich der Informationsstatus intonatorisch gut aus den abgeleiteten Parametern *F0-Maximum* und *-Spannweite* der lokalen Konturen festmachen. Abbildung 15.3 zeigt für diese Kenngrößen die Unterschiede für Original- und modellierte Konturen. Im Falle der modellierten Konturen fallen die Unterschiede weniger stark aus, sind aber wie für die Originalkonturen signifikant (**F0-Maxima:** Original: Welch-Test, $t_{245} = 7.11$, $p < 0.005$; PKS: Welch-Test, $t_{243} = 5.13$, $p < 0.005$; **Spannweite:** Original: t-Test, $t_{6002} = -4.75$, $p < 0.001$; PKS: Welch-Test, $t_{248} = 2.79$, $p = 0.01$)

Das allgemeine Niveau von F0-Maxima und -Spannweiten ist in den modellierten Konturen signifikant niedriger, was von einem gegenüber dem Original flacheren Konturverlauf zeugt (t-Test für abhängige Stichproben, F0-Maxima: $t_{6003} = 83.51$, $p < 0.001$; Spannweite: $t_{6003} = 161.56$, $p < 0.001$).

Informative Neuheit und Konturklassen

Da nur für Nomen, also Substantive und Eigennamen, der Status bezüglich ihrer Neuheit festgelegt wurde, wurden zur folgenden Untersuchung nur lokale Segmente mit Nomen als Kernwort herangezogen. Die Konturklassenwahrscheinlichkeiten weichen entsprechend von den Angaben in Tabelle 10.3 ab.

Die Extrahierung von Zusammenhängen zwischen Konturklassen und neuer beziehungsweise gegebener Information erfolgte für jede Klasse auf Grundlage eines χ^2-Tests. Tabelle 15.1 zeigt die Testergebnisse für die beiden Formen der Textsegmentierung.

Zusätzlich angegeben sind für jede Konturklasse ihre bedingte Auftretenswahrscheinlichkeiten $P(\text{Klasse}|\text{gegeben})$, wenn das segmentfinale Kernwort gegebene Information

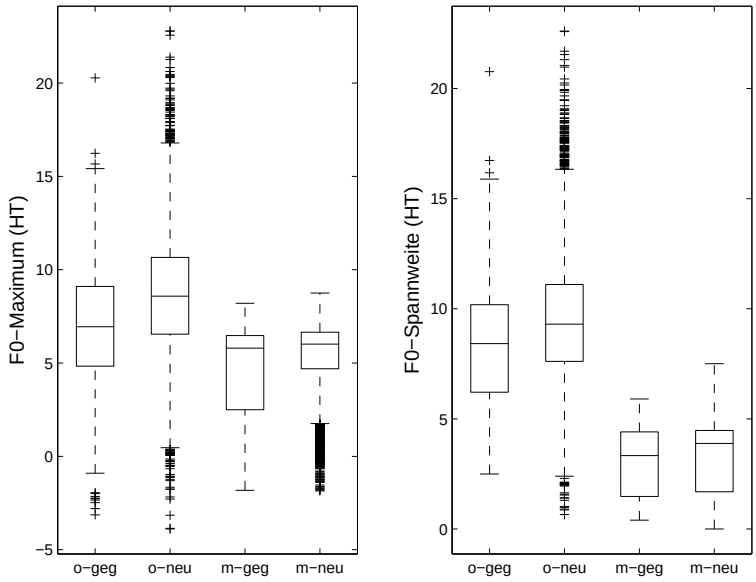

Abbildung 15.3: F0-Maxima und -spannweiten lokaler Konturen in Abhängigkeit informativer Neuheit (in Halbtönen, Basis 50 Hz); **o-**: Originalkontur, **m-**: modellierte Kontur, **geg**: gegebene Information, **neu**: neue Information.

trägt sowie P(Klasse|neu) für neue Information. Diese bedingten Wahrscheinlichkeiten sind der A-priori-Wahrscheinlichkeit P(Klasse) gegenübergestellt. Bei Überschreitung des kritischen χ^2-Werts gibt ein Vergleich dieser Wahrscheinlichkeiten Aufschluss darüber, ob die Konturklasse mit gegebener oder neuer Information assoziiert ist. Im ersten Fall gilt P(Klasse|gegeben) $> P$(Klasse), im zweiten P(Klasse|neu) $> P$(Klasse).

Beide Segmentierungen ergeben übereinstimmend Klasse c_1 als Träger gegebener Information. Die weiteren Befunde haben einen komplementären Klassenbezug und widersprechen sich daher nicht.

15.2.2 Hypothesen

Aus der Vereinigungsmenge der korpusstatistischen Befunde lassen sich zwei Hypothesen ableiten:

H3 Klassen c_1 und c_4 codieren die Übermittlung bereits gegebener Information.

H4 Klassen c_2, c_3 und c_5 codieren die Übermittlung neuer Information.

15.3 Perzeptive Validierung

15.3.1 Methode

Den Versuchspersonen wurden über Kopfhörer Stimuli der Form:

Textsegmentierung mittels TextTiling					
Klasse	Codierung	χ^2	P(Klasse|gegeben)	P(Klasse|neu)	P(Klasse)
c_1	gegeben	5.09*	0.22	0.18	0.21
c_2	–	0.92	0.18	0.19	0.18
c_3	neu	20.12*	0.15	0.20	0.19
c_4	gegeben	13.48*	0.21	0.25	0.22
c_5	–	1.07	0.20	0.21	0.20
Satzweise Textsegmentierung					
Klasse	Codierung	χ^2	P(Klasse|gegeben)	P(Klasse|neu)	P(Klasse)
c_1	gegeben	52.20*	0.38	0.19	0.21
c_2	neu	3.87*	0.14	0.19	0.18
c_3	–	2.72	0.15	0.19	0.19
c_4	–	0.13	0.21	0.22	0.22
c_5	neu	11.68*	0.12	0.21	0.20

Tabelle 15.1: Zusammenhang zwischen Konturklassen und Informationsstatus. **Oben:** bei Textsegmentierung mittels des TextTiling-Verfahrens, **unten:** bei satzweise Textsegmentierung. *: Zusammenhänge signifikant ($\alpha = 0.05$).

Ja, eine X (z. B. *Ja, eine Blume*)

mit variierten Zielwörtern und Intonationsklassen präsentiert, wie in den Abschnitten 13.4.2 und 13.4.3 beschrieben. Dazu wurden ihnen als Endpunkte einer fünfstufigen Likert-Skala visuell zwei mögliche Fragen gezeigt, auf die der Stimulus als Antwort verstanden werden kann:

- *Ist das eine X?* (*Ist das eine Blume?*)
- *Ist das ein Hyperonym(X)?* (*Ist das eine Pflanze?*)

Bezogen auf die erste Alternative enthält die Antwort über eine reine Bestätigung hinaus keine zusätzliche Information. Bezüglich der zweiten Alternative besteht die zusätzliche (neue) Information darin, einen Oberbegriff (*Pflanze*) zu konkretisieren (als *Blume*).

Die Aufgabe der Versuchspersonen bestand nun darin, auf der fünfstufigen Skala zu beurteilen, zu welcher der beiden Fragen die gegebene Antwort hinsichtlich ihrer Intonation besser passt, ob sie die Kontur also eher beim Auftreten neuer Information erwarten wie im Frage-Antwort-Paar

Ist das eine Pflanze? – Ja, eine Blume,

oder bei der Bestätigung gegebener Information, wie dies in

Ist das eine Blume? – Ja, eine Blume

geschieht.

Die präsentierten Antworten wurden mit der Antwortpartikel *Ja* eingeleitet, um ihre etwaige Deutung als Kontrast zu verhindern, das heißt im konkreten Beispiel, *Blume* sollte nicht als Kontrast zu *Pflanze* verstanden werden, sondern als Konkretisierung.

Stimulusbeispiele sind in Abbildung 15.4 zu finden. Die modellierte Kontur wurde nur über den zweiten Teil der Antwort, also über *eine X*, gelegt. Der lineare, von 90 auf 80 Hz fallende F0-Verlauf über der vorangestellten Antwortpartikel war über alle Stimuli konstant. Die Pause zwischen Antwortpartikel und restlicher Antwort betrug 300 ms.

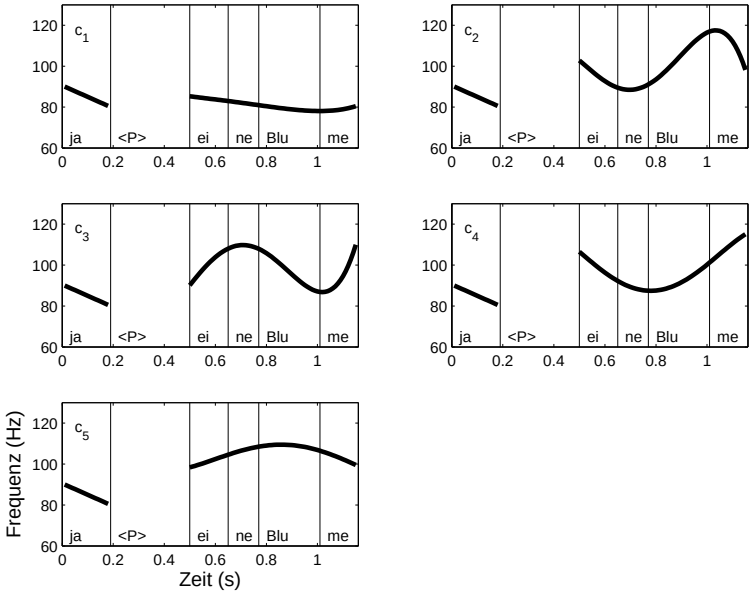

Abbildung 15.4: Stimulusbeispiel für jede Konturklasse zur Untersuchung der perzipierten Neuheit.

15.3.2 Ergebnisse

Klasse	Median	arithm. Mittel	Interquartilsabstand	Standardabweichung
c_1	1	2.02	2	1.41
c_2	4	4.03	1	1.14
c_3	4	3.36	2	1.38
c_4	4	3.50	1	1.07
c_5	4	3.46	2	1.22

Tabelle 15.2: Mittelwerte und Streuungsmaße der Beurteilung der Konturklassen hinsichtlich informativer Neuheit.

Klassenabhängige Neuheitscodierung

Abbildung 15.5 zeigt die Boxplots der klassenabhängigen Urteile, sowie deren relative klassenabhängige Häufigkeiten.

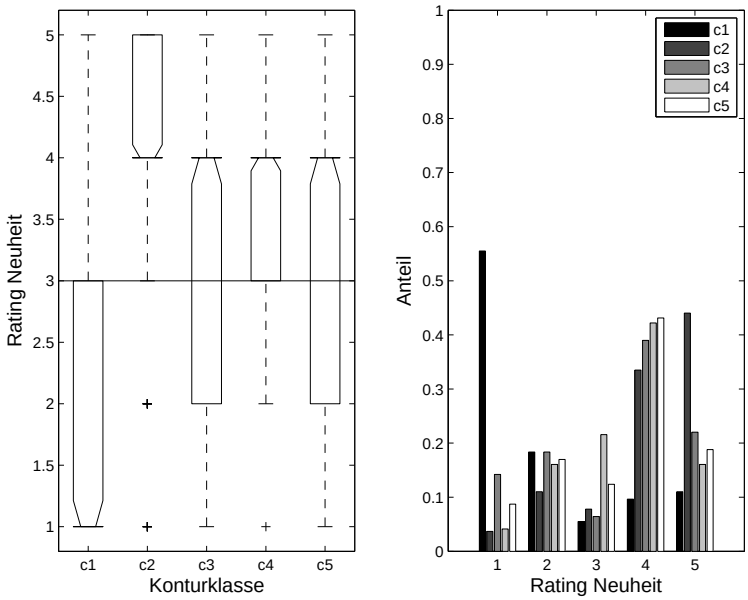

Abbildung 15.5: **Links:** Boxplots zur Beurteilung der Neuheit in Abhängigkeit der lokalen Konturklasse. **Rechts:** Relative Häufigkeiten der Urteile 1 – 5 für jede Konturklasse.

Es konnten signifikante klassenabhängige Unterschiede in der Neuheitsbeurteilung festgestellt werden (Kruskal-Wallis-Test, $\chi^2_4 = 217.12$, $p < 0.001$). Die mit Klasse c_1 verknüpfte mittlere Neuheit war signifikant niedriger als die der restlichen Klassen, die mit c_2 verknüpfte Neuheit gegenüber allen anderen Klassen signifikant höher (Dunnett-Post-Hoc-Test, $\alpha = 0.05$). Zwischen den Klassen c_3, c_4 und c_5 gab es keine signifikanten Unterschiede, wobei c_3 in dieser mittleren Gruppe die niedrigsten Werte aufwies. Für die Abstufung der Neuheitsurteile $\mathrm{nov}(c_n)$ ergab sich damit folgendes Bild:

$$\mathrm{nov}(c_1) < \mathrm{nov}(c_3), \mathrm{nov}(c_4), \mathrm{nov}(c_5) < \mathrm{nov}(c_2)$$

Alle Urteilsmittelwerte unterschieden sich signifikant von der *Unentschieden*-Stufe 3 (einseitige Vorzeichentests für eine Stichprobe zum Medianvergleich, $|z| > 4.27$, $p < 0.001$). Außer für c_1 lagen sie für alle Klassen darüber.

Urteilskonsistenz

Vergleich mit Zufallsniveau Die Interquartilsabstände (IQR) der Urteile fielen in diesem Experiment allgemein etwas höher aus als im vorangegangenen zur Beurteilung der Bedeutsamkeit, bewegten sich aber ebenfalls signifikant unterhalb des Zufallsniveaus

von 2.4 (einseitige Vorzeichentests für eine Stichprobe zum Medianvergleich, $z < -2.65$, $\alpha = 0.05$ Bonferroni-korrigiert, $p < 0.005$).

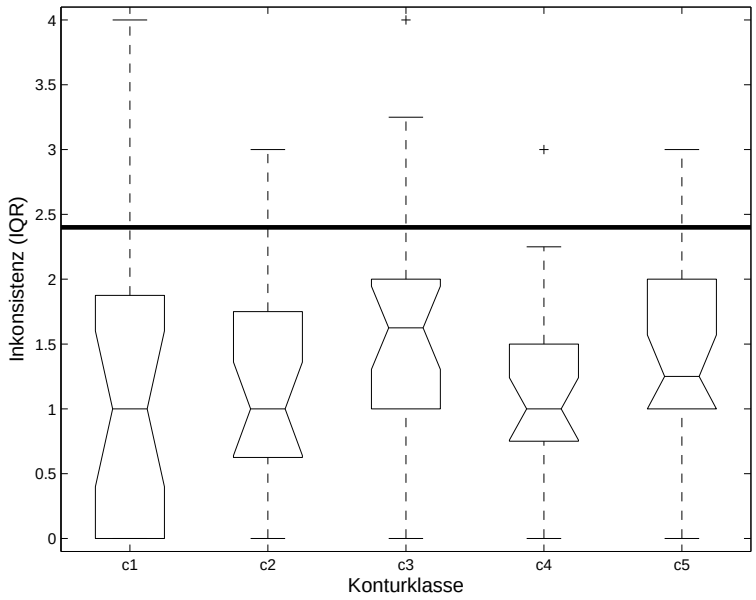

Abbildung 15.6: Klassenabhängige Urteilsinkonsistenz in Form von Interquartilsabständen (IQR) bei der Beurteilung informativer Neuheit.

Paarweiser Vergleich Der paarweise Vergleich der klassenabhängigen Urteilsstreuungen ergab eine signifikant höhere Inkonsistenz bei c_1 und c_3 gegenüber c_2 und c_4 (paarweiser Levene-Test, $\alpha = 0.05$ Dunn-Sidak-korrigiert, $p \leq 0.001$).

Schlussfolgerungen

Festhalten lässt sich an dieser Stelle:

- Auf Parameterebene erwiesen sich Polynomkoeffizient s_0, sowie F0-Maxima und Spannweite mit signifikant erhöhten Werten bei informativer Neuheit als geeignete Entsprechungen des Informationsstatus.
- Auf Symbolebene der Konturklassen wurde Hypothese **H3** im Hinblick auf Klasse c_1 bestätigt, aber bezüglich c_4 widerlegt. c_1 wird perzeptiv mit gegebener Information verbunden, c_4 tendenziell eher mit neuer Information.
- Hypothese **H4** wurde bestätigt: Klassen c_2 und c_5 werden perzeptiv mit neuer Information verknüpft, wobei diese Tendenz vor allem bei c_2 sehr stark zu Tage tritt.
- Auch die verbleibende Konturklasse c_3 konnte als Übermittler neuer Information eingeordnet werden.

- Die Urteile fielen für alle Konturklassen konsistent aus, am wenigsten konsistent für Klassen c_1 und c_3.

Kapitel 16

Äußerungsfinalität

Wie schon für die linguistischen Konzepte Bedeutsamkeit und Neuheit erfolgte auch hier eine korpusstatistische Untersuchung zum Zusammenhang zwischen Äußerungsfinalität und Intonation, woraus sich Hypothesen ableiten ließen, deren experimentelle Überprüfung am Ende dieses Kapitels beschrieben wird.

16.1 Modellierung

Wie in Kapitel 9 angesprochen, wurden die Nachrichtensätze einzeln aufgenommen. Daraus ergibt sich, dass jedes Satzende zugleich ein Äußerungsende darstellt. Somit wurden die jeweils letzten lokalen Segmente im Satz als äußerungsfinal festgelegt und die restlichen Segmente als non-final.

16.2 Korpusstatistik und Hypothesen

16.2.1 Befunde

Interpretation der Stilisierungsparameter

Hinsichtlich der Finalität wiesen alle Koeffizienten signifikante Unterschiede auf ($\alpha = 0.05$; s_0: zweiseitiger t-Test für unabhängige Stichproben, $t_{9217} = 17.33$, $p < 0.001$; s_1: zweiseitiger Welch-Test , $t_{9217} = 15.04$, $p < 0.005$; s_2: zweiseitiger Welch-Test, $t_{9217} = 9.85$, $p < 0.005$; s_3: zweiseitiger Welch-Test, $t_{9217} = -1.80$, $p < 0.05$; siehe Abbildung 16.1).

Wie schon der Informationsstatus lässt sich auch die Finalität intonatorisch an F0-Maxima und -Spannweiten der lokalen Konturen ablesen. In Abbildung 16.2 sind die Unterschiede für Original- und modellierte Konturen hinsichtlich dieser Kenngrößen zu sehen: Non-finale Konturen haben höhere F0-Maxima und eine höhere Spannweite, beides Kennzeichen eines progredienten F0-Verlaufs. Für Original- und modellierte Konturen fallen diese Unterschiede gleichermaßen deutlich aus (Welch-Tests; **F0-Maxima:** Original $t_{1207} = 25.60$, $p < 0.005$; PKS: $t_{1148} = 16.36$, $p < 0.005$; **Spannweite:** Original

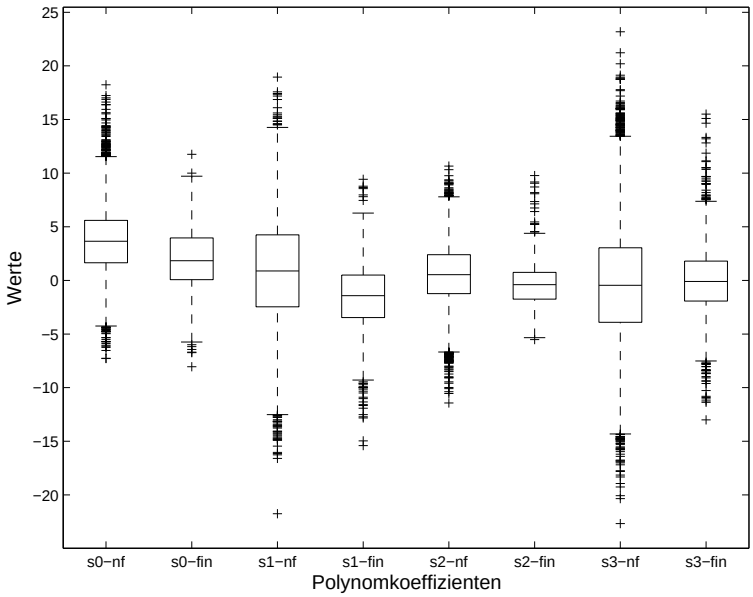

Abbildung 16.1: Polynomkoeffizienten s_j in Abhängigkeit der Finalität; **nf**: nicht-final, **fin**: final.

$t_{1284} = 18.74$, $p < 0.005$; PKS: $t_{1350} = 26.80$, $p < 0.005$).

Finalität und Konturklassen

Tabelle 16.1 enthält die anhand des χ^2-Tests und des Vergleichs zwischen bedingten und A-priori-Konturklassen-Wahrscheinlichkeiten extrahierten Zusammenhänge zwischen Intonation und Finalität.

Klasse	Codierung	χ^2	P(Klasse|final)	P(Klasse|non-final)	P(Klasse)
c_1	final	458.32*	0.49	0.19	0.22
c_2	non-final	135.68*	0.04	0.19	0.18
c_3	non-final	19.59*	0.12	0.18	0.17
c_4	non-final	131.94*	0.06	0.21	0.20
c_5	final	17.51*	0.28	0.22	0.23

Tabelle 16.1: Zusammenhang zwischen Konturklassen und Finalität. *: Zusammenhänge signifikant ($\alpha = 0.001$).

16.2.2 Hypothesen

Aus den angeführten Befunden ergeben sich folgende Hypothesen:

H5 Die Klassen c_1 und c_5 codieren Äußerungsfinalität.

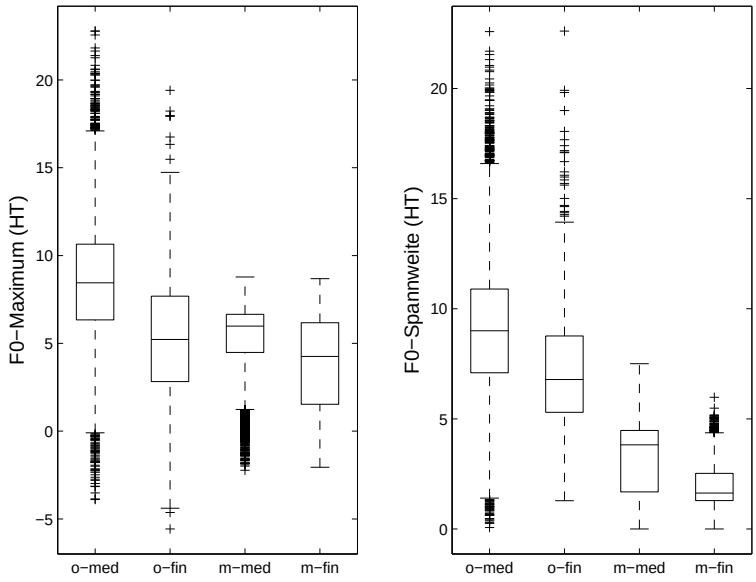

Abbildung 16.2: F0-Maxima und -spannweiten lokaler Konturen in Abhängigkeit der Finalität (in Halbtönen, Basis 50 Hz); **o-**: Originalkontur, **m-**: modellierte Kontur, **nf**: nicht-final, **fin**: final.

H6 Die Klassen c_2, c_3 und c_4 repräsentieren nicht-äußerungsfinale Konturen.

16.3 Perzeptive Validierung

16.3.1 Methode

Den Versuchspersonen wurden auf die visuell gestellte Frage "Was siehst Du?" als Endpunkte einer fünfstufigen Likert-Skala zwei Antwortalternativen der folgenden Form gezeigt:

- *Eine X.* (Z. B. *eine Blume.*)
- *Eine X und eine Y.* (Z. B. *Eine Blume und eine Birne.*)

Dazu wurden über Kopfhörer Stimuli der Form

eine X (z. B. *eine Blume*)

präsentiert, wobei erneut Zielwörter und Intonationsklassen wie in den Abschnitten 13.4.2 und 13.4.3 beschrieben variiert wurden. Stimulusbeispiele sind in Abbildung 16.3 zu finden.

Die Versuchspersonen hatten nun auf der Skala zu beurteilen, zu welcher der beiden Antworten der akustisch präsentierte Ausschnitt im Hinblick auf seinen Intonationsverlauf besser passt. Im Falle der Wahrnehmung eines progredienten non-finalen F0-Verlaufs

wäre eine Antworttendenz Richtung *Eine Blume und eine Birne* zu erwarten, im Falle eines finalen Verlaufs eine Tendenz Richtung *Eine Blume*.

In den visuell dargebotenen *Eine X und eine Y*-Antworten wurde neben X auch Y variabel belegt, um eine bei konstantem Y mögliche, aber unerwünschte kontrastive Deutung von X zu verhindern. Für Y wurden hierzu in randomisierter Reihenfolge die zur Verfügung stehenden Wörter aus der Zielwortmenge eingesetzt, jedes nur einmal und unter Berücksichtigung, dass X und Y nicht durch dasselbe Wort ersetzt wurden.

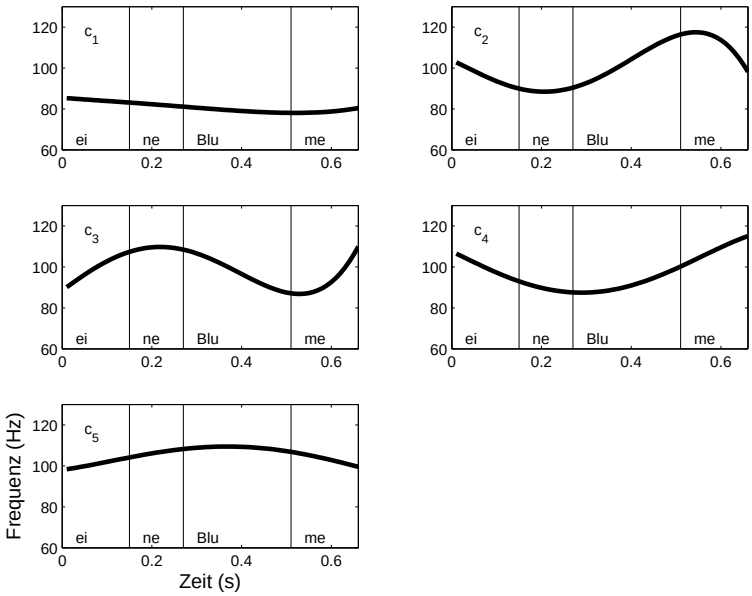

Abbildung 16.3: Stimulusbeispiel für jede Konturklasse zur Untersuchung der perzipierten Finalität.

16.3.2 Ergebnisse

Klasse	Median	arithm. Mittel	Interquartilsabstand	Standardabweichung
c_1	5	4.42	1	0.96
c_2	2	2.56	3	1.39
c_3	2	2.67	3	1.44
c_4	2	2.00	1	1.08
c_5	2	2.62	2	1.24

Tabelle 16.2: Mittelwerte und Streuungsmaße der Beurteilung der Konturklassen hinsichtlich Finalität.

Klassenabhängige Finalitätscodierung

Abbildung 16.4 zeigt die Versuchspersonenurteile im Hinblick auf die perzipierte Finalitätsmarkierung der Konturklassen in Form von relativen Häufigkeiten und Boxplots. Es sind hochsignifikante klassenabhängige Unterschiede festzustellen (Kruskal-Wallis-Test, $\chi^2_4 = 316.92$, $p < 0.001$). Der Dunnett-Post-hoc-Test liefert folgende signifikante Abstufung ($\alpha = 0.01$) hinsichtlich der Finalitätsurteile $\mathrm{fin}(c_n)$:

$$\mathrm{fin}(c_4) < \mathrm{fin}(c_2), \mathrm{fin}(c_3), \mathrm{fin}(c_5) < \mathrm{fin}(c_1)$$

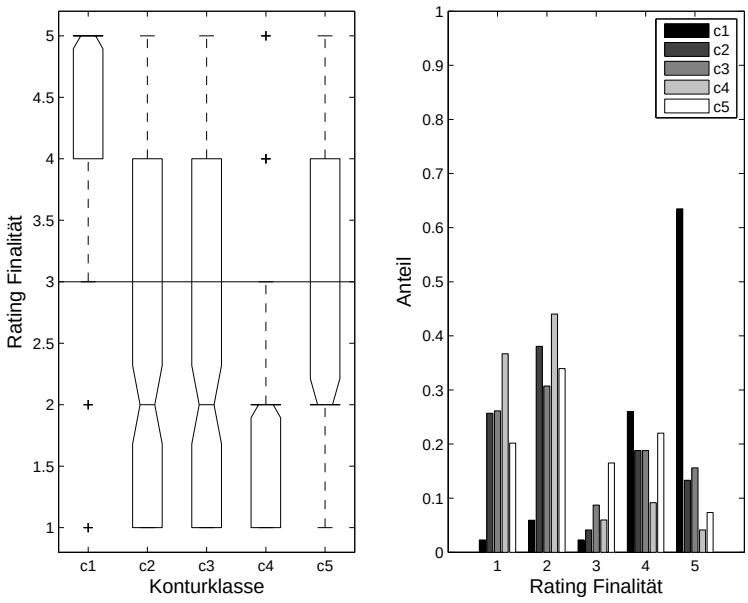

Abbildung 16.4: **Links:** Boxplots zur Beurteilung der Finalität in Abhängigkeit der lokalen Konturklasse. **Rechts:** Relative Häufigkeiten der Urteile 1 – 5 für jede Konturklasse.

Die perzipierte Finalitätsmarkierung erreicht für alle Klassen signifikant von 3 (*unentschieden*) verschiedene Mittelwerte (einseitige Vorzeichentests für eine Stichprobe zum Medianvergleich, $\alpha = 0.05$ Bonferroni-korrigiert, $|z| > 3.40$, $p < 0.001$), wobei einzig c_1 als äußerungsfinal perzipiert wird, die anderen Klassen als progredient.

Urteilskonsistenz

Vergleich mit Zufallsniveau Alle Klassen ließen sich konsistenter beurteilen als Zufallsniveau (einseitige Vorzeichentests für eine Stichprobe zum Medianvergleich, $\alpha = 0.05$ Bonferroni-korrigiert, $z < -3.47$, $p < 0.001$; siehe Abbildung 16.5).

Paarweiser Vergleich Der Vergleich der Klassen untereinander ergab eine signifikant niedrigere Inkonsistenz für c_1 und c_4 gegenüber den anderen Klassen (paarweiser

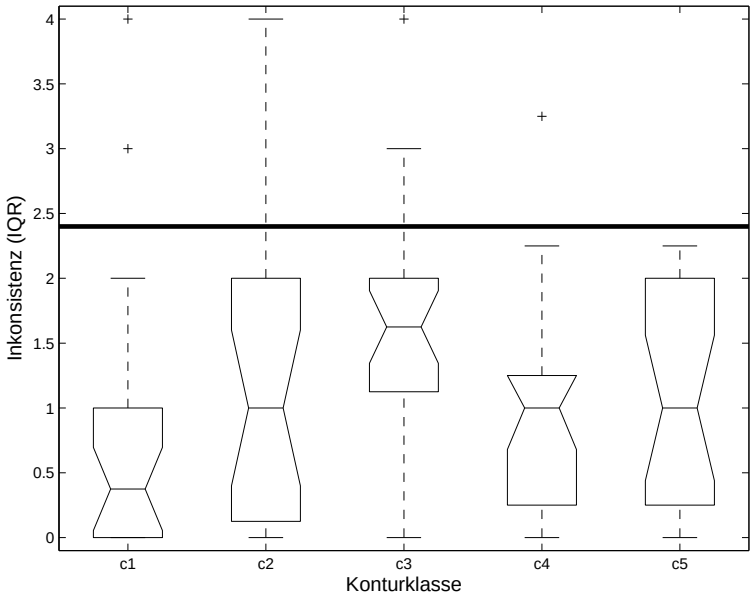

Abbildung 16.5: Klassenabhängige Urteilsinkonsistenz in Form von Interquartilsabständen (IQR) bei der Beurteilung von Finalität.

Levene-Test, $\alpha = 0.05$ Dunn-Sidak-korrigiert, $p \leq 0.001$). c_3 zeigte die höchste Urteilsinkonsistenz, signifikant höher gegenüber c_1, c_4 und c_5.

Schlussfolgerungen

Festhalten lässt sich an dieser Stelle:

- Auf parametrischer Ebene sind alle Polynomkoeffizienten an der Codierung von Finalität beteiligt. Dasselbe gilt für F0-Maxima und Spannweiten, die beide im finalen Kontext niedrigere Werte aufweisen.

- Auf Symbolebene der Konturklassen konnte Hypothese **H5** nur zum Teil bestätigt werden: Während Kontur c_1 wie prädiziert als äußerungsfinal wahrgenommen wird, ist bei c_5 eine Non-final-Einschätzung zu beobachten. Allerdings ist die Tendenz der non-finalen Einordnung bei c_5 weniger stark als bei den von Hypothese **H6** betroffenen Klassen.

- Hypothese **H6** wurde bestätigt: Die Klassen c_2, c_3 und c_4 wurden perzeptiv mit einer Äußerungsweiterführung in Verbindung gebracht.

- Die Urteile fielen für alle Konturklassen konsistent aus, erneut am wenigsten konsistent bei Klasse c_3.

Kapitel 17

Linguistische Modellierung: Das PKS-EB-Modell

Inhalt dieses Kapitels ist die Zusammenfügung der nun gewonnenen perzeptiven Befunde zur Entwicklung eines Konturvorhersagemodells.

Voraussetzung einer solchen linguistischen Unterfütterung des PKS-Modells anhand der Perzeptionsexperimente ist eine stabile Beurteilung der Konturklassen. Im nächsten Abschnitt werden die hierzu gewonnenen Ergebnisse zur Urteilskonsistenz zusammengefasst, bevor linguistische Einzelinterpretationen der Konturklassen versucht werden, die sich schließlich zu einem Entscheidungsbaum zur Auswahl der passenden Kontur zusammenfügen.

17.1 Voraussetzungen

Urteilskonsistenz der Versuchspersonen

Abbildung 17.1 zeigt für alle Versuchspersonen in Form von Boxplots die IQRs ihrer Urteile, wobei jeweils für jedes Teilexperiment und jede Konturklasse ein solcher Inkonsistenzwert berechnet wurde. Mit vier Ausnahmen urteilten alle Versuchspersonen signifikant konsistenter als Zufallsniveau (einseitige Vorzeichentests für eine Stichprobe zum Medianvergleich, $\alpha = 0.05$ Bonferroni-korrigiert, $z < -3.1$, $p \leq 0.002$; ohne Korrektur nur zwei Ausnahmen).

Eine mit den Faktoren *Geschlecht, Herkunft (Nord-, Mittel-, Süddeutschland)* und *musikalische Vorbildung (ja, nein)* durchgeführte ANOVA hinsichtlich möglicher Unterschiede in der gemessenen Inkonsistenz erbrachte für keine der untersuchten Gruppierungen signifikante Unterschiede.

Über mögliche Performanzunterschiede in Abhängigkeit der phonetischen Vorbildung ist keine Aussage möglich, da alle Versuchspersonen wie in Abschnitt 11.2.1 motiviert phonetisch vorgebildet waren.

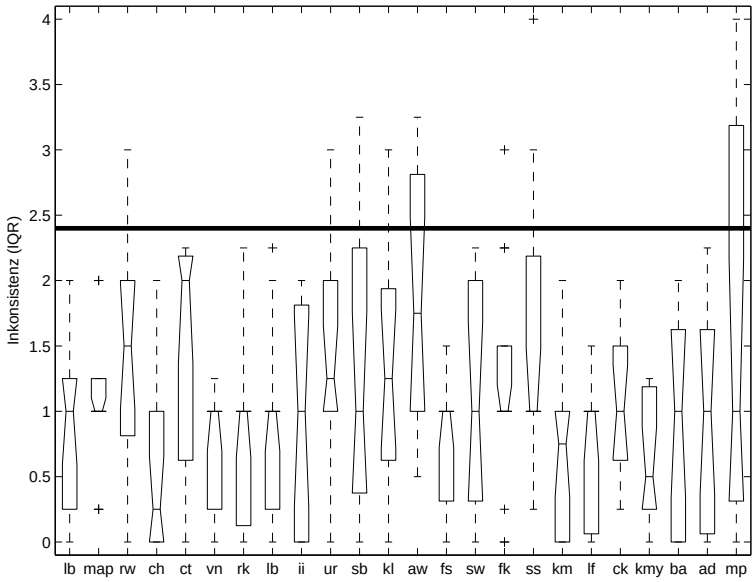

Abbildung 17.1: Urteilsinkonsistenz der Versuchspersonen in Form von Interquartilsabständen (IQR). Inkonsistenz auf Zufallsniveau bei 2.4 (horizontale Linie).

Konsistenz der Konturklassenbeurteilung

Wie in den vorangehenden Kapiteln gezeigt wurde, bewegten sich die IQRs der linguistischen Beurteilungen der Konturklassen signifikant unterhalb der Zufallsniveaus (siehe hierzu Abbildungen 14.4, 15.6 und 16.5). Die Beurteilungskonsistenz der Konturklassen seitens der Versuchspersonen spricht für eine ausreichende Validität der Befunde, die es erlaubt, die Ergebnisse zur linguistischen Intonationsmodellierung heranzuziehen.

17.2 Bedeutung lokaler Konturklassen

Abbildung 17.2 fasst die perzeptiven Beurteilungen der lokalen Konturklassen in Abhängigkeit der linguistischen Konzepte zusammen.

17.2.1 Multiple Beziehungen

Wie in Abbildung 17.2 zu sehen ist, bestehen zwischen der Ebene der Intonation und der Ebene der linguistischen Konzepte keine eindeutigen Abhängigkeiten:

1. Dieselbe Intonationsklasse kann die Ausprägungen mehrerer linguistischer Konzepte codieren.

2. Die Ausprägungen desselben linguistischen Konzepts können mit mehr als einer Intonationsklasse codiert werden.

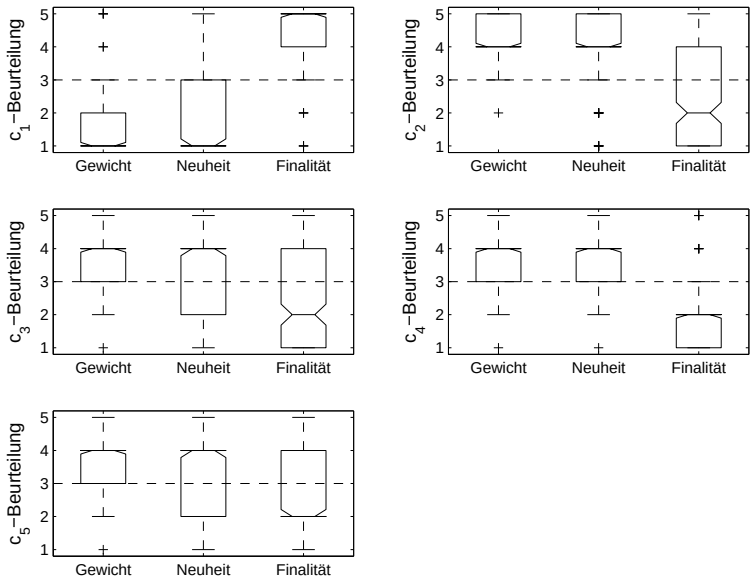

Abbildung 17.2: Zusammenfassung der linguistischen Beurteilungen der lokalen Konturklassen.

So codiert beispielsweise Klasse c_2 Ausprägungen der Konzepte *Neuheit* und *Bedeutsamkeit*, während zugleich das Konzept *Neuheit* neben c_2 auch durch Klasse c_4 codiert werden kann. Diese mehrdeutigen Beziehungen sind schematisch in Abbildung 17.3 dargestellt.

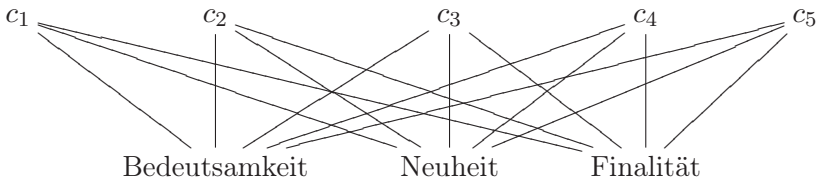

Abbildung 17.3: Multiple Beziehungen zwischen Intonations- und linguistischer Konzeptebene.

Korrelationen

Zusätzlich lassen sich moderate, aber signifikante Korrelationen ($p < 0.05$) feststellen:

1. zwischen Konturklassen bezüglich der Codierung von Konzeptausprägungen (**Klassenkorrelationen**), und

2. zwischen Konzepten bezüglich der ihnen zugeordneten Klassen (**Konzeptkorrelationen**).

Klassenkorrelationen Die Stichproben – eine je Klasse c_i –, zwischen denen die Klassenkorrelationen gemessen wurden, setzen sich zusammen aus den Urteilsmittelwerten jeder Versuchsperson in jedem der Teilexperimente 1–3 für Klasse c_i.

In Tabelle 17.1 sind die Spearman-Rangkorrelationen zwischen den klassenabhängigen Urteilsmedianen zu finden, in Abbildung 17.4 Scatterplots und paarweise Korrelationen nach Pearson zwischen den klassenabhängigen arithmetischen Urteilsmittelwerten.

Während einzig Klasse c_1 mit allen anderen Klassen eine negative Korrelation aufweist, sind die restlichen Klassen untereinander positiv korreliert.

	c_2	c_3	c_4	c_5
c_1	-0.60	-0.38	-0.59	-0.40
c_2		0.47	0.53	0.47
c_3			0.44	0.28
c_4				0.50

Tabelle 17.1: Klassenkorrelationen. Spearman-Rangkorrelationen zwischen den konturklassenbezogenen Urteilsmedianen. Verglichene Stichproben (eine je Klasse): Urteilsmediane jeder Versuchsperson in allen Teilexperimenten 1–3. Alle Werte sind signifikant von 0 verschieden (t-Test, $p < 0.001$).

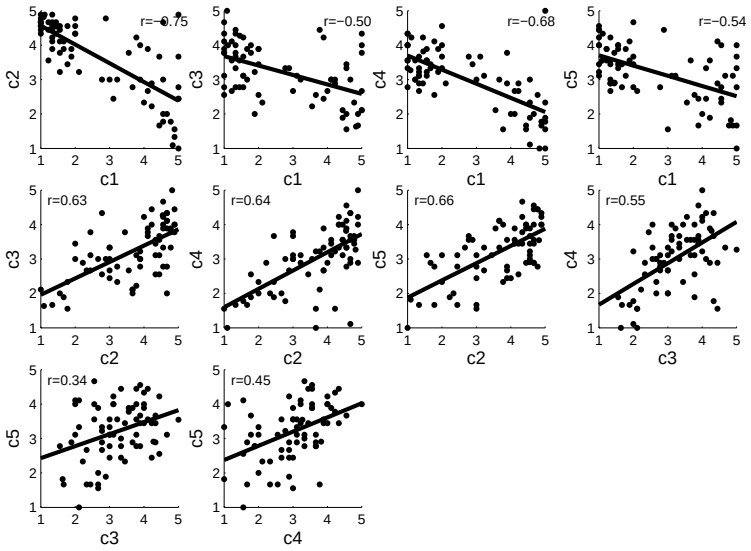

Abbildung 17.4: Klassenkorrelationen. Korrelationen nach Pearson zwischen den konturklassenbezogenen arithmetischen Urteilsmittelwerten. Verglichene Stichproben (eine je Klasse): arithmetische Urteilsmittelwerte jeder Versuchsperson in allen Teilexperimenten 1–3. Alle Werte sind signifikant von 0 verschieden (t-Test, $p < 0.001$).

Konzeptkorrelationen Die Stichproben (eine je Konzept k_i) zur Berechnung der paarweisen Konzeptkorrelationen bestehen aus den Urteilsmittelwerten jeder Versuchsperson für jede der Konturklassen c_1 bis c_5 für Konzept k_i.

Tabelle 17.2 enthält die Spearman-Rangkorrelationen zwischen den konzeptabhängigen Urteilsmedianen, Abbildung 17.5 zeigt Scatterplots sowie paarweise Korrelationen nach Pearson zwischen den konzeptabhängigen arithmetischen Urteilsmittelwerten.

Während *Finalität* mit den beiden anderen Konzepten negativ korreliert ist, sind diese untereinander positiv korreliert.

	Neuheit	Finalität
Bedeutsamkeit	0.51	-0.33
Neuheit		-0.28

Tabelle 17.2: Konzeptkorrelationen. Spearman-Rangkorrelationen zwischen den konzeptbezogenen Urteilsmedianen. Verglichene Stichproben (je eine für Konzept Bedeutsamkeit, Neuheit und Finalität): Urteilsmediane jeder Versuchsperson zu allen Konturklassen. Alle Werte sind signifikant von 0 verschieden (t-Test, $p < 0.001$).

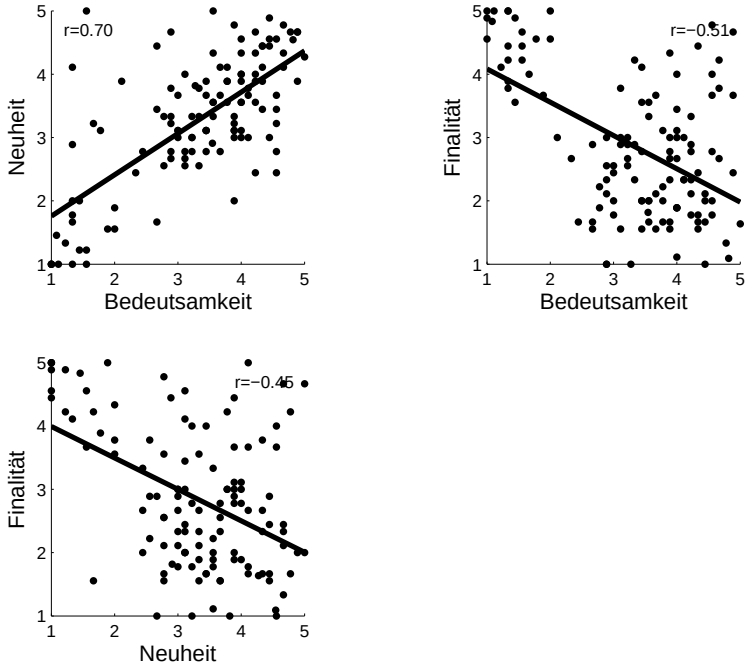

Abbildung 17.5: Konzeptkorrelationen. Korrelationen nach Pearson zwischen den konzeptbezogenen arithmetischen Urteilsmittelwerten. Verglichene Stichproben (je eine für Konzept Bedeutsamkeit, Neuheit und Finalität): arithmetische Urteilsmittelwerte jeder Versuchsperson zu allen Konturklassen. Alle Werte sind signifikant von 0 verschieden (t-Test, $p < 0.001$).

Folgerung

Festzuhalten ist also, dass weder die Intonationsklassen noch die linguistischen Konzepte orthogonal partitioniert werden können. Die mehrdeutigen Beziehungen zwischen Intonation und Konzeptebene sowie die gefundenen Korrelation zeugen vielmehr von einer variablen intonatorischen Codierung zusammenhängender linguistischer Konzepte.

17.2.2 Klassenzuordnung

Tabelle 17.3 zeigt die perzeptive Zuordnung der Konturklassen zu Ausprägungen linguistischer Attribute. Dieser Zuordnung sind die korpusstatistischen Befunde gegenübergestellt.

Klasse	Bedeutsamkeit		Neuheit		Finalität	
	Korpus	Perzeption	Korpus	Perzeption	Korpus	Perzeption
c_1	gering	gering	gegeben	gegeben	final	final
c_2	hoch	hoch	neu	neu	non-final	non-final
c_3	–	eher hoch	neu	eher neu	non-final	non-final
c_4	–	eher hoch	gegeben	neu	non-final	non-final
c_5	–	eher hoch	neu	eher neu	final	eher non-final

Tabelle 17.3: Linguistische Funktionen der Intonationskonturklassen auf Basis der Korpusuntersuchung und der Perzeptionsexperimente.

Mit Ausnahme von c_4 bezüglich des Konzepts *Neuheit* und von c_5 bezüglich *Finalität* widersprechen sich korpusstatistische und experimentelle Befunde nicht.

17.2.3 Das PKS-EB-Modell zur Intonationsvorhersage

Auf Grundlage der im vorangegangenen Abschnitt vorgenommenen Zuordnung von lokalen Konturklassen zu Merkmalsausprägungen linguistischer Konzepte (vergleiche Tabelle 17.3) lässt sich die Wahl der Konturklasse in Form eines Entscheidungsbaums modellieren, der im Folgenden auch als **PKS-EB-Modell** bezeichnet wird; *EB* steht hierbei für *Entscheidungsbaum*. Ein denkbarer binär verzweigender Baum ist in Abbildung 17.6 zu sehen. Jeder Pfad repräsentiert eine geordnete Sequenz von Entscheidungen, an deren Ende die zu wählende Konturklasse steht.

Der asymmetrischen Konstruktion des Baums liegt die auf Plausibilität und entsprechenden Korrelationen gestützte Annahme zugrunde, dass *gegebene Information* (linker Hauptast) bereits *geringe Bedeutsamkeit* impliziert. Plausibel erscheint die Annahme deshalb, da gegebene Information aus dem Kontext gut vorhersagbar ist und daher nach Bolinger (1972) geringes semantisches Gewicht trägt. Eine weitere Verzweigung in Abhängigkeit der Bedeutsamkeit erübrigt sich in diesem Sinne also bei gegebener Information.

Während sich die Klassen c_1, c_2 und c_3 auf korpusstatistischer wie perzeptiver Grundlage problemlos Blättern des Baums zuordnen lassen, müssen zum aktuellen Forschungs-

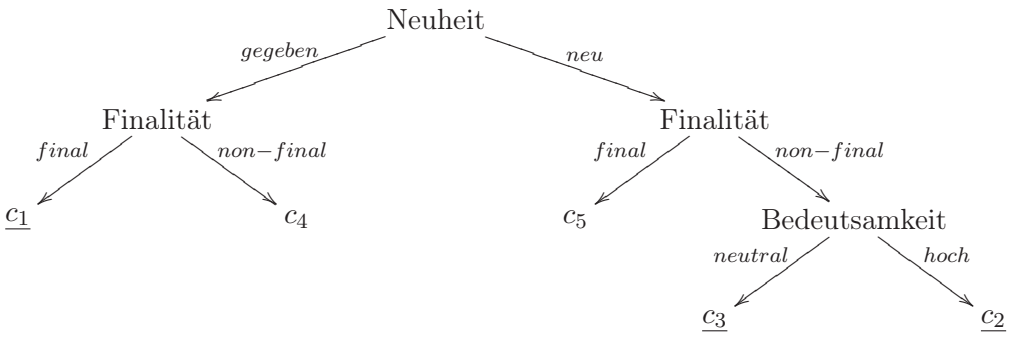

Abbildung 17.6: PKS-EB-Modell: Entscheidungsbaum zur Konturauswahl basierend auf den ermittelten Korpusstatistiken und perzeptiver Validierung (Übereinstimmungen sind unterstrichen).

stand bei den verbleibenden Klassen noch Kompromisse eingegangen werden: Um das zum Pfad *gegeben&non-final* gehörige Blatt zu besetzen, wurde für c_4 nur die korpusstatistisch ermittelte Neuheitscodierung berücksichtigt. Gleiches gilt für Klasse c_5 am Ende des Pfads *neu&final* hinsichtlich der Finalitätscodierung.

17.3 Perzeptive Validierung des PKS-EB-Modells

In einem Perzeptionsexperiment wurde die perzeptive Adäquatheit der Konturvorhersagen des durch den Entscheidungsbaum in Abbildung 17.6 repräsentierten PKS-EB-Modells für die Faktoren *Finalität* und *Neuheit* überprüft.[1]

Hierzu wurden Versuchspersonen Intonationskonturen präsentiert, wie sie vom PKS-EB-Modell auf Grundlage des Diskurskontexts vorhergesagt werden, sowie davon abweichende Konturen. Getestet werden sollten hierbei die folgenden Hypothesen:

H7 Das PKS-EB-Modell ist geeignet, perzeptiv akzeptable Konturen vorherzusagen.

H8 Die Modellvorhersagen sind perzeptiv adäquater als von den Vorhersagen abweichende Konturalternativen.

17.3.1 Versuchspersonen

Es nahmen zehn phonetisch vorgebildete Versuchspersonen im Alter zwischen 24 und 39 Jahren am Experiment teil, neun davon mit deutscher Muttersprache, eine ungarischer Herkunft mit über zehnjährigem Wohnsitz in Deutschland. Alle Versuchspersonen hatten bereits an den Teilexperimenten 1–5 teilgenommen. Der Autor nahm nicht am Experiment teil.

[1] Eine Begründung für das Weglassen des Faktors *Bedeutsamkeit* findet sich im folgenden Kapitel 18.

17.3.2 Methode

Versuchspersonen sollten die Adäquatheit des Intonationsverlaufs auf Zielsätzen im jeweiligen Diskurskontext bewerten, der durch einen vorangehenden Satz gegeben war. Die ihnen vorgelegte Anleitung ist in Anhang D.2 abgedruckt.

In jedem Trial wurden vier intonatorische Varianten eines Zielsatzes im Zusammenhang mit einem vorangehenden Diskurskontext-Satz präsentiert, die beliebig oft angehört werden konnten und hinsichtlich der Adäquatheit für Neuheit und Finalität auf einer fünfstufigen Skala mit den Endpunkten *adäquat* – *inadäquat* zu beurteilen waren (siehe den Screenshot in Anhang E).

Die Zielsätze in den präsentierten Satzpaaren waren so gestaltet, dass zwei lokale Segmente zur Variierung der lokalen Kontur in Frage kamen, eines in non-finaler, das andere in finaler Position. Da in jedem Trial nur ein lokales Segment behandelt wurde, waren für jedes Satzpaar zwei Trials angesetzt. Das Vorgehen sei an einem Beispiel illustriert (eine Liste aller Satzpaare mit den zugehörigen Variationen findet sich in Anhang C.2).

Dort steht eine Buche. $[$*Die Buche*$]_{s1}$ *verliert* $[$*ihre Blätter*$]_{s2}$.

Der Diskurskontext ist durch den ersten Satz gegeben. Im zweiten Satz wurden in zwei Trials – einen für jedes der lokalen Segmente s_1 und s_2 – jeweils vier verschiedene Konturvarianten erzeugt:

- V: die durch das Modell vorhergesagte Konturklasse c_v,
- V_n: eine Konturklasse, die nur hinsichtlich der Neuheitscodierung mit c_v übereinstimmt, also hinsichtlich Finalität kontrastiert,
- V_f: eine Konturklasse, die nur hinsichtlich der Finalitätscodierung mit c_v übereinstimmt, also hinsichtlich Neuheit kontrastiert,
- V_0: eine Konturklasse, die weder hinsichtlich der Neuheits- noch der Finalitätscodierung mit c_v übereinstimmt.

Auf das obige Beispiel lässt sich dies folgendermaßen beziehen:

Diskurskontext	*Dort steht eine Buche.*	
Zielsatz	$[$*Die Buche*$]_{s1}$ *verliert* $[$*ihre Blätter*$]_{s2}$.	
s_1	Status	gegeben, non-final
	Varianten	V: c_4, V_n: c_1, V_f: c_2, V_0: c_5
s_2	Status	neu, final
	Varianten	V: c_5, V_n: c_2, V_f: c_1, V_0: c_4

Das lokale Segment s_1 befindet sich in *non-finaler* Position und trägt *gegebene Information*. Die PKS-EB-Vorhersage lautet, wie im Entscheidungsbaum in Abbildung 17.6 ablesbar, c_4. Zu dieser Intonationsvariante V wurden kontrastive Varianten generiert: V_0

ist gegeben durch die Konturklasse, die sich von c_4 im PKS-EB-Modell sowohl im Hinblick auf Neuheits- als auch Finalitätscodierung unterscheidet, also durch c_5. V_f kontrastiert bezüglich Neuheit (Klasse c_2) und V_n bezüglich Finalität (c_1).

Die Stimuli wurden wie in Kapitel 13 beschrieben ausgehend von den modellierten F0-Verläufen und den regressionsbaumbasierten Dauervorhersagen mit Mbrola resynthetisiert. Alle globalen Konturen wurden konstant mit einem initialen Niveau von 80 Hz und einem Deklinationsfaktor von -1.5 modelliert. Sowohl die Trials als auch die Darbietung der Intonationsalternativen erfolgten in randomisierter Reihenfolge.

17.3.3 Ergebnisse

Mittelwerte und Streuungen der Adäquatheitsurteile für jede der Konturvarianten sind in Tabelle 17.4 aufgelistet und in Abbildungen 17.7 und 17.8 in Form von Boxplots und Balkendiagrammen graphisch dargestellt.

Vergleicht man die Vorhersagen des Modells mit den restlichen Varianten zusammengenommen, lässt sich Folgendes feststellen:

- Die PKS-EB-Vorhersagen werden allgemein akzeptiert. Der Urteilsmedian liegt bei 4 und ist damit signifikant höher als die mittlere Bewertungsstufe 3 (einseitiger Vorzeichentest für eine Stichprobe zum Medianvergleich, $z = 7.12$, $p < 0.001$).

- Die PKS-EB-Vorhersagen werden signifikant besser beurteilt als die Konturalternativen zusammengenommen (Mann-Whitney-Test, $p < 0.001$).

- Betrachtet man die Konturalternativen getrennt, so sind ebenfalls signifikante Adäquatheitsunterschiede festzustellen (Kruskal-Wallis-Test, $\chi^2_3 = 88.45$, $p < 0.001$). Die PKS-EB-Vorhersage führte gegenüber allen anderen Konturalternativen zu signifikant besseren Bewertungen (Dunnett-Post-hoc-Test, $\alpha = 0.05$).

- Während auch die finalitätserhaltende Konturvariante als signifikant besser gegenüber der neuheitserhaltenden und komplett kontrastiven bewertet wurde, war kein signifikanter Unterschied zwischen letzteren beiden Varianten festzustellen (Dunnett-Post-hoc-Test, $\alpha = 0.05$).

17.3.4 Schlussfolgerung

Anhand der ersten drei obigen Befunde konnten also die Hypothesen **H7** und **H8** zur Güte des PKS-EB-Modells bestätigt werden: das PKS-EB-Modell ist geeignet, perzeptiv akzeptable Konturen vorherzusagen, und die Modellvorhersagen sind perzeptiv adäquater als von den Vorhersagen abweichende Konturalternativen.

Variante	Median	arithm. Mittel	Interquartilsabstand	Standardabweichung
V	4	4.11	1	0.89
V_f	4	3.50	1	1.06
V_n	2	2.58	3	1.33
V_0	2	2.26	2	1.36
$\neg V$	3	2.78	2	1.36

Tabelle 17.4: Mittelwerte und Streuungsmaße der Beurteilung der Konturvarianten hinsichtlich Adäquatheit. V: Modellvorhersage, V_f: nur hinsichtlich Finalitätscodierung übereinstimmende Variante; V_n: nur hinsichtlich Neuheitscodierung übereinstimmende Variante; V_0: vollständig kontrastive Variante; $\neg V$: $\{V_f, V_n, V_0\}$.

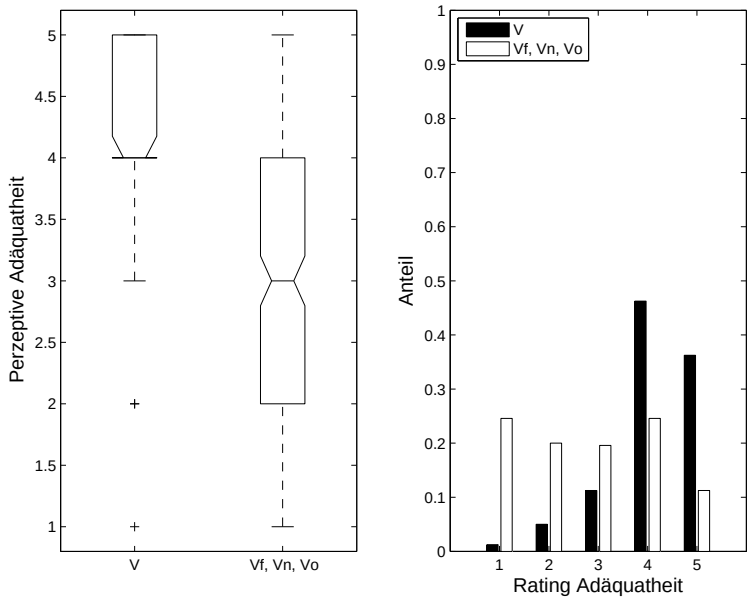

Abbildung 17.7: **Links:** Boxplots der perzeptiven Adäquatheitsurteile für die Vorhersagen V des PKS-EB-Modells und die intonatorischen Varianten V_f, V_n und V_0 zusammengenommen. **Rechts:** relative Häufigkeiten der Adäquatheitsurteile.

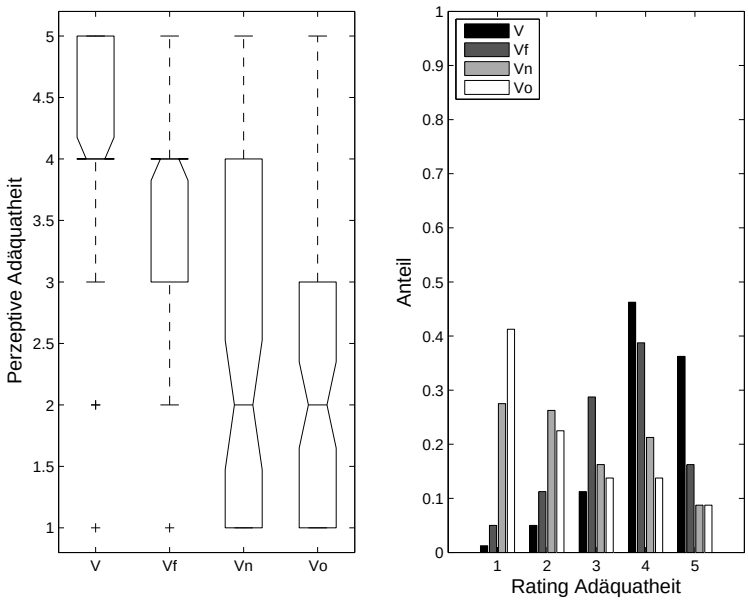

Abbildung 17.8: **Links:** Boxplots der perzeptiven Adäquatheitsurteile für die Vorhersagen V des PKS-EB-Modells und die intonatorischen Varianten V_f, V_n und V_0. **Rechts:** relative Häufigkeiten der Adäquatheitsurteile.

Kapitel 18

Diskussion und Zusammenfassung des Teils III

Teil III dieser Arbeit hatte die linguistische Interpretation des PKS-Modells zum Inhalt, sowie die Entwicklung des auf diesen Untersuchungen basierenden PKS-EB-Modells zur Konturvorhersage. Einige damit verbundene Aspekte seien in diesem Kapitel noch einmal zur Diskussion aufgegriffen.

18.1 Analyseverfahren

Ziel der linguistischen Analyse war es zu prüfen, ob eine empirisch abgesicherte, also nicht rein impressionistische, linguistische Interpretation, der *bottom-up* gewonnenen Intonationsrepräsentation möglich ist. Hierfür wurde das Korpus mittels automatisierter Verfahren linguistisch analysiert, der statistische Zusammenhang zwischen den gewonnenen linguistischen Parametern einerseits und den Stilisierungsparametern sowie den lokalen Konturklassen andererseits ermittelt, um daraus schließlich Hypothesen über die linguistische Funktion der Konturklassen zu gewinnen. Die Hypothesen wurden im Anschluss durch Perzeptionsexperimente überprüft.

18.1.1 Korpusanalysen

Wie bereits erwähnt, bestand der primäre Zweck der automatisierten Analysen in der Datenaufbereitung zur Hypothesengenerierung, weshalb eine gewisse Fehlerquote in Kauf genommen werden konnte und auf manuelle Analysen somit verzichtet wurde.

18.1.2 Perzeptive Untersuchung

In Ermangelung an etablierten Standardverfahren zur Untersuchung der Intonationsperzeption soll das experimentelle Vorgehen dieser Studie an dieser Stelle kurz begründet werden.

Experiment-Design

Ein zu erwartender Mangel an Sensibilität gegenüber Intonation bei einem Teil der Versuchspersonen wurde durch vorangehende Trainingsphasen sowie die Möglichkeit der beliebig häufigen Stimulus-Wiederholung zu entschärfen versucht.

Auf Grund der großen Menge an Teilexperimenten in nur einer Session bestand ein weiteres Ziel der Experimentgestaltung darin, möglichst viele Trials mit möglichst geringer Dauer zu erhalten. Dies wurde für die Teilexperimente 1–3 zur linguistischen Interpretation der Konturklassen im Wesentlichen erreicht durch:

- kompaktes Stimulus-Design,
- Verwendung bipolarer Skalen.

Stimuli Anders als in Intonationsstudien mit relativ aufwendiger Gestaltung der Kontexte, beispielsweise in Form längerer Textabschnitte, innerhalb derer Intonationskonturen zu beurteilen sind (Niebuhr, 2007b; Welby, 2003) wurden hier die Kontexte möglichst kompakt gehalten, zum Beispiel in Form eines wenige Wörter umfassenden Frage-Antwort-Paars. Keine der Versuchspersonen berichtete von Schwierigkeiten der Urteilsfällung in Abhängigkeit eines unklaren Kontexts.

Skalen Weiter wurden die Versuchspersonenurteile auf bipolaren Skalen des Typs

<Aussage 1> ... <Aussage 2>.

gemessen. Dadurch ließ sich die Anzahl der Trials für jede Aussage verdoppeln. Rechtzufertigen ist dieses Vorgehen aus den folgenden Gründen:

- Es handelt sich ausschließlich um Komplementäraussagen: *neu* vs. *gegeben, bedeutsam* vs. *belanglos, final* vs. *non-final*, es kommen daher nicht beide Endpunkte der Skala gleichzeitig als Antwort in Frage.
- Konturklassen, die sich weder zur Codierung von *<Aussage 1>* noch von *<Aussage 2>* eignen, lassen sich auch mit bipolaren Skalen identifizieren, und zwar durch eine hohe Streuung oder auch bei verwendeter ungeradzahliger Stufenzahl durch das Nicht-Vorhandensein einer signifikanten Abweichung von der mittleren Urteilsstufe.

Die Aussagen wurden bezüglich Neuheit und Finalität, wie in den Screenshots in Anhang E zu sehen, nicht direkt, sondern beispielhaft übermittelt, da hier von einer damit verbundenen Erleichterung der Aufgabe ausgegangen wurde.

Interpretation der Streuung Ein Kriterium zur Feststellung, ob eine Konturklasse sich zur Codierung eines linguistischen Konzepts eignet, war der Blick auf die Streuung der Urteile, die im Falle der Eignung signifikant kleiner als eine zufällig zustande kommende Streuung zu sein hatte. Diese Referenzstreuung wurde auf Grund bias-freier und damit über die Skala gleichverteilter Zufallsantworten ermittelt. Der resultierende Interquartilsabstand (die IQR) beträgt hier 2.4. Geht man stattdessen von einer Normalverteilung der Zufallsantworten um die "Unentschieden"-Stufe 3 aus, so beträgt die Referenzstreuung in Form der IQR 2. Wie in den Abbildungen 14.4, 15.6 und 16.5 zu sehen ist unterschreiten die beobachteten Streuungen mit einer Ausnahme signifikant auch diesen Wert (einseitige Vorzeichentests für eine Stichprobe zum Medianvergleich, $p \leq 0.01$; Ausnahmen: c_3 bei Neuheit $p = 0.09$, bei Finalität $p = 0.04$). Wegen der gewissen Arbitrarität bei der Festlegung der Referenzstreuung zur Interpretation der Konturklasse wurde in allen Fällen zusätzlich der Urteilsmedian mit dem Referenzmedian (Stufe 3, *"Weiß nicht"*) herangezogen. Auch hier waren alle Abweichungen signifikant (einseitige Vorzeichentests für eine Stichprobe zum Medianvergleich, $p < 0.001$).

Bias- und Strategievermeidungsmanagement Auf Counter-Balancing der Aussage-Alternativen wurde wie schon erwähnt verzichtet, da drei Versuchspersonen eines informellen Vorexperiments sich dadurch stark abgelenkt sahen und von dadurch provozierten Fehlantworten berichteten. Wie aber in den Abbildungen 15.5, 14.3 und 16.4 zu sehen ist, wurde für jede der Teilaufgaben die gesamte Skala zur Einordnung der Konturklassen genutzt, wodurch ein größerer Einfluss eines etwaigen auf fehlendes Counter-Balancing zurückführbaren Bias ausgeschlossen werden kann.

Durch Verwendung von Distraktoren wurde der Entwicklung von Antwortstrategien entgegengewirkt. Die in Abschnitt 13.4.2 beschriebenen sorgfältigen Auswahl der Zielwörter hatte den Zweck, Einflüsse des Lexikons auf die Urteile weitestgehend zu reduzieren.

Auswertung der Ergebnisse

Die Verwendung von Likert-Skalen hat sich in der Intonationsforschung mittlerweile etabliert (Birch und Clifton, 1995; Welby, 2003). Nicht unumstritten ist allerdings ihre Auswertung (Jamieson, 2004; Rietveld und Chen, 2006). Während das ordinale Skalenniveau der Likert-Skala strenggenommen nur non-parametrische Tests erlaubt, rechtfertigen Befürworter die Anwendung effizienterer parametrischer Tests damit, dass jedenfalls bei ungeradzahligen Stufenzahlen Äquidistanz zwischen den Stufen und somit quasi eine Intervallskalierung angenommen werden könne. Zwar wurden auch in dieser Arbeit fünfstufige und damit ungeradzahlige Skalen verwendet, dennoch fällt es schwer, eine Äquidistanz der Stufen-Abstände nachzuweisen. Aus diesem Grund kamen hier weniger umstrittene und dafür konservativere non-parametrische Testverfahren zum Einsatz. Die Ergebnisse legen aber nahe, dass in diesem Fall auch diese Tests Aufschlussreiches zu Tage fördern konnten.

18.2 Linguistische Interpretation

18.2.1 Interpretierbarkeit der Stilisierungsparameter

Die Polynomkoeffizienten der Stilisierungsfunktion ließen sich nur teilweise linguistisch interpretieren.

Semantisches Gewicht Die Korrelationen zwischen Koeffizienten und semantischem Gewicht erwiesen sich allesamt als gering, so dass hier keine tragfähigen interpretativen Aussagen gemacht werden können.

Neuheit Unter den Polynomkoeffizienten stellte sich nur s_0 für das allgemeine F0-Niveau als Indikator informativer Neuheit heraus. Im Falle neuer Information nimmt er höhere Werte an, die zu einer erhöhten Prominenz führen. Entsprechend sind auch F0-Maximum und Spannweite bei neuer Information höher und tragen damit ebenso zur Steigerung der Prominenz bei.

Finalität Hinsichtlich der Äußerungsfinalität wiesen alle Koeffizienten signifikante Unterschiede auf. Finale Konturen sind hierbei gekennzeichnet durch ein allgemein niedrigeres F0-Niveau (s_0) und einen fallenden Verlauf (s_1 negativ). Auch F0-Maxima und Spannweiten sind gegenüber dem nicht-finalen Verlauf geringer. Die progrediente Form nicht-finaler Konturen spiegelt sich in höheren s_0-Werten wider, was sich auch als hoher Grenzton interpretieren lässt, sowie in positiven s_1-Werten, also einem steigenden F0-Verlauf. Koeffizient s_3, der unter anderem den Verlauf des postakzentuierten Abschnitts steuert, trägt dagegen wider Erwarten mit negativem Mittelwert nichts zur Progredienz bei.

In Anbetracht dieser nicht vollständig gegebenen direkten Interpretierbarkeit der Stilisierung auf parametrischer Ebene erscheint es sinnvoll, den Schwerpunkt auf eine „traditionellere" linguistische Analyse auf Symbolebene zu legen, also zur weiteren Untersuchung Konturklassen heranzuziehen.

18.2.2 Interpretierbarkeit der Konturklassen

Allgemein konnten perzeptiv alle Konturklassen jedem der untersuchten linguistischen Konzepte zugeordnet werden, was festzustellen ist anhand der oben festgelegten Kriterien:

- signifikant niedrigere Streuung als bei Zufallsantworten zu erwarten,
- signifikanter Unterschied zwischen Urteilsmedian und der mittleren "Unentschieden"-Stufe.

Übereinstimmung zwischen korpusstatistischen und perzeptiven Befunden In den meisten Fällen widersprachen die perzeptiven Klassenzuordnungen den korpusstatistischen Befunden nicht (zwölf Übereinstimmungen, drei Fälle ohne korpusstatistische Festlegung). Die beiden Ausnahmen betrafen c_4 hinsichtlich Neuheit und c_5 hinsichtlich Finalität, wobei hier für c_5 die perzeptiven Urteile weniger eindeutig ausfielen als in den Fällen der Übereinstimmung.

Klasse c_3 Am wenigsten konsistent wurde insgesamt die Klasse c_3 eingeschätzt, was an ihrem zu einem gewissen Grad ambigen F0-Verlauf liegen mag. In den Stimulusbeispielen in Abbildung 14.2 zur Beurteilung der Bedeutsamkeit ist zu sehen, dass die c_3-Kontur den Gipfel bereits auf der wortbetonten Silbe des dem Zielwort vorangehenden Artikels erreicht. Eine Versuchsperson berichtete auch von ihrem Eindruck, dass es sich hierbei um einen engen Fokus handeln könnte, was eine Uminterpretation des Artikels in ein Numerale zur Folge hätte (*das ist* **eine** *Blume – es handelt sich nicht um zwei*). Vermutlich in Abhängigkeit der wahlweisen Konzentration auf diesen vermeintlichen engen Fokus oder das dadurch deakzentuierte Kernwort haben Versuchspersonen den c_3-Stimuli höhere oder niedrigere Bedeutsamkeiten zugeordnet.

An der vergleichsweise großen Streuung der Finalitätsurteile mag die variierende Bewertung des kurzen Anstiegs am Ende der c_3-Kontur (siehe Abbildung 16.3) eine Rolle gespielt haben.

18.2.3 Modellierung

Beschränkungen

In dieser Arbeit wurde nur eine Auswahl linguistischer Konzepte untersucht, so stehen beispielsweise noch Intonationsanalysen zu Kontrastkonstruktionen aus. Auch Fragen sowie paralinguistische Funktionen wurden mangels Vorhandenseins im verwendeten Korpus nicht mitmodelliert.

Multiple Beziehungen

Allgemein wurden multiple Beziehungen zwischen Konturklassen und behandelten linguistischen Konzepten sichtbar, wodurch sich folgende Korrelationen ergeben:

- Klassenkorrelationen zwischen der Beurteilung der Konturklassen (Abbildung 17.4) und
- Konzeptkorrelationen zwischen den Beurteilungsaufgaben (Abbildung 17.5).

Klassenkorrelationen Dasselbe linguistische Konzept kann mit verschiedenen Konturklassen codiert werden. Dies erlaubt ein hohes Ausmaß an Variabilität in der Produktion und spiegelt unter Umständen die hohe Kompetenz des ausgebildeten Sprechers, von dem die Trainingsdaten stammen, wider.

Konzeptkorrelationen Konturklassen können verschiedene Konzepte gleichermaßen codieren. Dem liegt zugrunde, dass die Konzepte selbst nicht orthogonal sind. Ein starker Zusammenhang besteht zwischen *informativer Neuheit* und *Bedeutsamkeit*.

Wie bereits ausgeführt, wird Bedeutsamkeit hier über Vorhersagbarkeit definiert und zwischen drei Arten gegebener Information unterschieden:

(a) im Diskursverlauf bereits übermittelt,

(b) zum geteilten Weltwissen gehörig und

(c) aus dem situativen Kontext erschließbar.

All diesen Ausprägungen von Gegebenheit ist eine vergleichsweise hohe Vorhersagbarkeit gemein:

(a) Ein Diskursreferent, von dem bereits die Rede ist, wird mit höherer Wahrscheinlichkeit wieder erwähnt als ein beliebiges anderes Objekt.

(b) Zum geteilten Weltwissen gehören eher häufig übermittelte und somit vorhersagbare Informationen.

(c) Information, die aus dem situativen Kontext erschließbar ist, ist trivialerweise aus diesem auch eher vorhersagbar, als Sachverhalte, die in keinem Verhältnis zur aktuellen Situation stehen.

(a) und (c) repräsentieren hohe lokale Vorhersagbarkeit, (b) eine hohe globale Vorhersagbarkeit. Festzustellen ist also ein starker Zusammenhang zwischen gegebener Information und hoher Vorhersagbarkeit (geringer Bedeutsamkeit) sowie umgekehrt zwischen neuer Information und geringer Vorhersagbarkeit (hoher Bedeutsamkeit).

Der starke Zusammenhang zwischen den Konzepten *Bedeutsamkeit* und *Neuheit* wirft erneut die in Abschnitt 5.1 behandelte Frage nach der adäquaten Abstraktionsebene der semantischen Beschreibung von Intonation auf (Pike, 1945; Gussenhoven, 1984; Peters, 2006). So ließe sich beispielsweise motivieren, diese voneinander abhängigen Konzepte unter dem Begriff *Relevanz* zusammenzufassen.

Gegen eine Zusammenfassung würde allerdings sprechen, dass in Kontrastkonstruktionen, die in dieser Arbeit auf Grund der unzureichenden Datenlage nicht untersucht wurden, mitunter zwischen Neuheit und Bedeutsamkeit zu trennen ist. So ist in:

> *Ich dachte, das Parkett knarzt. Es ist aber nicht das* **Parkett***, sondern der Stuhl.*

Parkett beim zweiten Auftreten in der Kontrastierung gleichzeitig gegeben und bedeutsam. Kontrastkonstruktionen fügen sich also nicht in das Abhängigkeitsmuster zwischen Neuheit und Bedeutsamkeit.

Intoneme

Anders als in kompositionalen Modellen beispielsweise von Pierrehumbert und Hirschberg (1990), in denen zwischen Tonakzenten zur Markierung des Informationsstatus und Grenztönen zur Orientierung der Phrase im Diskursverlauf unterschieden wird, ist eine entsprechende Einteilung der Konturklassen im PKS-Modell nicht möglich, da die Klassen, wie in Tabelle 17.3 zu sehen ist, zugleich Informationsstatus in Form von Neuheit und Orientierung in Form von Finalität codieren.

Alternativ ließe sich der Begriff des Intonems als diskrete bedeutungstragende Einheit (Isačenko und Schädlich, 1964; Stock und Zacharias, 1982) heranziehen, siehe auch Abschnitt 5.3. In dieser Tradition wäre es grundsätzlich denkbar, den Entscheidungsbaum in Abbildung 17.3 zu bemühen, um die gefundenen Konturklassen als Intoneme in Form von Bündeln distinktiver semantischer Merkmale darzustellen. So ließen sich "Intoneme" c_2 und c_3 charakterisieren als

$$c_2 = [\textit{neu, non-final, hohe Bedeutsamkeit}]$$
$$c_3 = [\textit{neu, non-final}]$$

und entsprächen damit funktional partiell dem Nonterminalitäts-Intonem $N \uparrow$ nach Stock und Zacharias (1982).

Eine besprochene Verschmelzung der hoch korrelierten Konzepte *Bedeutsamkeit* und *Neuheit* ergäbe aber einen Zusammenfall der angenommenen Intoneme, c_2 und c_3 wären dann „Allointone" desselben Intonems

$$c_{2,3} = [\textit{relevant, non-final}].$$

Dieses Beispiel zeigt, dass eine Intonemanalyse bei der aktuellen Befundlage letztlich der Willkür unterliegt.

Vorhersage der Intonationsklassen

Es konnte experimentell nachgewiesen werden, dass sich das PKS-EB-Modell zur Vorhersage perzeptiv adäquater Konturverläufe eignet. Interessant für zukünftige Studien wäre nun die Prüfung eines Modelleinsatzes in der textbasierten Intonationsvorhersage.

Vergleichsweise hohe Akzeptanzwerte erhielten auch Konturklassen, die sich hinsichtlich der Vorhersagen nicht in der Finalitäts-, sondern nur in der Neuheitscodierung unterschieden, was zwei mögliche Schlussfolgerungen zulässt:

- Fehlerhafte Vorhersagen der Neuheitscodierung werden allgemein als weniger fatal beurteilt als fehlerhafte Vorhersagen zur Finalität. Oder:
- Die hinsichtlich Neuheit kontrastierten Konturklassen sind perzeptiv weniger distinkt als die Klassen, die in Finalitätsopposition zueinander stehen.

Zum jetzigen Kenntnisstand ist nicht entscheidbar, welcher dieser Schlüsse eher zutrifft.

18.2.4 Kontexteinflüsse

Funktionale Kontextunabhängigkeit

Die in den Teilexperimenten 1–3 vorgenommenen perzeptiven Untersuchungen beschränkten sich auf Äußerungen mit nur einem lokalen Segment und eignen sich daher nicht für Aussagen über Kontextabhängigkeiten der Funktionen lokaler Konturen als Teil längerer Äußerungen.

Im Gegensatz dazu waren bei der perzeptiven PKS-EB-Evaluierung lokale Segmente in größere Kontexte eingebettet. Hier war über diverse Konturkontexte hinweg eine global höhere Adäquatheit der vorhergesagten gegenüber alternativen Konturklassen festzustellen, was letztlich für eine Kontextunabhängigkeit der Funktion einer Konturklasse im Hinblick auf Neuheits- und Finalitätscodierung spricht, also für eine stabile wenn auch multiple Form-Funktion-Beziehung.

Einzuräumen ist jedoch, dass auf Grund der Beschränkung auf diese beiden linguistischen Konzepte keine Aussagen über eine Kontextabhängigkeit pragmatischer Deutungen von Intonationskonturen möglich sind, so wie sie beispielsweise Ward und Hirschberg (1985) vorgefunden hatten (siehe Abschnitt 5.1 dieser Arbeit).

Syntagmatische Kontextunabhängigkeit

Auf korpusstatistischer Ebene sind nur geringe Abhängigkeiten der Konturklassen untereinander festzustellen. Die Trigrammwahrscheinlichkeiten der Konturklassen belaufen sich allesamt auf Werte kleiner 0.34 und sind damit deutlich niedriger als beispielsweise in Tonakzent-Grenzton-Sequenzen (Dainora, 2002); vergleiche auch Abschnitt 7.3. Dies lässt den Schluss zu, dass die hier gefundenen lokalen Konturen durch den intonatorischen Kontext vergleichsweise wenig determiniert sind, also keiner restriktiven Intonotaktik (Noteboom, 1997) unterworfen sind.

18.3 Zusammenfassung des Teils III

Die lokalen Konturklassen des PKS-Modells konnten über statistische Korpusanalysen und die perzeptive Validierung der daraus resultierenden Hypothesen mit den linguistischen Konzepten *Bedeutsamkeit, Neuheit* und *Finalität* verknüpft werden. Bis auf zwei Ausnahmen standen die korpusstatistischen und perzeptiven Befunde nicht im Widerspruch zueinander.

Es wurden multiple Beziehungen zwischen Intonations- und linguistischer Konzeptebene festgestellt. Dies zeugt zum einen von einer gewissen Variabilität in der Wahl intonatorischer Mittel zur Codierung eines Konzepts sowie von einer Abhängigkeit der Konzepte untereinander.

Auf Grundlage der korpusstatistischen und perzeptiven Ergebnisse wurde ein Entscheidungsbaum zur linguistischen Vorhersage der lokalen Konturklassen entwickelt (das PKS-EB-Modell). Eine perzeptive Evaluierung dieses Modells ergab allgemein hohe Na-

türlichkeitsurteile für die vom Modell vorhergesagten Konturen und eine bessere Beurteilung der Vorhersagen im Vergleich mit nicht vorhergesagten Alternativen.

Zusätzlich zur im Teil II herausgestellten Signalnähe des PKS-Modells ist nun also auch sein Potential zur linguistischen Verankerung festzustellen.

Teil IV

Abschließende Zusammenfassung und Ausblick

Gegenstand dieser Arbeit war die Entwicklung eines datenbasierten Intonationsmodells, dass zur automatischen Analyse und Synthese von F0-Konturen herangezogen werden kann und dabei linguistisch interpretierbar ist.

Das PKS-Intonationsmodell

Modellcharakteristika Das in dieser Arbeit entwickelte PKS-Intonationsmodell lässt sich charakterisieren als parametrisch, konturbasiert und superpositional. Intonation wird als Superposition von polynomial stilisierten globalen und lokalen F0-Konturen repräsentiert. Die streng hierarchische prosodische Struktur zur Verankerung der Konturen besteht aus zwei Ebenen: aus zeitnormalisierten globalen und lokalen Segmenten, die anhand von Sprechpausen, Interpunktion und Wortartinformation ermittelt werden. Globale Segmente werden hierbei durch Sprechpausen und Interpunktion begrenzt. Lokale Segmente umspannen als Akzentgruppe in Anlehnung an Arbeiten zur Prosodie-Syntax-Schnittstelle (siehe hierzu Abschnitt 10.1) eine Folge von Funktionswörtern mit abschließendem Inhaltswort.

Globale Konturen werden linear, lokale Konturen mit Polynomen dritter Ordnung stilisiert. Diese Konturen werden mittels Kmeans-Clustering jeweils zu einer geringen Anzahl von diskreten Konturklassen zusammengefasst. Mit den durch numerische Optimierung gewonnenen Clusterparametern ergaben sich drei globale und fünf lokale Klassen.

Phonetische Regressionsmodelle dienen der Überführung dieser abstrakten Konturklassen in konkrete kontextabhängige Realisierungen. Regressionsmodelle wurden hierbei entwickelt zur Vorhersage des Pitch Resets, zur kontextabhängigen Anpassung der Deklinations-Baseline, sowie zur Anpassung der Polynomkoeffizienten der lokalen Konturen. Durch letztere Operation ergibt sich indirekt die Modellierung der Deklinations-Topline.

Konzeptuelle Aspekte Die Entscheidung für ein parametrisches und konturbasiertes Modell war im Wesentlichen durch die damit verbundene Signalnähe und Automatisierbarkeit der Modellierung motiviert und ist durch phonetische Befunde zu rechtfertigen (siehe hierzu Abschnitt 12.2). Auch der superpositionale Aufbau trägt phonetischen Befunden zur intonatorischen Vorausplanung Rechnung.

Hinsichtlich der parametrischen und konturbasierten Beschreibung steht das Modell u.a. in der Tradition von Fujisaki (1987), Möhler (1998b) und Taylor (2000), hinsichtlich Superposition in der Tradition von Fujisaki (1987) und Möbius (1993a). Wie bei Möhler (2001) werden durch Clustern der Stilisierungsparameterwerte Konturklassen gewonnen. Die die hier gewählte polynomiale Stilisierung niedriger Ordnung garantiert (a) eine akzeptable Reproduzierbarkeit des Signals, also die Erfassung von F0-codierter Prominenz und Progredienz sowie unterschiedlicher F0-Gipfel- und -Taltypen ohne Mitmodellierung von Rauschen sowie (b) anders als die komplexeren Stilisierungsfunktionen der oben genannten Modelle auf Grund der analytischen eineindeutigen Anpassung eine vollständige Reproduzierbarkeit der Abstrahierung. Letzteres besagt, dass eine konkrete F0-Kontur genau eine abstrakte Form (eine konkrete Belegung der Polynomkoeffizienten) besitzt,

und dass sich aus einem bestimmten Koeffizientenbelegung genau eine F0-Kontur erzeugen lässt. Diese Eigenschaft ist essentiell für eine Partitionierung der F0-Stilisierungen in Intonationsklassen, die sich an der Ähnlichkeit der F0-Konturen orientiert, sowie für linguistische Interpretationsversuche, wie sie sowohl für die Stilisierungskoeffizienten als auch für die Konturklassen durchgeführt wurden.

Reichweite Das Modell wurde für die Intonation von Deklarativsätzen anhand der gelesenen Texte eines professionellen Nachrichtensprechers entwickelt. Frageintonation, spontansprachliche Phänomene, Variation zwischen Sprechern, sowie paralinguistische Einflüsse auf die Sprechmelodie wurden ausgeklammert.

Datenanforderungen Ein wesentliches Ziel bei der Gestaltung des PKS-Modells war, möglichst geringe Anforderungen an eine Vorabpräparierung des Korpus zu stellen. Alle benötigten Informationen sind mit automatischen Verfahren zur Pausen- und Silbenkerndetektion auf Signalebene sowie zu Part-of-Speech-Tagging und Wortbetonungszuweisung auf Symbolebene extrahierbar. Die Signal-Text-Alinierung in Form einer Zuordnung der detektierten Silbenkerne zum Text wurde unter Vermittlung einer automatischen Lautsegmentierung bewerkstelligt.

Zur F0-Stilisierung wurden nur F0-Abschnitte über Silbenkernen herangezogen, was eine exakte Silbensegmentierung überflüssig macht, ebenso wie eine Gewichtung von F0-Abschnitten in Abhängigkeit ihrer Nähe zu Silbenkernen. Die Ermittlung der prosodischen Struktur beschränkt sich auf die automatische Ermittlung prosodischer Phrasengrenzen, eine Lokalisierung oder gar Klassifizierung von Akzenten entfällt.

Auf diese Weise ist eine vollständige Automatisierbarkeit der Vorverarbeitung auf Signal- und Symbolebene mit hinreichender Güte möglich, so dass auf manuelle Aufbereitung der Daten durch Experten verzichtet werden kann. Dies erlaubt eine schnelle Adaption des Modells auf beliebige Sprachdaten und vermeidet Inkonsistenzen durch unvollständiges Inter-Labeller-Agreement. Somit entfällt auch die expertengeleitete Anpassung des prosodischen Inventars an neue Dialekte oder Sprachen, das Inventar lässt sich unmittelbar aus den Daten heraus ableiten.

Evaluierung

Mathematisch Die objektiv-mathematische Evaluierung von zwei Varianten des PKS-Modells mit unterschiedlicher Anzahl lokaler Konturklassen ergab, dass eine Erhöhung der Anzahl von Konturklassen zu erhöhter Form-Ähnlichkeit und geringeren Distanzen zwischen Originalkonturen und ihren modellierten Entsprechungen führt. Die vernachlässigbaren Performanzunterschiede zwischen Trainings- und unabhängigen Testkorpora zeugen von einem hohen Grad an Generalisierbarkeit des Modells auf ungesehene Daten.

Perzeptiv Eine perzeptive Evaluierung erbrachte, dass PKS-modellierte Konturen gegenüber den Originalkonturen als weniger natürlich empfunden, aber immer noch oberhalb der Durchschnittsstufe bewertet werden. Aus der Variation der F0-Maxima und F0-

Spannweiten in lokalen Segmenten sowie den Beurteilungen der Sprecherintentionen lässt sich ablesen, dass die resynthetisierten Konturen dem Original entsprechend Finalität und Progredienz abbilden und auch Aspekte wie neue Information oder Bedeutsamkeit zum Ausdruck bringen können, wenn auch in weniger starker Form. Diese Abschwächung ist sehr wahrscheinlich auf die Verwendung von Zentroiden als Konturklassenrepräsentanten zurückzuführen, was allgemein zu flacheren Konturverläufen führt.

Linguistische Interpretation

Vorgehen Untersucht wurde auf Parameterebene und Symbolebene die linguistische Interpretierbarkeit von Stilisierungskoeffizienten und lokalen Konturklassen bezüglich der Konzepte *Bedeutsamkeit, informative Neuheit* und *Finalität*. Um der datenbasierten Bottom-up-Entwicklung des Intonationsmodells Rechnung zu tragen, war hierzu ein neuer Untersuchungsansatz vonnöten, der über rein impressionistische Korpusanalysen oder das Testen gegebener Hypothesen durch Perzeptionsexperimente hinausgeht. Der Ansatz in dieser Arbeit bestand darin, anhand automatisierter linguistischer Korpusanalysen Hypothesen darüber zu gewinnen, welche Konturklassen zur Codierung welcher Konzepte herangezogen werden, und diese Hypothesen im Anschluss durch Perzeptionsexperimente zu überprüfen. Auf diese Weise gelang eine systematische linguistische Verankerung des PKS-Modells in Form eines Entscheidungsbaums zur Vorhersage der linguistisch passenden Konturklasse. Die Adäquatheit der Vorhersagen wurde wiederum mit einem Perzeptionsexperiment sichergestellt.

Modellierung linguistischer Einflussfaktoren Bedeutsamkeit wurde als Vorhersagbarkeit in Form von Trigrammwahrscheinlichkeiten modelliert. Zur Bestimmung, ob ein Wort neue oder gegebene Information trägt, wurde eine Diskurssegmentierung und innerhalb der entstandenen Themenblöcke eine Koreferenzresolution durchgeführt. Finalität fiel mit Satzgrenzen zusammen.

Interpretation der Stilisierungsparameter Bezüglich des semantischen Gewichts ließen sich die Polynomkoeffizienten auf Grund der geringen Korrelationen nicht tragfähig interpretieren. Informative Neuheit ging auf parametrischer Ebene einher mit einer zur Prominenzsteigerung dienlichen Erhöhung des F0-Niveaus, des F0-Maximums sowie der Spannweite. Finalität und Progredienz waren anhand der Koeffizienten für das allgemeine F0-Niveau und die Steigung gut voneinander zu trennen, wie zu erwarten dahingehend, dass Progredienz sich durch einen F0-Anstieg und ein höheres F0-Niveau auszeichnet, Finalität dagegen durch eine absinkende F0.

Form-Funktion-Beziehung zwischen Konturklassen und linguistischen Konzepten Festzustellen ist, dass alle lokalen Konturklassen, obwohl rein datenbasiert und nicht durch Expertenurteile gewonnen, post hoc linguistisch interpretierbar sind. Dies ergaben sowohl korpusstatistische als auch perzeptive Befunde, die mit zwei Ausnahmen nicht im Widerspruch zueinander standen.

Alle Konturen codieren hierbei mehrere der untersuchten Konzepte, und auch jedes Konzept wird durch mehrere Konturen repräsentiert, was auf Variabilität in der intonatorischen Codierung von Semantik und Diskurs sowie auf eine Abhängigkeit der linguistischen Konzepte untereinander schließen lässt.

Ferner war eine Kontextunabhängigkeit der intonatorischen Codierung der untersuchten linguistischen Konzepte festzustellen. Daraus kann gefolgert werden, dass multiple aber stabile Beziehungen zwischen Form und Funktion der Intonation bestehen.

Modellierung linguistischer Intonationsvorhersage Auf Grundlage der gewonnenen Erkenntnisse über die linguistische Funktionalität der Konturklassen wurde das PKS-EB-Modell zur diskursbasierten Intonationsvorhersage entwickelt. Unter diesem Modell ist ein Entscheidungsbaum zu verstehen, der anhand der linguistischen Kenngrößen die passende lokale Konturklasse auswählt.

Die durch dieses Modell vorhergesagten Konturen wurden perzeptiv als in hohem Maße adäquat beurteilt und deutlich besser als nicht vom Modell vorhergesagte Intonationsvarianten. Diese Befunde lassen den Schluss zu, dass PKS-EB-Modell für die Aufgabe der Intonationsvorhersage qualifiziert ist.

Weitere Anwendungsmöglichkeiten

Sprachsynthese Auf Grund seiner erwiesenen linguistischen Verankerung ist eine Einbindung des PKS-Modells in ein Prosodie-Modul eines Sprachsynthesesystems, das diskursanalytische Analysen miteinbezieht, grundsätzlich denkbar. Auf Grund seiner geringen Anforderungen an eine vollständig automatisierbare Korpusaufbereitung lässt es sich dabei ohne großen Aufwand auf unterschiedliche Korpora und damit auf unterschiedliche Sprecher, Dialekte oder Sprachen adaptieren.

Dialektidentifikation Eine kürzlich vom Autor dieser Arbeit durchgeführte aber bislang unveröffentlichten Studie ergab, dass im *RVG 1*-Korpus (Burger und Schiel, 1998) drei deutsche Dialekte allein anhand von Sequenzen lokaler Konturklassen, die mit dem PKS-Modell ermittelt wurden, mit einer Performanz von derzeit etwa 60 % richtig identifiziert werden konnten. Hierbei wurde ein Bayes'scher Klassifikator eingesetzt, der die Identifizierung anhand der dialektabhängigen Wahrscheinlichkeiten der Konturklassensequenzen vornahm. Diese vorläufigen Resultate machen eine Nutzung des PKS-Modells als ergänzenden intonatorischen Merkmalsextraktor zur Dialektidentifizierung interessant. Zu prüfen wäre auch ein Einsatz des Modells in der Sprachen- und Sprechererkennung.

Andere Domänen Das PKS-Modell ist hinreichend abstrakt, um grundsätzlich auch in anderen Domänen als der Intonation Einsatz finden zu können, nämlich überall dort, wo es darum geht, Konturen zu segmentieren, zu parametrisieren und zu klassifizieren. Denkbar wäre hier im Bereich der Prosodie die Modellierung von Intensitätsverläufen sowie von Sprechgeschwindigkeit und damit Rhythmus.

Schlussfolgerungen

In dieser Arbeit konnte gezeigt werden, dass es möglich ist, mit einer rein datenbasierten automatischen Modellierung eine perzeptiv akzeptable Intonationsrepräsentation zu schaffen, die sich linguistisch verankern und somit sowohl aus dem Signal als auch aus dem Text heraus gewinnen lässt. Diese Vernetzung befähigt das Modell sowohl zur Intonationsanalyse als auch -synthese, wodurch es in sprachtechnologischen Anwendungen Einsatz finden kann, ebenso wie in der phonetischen Grundlagenforschung bei der automatischen Analyse nicht manuell aufbereiteter Sprachdaten.

Literaturverzeichnis

S. Abney. Parsing By Chunks. In R. Berwick, S. Abney, und C. Tenny (Hrsg.), *Principle-Based Parsing*, S. 257–278. Kluwer Academic Publishers, Dordrecht, 1991.

L.M.H. Adriaens. *Ein Modell deutscher Intonation: eine experimentell-phonetische Untersuchung nach den perzeptiv relevanten Grundfrequenzänderungen in vorgelesenem Text*. Doktorarbeit, University of Technology, Eindhoven, 1991.

P.D. Agüero und A. Bonafonte. Consistent Estimation of Fujisaki's Intonation Model Parameters. In *Proc. SPECOM*, Patras, 2005.

P.D. Agüero, K. Wimmer, und A. Bonafonte. Joint Extraction and Prediction of Fujisaki's Intonation Model Parameters. In *Proc. Interspeech*, S. 757–760, Jeju Island, Korea, 2004.

S. Ananthakrishnan und S. S. Narayanan. Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. *IEEE Transactions on Audio, Speech & Language Processing*, 16(1):216–228, 2008.

M.D. Anderson, J.B. Pierrehumbert, und M.Y. Liberman. Synthesis by rule of English inonation patterns. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, S. 281–284, New York, 1984.

M. Atterer und D.R. Ladd. On the phonetics and phonology of ßegmental anchoringöf F0: evidence from German. *Journal of Phonetics*, 32:177–197, 2004.

R.H. Baayen, R. Piepenbrock, und L. Gulikers. The CELEX Lexical Database. CD-ROM, 1995. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

J. Bachenko und E. Fitzpatrick. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170, 1990.

A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, und H. Niemann. Prosodic Feature Evaluation: Brute Force or Well Designed. In *Proc. ICPhS*, S. 2315–2318, San Francisco, 1999.

S. Baumann, M. Grice, und S. Steindamm. Prosodic Marking of Focus Domains – Categorical or Gradient. In *Proc. Speech Prosody*, S. 301–304, Dresden, 2006.

G.W. Beattie, A. Cutler, und M. Pearson. Why is Mrs Thatcher interrupted so often? *Nature*, 300:744–747, 1982.

M.E. Beckman. *Stress and Non-Stress Accent*. Foris, Dordrecht, 1986.

M.E. Beckman und J. Pierrehumbert. Intonational structure in English and Japanese. In *Phonology Yearbook*, Band 3, S. 255–310. Cambridge University Press, 1986b.

M. Bierwisch. Regeln für die Intonation deutscher Sätze. Untersuchungen über Akzent und Intonation im Deutschen. In *Studia Grammatica VII*, S. 99–199. Akademieverlag, Berlin, 1966.

S. Birch und C.Jr. Clifton. Focus, accent, and argument structure: Effects on language comprehension. *Language and Speech*, 38:365–391, 1995.

E. Blaauw. *On the perceptual classification of spontaneous and read speech*. Doktorarbeit, Research Institute for Language and Speech (OTS), Utrecht University, 1995.

A. Black und N. Campbell. Predicting the Intonation of Discourse Segments From Examples in Dialogue Speech. In *Proc. ESCA Workshop on spoken dialogue systems*, S. 197–200, Aalborg, 1995.

A. Black und A. Hunt. Generating F0 contours from ToBI labels using linear regression. In *Proc. ICSLP*, Band 3, S. 1385–1388, Philadelphia, 1996.

A.W. Black und P. Taylor. CHATR: A generic speech synthesis system. In *Proc. COLING94*, S. 983–986, 1994.

A.W. Black und P.A. Taylor. The Festival Speech Synthesis System: System documentation. Technical Report HCRC/TR–83. Forschungsbericht, Human Communciation Research Centre, University of Edinburgh, 1997.

D. Bolinger. Intonation: Levels Versus Configurations. *Word*, 7:199–210, 1951.

D. Bolinger. Accent is predictable (if you're a mind reader). *Language*, 48:633–644, 1972.

B. Braun, G. Kochanski, E. Grabe, und B.S. Rosner. Evidence for attractors in English intonation. *J. Acoustical Society of America*, 119(6):4006–4015, 2006.

N. Braunschweiler. Integrated cues of voicing and vowel length in German: A production study. *Language and Speech*, 40:353–376, 1997.

L. Breiman, J. Friedman, C.J. Stone, und R.A. Olshen. *Classification and Regression Trees*. Wadsworth & Brooks, Pacific Grove, CA., 1984.

C. Brinckmann und J. Trouvain. The role of duration models and symbolic representation for timing in synthetic speech. *International Journal of Speech Technology*, 6:21–31, 2003.

C. Brindöpke, G.A. Fink, F. Kummert, und G. Sagerer. A HMM-based recognition system for perceptive relevant pitch movements of spontaneous German speech. In *Proc. ICSLP*, S. 2895–2898, Sydney, 1998.

G. Brown, K.L. Currie, und J. Kenworthy. *Questions of Intonation.* Croom Helm, London, 1980.

I. Bulyko und M. Ostendorf. Joint Prosody Prediction and Unit Selection for Concatenative Speech Synthesis. In *Proc. ICASSP*, S. 781–784, 2001.

S. Burger und F. Schiel. RVG 1 - A Database for Regional Variants of Contemporary German. In *Proc. LREC*, S. 1083–1087, Granada, Spain, 1998.

H. Bußmann. *Lexikon der Sprachwissenschaft.* Alfred Kröner Verlag, 2 edition, 1990.

D.E. Carlson. *Some acoustical and perceptual correlates of speaker gender identification.* Doktorarbeit, University of Florida, Gainesville, 1981.

S. Cassidy und J. Harrington. EMU: an enhanced hierarchical speech database management system. In *Proc. 6th Australian International Conference on Speech Science and Technology*, S. 361–366, 1996.

W. L. Chafe. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. Li (Hrsg.), *Subject and topic*, S. 25–55. Academic Press, New York, 1976.

F. Charpentier und E. Moulines. Pitch-Synchronous Waveform Prosessing Techniques for Text-to-Speech Synthesis Using Diphones. In *Proc. Eurospeech*, S. 13–19, 1989.

S. Chiu. Fuzzy Model Identification Based on Cluster Estimation. *J. Intelligence & Fuzzy Systems*, 2(3):267–278, 1994.

T. Cho. Prosodically-conditioned strengthening and vowel-to-vowel coarticulation. *Journal of Phonetics*, 32:141–176, 2004.

N. Chomsky und M. Halle. *The Sound Pattern of English.* Harper & Row, New York, 1968.

K.W. Church und P. Hanks. Word association norms, mutual information and lexicography. *ACL*, 27:76–83, 1989.

G Cinque. A null theory of phrase and compound stress. *Linguistic Inquiry*, 24:239–297, 1993.

R. Clark und S. King. Joint Prosodic and Segmental Unit Selection Speech Synthesis. In *Proc. Interspeech*, 2006. paper 1262.

R.A.J. Clark und K.E. Dusterhoff. Objective methods for evaluating synthetic intonation. In *Proc. Eurospeech*, Band 4, S. 1623–1626, Budapest, 1999.

A. Cohen, R. Collier, und J. t'Hart. Declination: construct or intrinsic feature of speech pitch. *Phonetica*, 39:254–273, 1982.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

J. Cole, H. Kim, H. Choi, und M. Hasegawa-Johnson. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics*, 35:180–209, 2007.

R. Collier. Physiological correlates of intonation patterns. *JASA*, 58:249–255, 1975.

R. Collier und C.E. Gelfer. Physiological explanations of F0 Declination. In *Proc. ICPhS*, S. 440, 1983.

R. Collier und J. t'Hart. Perceptual experiments on Dutch intonation. In *Proc. ICPhS*, S. 880–884, The Hague, Paris, 1972. Mouton.

B. Connell und D.R. Ladd. Aspects of pitch realisation in Yoruba. *Phonology*, 7:1–30, 1990.

W.E. Cooper und J.M. Sorensen. *Fundamental frequency in sentence production*. Springer, New York, 1981.

Boersma P.and Weenink D. PRAAT, a system for doing phonetics by computer. Forschungsbericht, Institute of Phonetic Sciences of the University of Amsterdam, 1999. 132–182.

A. Dainora. Eliminating downstep in prosodic labeling of American English. In *Proc. ISCA Workshop on Prosody, Speech Recognition and Understanding*, S. 41–46, 2001.

A Dainora. Does intonational meaning come from tones or tunes? evidence against a compositional approach. In *Proc. Speech Prosody*, S. 235–238, Aix-en-Provence, France, 2002.

C. d'Alessandro und M. Castellengo. The pitch of short-duration vibrato tones. *JASA*, 95(3):1617–1630, 1994.

C. d'Alessandro und P. Mertens. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language*, 9(3):257–288, 1995.

J.R. de Pijper. *Modelling British English intonation: An analysis by re-synthesis of British English intonation*. Foris, Dordrecht, 1983.

J.R. de Pijper und A.A. Sandermann. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, 96:2037–2047, 1994.

A. M. C. de Sluijter und V. J. van Heuven. Spectral Balance as an acoustic correlate of linguistic stress. *JASA*, 100(4):2471–2485, 1996.

A.P. Dempster, N.M. Laird, und D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

A. Di Cristo. *De la microprosodie á l'intonosyntaxe*. Doktorarbeit, Université de la Provence, 1985.

A. Di Cristo und D.J. Hirst. Modelling French micromelody: Analysis and Synthesis. *Phonetica*, 43:11–30, 1986.

G. Dogil. Phonetic correlates of word stress. In *Word stress*, Band 2 aus *AIMS*, S. 1–59. 1995.

J. Doughty und W. Garner. Pitch characteristics of short tones. II. Pitch as a function of tonal duration. *J. Experimental Psychology*, 38:478–494, 1948.

K. Dusterhoff und A. Black. Generating F0 contours for speech synthesis using the Tilt intonation theory. In *Proc. ESCA Workshop of Intonation*, S. 107–110, Athens, Greece, 1997.

K.E. Dusterhoff, A.W. Black, und P. Taylor. Using Decision Trees within the Tilt Intonation Model to Predict F0 Contours. In *Proc. European Conf. on Speech Communication and Technology*, S. 1627 – 1630, Budapest, 1999.

T. Dutoit, F. Bataille, V. Pagel, N. Pierret, und O. van der Vreken. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In *Proc. ICSLP*, S. 1393–1396, Philadelphia, 1996.

Elsnet. European Corpus Initiative Multilingual Corpus I (ECI/MCI). http://www.elsnet.org/eci.html, 2008.

Y. Erikson und M. Alstermark. Fundamental Frequency correlates of the grave word accent in Swedish: the effect of vowel duration. In *Speech Transmission Laboratory, Quarterly Progress and Status Report*, Band 2–3, S. 53–60. KTH, Sweden, 1972.

C. Fabricius-Hansen, P. Gallmann, P. Eisenberg, R. Fiehler, und J. Peters. *Duden: Die Grammatik*. Dudenverlag, Mannheim, Leipzig, Wien, Zürich, 2009.

C. Féry. *German intonational patterns*. Niemeyer, Tübingen, 1993.

C. Féry, E. Kaiser, R. Hörnig, T. Weskott, und R. Kliegl. Perception of intonation contours on given and new referents: a completion study and an eye-movement experiment. In P. Boersma und S. Hamann (Hrsg.), *Phonology in Perception*, Phonology & Phonetics. Mouton de Gruyter, Berlin, New York, 2009.

J.L. Flanagan und M.G. Saslow. Pitch Discrimination for Synthetic Vowels. *JASA*, 30 (5):435–442, 1958.

D. B. Fry. Duration and intensity as physical correlates of linguistic stress. *JASA*, 27: 765–768, 1955.

D. B. Fry. Experiments in the perception of stress. *Language and speech*, 1:126–152, 1958.

H. Fujisaki. A note on physiological and physical basis for the phrase and the accent components in the voice fundamental frequency contour. In O. Fujimura (Hrsg.), *Vocal physiology: voice production, mechanisms, and functions*, S. 165–175. Raven, New York, 1987.

H. Fujisaki. Modeling the generation process of F0 contours as manifestation of linguistic and paralinguistic information. In *Proc. ICPhS*, S. 1–10, Aix-en-Provence, 1991.

H. Fujisaki und K. Hirose. Analysis of voice fundamental frequency contours for declrataive sentences of Japanese. *JASA*, 5(4):233–241, 1984.

W.A. Gale und G. Sampson. Good-turing frequency estimation without tears. *J. Quantitative Linguistics*, 2(3):217–237, 1995.

J.P Gee und F. Grosjean. Performance structures: a psycholinguistic and a linguistic appraisal. *Cognitive Psychology*, 15:411–458, 1983.

J. Goldsmith. *Autosegmental Phonology*. Doktorarbeit, MIT, Cambridge, 1976.

I.J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264, 1953.

E. Grabe. Pitch accent realisation in English and German. *Journal of Phonetics*, 26: 129–144, 1998.

M. Grice und R. Benzmüller. Transcription of German Intonation using ToBI tones; The Saarbruecken System. In *Phonus*, Band 1, S. 33–51. University of the Saarland, 1995.

M. Grice, M. Reyelt, R. Benzmüller, J. Mayer, und A. Batliner. Consistency in Transcription and Labelling of German Intonation with GToBI. In *Proc. ICSLP*, S. 1716–1719, New Castle, Delaware, 1996.

N. Grønnum. Prosodic parameters in a variety of Danish standard languages, with a view towards Swedish and German. *Phonetica*, 47:182–214, 1990.

N. Grønnum. Superposition and subordination in intonation: a non-linear approach. In *Proc. ICPhS*, Band 2, S. 124–131, Stockholm, 1995.

F. Grosjean und M. Collins. Breathing, pausing and reading. *Phonetica*, S. 98–114, 1979.

B. Grosz, A. Joshi, und S. Weinstein. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225, 1995.

B.J. Grosz und C.L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

P.D. Grünwald. *The Minimum Description Length Principle*. MIT press, Cambridge, MA, 2007.

H. Günther. *Schriftliche Sprache: Strukturen geschriebener Wörter und ihre Verarbeitung beim Lesen*, Band 40 aus *Konzepte der Sprach- und Literaturwissenschaft*. Niemeyer, Tübingen, 1988.

C. Gussenhoven. Focus, mode and nucleus. *J. of Linguistics*, 19:377–419, 1983.

C. Gussenhoven. *On the grammar and semantics of sentence accents*. Foris, Dordrecht, 1984. Neudruck von A semantic analysis of the nuclear tones of English (1983).

C. Gussenhoven. On the limits of focus projection in English. In P. Bosch und R. van der Sandt (Hrsg.), *Focus: Linguistic, cognitive, and computational perspectives*, S. 43–55. Cambridge University Press, Cambridge, 1999.

C. Gussenhoven. Intonation and interpretation: Phonetics and Phonology. In *Proc. Speech Prosody*, S. 47–57, Aix-en-Provence, 2002.

C. Gussenhoven. Experimental Approaches to Establishing Discreteness of Intonational Contrasts. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, und J. Schließer (Hrsg.), *Methods in Empirical Prosody Research*, Language, Context, and Cognition, S. 321–334. Mouton de Gruyter, Berlin, New York, 2006.

U. Gut und P.S. Bayerl. Measuring the Reliability of Manual Annotations of Speech Corpora. In *Proc. Speech Prosody*, S. 565–568, Nara, 2004.

Judith Haan. *Speaking of Questions. An Exploration of Dutch Question Intonation*. Doktorarbeit, Netherlands Graduate School of Linguistics, 2001.

M. A. K. Halliday. *Intonation and Grammar in British English*. Mouton, Den Haag, 1967a.

M. A. K. Halliday. Notes on transitivity and theme in English, part II. *Journal of Linguistics*, 3:199–244, 1967b.

J. Harrington, S. Palethorpe, und C.I. Watson. Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In *Proc. Interspeech*, S. 2753–2756, Antwerp, 2007.

M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proc. International Conference on Computational Linguistics*, Band 2, S. 539–545, Nantes, 1992.

M. Hearst. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64, 1997.

B.G. Henning. Frequency discrimination of random-amplitude tones. *JASA*, 39:336–339, 1966.

D.J. Hermes. Measuring the perceptual similarity of pitch contours. *Journal for Speech, Language, and Hearing Research*, 41:73–82, 1998.

D.J. Hermes und J.C. van Gestel. The frequency scale of speech intonation. *JASA*, 90: 97–102, 1991.

W. Hess. Grundlagen der Phonetik 4: Suprasegmentale Merkmale und Prosodie. Vorlesungsskript, 2003.

B. Heuft, T. Portele, F. Höfer, J. Krämer, H. Meyer, M. Rauth, und G. Sonntag. Parametric Description of F0-Contours in a Prosodic Database. In *Proc. ICPhS*, Band 2, S. 378–381, Stockholm, 1995.

B. Heuft, B. Streefkerk, und T. Portele. Evaluierung der automatischen Parametrisierung von Grundfrequenzkonturen. In *Proc. Elektronische Sprachsignalverarbeitung 7*, S. 170–175, Berlin, 1996.

M. Higashikawa und F.D. Minifie. Acoustical-Perceptual Correlates of Whisper Pitch in Synthetically Generated Vowels. *Speech, Language, and Hearing Research*, 42:583–591, 1999.

N. Higuchi, T. Hirai, und Y. Sagisaka. Effects of Speaking Style on Parameters of Fundamental Frequency Contour. In J.P.H. van Santen, R.W. Sproat, J.P. Olive, und J. Hirschberg (Hrsg.), *Progress in Speech Synthesis*, S. 417–428. Springer-Verlag, Berlin, 1997.

J. Hirschberg. Pitch Accent in Context: Predicting Intonational Prominence from Text. *Artificial Intelligence*, 63:305–340, 1993.

J. Hirschberg. A corpus-based approach to the study of speaking style. In M. Horne (Hrsg.), *Prosody: Theory and Experiment. Studies Presented to Gösta Bruce*, S. 335–350. Kluwer Academic Publishers, Dordrecht, 2000.

J. Hirschberg und J. Pierrehumbert. The intonational structuring of discourse. In *Proc. 24th Annual Meeting, Association for Computational Linguistics*, S. 136–144, New York, 1986.

J. Hirschberg, D. Litman, J. Pierrehumbert, und G. Ward. Intonation and the intentional structure of discourse. In *Proc. 10th international joint conference on Artificial intelligence*, S. 636–639, Mailand, 1987.

Daniel Hirst. Detaching intonational phrases from syntactic structure. *Linguistic Inquiry*, 24:781–788, 1993.

D.J. Hirst und A. Di Cristo (Hrsg.). *Intonation Systems. A survey of Twenty Languages.* Cambridge University Press, Cambridge, 1998.

D.J. Hirst und R. Espesser. Automatic modelling of fundamental frequency using a quadratic spline function. In *Travaux de l'Institut de Phonetique d'Aix*, Band 15, S. 71–85. 1993.

D. House. *Tonal Perception in Speech.* Lund University Press, Lund, 1990.

A. Isačenko und H.-J. Schädlich. *Untersuchung über die deutsche Satzintonation.* Akademie-Verlag, Berlin, 1964.

A.V. Isačenko und H.-J. Schädlich. *A model of standard German intonation.* Janua Linguarum, Series Practica. Mouton, The Hague Paris, 1970.

S. Jamieson. Likert scales: how to (ab)use them. *Medical Education*, 38(12):1217–1218, 2004.

J. Jan Van Santen, T. Mishra, und E. Klabbers. Estimating Phrase Curves in the General Superpositional Intonation Model. In *Proc. ISCA Speech Synthesis Workshop*, Pittsburgh, 2004.

M. Jilka. Regelbasierte Generierung natürlich klingender Intonation des Amerikanischen Englisch. Magisterarbeit, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 1996.

M. Jilka, G. Möhler, und G. Dogil. Rules for the Generation of ToBI-based American English Intonation. *Speech Communication*, 28:83–108, 1999.

K. Johnson. Speech perception without speaker normalization: An exemplar model. In Keith Johnson und John W. Mullennix (Hrsg.), *Talker Variability in Speech Processing*, S. 145–166. Academic Press, San Diego, 1997.

L. Karttunen. Discourse referents. In *Notes from the linguistic underground*, Band 7 aus *Syntax and semantics*, S. 363–586. Academic Press, London, 1976.

A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Shaker, Aachen, 1997.

D.H. Klatt. Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *JASA*, 53(1):8–16, 1973.

D.H. Klatt. Synthesis by rule of Segmental Durations in English Sentences. In B. Lindblom und S.E.G. Öhman (Hrsg.), *Frontiers of Speech Communication Research*, S. 287–299. Academic Press, 1979.

D.H. Klatt und L.C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *JASA*, 87(2):820–857, 1990.

F. Kleber. Form and function of falling pitch contours in English. In *Proc. Speech Prosody*, S. 61–64, Aix-en-Provence, 2006.

K. Kohler. PROLAB – the Kiel system of prosodic labelling. In *Proc. ICPhS*, S. 162–165, Stockholm, 1995a.

K. Kohler. *Einführung in die Phonetik des Deutschen.* Erich Schmidt Verlag, Berlin, 1995b.

K.J. Kohler. Categorical pitch perception. In *Proc. ICPhS*, S. 331–333, Tallinn, 1987.

K.J. Kohler. A model of German intonation. In *AIPUK*, Band 25, S. 295–360. 1991.

Klaus J. Kohler. F0 in the production of lenis and fortis plosives. *Phonetica*, 39:199–218, 1982.

C. Kuzla. *Prosodic Structure in Speech Production and Perception.* Doktorarbeit, Max Planck Institute for Psycholinguistics, Nijmegen, 2009.

D. R. Ladd. *The structure of intonational meaning: Evidence from English.* Indiana University linguistic Club, Bloomington, 1980.

D. R. Ladd. Intonational Phrasing: the case for recursive prosodic structure. In C.J. Ewen und J.M. Anderson (Hrsg.), *Phonology Yearbook*, Band 3, S. 311–340. Cambridge University Press, Cambridge, 1986.

D.R. Ladd. Declination: A review and some hypotheses. *Phonology Yearbook*, 1:53–74, 1984.

D.R. Ladd. An introduction to intonational phonology. In *Papers in Laboratory Phonology II: Gesture, segment, prosody*, S. 321–334. Cambridge University Press, Cambridge, 1992.

D.R. Ladd. *Intonational Phonology.* Cambridge University Press, Cambridge, MA, 1996.

R.D. Ladd und R. Morton. The perception of intonational emphasis: continuous or categorical? *Journal of Phonetics*, 25(3):313–342, 1997.

J.C. Lagarias, J.A. Reeds, M.H. Wright, und P.E. Wright. Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, 9 (1):112–147, 1998.

J. Laver. *The phonetic description of voice quality.* Cambridge University Press, 1980.

I. Lehiste. *Suprasegmentals.* MIT Press, Cambridge, MA, 1970.

W.J.M. Levelt. *Speaking: From Intention to Articulation.* ACL-MIT Press Series in Natural-Language Processing. MIT Press, Cambridge, MA, 1989.

M. Liberman. *The Intonation System of English.* Doktorarbeit, MIT, Cambridge, 1975.

M. Liberman und K. Church. Text analysis and word pronunciation in text-to-speech synthesis. In S. Furui und Sondhi M.M. (Hrsg.), *Advances in Speech Signal Processing*, S. 791–832. Dekker, New York, 1992.

M. Liberman und J. Pierrehumbert. Intonational Invariance under Changes in Pitch Range and Length. In M. Aronoff und R. Oehrle (Hrsg.), *Language Sound Structure*, S. 157–233. MIT Press, Cambridge, MA, 1984.

M. Liberman und A. Prince. On Stress and Linguistic Rhythm. *Linguistic Inquiry*, 8: 249–336, 1977.

B. Lindblom. Spectographic study of vowel reduction. *JASA*, 35(11):1773–1781, 1963.

B.E.F. Lindblom. Exolaining phonetic variation: A sketch of the H&H theory. In W. Hardcastle und A. Marchal (Hrsg.), *Speech Production and Speech Modeling*, S. 403–439. Kluwer, Dordrecht, 1990.

S.E. Linville. *Vocal Aging*. Singular Thomson Learning, San Diego, 2001.

A. Löfquist. Intrinsic and extrinsic f0 variations in Swedish tonal accents. *Phonetica*, 31: 228–247, 1975.

C.D. Manning und H. Schütze. *Foundations of statistical natural language processing*. MIT, Cambridge, Massachusetts, 2001.

J. Mayer. Transcribing German Intonation – The Stuttgart System. Forschungsbericht, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 1995.

J. Mayer. *Intonation und Bedeutung: Aspekte der Prosodie-Semantik-Schnittstelle im Deutschen*. Doktorarbeit, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 1997.

P. Mertens und C. d'Allessandro. Pitch Contour Stylization Using A Tonal Perception Model. In *Proc. ICPhS*, Band 4, S. 228–231, Stockholm, 1995.

H. Mixdorff. *Intonation Patterns of German – Model-based Quantitative Analysis and Synthesis of F0-Contours*. Doktorarbeit, TU Dresden, 1998.

H. Mixdorff. *An Integrated Approach to Modeling German Prosody*. Doktorarbeit, TU Dresden, 2002.

H. Mixdorff und H.R. Pfitzinger. A quantitative study of F0 peak alignment and sentence modality. In *Proc. Interspeech*, S. 1003–1006, Brighton, 2009.

Y. Mo, J. Cole, und M. Hasegawa-Johnson. Prosodic effects on vowel production: evidence from formant structure. In *Proc. Eurospeech*, S. 2535–2538, Brighton, 2009.

B. Möbius. *Ein quantitatives Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen*. Niemeyer-Verlag, Tübingen, 1993a.

B. Möbius. Perceptual evaluation of rule-generated intonation contours for German interrogatives. In *Proc. ESCA Workshop on Prosody*, S. 216–219, 1993b.

B. Möbius. Components of a quantitative model of German intonation. In *Proc. ICPhS*, Band 2, S. 108–115, Stockholm, 1995.

B. Möbius und M. Pätzold. F0 synthesis based on a quantitative model of German intonation. In *Proc. ICSLP*, S. 361–364, 1992.

G. Möhler. *Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese*. Doktorarbeit, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 1998a.

G. Möhler. Describing intonation with a parametric model. In *Proc. ICSLP*, S. 2851–2854, Sydney, 1998b.

G. Möhler. *Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese*. Doktorarbeit, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 1998c.

G. Möhler. Improvements of the PaIntE model. Forschungsbericht, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2001.

G. Möhler und A. Conkie. Parametric modeling of intonation using vector quantization. In *Proc. 3rd ESCA Workshop on Speech Synthesis*, S. 311–316, 1998.

B.C.J. Moore und B.R. Glasberg. A revision of Zwicker's loudness model. *Acta Acustica*, 82:335–345, 1996.

D. Mücke, M. Grice, J. Becker, A. Hermes, und S. Baumann. Articulatory and Acoustic Correlates of Prenuclear and Nuclear Accents. In *Proc. Speech Prosody*, S. 297–300, Dresden, 2006.

I. Nabelek und I.J. Hirsh. On the Discrimination of Frequency Transitions. *JASA*, 45 (6):1510–1519, 1969.

M.A. Nascimento und A.C.R. da Cunha. An Experiment Stemming Non-Traditional Text. In *Proc. SPIRE'98*, S. 75–80, Santa Cruz de La Sierra, Bolivien, 1998.

J.A. Nelder und R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

M. Nespor und I. Vogel. *Prosodic Phonology*. Foris, Dordrecht, 1986.

C.G. Nevill-Manning und I.H. Witten. Identifying hierarchical structure in sequences: A linear-time algorithm. *J. Artificial Intelligence Research*, 7:67–82, 1997.

O. Niebuhr. Categorical perception in intonation: a matter of signal dynamics? In *Proc. Interspeech*, S. 109–112, Antwerpen, 2007a.

O. Niebuhr. *Perzeption und kognitive Verarbeitung der Sprechmelodie: Theoretische Grundlagen und empirische Untersuchungen.* Doktorarbeit, IPDS, Christian-Albrechts-Universität zu Kiel, 2007b.

O. Niebuhr. Intonation segments and segmental intonations. In *Proceedings 10th Interspeech*, S. 2435–2438, Brighton, 2009.

O. Niebuhr und K.J. Kohler. Perception and cognitive processing of tonal alignment in German. In *Proc. International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages (TAL)*, S. 155–158, Beijing, 2004.

H. Niemann, E. Nöth, A. Kießling, R. Kompe, und A. Batliner. Prosodic Processing and its Use in Verbmobil. In *Proc. ICASSP*, S. 75–78, München, 1997.

F. Nolan. Intonational equivalence: an experimental evaluation of pitch scales. In *Proc. 15th ICPhS*, S. 771–774, Barcelona, 2003.

S.G. Nooteboom, J.P.L. Brokx, und J.J. De Rooij. Contributions of prosody to speech perception. In G.B. Levelt, W.J.M.and Flores d'Arcais (Hrsg.), *Studies in the Perception of Language*, S. 75–107. Wiley & Sons, New York, 1978.

J.O. Nordmark. Mechanisms of frequency discrimination. *JASA*, 44(6):1533–1540, 1968.

S. Noteboom. The Prosody of Speech: Melody and Rhythm. In W.J. Hardcastle und J. Laver (Hrsg.), *The Handbook of Phonetic Sciences*, S. 653–668. Blackwell, Oxford, 1997.

J.J. Ohala. Production of tone. In V. Fromkin (Hrsg.), *Tone: A linguistic survey*, S. 5–39. Academic Press, New York, 1978.

J.J. Ohala. Respiratory activity in speech. In W.J. Hardcastle und A. Marchal (Hrsg.), *Speech production and speech modelling*, S. 23–53. Kluwer Academic Publishers, Netherlands, 1990.

S.E.G. Öhman. Word and sentence intonation: a quantitative model. *Speech Transmission Laboratory—Quarterly Progress and Status Report*, 2–3:20–54, 1967.

S.E.G. Öhman. A model of word and sentence intonation. *Speech Transmission Laboratory—Quarterly Progress and Status Report*, 2–3:6–11, 1968.

S.E.G. Öhman und J. Lindqvist. Analysis-by-Synthesis of Prosodic Pitch Contours. In *STL-QPSR*, Band 4, S. 1–6. 1965.

H. Palmer. *English Intonation with Systematic Exercises.* Cambridge University Press, 1922.

S. Pan und J. Hirschberg. Modeling local context for pitch accent prediction. In *Proc. ACL*, S. 233–240, Hong Kong, 2000.

S. Pan und K.R. McKeown. Word Informativeness and Automatic Pitch Accent Modeling. In *EMNLP/VCL*, S. 148–157, 1999.

J. Peters. *Intonation deutscher Regionalsprachen*. Mouton de Gruyter, Berlin, New York, 2006.

C. Petrone und M. D'Imperio. Tonal structure and constituency in Neapolitan Italian: Evidence for the Accentual Phrase in statements and questions. In *Proc. Speech Prosody*, S. 301–304, Campinas, Brazil, 2008.

H.R. Pfitzinger. Five Dimensions of Prosody: Intensity, Intonation, Timing, Voice Quality, and Degree of Reduction. In *Speech Prosody Abstract Book*, S. 6–9, Dresden, 2006.

H.R. Pfitzinger, S. Burger, und S. Heid. Syllable Detection in Read and Spontaneous Speech. In *Proc. ICSLP*, Band 2, S. 1261–1264, Philadelphia, 1996.

H.R. Pfitzinger, H. Mixdorff, und J. Schwarz. Comparison of Fujisaki-model extractors and F0 stylizers. In *Proc. Interspeech*, S. 2455–2458, Brighton, 2009.

J. Pierrehumbert. *The phonology and phonetics of Englisch intonation*. Doktorarbeit, MIT, Cambridge, MA, 1980.

J. Pierrehumbert und M. Beckman. *Japanese tone structure*. MIT Press, Cambridge, Massachusetts, 1988.

J. Pierrehumbert und J. Hirschberg. The Meaning of Intonational Contours in the Interpretation of Discourse. In P.R. Cohen, J. Morgan, und M.E. Pollack (Hrsg.), *Intentions in Communication*, S. 271–311. MIT Press, Cambridge, 1990.

J. Pierrehumbert und S.A. Steele. Categories of tonal alignment in English. *Phonetica*, 46:181–196, 1989.

K.L. Pike. *The intonation of American English*, Band 1 aus *University of Michigan publications*. University of Michigan Press, Ann Arbor, 1945.

J. Pitrelli, M. Beckman, und J. Hirschberg. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proc. ICSLP*, S. 123–126, Yokohama, 1994.

I. Pollack. Detection of rate of change of auditory frequency. *J. Experimental Psychology*, 77(4):535–41, 1968.

T. Portele, B. Steffan, R. Preuss, W.F. Sendlmeier, und W. Hess. HADIFIX - a speech synthesis system for German. In *Proc. ICSLP*, S. 1227–1230, Banff, 1992.

S. A. Prevost. *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. Doktorarbeit, University of Pennsylvania, 1995.

E. Prince. Toward a taxonomy of given-new information. In *Radical Pragmatics*, S. 223–255. Academic Press, New York, 1981.

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, 1993.

L.R. Rabiner, M.R. Sambur, und C.E. Schmidt. Applications of nonlinear smoothing algorithm to speech processing. *IEEE Trans. ASSP*, S. 552–557, 1975.

S. Rapp. Automatic labelling of German prosody. In *Proc. ICSLP*, S. 1267–1270, 1998a.

S. Rapp. *Automatisierte Erstellung von Korpora für die Prosodieforschung*. Doktorarbeit, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 1998b.

T. Rathcke. *Komparative Phonetik und Phonologie der Intonationssysteme des Deutschen und des Russischen*. Doktorarbeit, Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität, München, 2008.

D.R. Reddy. Pitch period determination of speech sounds. *Commun. ACM*, 20:343–348, 1967.

U.D. Reichel. Textbasierte Vorhersage prosodischer Strukturierung. Magisterarbeit, University of Munich, 2002.

U.D. Reichel. Improving Data Driven Part-of-Speech Tagging by Morphologic Knowledge Induction. In *Proc. AST Workshop*, S. 65–73, Maribor, 2005a.

U.D. Reichel. Balloon. Vortragsfolien, 2005b.

U.D. Reichel. Data-driven Extraction of Intonation Contour Classes. In *Proc. 6th ISCA Workshop on Speech Synthesis*, S. 240–245, Bonn, 2007a.

U.D. Reichel. Text-based prediction of automatically extracted intonation contour classes. In *Proc. AST Workshop*, Maribor, 2007b.

U.D. Reichel und H.R. Pfitzinger. Text Preprocessing for Speech Synthesis. In *Proc. TC-Star Speech to Speech Translation Workshop*, S. 207–212, Barcelona, Spain, 2006.

U.D. Reichel und F. Schiel. Using Morphology and Phoneme History to improve Grapheme-to-Phoneme Conversion. In *Proc. Eurospeech*, S. 1937–1940, Lisboa, 2005.

U.D. Reichel und K. Weilhammer. Automated Morphological Segmentation and Evaluation. In *Proc. LREC*, S. 503–506, Lisbon, Portugal, 2004.

U.D. Reichel und R. Winkelmann. Removing micromelody from fundamental frequency contours. In *Proc. Speech Prosody*, Chicago, 2010.

U.D. Reichel, F. Kleber, und R. Winkelmann. Modelling similarity perception of intonation. In *Proc. Eurospeech*, S. 1711–1714, 2009.

M. Reyelt und A. Batliner. Ein Inventar prosodischer Etiketten für VERBMOBIL. Forschungsbericht, Verbmobil Memo 33, 1994.

M. Reyelt, M. Grice, R. Benzmüller, J. Mayer, und A. Batliner. Prosodische Etikettierung des Deutschen mit ToBI. In D. Gibbon (Hrsg.), *Natural Language and Speech Technology, Results of the third KONVENS conference*, S. 144–155. Mouton de Gruyter, Berlin, New York, 1996.

A. Riester. A Semantic Explication of Information Status and the Underspecification of the Recipients' Knowledge. In A. Grønn (Hrsg.), *Proc. Sinn und Bedeutung*, Band 12, S. 507–522, Oslo, 2008.

T. Rietveld und A. Chen. How to obtain and process perceptual judgements of intonational meanings. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, und J. Schließer (Hrsg.), *Methods in Empirical Prosody Research*, Language, Context, and Cognition, S. 283–320. Mouton de Gruyter, Berlin, New York, 2006.

T. Rietveld und C. Gussenhoven. Aligning pitch targets ins speech synthesis: Effects of syllable structure. *Journal of Phonetics*, 23:375–285, 1995.

T. Rietveld und P. Vermillion. Cues for Perceived Pitch Register. *Phonetica*, 60:261–272, 2003.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

R.J. Ritsma. Pitch discrimination and frequency discrimination. In *Proc. 5th International Congress on Acoustics*, Liège, 1965. paper B22.

R.J. Ritsma. Frequencies dominant in the perception of the pitch of complex sounds. *JASA*, 42:191–198, 1967.

M. Rossi. Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica*, 23:1–33, 1971.

A. Sakurai, K. Hirose, und N. Minematsu. Data-driven generation of F0 contours using a superpositional model. *Speech Communication*, S. 535–549, 2003.

A. Savitzky und M.J.E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

K. Schaefer-Vincent. Pitch period detection and chaining: Method and evaluation. *Phonetica*, 40:177–202, 1983.

F. Schiel. Bavarian Archive for Speech Signals, Siemens Synthesis Corpus - SI1000P. http://www.phonetik.uni-muenchen.de/forschung/Bas/BasSI1000Peng.html, 1998.

F. Schiel. Automatic Phonetic Transcription of Non-Prompted Speech. In *Proc. ICPhS*, S. 607–610, San Francisco, 1999.

F. Schiel, Chr. Draxler, Ph. Hoole, und H.G. Tillmann. New ressources at BAS: acoustic, multimodal, linguistic. In *Proc. Eurospeech*, S. 2271–2274, Budapest, 1999.

B. Schouten, E. Gerrits, und A. van Hessen. The end of categorical perception as we know it. *Speech Communication*, 41:71–80, 2003.

M. Schröder und J. Trouvain. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6:365–377, 2003.

A. Schweitzer und B. Möbius. Experiments in Automatic Prosodic Labeling. In *Proc. Eurospeech*, S. 2515–2518, Brighton, 2009.

A. Schweitzer, N. Braunschweiler, und E. Morais. Prosody generation in the Smartkom project. Forschungsbericht, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2002.

E.O. Selkirk. *Phonology and syntax: the relation between sound and structure*. MIT Press, Cambridge, MA, 1984.

R.L. Sergeant und J.D. Harris. Sensitivity to Unidirectional Frequency Modulation. *JASA*, 34(10):1625–1628, 1962.

D.F. Shanno. Conditioning of Quasi-Newton Methods for Function Minimization. *Mathematics of Computing*, 24:647–656, 1970.

K. Silverman und J. Pierrehumbert. The timing of prenuclear high accents in English. In J. Kingston und M.E. Beckman (Hrsg.), *Papers in Laboratory Phonology*, S. 72–106. Cambridge University Press, Cambridge, 1990.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, und J. Hirschberg. TOBI: A standard for labeling English prosody. In *Proc. ICSLP'92*, S. 867–870, 1992.

K.E.A. Silverman. What causes vowels to have intrinsic fundamental frequency? *Cambridge Papers in Phonetics and Experimental Linguistics*, 3:1–15, 1984.

J.O. Smith und J.S. Abel. The Bark and ERB Bilinear Transforms. *IEEE Transactions on Speech and Audio Processing*, 7(6):697–708, 1999.

P. Specker. A powerful postprocessing algorithm for time-domain pitch trackers. In *Proc. ICASSP*, New York, 1984. paper 28B.2.

J.M. Stewart. The typology of the Twi tone system. In *Bulletin of the Institute of African Studies*, S. 1–27. University of Ghana, 1965.

E. Stock und C. Zacharias. *Deutsche Satzintonation*. VEB Verlag Enzyklopädie, Leipzig, 1982.

G. Stoll. Pitch of vowels: experimental and theoretical investigation of its dependence on vowel quality. *Speech Communication*, 3:137–150, 1984.

H. Strik und L. Boves. Downtrend in F0 and Psb. *J. of Phonetics*, 23:203–220, 1995.

M. Swerts und R. Geluykens. Prosody as a marker of information flow in spoken discourse. *Language and Speech*, 37(1):21–43, 1994.

A.K. Syrdal, G. Möhler, K. Dusterhoff, A. Conkie, und A.W. Black. Three Methods of Intonation Modeling. In *Proc. Third International Workshop on Speech Synthesis*, S. 305–310, Jenolan Caves, 1998.

A.K. Syrdal, J. Hirschberg, J. McGory, und M. Beckman. Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication*, 33(1–2): 135–151, 2001.

P. Taylor. Analysis and Synthesis of Intonation using the Tilt Model. *Journal of the Acoustical Society of America*, 107:1697–1714, 2000.

P. Taylor und A.W. Black. Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language*, 12:99–117, 1998.

P.A. Taylor. The rise/fall/connection model of intonation. *Speech Communication*, 15: 169–186, 1995.

E. Terhardt. Calculating virtual pitch. *J. Hearing Research*, 1:155–182, 1979.

E. Terhardt. *Akustische Kommunikation*. Springer, Berlin/Heidelberg, 1998.

J. Terken. Fundamental Frequency and perceived prominence of accented syllables. *JASA*, 89(4):1768–1776, 1991.

J. Terken. Fundamental Frequency and perceived prominence of accented syllables II: Nonfinal accents. *JASA*, 95:3662–3665, 1994.

J. t'Hart, R. Collier, und A. Cohen. *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge University Press, Cambridge, 1990.

K. Thomassen. Categoriale perceptie bij 2IFC en ABX. Magisterarbeit, Utrecht University, 1993.

N.G. Thorsen. Intonation and text in Standard Danish. *JASA*, 77:1205–1216, 1985.

B. Tischer. *Die vokale Kommunikation von Gefühlen*. Fortschritte der psychologischen Forschung. Beltz Psychologie Verlags Union, Weinheim, 1993.

I.R. Titze. Physiologic and acoustic differences between male and female voices. *JASA*, 85:1699–1707, 1989a.

I.R. Titze. On the relation between subglottal pressure and fundamental frequency in phonation. *JASA*, 85(2):901–906, 1989b.

H. Traunmüller. Some aspects of the sound of speech sounds. In M.E.H. Schouten (Hrsg.), *The Psychophysics of Speech Perception*, S. 293–305. Martinus Nijhoff Publishers, Dordrecht, 1987.

E. Uldall. Attitudinal meanings conveyed by intonation contours. *Langiage and Speech*, 3:223–234, 1960.

E. Vallduví. Information packaging: A survey. Forschungsbericht HCRC/RP-44, 1993.

K. van Deemter. Towards a blackboard model of accenting. *Computer Speech and Language*, 12(3):143–164, 1998.

J. van Santen und J. Hirschberg. Segmental effects on timing and height of pitch contours. In *Proc. ICSLP*, S. 719–722, Yokohama, 1994.

J.P.H. van Santen, B. Möbius, J. Venditti, und C. Shih. Description of the Bell Labs intonation system. In *Proc. Third International Workshop on Speech Synthesis*, S. 293–298, Jenolan Caves, Australia, 1998.

N. M. Veilleux. *Computational models of the prosody/syntax mapping for spoken language systems*. Doktorarbeit, College of Engineering, Boston University, Boston, 1994.

H. Vereecken, J.-P. Martens, C. Grover, J. Fackrell, und B. Van Coile. Automatic Prosodic Labeling of 6 Languages. In *Proc. ICSLP*, S. 1399–1402, Sydney, 1998.

A. Wagner. Analysis and Recognition of Accentual Patterns. In *Proc. Eurospeech*, S. 2427–2430, Brighton, 2009.

R.A. Wagner und M.J. Fischer. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173, 1974.

G. Ward und J. Hirschberg. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, S. 747–776, 1985.

P. Welby. Effects of Pitch Accent Position, Type, and Status of Focus Projection. *Language and Speech*, 46(1):53–81, 2003.

A.D. Well. Perceptual factors in reading. In *Eye movements in reading – Perceptual and language processes*, S. 41–51. Rayner, K., 1983.

R.S. Wells. The Pitch Phonemes of English. *Language*, S. 27–39, 1945.

W.E. Welmers. Tonemics, morphotonemics, and tonal morphemes. *General Linguistics*, 4:1–9, 1959.

E.G. Wever. Action currents in the auditory nerve in response to acoustical stimulations. *Proc. National Acadademy of Science*, 16:344–350, 1930.

D.H. Whalen und A.G. Levitt. The universality of intrinsic F0 of vowels. *J. Phonetics*, 23:349–366, 1995.

C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, und P. Price. Segmental Durations in the Vicinity of Prosodic Phrase Boundaries. *JASA*, 91(3):1707–1717, 1992.

C.W. Wightman. ToBI Or Not ToBI? In *Proc. Speech Prosody*, S. 25–29, Aix-en-Provence, 2002.

S. Winkler. *Focus and Secondary Predication*. Mouton de Gruyter, Berlin, New York, 1997.

S.A. Xue und D. Deliyski. Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications. *Educational Gerontology*, S. 159–168, 2001.

M. L. Zubizarreta. *Prosody, Focus, and Word Order*. MIT Press, Cambridge, Massachusetts, 1998.

Anhang A

Parameter der phonetischen Regressionsmodelle

Alle Modellgleichungen auf die hier im Folgenden referiert wird, sind in Abschnitt 10.4 zu finden. Die Gewichte der nachfolgenden Tabellen resultieren aus der Entwicklung der Regressionsmodelle anhand des kompletten Korpus.

Anpassung globaler Konturen Die vier Prädiktoren aus dem linearen Regressionsmodell in Gleichung 10.8 wurden mittels Hauptkomponentenanalyse orthogonalisiert. Tabelle A.1 zeigt die Gewichtung der resultierenden Hauptkomponenten.

w_0	-0.0750
w_1	-0.0748
w_2	-0.1608
w_3	0.0057

Tabelle A.1: Gewichte im Regressionsmodell zur Anpassung globaler Konturen.

Anpassung lokaler Konturen Zur Vorhersage jedes der vier Polynomkoeffizienten der lokalen Konturen wurde jeweils ein lineares Regressionsmodell trainiert (siehe Gleichung 10.9). In Tabelle A.2 finden sich die jeweiligen Gewichte der Hauptkomponenten, die durch Orthogonalisierung der acht Prädiktoren aus den Regressionsmodellen zur Vorhersage der Polynomkoeffizienten in Gleichung 10.9 ermittelt wurden.

Pitch Reset Auch für die Prädiktoren aus Gleichung 10.10 zur Vorhersage des Pitch Resets wurde eine Hauptkomponentenanalyse durchgeführt. Die Gewichte der daraus hervorgegangenen Hauptkomponenten finden sich in Tabelle A.3.

	s_0	s_1	s_2	s_3
w_0	-0.0459	0.0377	0.1513	-0.0994
w_1	-0.0012	-0.0043	0.0174	0.0515
w_2	0.0081	-0.0255	0.0082	0.1146
w_3	-0.1182	0.1017	0.1236	-0.0724
w_4	0.0723	0.0332	-0.0628	0.0208
w_5	0.1028	0.0960	0.1201	0.0638
w_6	0.0106	0.0688	-0.0384	-0.0583
w_7	0.0304	-0.0514	0.0026	0.0076

Tabelle A.2: Gewichte im Regressionsmodell zur Anpassung lokalen Konturen. Ein Satz je Polynomkoeffizient s_n.

w_0	0.0020
w_1	-0.1821
w_2	0.0718
w_3	-0.0493
w_4	-0.1464

Tabelle A.3: Gewichte im Regressionsmodell zur Vorhersage des Pitch Resets.

Anhang B

Lautdauernmodellierung

B.1 Intrinsische Lautdauern

Klasse	**Laute**	**intrinsische Dauer** [ms]
hohe Langvokale	i: y: u:	90
mittelhohe Langvokale	e: 2: E:	86
tiefe Langvokale	a:	121
hohe Kurvokale	I Y U	69
mittelhohe Kurvokale	9 E O	75
tiefe Kurzvokale	a	88
Standard-Diphtonge	aI aU OY	132
lange 6-Diphtonge	2:6 E:6 a:6 e:6 i:6 o:6 u:6 y:6	130
kurze 6-Diphtonge	E6 96 I6 O6 U6 Y6 a6	92
Schwa	@ 6	59
stimmlose Plosive	p t k	64
stimmhafte Plosive	b d g	49
Glottal Stop	Q	50
Stimmlose Frikative	f s S C x h	71
Stimmhafte Frikative	v z Z	64
Nasale	m n N	71
Laterale	l	64
Approximanten	j	67
Trills	r R	54
Pausen	<P>	420

Tabelle B.1: Intrinsische Lautdauern $\bar{d}$ (in Millisekunden) als arithmetische Dauermittelwerte innerhalb der entsprechenden Lautklassen.

B.2 Modell zur Vorhersage des Daueranpassungsfaktors

- Modell: $\hat{d}_x = \bar{d}_x \cdot f$.
- $\bar{d}_x$: inhärente Dauer; $\hat{d}_x$: kontextabhängige Dauer von Laut x.
- kontextabhängige Vorhersage von f anhand eines Regressionsbaums:

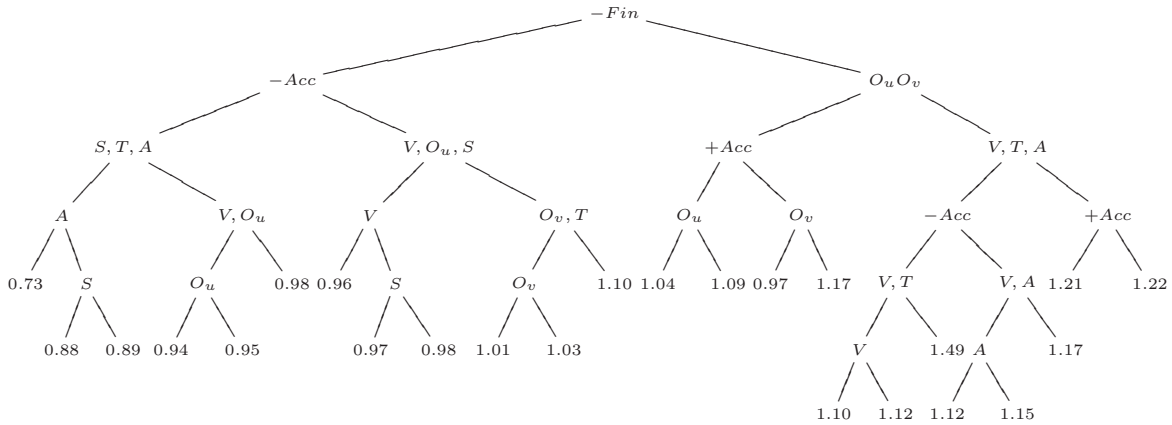

Abbildung B.1: Regressionsbaum zur Vorhersage des Faktors f im Dauermodell. An Verzweigungen bedeutet der linke Ast „Attributwert gegeben" und der rechte Ast „Attributwert nicht gegeben".

Attribut	**Werte**	**Erläuterung**
Akzentuierung	*+Acc*	in akzentuierter Silbe
	–Acc	in nicht akzentuierter Silbe
Phrasenfinalität	*+Fin*	phrasenfinal
	–Fin	nicht phrasenfinal
Lautklasse	V	Vokal
	A	Approximant
	S	Sonorant (Nasal,Lateral)
	T	Trill
	O_v	stimmhafter Obstruent
	O_u	stimmloser Obstruent

Tabelle B.2: Attribute des Regressionsbaums.

Anhang C

Stimuli

C.1 Zielwörter in den Perzeptionsexperimenten 1–3

Bahre	(Tragegestell)	Leier	(Musikinstrument)
Beere	(Frucht)	Liege	(Schlafstätte)
Beule	(Verletzung)	Lilie	(Blume)
Bibel	(Buch)	Made	(Insekt)
Biene	(Insekt)	Mauer	(Hindernis)
Birke	(Baum)	Meise	(Vogel)
Birne	(Obst)	Mine	(Bauwerk)
Blase	(Organ)	Möhre	(Gemüse)
Blume	(Pflanze)	Möwe	(Vogel)
Bluse	(Kleidungsstück)	Mühle	(Gebäude)
Bohne	(Gemüse)	Murmel	(Spielzeug)
Börse	(Marktplatz)	Mumie	(Ausgrabungsfund)
Brise	(Luftstrom)	Nadel	(Nähwerkzeug)
Brühe	(Flüssigkeit)	Narbe	(Verletzung)
Bude	(Gebäude)	Nase	(Organ)
Bühne	(Plattform)	Niere	(Organ)
Diele	(Zimmer)	Nudel	(Teigware)
Dose	(Gefäß)	Rebe	(Pflanze)
Droge	(Substanz)	Robe	(Kleidungsstück)
Drüse	(Organ)	Röhre	(Hohlraum)
Düse	(Rohr)	Rose	(Blume)
Gabel	(Essgerät)	Sahne	(Milchprodukt)
Garde	(Truppe)	Sauna	(Raum)
Geige	(Musikinstrument)	Sohle	(Unterlage)
Geisel	(Gefangene)	Vase	(Gefäß)
Grube	(Vertiefung)	Vene	(Blutgefäß)
Gurke	(Gemüse)	Waage	(Messinstrument)
Kiwi	(Obst)	Wade	(Körperteil)
Laube	(Gebäude)	Waise	(Kind)
Leber	(Organ)	Weide	(Baum)

Tabelle C.1: Zielwörter (mit zugehörigen Hyperonymen)

C.2 Satzpaare für das Perzeptionsexperiment 6

In den folgenden Satzpaaren bildet jeweils der erste Satz den Diskurskontext und der zweite Satz den Zielsatz, in den die Intonationskontur auf zwei der lokalen Segmente getrennt variiert wurden.

Für jedes der lokalen Segmente wird im Zielsatz getrennt der Diskursstatus hinsichtlich Neuheit und Finalität angegeben, die darauf basierende PKS-EB-Vorhersage V der adäquaten lokalen Konturklasse c_v, sowie die zum Vergleich erzeugten Varianten V_n, V_f und V_0, unter denen ebenfalls lokale Konturklassen zu verstehen sind, die mit c_v nur in der Neuheitscodierung (V_n) oder nur in der Finalitätscodierung (V_f), oder in keiner von beiden (V_0) übereinstimmen.

Satzpaar 1

Diskurskontext	*Dort steht eine Buche.*	
Zielsatz	[*Die Buche*]$_{s1}$ *verliert* [*ihre Blätter*]$_{s2}$.	
s_1	Status	gegeben, non-final
	Varianten	V: c_4, V_n: c_1, V_f: c_2, V_0: c_5
s_2	Status	neu, final
	Varianten	V: c_5, V_n: c_2, V_f: c_1, V_0: c_4

Satzpaar 2

Diskurskontext	*Dort steht eine Buche.*	
Zielsatz	[*Auch ein Traktor*]$_{s1}$ [*und ein Ochse*]$_{s2}$.	
s_1	Status	neu, non-final
	Varianten	V: c_2, V_n: c_5, V_f: c_4, V_0: c_1
s_2	Status	neu, final
	Varianten	V: c_5, V_n: c_2, V_f: c_1, V_0: c_4

Satzpaar 3

Diskurskontext	*Dort steht eine Buche.*	
Zielsatz	[*Die Kinder*]$_{s1}$ *bewundern* [*die Buche*]$_{s2}$.	
s_1	Status	neu, non-final
	Varianten	V: c_2, V_n: c_5, V_f: c_4, V_0: c_1
s_2	Status	gegeben, final
	Varianten	V: c_1, V_n: c_4, V_f: c_5, V_0: c_2

Satzpaar 4

Diskurskontext	*Dort stehen eine Buche und eine Scheune.*	
Zielsatz	*Die* $[Buche]_{s1}$ *verdunkelt die* $[Scheune]_{s1}$.	
s_1	Status	gegeben, non-final
	Varianten	V: c_4, V_n: c_1, V_f: c_2, V_0: c_5
s_2	Status	gegeben, final
	Varianten	V: c_1, V_n: c_4, V_f: c_5, V_0: c_2

Die lokalen Konturen im Diskurskontext, sowie in den verbleibenden lokalen Segmenten der Zielsätze wurden impressionistisch mit dem Ziel größtmöglicher Natürlichkeit wie folgt festgelegt:

Lokales Segment	**Konturklasse**
[*Dort steht eine Buche*]	c_5
[*Dort stehen eine Buche*] ...	c_4
... [*und eine Scheune*]	c_5
... [*verliert*] ...	c_1
... [*bewundern*] ...	c_5
... [*verdunkelt*] ...	c_5

Anmerkung: Das Verb *steht* wurde hier wie ein Auxiliar behandelt, dominiert also kein eigenes lokales Segment.

Anhang D

Versuchspersonenanleitungen für die Perzeptionsexperimente

D.1 Anleitung für Perzeptionsexperimente 1–5

Allgemeine Vorbemerkungen

1. Es handelt sich um künstlich erzeugte Äußerungen, deren Lautqualität Schwächen aufweisen mag. Versuchen Sie, diese Schwächen möglichst zu ignorieren und nur auf die Sprechmelodie zu achten.

2. In jedem der Teilexperimente haben Sie die Aufgabe, die präsentierte Sprechmelodie hinsichtlich einer bestimmten Fragestellung zu beurteilen. Dafür steht Ihnen eine 5-stufige Skala in Form einer Knopfreihe zur Verfügung. Die beiden Enden der Skala sind mit Urteilsalternativen versehen. Klicken Sie auf die am weitesten außen befindlichen Knöpfe, wenn Sie sich relativ sicher bei der Beurteilung sind. Wählen Sie den zweiten Knopf von links oder rechts, wenn Sie zu einer der beiden Alternativen tendieren. Wählen Sie den mittleren Knopf, wenn Sie sich nicht entscheiden können.

3. Lesen Sie bitte vor der Durchführung jedes der Teilexperimente den entsprechenden Abschnitt dieser Anleitung durch. Jedem der Teilexperimente geht eine kurze Eingewöhnung voraus.

1. Teilexperiment

Ziel dieses Experiments ist es, herauszufinden, mit welchen Sprechmelodien bereits bekannte und neue Informationen übermittelt werden können.

Hierzu werden Ihnen schriftlich zwei Fragen wie die folgenden:

Ist das eine Harfe?
Ist das ein Musikinstrument?

präsentiert und über Kopfhörer eine Antwort der Form

Ja, eine Harfe.

Bezogen auf die Frage *"Ist das eine Harfe?"* enthält die Antwort *"Ja, eine Harfe."* über eine Bestätigung hinaus keine zusätzliche Information. Anders in Bezug auf die Frage *"Ist das ein Musikinstrument?"*. Hier besteht die neue Information in der Konkretisierung, dass es sich bei dem Musikinstrument um eine Harfe handelt.

Ihre Aufgabe besteht nun darin, anhand der Sprechmelodie der Antwort zu beurteilen, zu welcher der beiden Fragen die Antwort eher passt, ob sie also nur eine einfache Bestätigung darstellt oder darüber hinaus zusätzliche Information beinhaltet.

Für Ihr Urteil haben Sie eine 5-stufige Skala zur Verfügung, an deren Enden die beiden Frage-Alternativen zu finden sind, links die einfache Bestätigung rechts die Hinzufügung neuer Information. Bitte verfahren Sie bei der Beurteilung so wie in Vorbemerkung 2 angegeben.

2. Teilexperiment

In diesem Teilexperiment geht es um die Frage, wie der Sprecher durch die Sprechmelodie zeigt, wieviel **Bedeutsamkeit** er seiner Aussage beimisst.

Ihnen werden über Kopfhörer Aussagen wie:

Das ist eine Flasche.

präsentiert, mit der Aufgabe, diese hinsichtlich der beigemessenen Relevanz zu beurteilen. Hierzu steht Ihnen wieder eine 5-stufige Skala zur Verfügung, diesmal mit den Endpunkten

belanglos
bedeutsam.

Bitte verfahren Sie bei der Beurteilung so wie in Vorbemerkung 2 angegeben.

3. Teilexperiment

Ziel dieses Teilexperiments ist es, herauszufinden, wie der Sprecher mit seiner Sprechmelodie markiert, ob er **weitersprechen** möchte oder am **Ende seines Redebeitrags** angelangt ist.

Hierzu sehen Sie im Display zwei alternative Antworten auf die Frage *"Was siehst Du?"*. Zum Beispiel:

Eine Flasche und eine Seife.
Eine Flasche.

Über Kopfhörer hören Sie:

Eine Flasche

Ihre Aufgabe besteht nun darin, anhand der Sprechmelodie zu beurteilen, ob das Gehörte eher Teil der längeren Antwort *"Eine Flasche und eine Seife."* ist, oder eine abgeschlossene Antwort darstellt (*"Eine Flasche."*). Gemeint sind in beiden Fällen "neutrale" Antworten, das heißt: keine Gegenfragen, ohne Ausdruck von Überraschung und ohne emotionale Markierung.

Zur Beurteilung steht Ihnen wieder eine 5-stufige Skala zwischen den beiden Antwort-Alternativen zur Verfügung. Bitte verfahren Sie bei der Beurteilung so wie in Vorbemerkung 2 angegeben.

4. Teilexperiment

Ihre Aufgabe besteht hier in der Beurteilung der **Natürlichkeit** der Sprechmelodie. Hierzu werden Ihnen kurze Sprachausschnitte eines Nachrichtensprechers vorgespielt, die Sie auf einer 5-stufigen Skala mit den Endpunkten

natürlich
sehr unnatürlich

bewerten können. Zur Erinnerung **Es handelt sich um künstlich erzeugte Äußerungen, deren Lautqualität Schwächen aufweisen mag. Versuchen Sie diese Schwächen möglichst zu ignorieren und nur auf die Sprechmelodie zu achten.**

5. Teilexperiment

Abschließend geht es hier noch einmal um die Beurteilung der Sprechmelodie hinsichtlich der behandelten Gesichtspunkte:

- Übermittlung **bereits bekannter gegenüber neuer Information**.
- Markierung der **Bedeutsamkeit** des Gesagten.
- **Fortführung gegenüber Abschluss** des Redebeitrags.

Ihnen werden hierzu Ausschnitte aus Äußerungen eines Nachrichtensprechers präsentiert. Versuchen Sie bitte, den **Inhalt** der Äußerungen so gut es geht zu **ignorieren** und sich auf die Sprechmelodie zu konzentrieren.

Die Ausschnitte stammen aus Aussagesätzen, sollten also nicht als Fragen interpretiert werden.

Zu jedem der Ausschnitte wird eingeblendet, in welcher der drei genannten Dimensionen die Sprechmelodie beurteilt werden soll. Bitte nutzen Sie wie in den Teilexperimenten 1–3 die 5-stufigen Skalen für Ihre Einschätzung.

D.2 Anleitung für Perzeptionsexperiment 6

Ihre Aufgabe: Sie sehen im Display ein Satzpaar, von dem Sie sich mittels der Knöpfe auf der linken Seite vier Versionen beliebig oft anhören können. Die Versionen unterscheiden sich nur im Hinblick auf die Sprechmelodie des zweiten Satzes. Ihre Aufgabe besteht nun darin, für jede der vier Versionen die Sprechmelodie **auf dem mit spitzen Klammern gekennzeichneten Abschnitt des zweiten Satzes** zu beurteilen, so geht es also beispielsweise im Satz:

>>Die Buche<< verliert ihre Blätter.

um die Sprechmelodie auf *Die Buche*. Hierzu sollten Ihnen folgende Gütekriterien dienen:

- Befindet sich der Abschnitt **am Ende des Satzes**, so sollte die Sprechmelodie den Satz als eine abgeschlossene Aussage kennzeichnen. Es sollte also nicht der Eindruck aufkommen, dass noch ein Teil der Äußerung fehlt, oder dass es sich hierbei um eine Frage handelt.
- Befindet sich der Abschnitt **mitten im Satz**, sollte anhand der Sprechmelodie erkennbar sein, dass noch ein Teil der Äußerung aussteht.
- Die Sprechmelodie sollte **Wortwiederholungen und neu hinzukommende Wörter** auf geeignete Weise kennzeichnen.

Für Ihre Urteile stehen Ihnen fünfstufige Skalen von **adäquat (links)** bis **inadäquat (rechts)** zur Verfügung.

Bitte beachten Sie: Es handelt sich um künstlich erzeugte Äußerungen, deren Lautqualität Schwächen aufweisen mag. Versuchen Sie, diese Schwächen möglichst zu ignorieren und nur auf die Sprechmelodie zu achten.

Anhang E

Screenshots der Experiment-Oberflächen

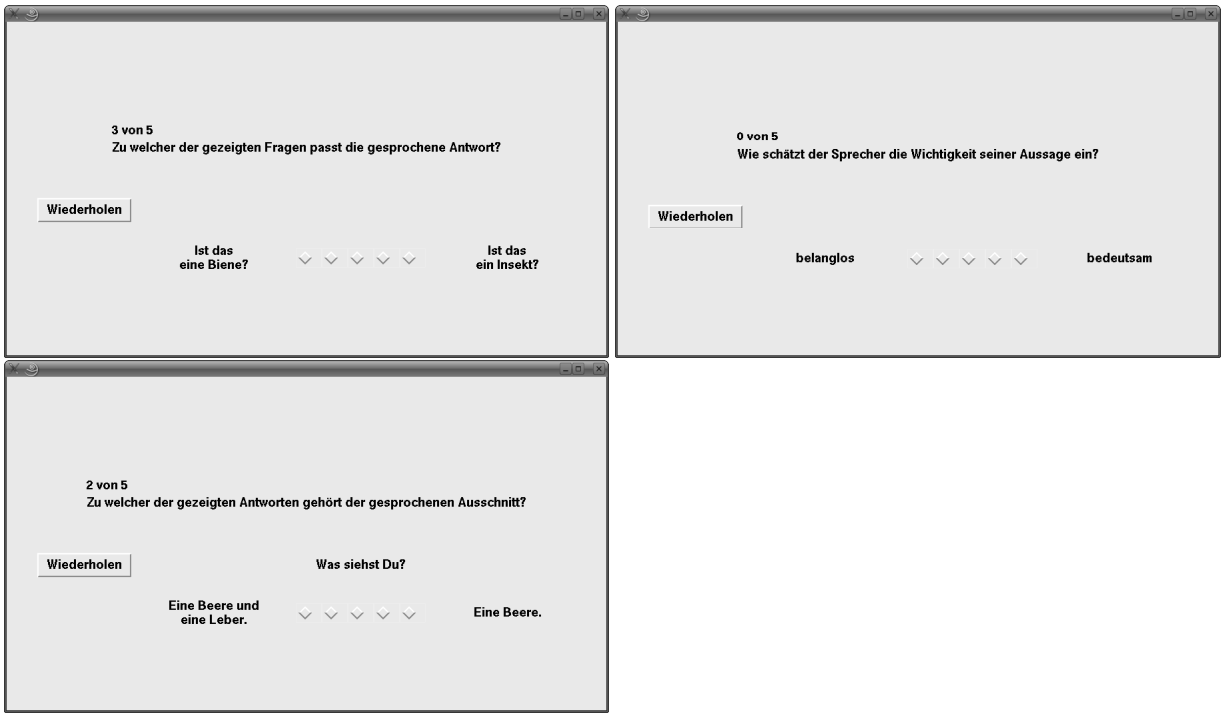

Abbildung E.1: Screenshots der Experimentoberflächen zur Beurteilung der informativen Neuheit (**oben links**), der Bedeutsamkeit (**oben rechts**) sowie der Finalität (**unten**).

Abbildung E.2: Screenshots der Experimentoberflächen zur vergleichenden Beurteilung der Sprecherintention bei Original- und modellierter F0-Konturen hinsichtlich informativer Neuheit (**oben links**), Bedeutsamkeit (**oben rechts**) sowie Finalität (**unten**).

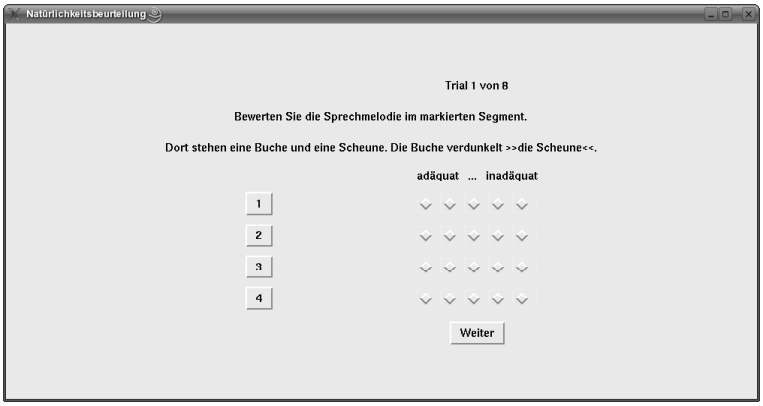

Abbildung E.3: Screenshot der Experimentoberfläche zur Beurteilung der Adäquatheit der durch den Entscheidungsbaum in Abbildung 17.6 vorhergesagten Intonationskonturen.