

Studiengang: Informatik
Prüfer: Prof. Dr. C. Rohrer
Betreuer: PD Dr. phil. B. Möbius

begonnen am: 13. Juni 2001
beendet am: 13. Dezember 2001
CR-Klassifikation: G.3, I.2.7, J.5

Diplomarbeit Nr. 1949

Modellierung der Lautdauer für die Sprachsynthese

Ursula Vollmer

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Azenbergstr. 12
D-70174 Stuttgart

Inhaltsverzeichnis

1	Einleitung	4
1.1	Motivation	4
1.2	Aufgabenstellung	4
2	Entwicklung der Lautdauerbestimmung	5
2.1	Lautdauermodell von Klatt	5
2.2	Silbenbasiertes Segmentdauermodell von Campbell	6
2.3	Segmentdauer und Sprachtiming nach van Santen	7
3	Definition der Faktoren	8
4	Sprachdatenbank	11
5	Aufbereitung der Sprachdaten	12
5.1	Einfügen der Silbentrennung	12
5.2	Trennung von Sequenzen aus Vokal und tiefem Schwa	12
5.3	Burst-Detektor	13
5.3.1	Feststellen des Plosionszeitpunktes	13
5.3.2	Algorithmus	15
5.3.3	Evaluierung	16
6	Statistische Auswertung	16
6.1	Produktsummenmodelle	16
6.2	Durstat	18
6.3	Auswertung der Daten mit Durstat	22
6.3.1	Analyse der erzeugten Modelle	22
6.3.2	Analyse der intrinsischen Lautauern	24
6.4	Bestimmung der Lautdauer	25
7	Implementierung und Integration in Festival	28
7.1	Das Sprachsynthesystem Festival	28
7.2	Implementierung	28
7.2.1	Kompromisse	30
7.2.2	Einfügen einer neuen Modellierung	30
8	Schlußfolgerung	31
8.1	Zukünftige Arbeiten	31
8.2	Verbesserungsvorschläge	31
8.3	Ausblick	31

Abbildungsverzeichnis

1	Trennung e: und 6	13
2	Trennung o: und 6	14
3	Trennung u: und 6	15
4	Trennung Y und 6	16
5	Aufspaltung eines b	17
6	Aufspaltung eines d	18
7	Aufspaltung eines g	19
8	Aufspaltung eines p	20
9	Aufspaltung eines t	21
10	Aufspaltung eines k	22
11	Kategorienbaum	24
12	Binärbaum	29

Tabellenverzeichnis

1	Ausschnitt aus den Daten für <i>Durstat</i>	23
2	Ergebniswerte der Konsonanten	25
3	Ergebniswerte der Vokale	26
4	Ergebniswerte für weitere Schwa-Aufspaltungen	26
5	Ausschnitt aus nmult.out	27

1 Einleitung

1.1 Motivation

Bei vielen Menschen existieren Vorbehalte gegenüber synthetisierter Sprache. Der häufig etwas unnatürliche Klang kann als ein Grund dafür angesehen werden. Ein möglichst natürlicher Klang ist nicht nur in psychologischer Hinsicht wünschenswert, er trägt auch zu einem besseren Verständnis des gesprochenen Textes bei. Die Lautdauer spielt bei der Modellierung eines natürlichen Klangs eine wichtige Rolle.

1.2 Aufgabenstellung

Gegenstand der Diplomarbeit soll die Erstellung eines Modells der Lautdauern des Deutschen, sein sowie die Implementierung und Integration dieses Modells in das Sprachsynthesesystem *IMS-Festival*. Hierzu sollen zunächst die linguistischen und phonetischen Faktoren definiert werden, die einen Einfluß auf die Lautdauer haben. Der zweite Schritt soll in der Auswahl oder Konstruktion eines Sprachkorpus bestehen, das diese Faktoren abdeckt. Mit Hilfe statistischer Verfahren werden die Effekte der Faktoren anschließend quantitativ bestimmt. Hierzu soll nach Möglichkeit das Produktsummenmodell von van Santen eingesetzt werden. Bei Abschluß der Arbeit soll eine neue Lautdauersteuerungskomponente für *IMS-Festival* vorliegen.

2 Entwicklung der Lautdauerbestimmung

2.1 Lautdauermodell von Klatt

Eine sehr frühes Lautdauermodell stammt aus den siebziger Jahren des zwanzigsten Jahrhunderts von Dennis H. Klatt ([Kla76]). Er beobachtete den Stellenwert der Dauer beim Unterscheiden zwischen unterschiedlich langen Vokalen, stimmhaften und stimmlosen Lauten, An- und Abwesenheit von Betonung oder auch phrasenfinalen und nicht finalen Silben in der englischen Sprache. Aufgrund durchgeführter Perzeptionstests stellt er die Abhängigkeit der Segmentdauer von Semantik, Syntax und segmentaler Zusammensetzung fest. Klatt definiert schließlich vier Ebenen, auf denen die Modifikation des Sprachtimings stattfindet. In der psychologischen oder semantischen Ebene spielen Faktoren wie die mittlere Sprachgeschwindigkeit und die Betonung eine Rolle. Die syntaktische Ebene enthält die phrasischen Akzentmuster und die Satz- und Phrasengrenzen. Auf lexikalischer Ebene kommen die segmentale Repräsentation und das lexikalische Akzentmuster dazu und auf der phonologischen Ebene werden schließlich die Dauerregeln definiert.

Neben extralinguistischen Faktoren, wie der psychologische und physikalische Zustand des Sprechers und die - davon beeinflusste - Sprechgeschwindigkeit, beeinflussen nach Klatt die folgenden Faktoren die Lautdauer: die Position in Satz und Wort, die Betonung, die semantische Neuheit eines Wortes, die Phrasenstruktur, die innere phonologische Segmentdauer, Interaktionen mit anderen Segmenten und physiologische Faktoren, wie Inkompressibilität. Inkompressibilität steht für die mögliche Existenz einer absoluten Minimaldauer für betonte Vokale, die zur Ausführung der Artikulationsbewegung benötigt wird.

Aufgrund der Regelmäßigkeit können Daueränderungen zu Regeln zusammengefaßt werden. Klatt stellt für Vokale und Konsonanten getrennte Regeln auf. Die Regeln für die Vokaldauer berücksichtigen Unterschiede der inneren phonologischen Dauer, die phrasale Position, den Einfluß postvokalischer Konsonanten und stimmhafter Frikative und die Betonung. Die Konsonantenregeln beziehen sich auf die Position in Wort und Phrase und auf die Betonung.

Da noch keine ausreichenden Daten zu syntaktischen Einflüssen vorlagen und semantische Variablen nicht aus den Klatt vorliegenden Nonsense-Silben ermittelt werden können, ist klar, daß diese Regeln nur ein erster Schritt zu einer kompletten Theorie sein konnten. Außerdem war auch Klatt klar, daß nicht alle Vokale, Konsonanten und Konsonatengruppen das gleiche Verhalten zeigen.

Untersuchungen des rhythmischen Aspektes von Sprachtiming ergaben eine hohe Empfindlichkeit der Hörer, die abnimmt, falls die zeitlichen Intervalle zwischen Onsets betonter Vokale nicht zerstört werden. Änderungen im

Timing individueller phonetischer Segmente werden nicht so stark wahr genommen, wenn der Satzrhythmus erhalten bleibt.

2.2 Silbenbasiertes Segmentdauermodell von Campbell

Die Arbeiten von Campbell ([Cam91], [Cam92]) stammen aus den neunziger Jahren des vergangenen Jahrhunderts. Um rhythmische Effekte auf höheren Ebenen besser zu berücksichtigen, wurde eine silbenbasierte Dauermodellierung entwickelt. Die Silbendauer wurde hierbei aufgrund rhythmischer und phrasaler Faktoren bestimmt und die Segmentdauer anschließend hinzugefügt.

Für die Beobachtungen auf den beiden Ebenen wurden zwei unterschiedliche Korpora verwendet, da für die Silbenebene eine prosodische Natürlichkeit Voraussetzung war, während für die Segmentebene dichte, phonetisch ausgeglichene Daten notwendig waren.

Bei der Bestimmung der Silbendauer spielten die Anzahl der Phoneme, die Art des Silbennukleus, die Silbenposition und -betonung und die Wortklasse eine Rolle. Mit diesen Faktoren wurde ein dreilagiges (Phrasen-, Silben- und Segmentebene) neuronales Netz trainiert. Nach erfolgtem Training ist dieses System sehr schnell und ein weiteres Training wird nur bei einem Sprecherwechsel oder bei der Verbesserung der Merkmalsbeschreibungen notwendig. Die Segmente wurden nach ihrer Position in der Silbe klassifiziert, wobei zusammenhängende Segmente zuerst zu Silben und Wörter zusammengefaßt werden mußten, da in den vorliegenden Daten keine Silbengrenzeninformationen vorhanden waren. Für jedes Phonem wurde der Mittelwert und die Varianz berechnet und die individuellen Segmentdauern wurden schließlich durch z-Transformation normalisiert. Campbell stellte eine Elastizitäts-Hypothese auf, welche den einzelnen Segmenten einen bestimmten Elastizitätswert zuordnet. Diese Elastizitätswerte dienen als Maß der Variation der Segmentdauern in kontinuierlicher Sprache. Er unterscheidet zwischen einer harten und einer weichen Form dieser Elastizitäts-Hypothese. Die harte Form besagt, daß Segmente einer Silbe in ihren zugehörigen Verteilungen auf die selbe Stelle fallen. Dies bedeutet, daß es für jede gegebene Silbe eine bestimmte Anzahl k an Standardabweichungen gibt, so daß die Länge eines jeden Segments in der Silbe der folgenden Summe entspricht: $\mu_{seg} + k\sigma_{seg}$. μ_{seg} steht hierbei für die mittlere Abweichung und σ_{seg} für die Standardabweichung der Dauern des jeweiligen Segmenttyps. Bei der weichen Form werden getrennte Statistiken für Silben in unterschiedlichen Satzpositionen und Segmenten in unterschiedlichen Silbenteilen bzw. phonetischen Kontexten aufgestellt.

2.3 Segmentdauer und Sprachtiming nach van Santen

Bei van Santen ([San94], [San95]) hängt die phonetische Lautdauer von kontextuellen Faktoren, wie die Identität der umgebenden Segmente, die Silbenbetonung, die Position in Wort und Phrase, der Bekanntheitsgrad des Wortes und die Phrasengrenzen, ab. Für natürlich klingende Sprache wird eine Nachahmung dieser Faktoren benötigt. Durch Fehler in anderen Komponenten könnte es jedoch auch notwendig werden, eine unnatürliche Lautdauer zu modellieren, um einen natürlichen Klang des Gesamtsystems zu erreichen. So benötigen zum Beispiel Systeme, die Silben am Satzende nicht abschwächen, an dieser Stelle eine kürzere Dauer als die natürliche Lautdauer. Dieser Fall wird hier jedoch nicht weiter betrachtet. Para-linguistische Faktoren, wie Sprachgeschwindigkeit und Sprachstil, wurden nicht berücksichtigt, da sie vermutlich durch die gesamte Datenbank konstant sind.

Van Santen machte die Beobachtung, daß die Anzahl der Faktoren mit starken Effekten in Segmentdauerdaten groß, aber auch eine sehr hohe Anzahl von seltenen - jedoch nicht zu vernachlässigenden - Merkmalsvektoren vorhanden ist. Bei der Betrachtung von Interaktionen zwischen den Faktoren wurden zwei Arten von Faktoren festgestellt: geordnete und kategoriale Faktoren. Die häufiger vorkommenden geordneten Faktoren gehen keine Umkehr- oder Vertauschungsinteraktionen ein, es werden lediglich die Effekte verstärkt. Kategoriale Faktoren unterscheiden zum Beispiel zwischen Vokalen und intervokalischen Konsonanten. Aufgrund dieser Eigenschaften von Segmentdauerdaten verwendete van Santen Produktsummenmodelle ([San93]) für die Lautdauermodellierung. Diese Modelle werden in einem eigenen Kapitel noch beschrieben.

Zur Faktorauswahl wurden die mittleren Dauern für jeden Faktorwert bestimmt und auf entscheidende Unterschiede untersucht. Bei vermischten Faktoren wurde eine systematisch ausgewählte Untermenge der Datenbank betrachtet. Für jeden Wert des kritischen Faktors wurden Mengen von Wertzusammensetzungen und, durch deren Kombination, Merkmalsvektoren bestimmt.

Van Santen stellte die Daten in einer Baumstruktur dar, um sowohl der Homogenität der Kategorien als auch der ausreichenden Anzahl an Beobachtungen Rechnung zu tragen. Er unterscheidet hierbei zwischen Vokalen, intervokalischen Konsonanten und Konsonantengruppen. Für die Bestimmung der Vokaldauer wurden die Vokalidentität, die Silbenbetonung, die umgebenden Konsonanten und die Position in Wort und Satz herangezogen. Die Dauer intervokalischer Konsonanten wurde aufgrund der Betonungswerte der umgebenden Vokale, der Position in Wort und Satz und dem Akzentstatus des Wortes bestimmt. Für die Analyse von Konsonantengruppen mußten dazu noch der segmentale Kontext und die Silben- und Wortgrenzen betrachtet werden.

3 Definition der Faktoren

Für die vorliegende Arbeit wurde eine Lautdauerbestimmung auf segmentaler Ebene gewählt. Die Gründe, welche für eine Lautdauerbestimmung auf höherer Ebene sprechen, liegen in der Bedeutung der Satz-, Phrasen- und Wortgrenzen für die Lautdauer. Diese Grenzen müßten daher mit einer höheren Genauigkeit vorhergesagt werden. Es wird jedoch bei Lautdauerbestimmung auf höherer Ebene keine höhere Genauigkeit erreicht. Eine Bestimmung der Silbendauer mit einer anschließenden Aufteilung der Dauer auf die Segmente der Silbe ist nicht ausreichend.

Der Einfluß der Segmente und ihrer segmentalen Umgebung auf die Lautdauer ist ausreichend groß, daß er nicht vernachlässigt werden darf. Ein möglicher Einfluß subsegmentaler Einheiten wird aufgrund fehlender Sprachdaten für eine Analyse auf dieser Ebene nicht berücksichtigt.

Da eine phrasenfinale Längung der Vokale und eine phraseninitiale Längung von Konsonanten beobachtet werden können, sollten Faktoren wie zum Beispiel die Grenze nach links bzw. rechts und die Phrasenposition in die Dauerberechnung einbezogen werden. Auch die Silbenbetonung spielt eine wichtige Rolle in der deutschen Sprache. Die Lautdauer wird außerdem von der unmittelbaren Umgebung des Segments beeinflusst. Für die Lautdauerbestimmung auf segmentaler Ebene wurden daher die folgenden Faktoren definiert:

1. Segmentidentität
2. Segmenttyp
3. Satzposition
4. Phrasenposition
5. Wortposition
6. Silbenposition
7. Phrasenlänge
8. Wortlänge
9. Silbenlänge
10. Identität des linken Segments
11. Identität des rechten Segments
12. Segmenttyp des linken Segments
13. Segmenttyp des rechten Segments
14. Grenze nach links

15. Grenze nach rechts

16. Wortklasse

17. Silbenbetonung

Die Werte des ersten Faktors sind die Phone der deutschen Sprache. Allerdings werden die Plosive jeweils in ihre Verschluss- und Öffnungsphase aufgespalten, um eine genauere Modellierung zu gewährleisten.

Faktor 2 hat die folgenden Segmenttypen als Werte:

- langer Vokal
- kurzer Vokal
- Diphthong
- Schwa
- stimmhafter Plosiv
- stimmloser Plosiv
- stimmhafter Frikativ
- stimmloser Frikativ
- Nasal
- Liquid
- Glide

Die Aufteilung der Segmente in Vokale und Konsonanten ist gebräuchlich und die unterschiedlichen Eigenschaften der beiden Gruppen sind wohl eindeutig.

Die Vokale werden anhand ihrer phonologischen Länge unterschieden, was bei der Lautdauerbestimmung ein offensichtlich sinnvoller Faktor ist. Außerdem bilden Diphthonge und Schwa aufgrund ihrer besonderen Eigenschaften jeweils eine getrennte Gruppe.

Die Konsonanten werden nach ihrer Artikulationsart in Plosive, Frikative, Nasale, Liquide und Glides aufgeteilt. Bei den Plosiven und den Frikativen wird außerdem noch zwischen stimmhaften und stimmlosen Lauten unterschieden.

Die Satzposition (Faktor 3) wird mit den Werten initial, medial oder final beschrieben. Sie drücken die Stellung der Phrase im Satz aus. Ein vierter Wert steht für den Fall, daß der Satz genau eine Phrase enthält.

Die Phrasenposition (Faktor 4) beschreibt die Stellung des Worts in der Phrase mit den Werten initial, medial und final.

Für die Position der Silbe im Wort (Faktor 5) werden die vier Werte initial, medial, final und einsilbig verwendet.

Die Silbenposition (Faktor 6) beschreibt die Stellung des Segments in der Silbe. Sie kann die Werte Onset, Nukleus, Coda und ambisyllabisch annehmen.

Die Faktoren sieben bis neun besitzen als Werte die Anzahl der in der Phrase enthaltenen Wörter (Faktor 7), die Anzahl der im Wort enthaltenen Silben (Faktor 8) bzw. die Anzahl der in der Silbe enthaltenen Segmente (Faktor 9).

Die Faktoren 10 und 11 enthalten die selben Werte wie der erste Faktor. Sie beschreiben die Segmentidentität des linken bzw. rechten Segments.

Entsprechend enthalten die Faktoren 12 und 13 die selben Werte wie der zweite Faktor. Sie stehen für den Segmenttyp des linken bzw. rechten Segments.

Die Faktoren 14 und 15 enthalten fünf Werte. Diese zeigen an, ob links bzw. rechts vom Segment eine Grenze vorhanden ist. Existiert eine derartige Grenze, so wird noch unterschieden, ob es sich um eine Satz-, Phrasen-, Wort- oder Silbengrenze handelt.

Der Faktor 16 hat zwei Werte, er unterscheidet zwischen Funktions- und Inhaltswörtern.

Der siebzehnte Faktor schließlich gibt an, ob die Silbe primär, sekundär oder überhaupt nicht betont ist.

4 Sprachdatenbank

Das von mir verwendete Kiel-Korpus entstand im Rahmen der PHONDAT-Projekte 1989-1992. Es hat den Vorteil, daß mit zwei Sprechern alle im Korpus enthaltenen Texte aufgenommen wurden. Mit den anderen Sprechern wurden nur von Teilen des verwendeten Textmaterials Aufnahmen gemacht. Aus den beiden Sprechern, deren Aufnahmen das gesamte Textmaterial umfassen, wurde ein männlicher Sprecher ausgewählt, da für die deutschsprachige Synthese des Sprachsynthesystems *Festival* auch eine männliche Stimme verwendet wird. Dadurch ist eine ausreichend große Datenmenge vorhanden, an der Untersuchungen zur Lautdauer vorgenommen werden können ohne daß unterschiedliche sprecherspezifische Eigenheiten verschiedener Sprecher die Ergebnisse beeinflussen.

Die Daten des Korpus beinhalten die Signal- und die Segmentationsdateien der PHONDAT-Projekte von 1990 und 1992. Diese bilden die Grundlage für ein Variantenlexikon gelesener Sprache ([Koh94]). Genauere Einzelheiten bezüglich der Aufarbeitung des in Kiel gesammelten Materials und der Segmentationskonventionen können in ([Koh92]) nachgelesen werden. Die von mir verwendeten Segmentationsdateien enthalten den Dateinamen, eine orthographische Repräsentation der Sätze (durch *oend* abgeschlossen), eine kanonische Transkription der Sätze in SAMPA (durch *kend* abgeschlossen), eine aus den Labels des Segmentierers konstruierte Transkription der Sätze (durch *hend* abgeschlossen), die Sample-Nummern und die Label. Satz-, Phrasen- und Wortgrenzen, Betonung und Wortklasse sind in den Daten annotiert und konnten daraus entnommen werden. Der Beginn eines Satzes wird dabei durch die Zeichenfolge *#c:* markiert, die Satz- und Phrasengrenzen durch *#* gefolgt von dem entsprechenden Satzzeichen. Der Beginn eines Wortes wird durch Voranstellen von *##* vor das erste Segment des Wortes gekennzeichnet. Die Silbenbetonung ist am ersten Segment der Silbe annotiert. Dabei steht *'* für die primäre und *''* für die sekundäre Betonung. Funktionswörter sind durch ein *+* am ersten Segment des Wortes gekennzeichnet.

Glottale Stops, in den Dateien durch *Q* bzw. *q* dargestellt, sind mit einer Lautdauer von Null eingetragen, da die Übergänge zum nachfolgenden Vokal fließend verlaufen. Sie werden daher zur Lautdauermodellierung nicht herangezogen. Für die Synthese muß daher für die glottalen Stops ein konstanter Wert festgelegt werden.

5 Aufbereitung der Sprachdaten

Da die vorliegenden Daten nicht vollständig den Bedürfnissen entsprachen, mußten noch einige zusätzliche Änderungen vorgenommen werden.

So mußten die im Korpus fehlenden Silbengrenzen noch eingefügt werden. Ein weiteres Problem war, daß von einem tiefen Schwa gefolgte Vokale als ein Segment gelabelt vorlagen. Außerdem mußten, wie bereits erwähnt, die Plosive in Verschluß- und Öffnungsphase aufgespaltet werden, wozu ein *Burst-Detektor* geschrieben wurde.

In den folgenden Abschnitten wird erläutert wie die notwendigen Änderungen durchgeführt wurden.

5.1 Einfügen der Silbentrennung

Da auch in dem dem Korpus beigefügten Lexikon keine Informationen zur Silbentrennung vorhanden sind, mußte ein anderes Lexikon gefunden oder konstruiert werden, welches die Wörter des Korpus - einschließlich der Silbengrenzen - enthält. Dazu wurde das vorhandene Lexikon mit Hilfe der Syllabifizierungsfunktion von *Festival* mit Silbengrenzen versehen. Dabei stellte sich jedoch heraus, daß ein ziemlich großer Teil der Wörter nicht korrekt syllabifiziert wurde. Der Grund hierfür liegt unter anderem in der Schwierigkeit, ambisyllabische Konsonanten einer bestimmten Silbe zuzuordnen. So wurde zum Beispiel das Wort *Anweisungen* phonetisch in der folgenden Weise getrennt: An-wei-su-ngen. Da dieses, mit *Festival* erzeugte, Lexikon somit auch nicht benutzbar war, wurde mir von Bernd Möbius ein Lexikon mit manuell eingefügten Silbengrenzen zur Verfügung gestellt, welches die Wörter des Kiel-Korpus enthält.

Die genaue Lautdauer der einzelnen Segmente konnte durch ein Skript gewonnen werden, welches die Sample-Nummern durch den Zeitpunkt, zu dem sie gesprochen wurden, ersetzt und außerdem die orthographische Repräsentation und die Transkriptionen der Sätze entfernt. Dieses Skript wurde mir von Wolfgang Wokurek zur Verfügung gestellt.

5.2 Trennung von Sequenzen aus Vokal und tiefem Schwa

Sequenzen, in denen einem tiefen Schwa ein Vokal folgt wurden im Korpus nicht getrennt, sondern als ein Segment gelabelt. Dies ist eine Folge dessen, daß das vorliegende Korpus automatisch gelabelt wurde. Der Übergang von einem Vokal zu einem tiefen Schwa verläuft jedoch fließend, weshalb die Trennstelle für einen automatischen Labeller nicht gut festzustellen ist. Die fehlenden Label mußten daher noch manuell eingefügt werden. Die Abbildungen 1 bis 4 zeigen einige Beispiele für die Übergänge zwischen Vokalen und tiefem Schwa und die vorgenommenen Trennungen zwischen den beiden

Phonen.

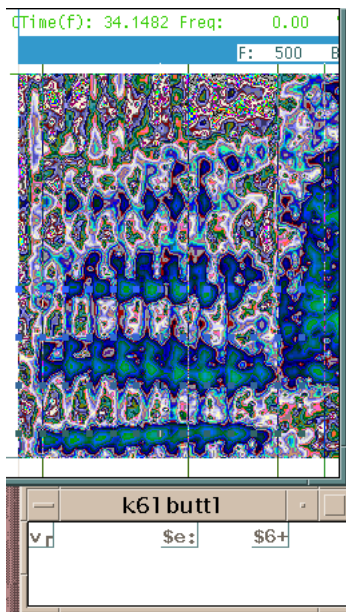


Abbildung 1: Beispiel zur Trennung einer Sequenz aus e: und 6

5.3 Burst-Detektor

5.3.1 Feststellen des Plosionszeitpunktes

Da aufgrund der hohen Anzahl von Plosiven im Korpus eine manuelle Aufspaltung in Verschuß- und Öffnungsphase zu aufwendig wäre, mußte ein *Burst-Detektor* geschrieben werden, welcher die Stelle der Plosion bestimmt. Die Stelle, an der zwischen Verschuß- und Öffnungsphase getrennt wird, liegt kurz vor dieser Plosionsstelle. Der Plosionszeitpunkt kann mit Hilfe der rms-Werten des f0-Verlaufs festgestellt werden. In diesen rms-Werten findet sich an der Plosionsstelle ein - je nach Art des Plosivs - mehr oder weniger starker Sprung.

Der folgende Ausschnitt aus einer rms-Datei zeigt die rms-Werte für einen stimmhaften Plosiv (*d*):

```
15.8869
11.8398
10.5817
10.9967
9.72451
240.054
```

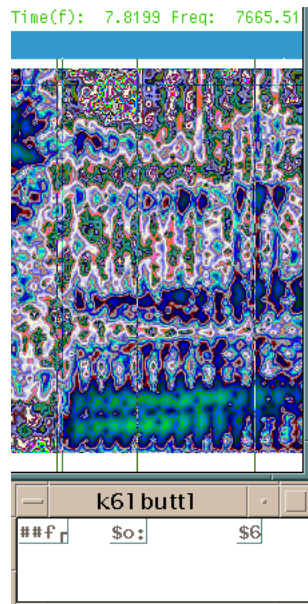


Abbildung 2: Beispiel zur Trennung einer Sequenz aus o: und 6

319.218

Betrachtet man die jeweiligen Differenzen zwischen zwei nacheinander stehenden rms-Werten, so bemerkt man einen Sprung um etwa 230 von 9.72451 auf 240.054. Alle anderen Differenzen sind erheblich kleiner. An dieser Stelle kann also der Plosionszeitpunkt angesetzt werden. Kurz vor dieser Stelle wird die Trennstelle zwischen Verschluß- und Öffnungsphase festgelegt. Schaut man sich zum Vergleich die rms-Werte eines stimmlosen Plosivs (t) an, so kann man einen erheblich höheren Sprung feststellen:

189.366
 85.6177
 45.1177
 33.0894
 27.6678
 14.9604
 13.9143
 14.4832
 16.3832
 530.688
 780.681

Der Sprung von 16.3832 auf 530.688 beträgt in etwa 514.

Beim Feststellen der Plosionsstelle muß also zwischen stimmhaften und stimmlosen Plosiven unterschieden werden.

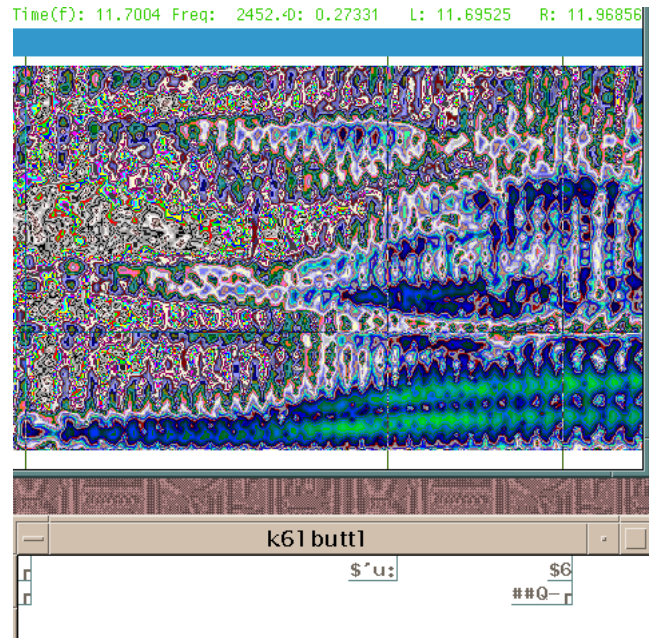


Abbildung 3: Beispiel zur Trennung einer Sequenz aus u: und 6

5.3.2 Algorithmus

Für die Bestimmung der Trennungsstelle zwischen Verschuß- und Öffnungsphase mußte zuerst der jeweilige Plosiv mit seiner Anfangs- und Endzeit aus den Dateien des Korpus extrahiert werden. Aus der Anfangs- und Endzeit des Plosivs und der bekannten Schrittweite (0.005 s), mit der die rms-Werte vorlagen, konnten Anfang und Ende des Plosivs in der entsprechenden rms-Datei bestimmt werden. Die zwischen diesen beiden Stellen liegenden rms-Werte wurden nun auf einen großen Sprung untersucht. Dabei wurde für die stimmhaften Plosive ein kleinerer Schwellwert festgelegt, da bei diesen der Sprung geringer ausfällt als bei den stimmlosen Plosiven (siehe dazu den letzten Abschnitt). Da es durchaus vorkommt, daß mehrere Sprünge in diesem Bereich der rms-Werte enthalten sind, muß in diesem Fall der letzte Sprung betrachtet werden, um die eigentliche Plosionsstelle zu erhalten. Ist der gesuchte Sprung gefunden, muß kurz vor dieser Stelle die Trennung zwischen Verschuß- und Öffnungsphase angesetzt werden. Der Trennungszeitpunkt wird durch Multiplikation der Stelle in der rms-Datei mit der Schrittweite berechnet. Wird kein Sprung gefunden, so wird dieser Zeitpunkt auf das Ende des gesamten Plosivs gesetzt.

Die Abbildungen 5 bis 10 zeigen Beispiele für die bestimmten Trennungszeitpunkte. In der rms-Kurve (dritte von oben) sieht man den Sprung, der an der Stelle der Plosion auftritt. Die Notwendigkeit für unterschiedliche

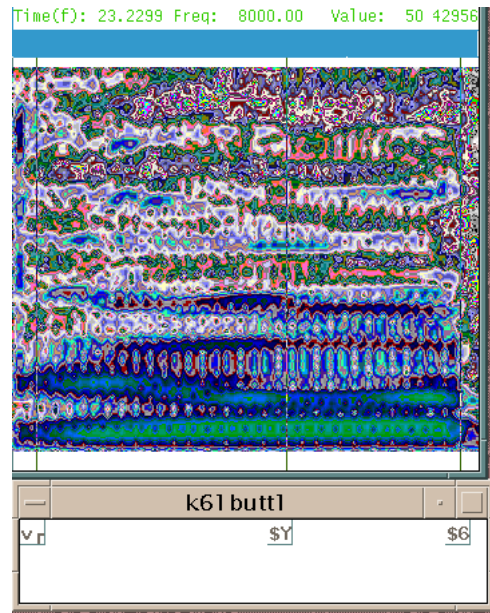


Abbildung 4: Beispiel zur Trennung einer Sequenz aus Y und 6

Schrittweiten bei stimmhaften und stimmlosen Plosiven kann in diesen Abbildungen nicht festgestellt werden, da die rms-Werte und damit auch der Unterschied zwischen den Werten nicht direkt abgelesen werden können.

5.3.3 Evaluierung

Zur Evaluierung der Funktionsweise des *Burst-Detektors* wurden stichprobenartig die zu Plosiven gehörenden Werte in den rms-Dateien betrachtet und auf die Korrektheit des Aufspaltungszeitpunktes untersucht.

Für diese untersuchten Daten wurde der Trennungszeitpunkt in etwa 90% der Fälle korrekt bestimmt. Bei den restlichen 10% wurde der Zeitpunkt etwas zu früh gewählt.

In etwa 14% der untersuchten Daten konnte keine Aufspaltung stattfinden, da kein erkennbarer Sprung in den rms-Werten vorhanden war. In diesen Fällen wurde der Öffnungsphase eine Dauer von Null zugeordnet.

6 Statistische Auswertung

6.1 Produktsummenmodelle

Produktsummenmodelle ([San93], [San94]) stellen eine Abänderung der Varianzmodell-Analyse dar. Die Interaktionsterme bestehen aus Produkten von Einzelfaktorskalen. Aufgrund der geordneten Struktur, die typisch für Pro-

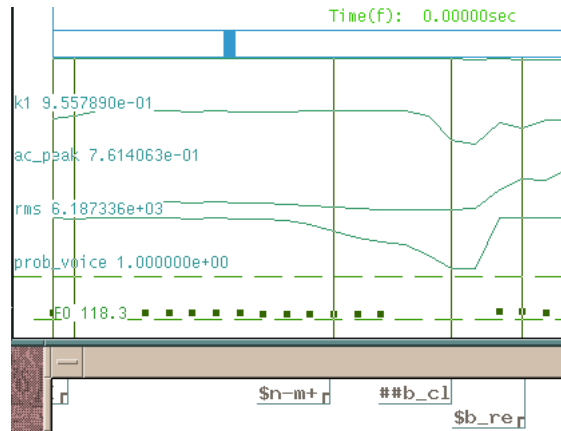


Abbildung 5: Beispiel zur Aufspaltung von Verschluß- und Öffnungsphase eines b

duktsammenmodelle ist, sind diese Modelle geeignet für Daten, in denen die Faktoren auf einer geordneten Skala und das abhängige Maß zumindest auf einer Intervallskala liegen. Desweiteren muß in diesen Daten eine Unabhängigkeit der einzelnen Faktoren gegeben sein und eventuell vorkommende Verletzungen der Verbindungsunabhängigkeit müssen verstärkend wirken. Dies bedeutet, daß die Auswirkungen eines beeinflussten Faktors in keinem Falle verringert, sondern höchstens verstärkt werden dürfen.

Aufgrund der großen Menge an Modellen, kann für ein gegebenes Interaktionsmuster zumindest ein Produktsammenmodell gefunden werden. Die Auswahl dieses Modells stellt jedoch eine kritische Phase bei der Konstruktion eines Vorhersagesystems dar. Für das beste Modell und die statistische Aussage über dessen Qualität müssen die benötigten Indexmengen gefunden und die Faktorskalen bestimmt werden.

In der Klasse der Produktsammenmodelle sind das additive Modell, das multiplikative Modell und das Modell von Klatt enthalten.

Die Gründe für die Verwendung von Produktsammenmodellen für die Lautdauermodellierung liegen in Sprachaspekten, wie Segmentdauer, Phrasengrenzen und Akzenten. Die Herausforderungen in Bezug auf die Segmentdauer liegen in der unregelmäßigen Frequenzverteilung im Merkmalsraum und in den Interaktionen, die zwischen den Faktoren existieren. Die unregelmäßige Frequenzverteilung bezieht sich dabei auf zwei Aspekte. Zum einen umfaßt der linguistische Raum nur einen Teil des Merkmalsraums, nämlich die Untermenge, welche die in einer Sprache tatsächlich vorkommenden Merkmalsvektoren enthält. Zum anderen decken die Trainingsproben lediglich eine kleine Untermenge des linguistischen Raums ab - und dies sehr ungleichmäßig. Allerdings enthalten auch Textproben von wenigen Sätzen mit Sicherheit sehr seltene Merkmalsvektoren. Diese seltenen Vek-

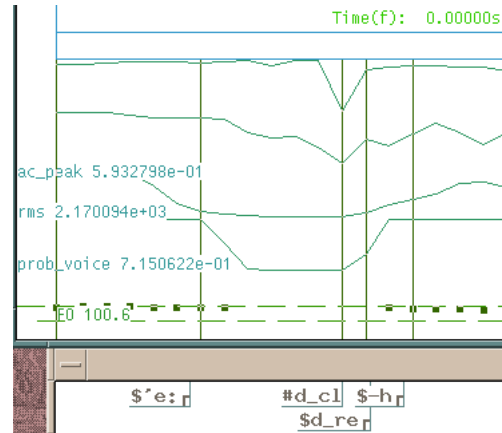


Abbildung 6: Beispiel zur Aufspaltung von Verschuß- und Öffnungsphase eines d

toren müssen also auf jeden Fall berücksichtigt werden - trotz der Schwierigkeit, Trainingsdaten zu finden, die all diese Vektoren abdecken. Zu den Interaktionen zwischen den Faktoren ist anzumerken, daß sie erfreulicherweise meist regelmäßig sind, das heißt die Effekte eines Faktors werden in der Regel nicht durch Effekte anderer Faktoren umgekehrt.

6.2 Durstat

Die statistische Auswertung wurde mit Hilfe des Statistikprogramms *Durstat* von Jan van Santen durchgeführt. Dieses Programm wurde dem IMS von Bell Labs zu Forschungszwecken zur Verfügung gestellt. Es stellt eine Shell-basierte Benutzerschnittstelle zur SoP-Modellierung und zur Benutzung statistischer Funktionen dar.

Zur Benutzung von *Durstat* werden einige Dateien benötigt. Zum einen muß eine Datei mit den Daten vorhanden sein. Diese Datei besteht aus einer Matrix, welche in jeder Spalte den Wert eines bestimmten Faktors enthält und in den beiden letzten Spalten außerdem noch die Segmentdauer und den Namen der Datei, aus welcher die Daten stammen (siehe Tabelle 1). Zum Erzeugen dieser Daten wurde ein Skript geschrieben, welches die benötigten Informationen aus den Dateien des Korpus extrahiert. In diesem Skript werden die Typen der Segmente bestimmt (siehe dazu auch Kapitel 3). Es wird die Position des Segments in der Silbe festgestellt, desweiteren die Position dieser Silbe im Wort, die Position des Worts in der Phrase und die Position der Phrase im Satz. Die Anzahl der Segmente in der Silbe, die Anzahl der Silben im Wort und die Anzahl der Wörter in der Phrase werden auch berechnet. Aus den Annotationen im Korpus werden die Grenzen, die

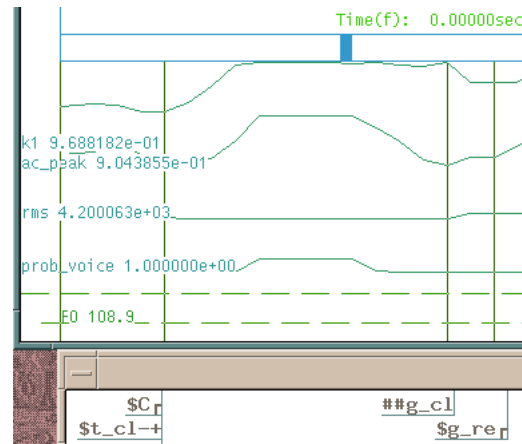


Abbildung 7: Beispiel zur Aufspaltung von Verschluß- und Öffnungsphase eines g

Wortklasse und die Betonung bestimmt. Schliesslich werden noch die Segmente und ihre Typen nach links und rechts festgestellt. Aus den im Korpus angegebenen Anfangszeiten der Segmente kann die Segmentdauer berechnet werden.

Außerdem wird die jeweils maximale Anzahl der Segmente in der Silbe, der Silben im Wort und der Wörter in der Phrase bestimmt. Mit diesen Daten wurden dann die Dateien fm.7, fm.8 und fm.9 (siehe nächster Abschnitt) erzeugt.

Desweiteren wird eine Datei mit dem Namen *label* benötigt, die eine Auflistung der Faktoren enthält. In der Datei *phondef* müssen die einzelnen Segmente ihren Typen (siehe Kapitel 3) zugeordnet werden. Schließlich wird für jeden Faktor noch eine Datei fm.X benötigt, wobei X der Spaltennummer in *label* bzw. in der Datendatei entspricht, die den Faktor enthält. In diesen *feature map*-Dateien werden die Level des jeweiligen Faktors aufgelistet.

Folgende Programme zur statistischen Auswertung stehen in *Durstat* zur Verfügung:

1. *install*: Mit diesem Programm werden die Dateien für eine neue Analyse vorbereitet. Es wird die Datei *data* erzeugt, welche die Daten in der für *Durstat* benötigten Darstellung enthält.
2. *select*: Hiermit wird eine Untermenge der Daten erzeugt. Dazu müssen die Faktorlevel bestimmt werden, welche in der Untermenge enthalten sein sollen. Die Untermenge wird in einem neuen Unterverzeichnis gespeichert.
3. *combine*: Dieses Programm kombiniert ausgewählte Verzeichnisse zu einem.

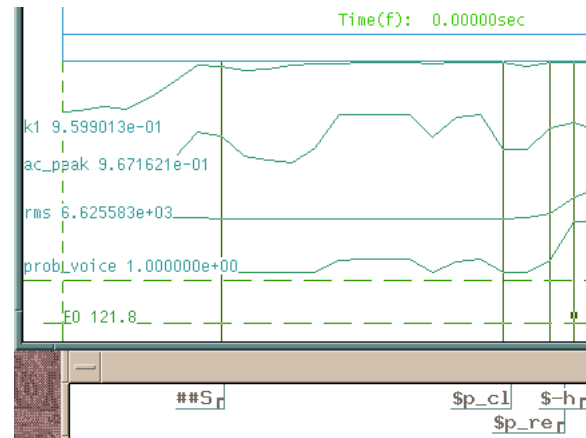


Abbildung 8: Beispiel zur Aufspaltung von Verschuß- und Öffnungsphase eines p

4. *compound*: Mit *compound* werden ausgewählte Faktoren kombiniert.
5. *equivalence*: Dieses Programm kann zur Feststellung äquivalenter Faktorpaare herangezogen werden.
6. *dependencies*: Zur Bestimmung weiterer Abhängigkeiten kann dieses Programm verwendet werden.
7. *raw_marginals*: Hiermit werden die Mittelwerte eines Faktors bestimmt.
8. *raw_2W_marginals*: Dieses Programm dient zur Bestimmung der Mittelwerte zweier, miteinander kombinierter, Faktoren.
9. *corrected_marginals*: Mit diesem Programm werden die korrigierten Mittelwerte eines Faktors bestimmt.
10. *corrected_2W_marginals*: Hiermit werden die korrigierten Mittelwerte zweier, miteinander kombinierter, Faktoren bestimmt.
11. *map*: Aufgrund der *fm.X*-Dateien wird die Datei *ndata* erzeugt.
12. *save_maps*: Die *fm.X*-Dateien werden gespeichert bzw. wiederhergestellt.
13. *addmodel*: Hiermit werden die Parameter des additiven Modells bestimmt.
14. *multmodel*: Dieses Programm erzeugt die Parameter des multiplikativen Modells.

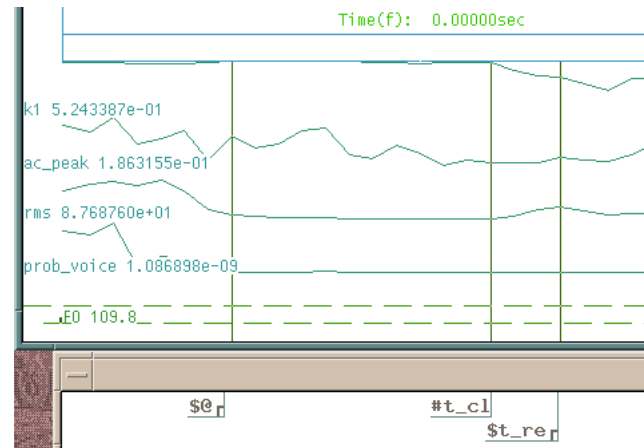


Abbildung 9: Beispiel zur Aufspaltung von Verschluß- und Öffnungsphase eines t

15. `multmodelimp`: Die Bedeutung der Faktoren des multiplikativen Modells wird mit diesem Programm bestimmt.
16. `residuals`: Die Reste des additiven bzw. multiplikativen Modells werden mit *residuals* bestimmt.
17. `var_expl`: Zur Erläuterung der Varianzen von Faktoren und Modellen kann dieses Programm herangezogen werden.
18. `extremes`: Hiermit können Extremwerte in der ausgewählten Datenuntermenge festgestellt werden.
19. `mk_H`: `mk_H` erzeugt X.h-Dateien für ein TTS-Dauermodul.
20. `clean`: `clean` löscht all Dateien, außer `data`, `label`, `phondef` und den `fm.X`-Dateien.
21. `sop`: Zur Bestimmung der Parameter für ein Produktsammenmodell kann dieses Programm verwendet werden.
22. `InterpretPath`: Dieses Programm interpretiert den Pfadnamen und zeigt somit die Zusammensetzung der Datenuntermenge an.
23. `Xout`: Das Gleichsetzen aller Level in einer `fm.X`-Datei kann mit diesem Programm erfolgen.
24. `aov`: Eine N-Wege-Analyse der Varianz kann mit diesem Programm durchgeführt werden.
25. `quit`: Hiermit wird *Durstat* beendet.

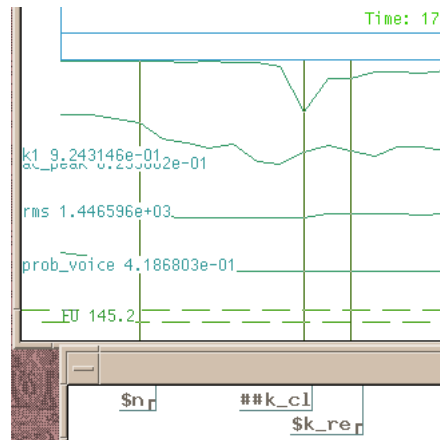


Abbildung 10: Beispiel zur Aufspaltung von Verschuß- und Öffnungsphase eines k

6.3 Auswertung der Daten mit Durstat

Eine Auswertung der aus dem Kiel-Korpus erhaltenen Daten mit *Durstat* führte zu dem in Abbildung 11 dargestellten Baum. Die Gesamtdaten wurden in Konsonanten und Vokale aufgespaltet.

Die Konsonanten wurden nach ihrer Position in der Silbe in ambisyllabische Konsonanten, im Onset stehende Konsonanten und in der Coda stehende Konsonanten aufgeteilt. In jeder dieser Gruppen wurden Sonoranten und Obstruenten getrennt, bei den Obstruenten wurde noch zwischen Frikativen und Plosiven unterschieden. Die ambisyllabischen Frikative, die Frikative im Onset und die Plosive in allen drei Gruppen wurden noch in stimmhaft und stimmlos aufgespaltet. Schließlich wurde sowohl in den Gruppen der stimmhaften als auch der stimmlosen Plosive noch zwischen Verschuß- und Öffnungsphase unterschieden.

Die Vokale wurden in die Segmenttypen Diphthong, Schwa, lange Vokale und kurze Vokale gespaltet. Die Gruppen der langen und der kurzen Vokale wurden nach der Betonung aufgeteilt in primär betonte, sekundär betonte und unbetonte Vokale. Für die Gruppe der Diphthonge lagen nicht genügend Daten vor, um eine weitere Aufteilung zu ermöglichen. Die Gruppe der Schwa wurde in zwei Gruppen geteilt, die jeweils eine Art Schwa (@ bzw. 6) enthalten.

6.3.1 Analyse der erzeugten Modelle

Abgesehen von den Werten der Gruppe der Schwa bestätigt die Analyse der Korrelationen und der rms-Werte in den Blättern des erzeugten Baumes (Ta-

h	u_fric	one	ini	ini	onset	4	2	2
none	OY	none	diph	utt	none	cont	un	
0.098625		k61be001_clean.lab						
OY	diph	one	ini	ini	nucleus	4	2	2
h	t_cl	u_fric	u_stop	none	syll	cont	prim	
0.141437		k61be001_clean.lab						
t_cl	u_stop	one	ini	fin	ambi	4	2	4
OY	t_re	diph	u_stop	syll	none	cont	un	
0.041938		k61be001_clean.lab						
t_re	u_stop	one	ini	fin	ambi	4	2	4
t_cl	schwa	u_stop	schwa	none	none	cont	un	
0.031000		k61be001_clean.lab						
schwa	schwa	one	ini	fin	nucleus	4	2	4
t_re	I	u_stop	short	none	word	func	un	
0.071813		k61be001_clean.lab						
			:					

Tabelle 1: Ausschnitt aus der für *Durstat* benötigten Datei, welche die Merkmalsvektoren enthält. Die Spalten enthalten die jeweiligen Werte der 17 Faktoren (Segmentidentität, Segmenttyp, Satz-, Phrasen-, Wort- und Silbenposition, Phrasen-, Wort- und Silbenlänge, Segmentidentität des linken/rechten Segments, linke/rechte Grenze, Wortklasse, Silbenbetonung). Spalte 18 enthält die Lautdauer und Spalte 19 den Namen der Datei, aus der die Daten stammen.

bellen 2 und 3) den konstruierten Baum. Diese Werte und die große Anzahl der vorhandenen Schwa würden eine weitere Aufspaltung der beiden Schwa-Gruppen nahelegen. Versuchsweise durchgeführte weitere Aufteilungen, die eventuell noch sinnvoll gewesen wären, führten jedoch auch zu keinen besseren Werten (siehe Tabelle 4), weshalb sie wieder verworfen wurden. Die Gesamtkorrelation zwischen beobachteten und vorhergesagten Segmentdauern der gesamten Daten beträgt 0.876. Dies bedeutet, daß beinahe 80% ($r^2 = 76.73$) der Varianz in den Dauerdaten vom Dauermodell erklärt werden. Es wurde gezeigt, daß mindestens 8% der Gesamtvarianz von Sprachtiming nicht aus dem Text vorhergesagt werden können. Die Gründe für die restliche Varianz liegen in lokalen Änderungen der Sprechgeschwindigkeit, einer zufälligen Schwankung in der Sprachproduktion oder unentdeckten systematischen Faktoren. Bei einer Erklärung von etwa 80% der Varianz kann jedoch davon ausgegangen werden, daß kein entscheidender Faktor, der sich auf die Segmentdauer auswirken könnte, bei der Modellkonstruktion übersehen wurde.

Betrachtet man die Korrelationen, die für andere Sprachen ([Moe00]) erhalten wurden, so bewegt sich der Wert für die vom Dauermodell erklärte

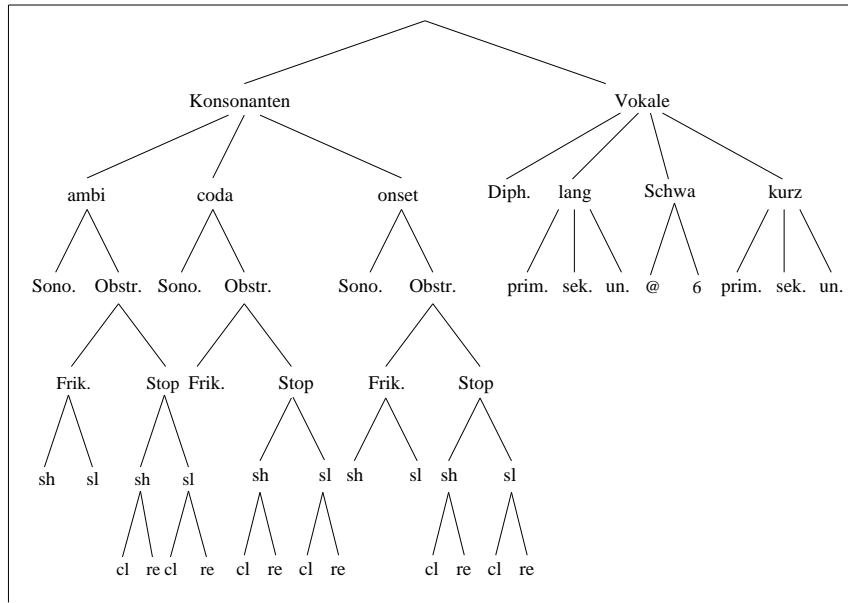


Abbildung 11: Mit Hilfe von *Durstat* erzeugter Kategorienbaum des deutschen Lautdauersystems. Zur Darstellung wurden die im Folgenden aufgeführten Abkürzungen verwendet: ambi = ambisyllabisch, Sono. = Sonorant, Obstr. = Obstruent, Frik. = Frikativ, sh = stimmhaft, sl = stimmlos, cl = Verschlußphase, re = Öffnungsphase, Diph. = Diphthong, prim. = primäre Betonung, sek. = sekundäre Betonung, un. = unbetont

Varianz zwischen 57.76% (Niederländisch mit einer Korrelation von 0.760) und 76.03% (Mandarin mit einer Korrelation von 0.872). Französisch liegt mit einer Korrelation von 0.847 dazwischen. Im Japanischen erhält man für die Vokale einen Korrelationswert von 0.880 und für die Konsonanten einen Wert von 0.940.

6.3.2 Analyse der intrinsischen Lautauern

Vergleicht man die intrinsischen Lautauern der Sonoranten in den drei unterschiedlichen Silbenstellungen, so stellt man gewisse Ähnlichkeiten in der Skalierung fest. *N* hat eine sehr hohe Dauer, wohingegen *r* eine sehr geringe Dauer hat.

Bei den Frikativen hat *h* immer die kürzeste Dauer. *S*, *s* und *f* zeigen die längsten intrinsischen Lautauern, während *C* immer in der Mitte liegt.

Auch bei den Verschlußphasen der Plosive kann eine regelmäßige Skalierung festgestellt werden. Dabei gilt sowohl für die stimmhaften, als auch für die stimmlosen Plosive, daß die Verschlußphasen der bilabialen Plosive am längsten, die der alveolaren am kürzesten.

S.pos.	Typ	Stimmh.	Phase	Korrelation	rms-Wert
ambi	Sonorant			0.698	0.011
	Frikativ	stimmlos		0.813	0.016
		stimmhaft		0.670	0.009
	Plosiv	stimmlos	Verschuß	0.564	0.016
	Plosiv	stimmlos	Öffnung	0.917	0.010
	Plosiv	stimmhaft	Verschuß	0.624	0.014
	Plosiv	stimmhaft	Öffnung	0.558	0.006
coda	Sonorant			0.569	0.026
	Frikativ			0.831	0.027
	Plosiv	stimmlos	Verschuß	0.613	0.019
	Plosiv	stimmlos	Öffnung	0.832	0.018
	Plosiv	stimmhaft	Verschuß	0.761	0.012
	Plosiv	stimmhaft	Öffnung	0.561	0.007
onset	Sonorant			0.694	0.020
	Frikativ	stimmlos		0.689	0.021
	Frikativ	stimmhaft		0.729	0.016
	Plosiv	stimmlos	Verschuß	0.671	0.017
	Plosiv	stimmlos	Öffnung	0.863	0.012
	Plosiv	stimmhaft	Verschuß	0.645	0.018
	Plosiv	stimmhaft	Öffnung	0.623	0.009

Tabelle 2: Diese Tabelle listet die Korrelationen und die rms-Werte für die Konsonanten des erstellten Baumes auf (S.pos. = Position des Segments in der Silbe, Stimmh. = Stimmhaftigkeit).

Bei einer Analyse der intrinsischen Lautauern der Vokale in den unterschiedlichen Gruppen, kann eine ähnliche Skalierung der Vokaldauern in jeder der Gruppen beobachtet werden.

Die tiefen Vokale haben dabei immer eine längere intrinsische Dauer als die hohen.

6.4 Bestimmung der Lautdauer

Für jedes Blatt des Kategorienbaumes wurde von *Durstat* ein Modell berechnet. Die dabei jeweils erzeugte *nmult.out*-Datei enthält die für die Berechnung benötigten Daten (vgl. Tabelle 5).

Die Formel zur Berechnung der vorhergesagten Lautdauer $Dur(i(f_2, \dots, f_n))$ eines gegebenen Segments i sieht folgendermaßen aus:

$$Dur(i(f_2, \dots, f_n)) = e^{F_1} * e^{F_2 f_2} * e^{F_3 f_3} * \dots * e^{F_n f_n}$$

Dabei stehen f_2, \dots, f_n für die Faktorwerte der Faktoren F_2, \dots, F_n . Mit F_1 wird die intrinsische Dauer des Segments in die Berechnung einbezogen. $F_2 f_2, \dots, F_n f_n$ sind die Koeffizienten der anderen Faktorwerte.

Typ	Betonung	Korrelation	rms-Wert
Diphthong		0.793	0.028
lang	primär	0.760	0.030
	sekundär	0.579	0.030
	unbetont	0.728	0.026
6 @		0.728	0.022
		0.523	0.033
kurz	primär	0.698	0.019
	sekundär	0.696	0.016
	unbetont	0.702	0.016

Tabelle 3: Diese Tabelle listet die Korrelationen und die rms-Werte für die Vokale des erstellten Baumes auf.

Faktor	Level	Korr. (6)	rms (6)	Korr. (@)	rms (@)
		0.728	0.022	0.523	0.033
Typ li.	Vokal	0.744	0.179	0.587	0.015
	Sono.	0.724	0.036	0.454	0.043
	Obst.	0.648	0.264	0.635	0.026
Typ re.	Vokal	0.596	0.040	0.707	0.159
	Sono.	0.732	0.016	0.616	0.017
	Obst.	0.686	0.023	0.446	0.040
	keiner	0.841	0.016	0.613	0.019
Gr. re.	keine	0.606	0.029	0.378	0.047
	Silbe	0.578	0.020	0.425	0.036
	andere	0.854	0.018	0.798	0.018
Klasse	Inh.	0.682	0.023	0.521	0.036
	Funk.	0.776	0.018	0.601	0.022
W.pos.	initial	0.628	0.017	0.926	0.008
	medial	0.567	0.022	0.420	0.050
	final	0.652	0.028	0.724	0.021
	mono	0.783	0.018	0.220	0.120

Tabelle 4: Diese Tabelle listet die Korrelationen und die rms-Werte für die versuchsweise durchgeführten weiteren Aufspaltungen der Schwa auf. Es wurden die folgenden Abkürzungen verwendet: Korr. = Korrelation, rms = rms-Wert, li. = links, re. = rechts, Gr. = Grenze, W.pos. = Position der Silbe im Wort, Sono. = Sonorant, Obst. = Obstruent, Inh. = Inhaltswort, Funk. = Funktionswort

"n"	1	-2.9715280813
"l"	1	-2.9340872221
"r"	1	-2.9719875430
"m"	1	-2.7471509016
"N"	1	-2.6521350652
"j"	1	-2.5326969702
"one"	3	0.0020849236
"fin"	3	0.0008815881
"ini"	3	-0.0261120507
"med"	3	0.0231455390
"med"	4	-0.0178558579
"fin"	4	0.0767971063
"ini"	4	-0.0589412483
"fin"	5	0.0061154651
"med"	5	-0.0224102110
"mono"	5	0.0099559610
"ini"	5	0.0063387849
	⋮	

Tabelle 5: Ausschnitt aus der nmult.out-Datei für ambisyllabische Sonoranten. Die erste Spalte enthält den Wert des Faktors, die zweite Spalte die Nummer des Faktors und die dritte Spalte die Koeffizienten des Faktors.

7 Implementierung und Integration in Festival

7.1 Das Sprachsynthesystem Festival

Das *Festival Speech Synthesis System* wurde an der Universität Edinburgh in Großbritannien entwickelt. Es bietet sowohl ein allgemeines Framework zum Erstellen von Sprachsynthesystemen, als auch Beispiele verschiedener Module. Es wird eine komplette Text-to-Speech Synthese angeboten. *Festival* ist multi-lingual. Neben den - am weitesten entwickelten - englischsprachigen (UK und US) Stimmen, wurden auch spanische und walisische Stimmen entwickelt.

Am Institut für maschinelle Sprachverarbeitung der Universität Stuttgart wird an einer deutschsprachigen Synthese für Festival gearbeitet. Die in dieser Arbeit beschriebene Lautdauermodellierung soll in diese Synthese eingefügt werden.

Festival wurde in C++ geschrieben und basiert auf den *Edinburgh Speech Tools*. Zur Steuerung existiert ein Scheme-basierter Kommando-Interpreter (SIOD).

Festival kann von wenigstens drei Benutzerklassen verwendet werden. Zum einen kann *Festival* von Benutzern angewendet werden, die lediglich - mit möglichst geringem Aufwand - eine qualitativ hochwertige Sprachsynthese aus beliebigem Text gewinnen wollen. Die zweite Benutzergruppe besteht aus Entwicklern, die Syntheseausgabe zu ihrem Sprachsystem hinzufügen möchten. Schließlich bietet *Festival* auch die Möglichkeit, neue Syntheseverfahren zu entwickeln und zu testen.

7.2 Implementierung

Die Implementierung des erstellten Modells wurde mit der Programmiersprache Scheme durchgeführt.

Festival bietet die Möglichkeit, CARTs (Classification and Regression Tree) zu erzeugen und zu benutzen.

Der mit Hilfe von *Durstat* erzeugte Kategorienbaum wird als CART implementiert. Dazu muß er jedoch zuerst in einen Binärbaum abgeändert werden (siehe Abb. 12). In dem implementierten CART werden die Entscheidungsfragen an den Knoten des Baumes beantwortet und beim Erreichen eines Blattes wird der Modellname des für dieses Blatt erstellten Modells zurückgegeben.

Die Daten aus zur Berechnung der Lautdauer aus den *nmult.out*-Dateien werden mit Hilfe eines Skripts zu einer Assoziationsliste umgeformt. In der Dauerfunktion wird aus dieser Liste mit dem Modellnamen aus dem CART und dem berechneten Faktorwert für jeden Faktor der entsprechende Koeffizient geholt. Mit diesen Koeffizienten wird schließlich die Lautdauer berechnet.

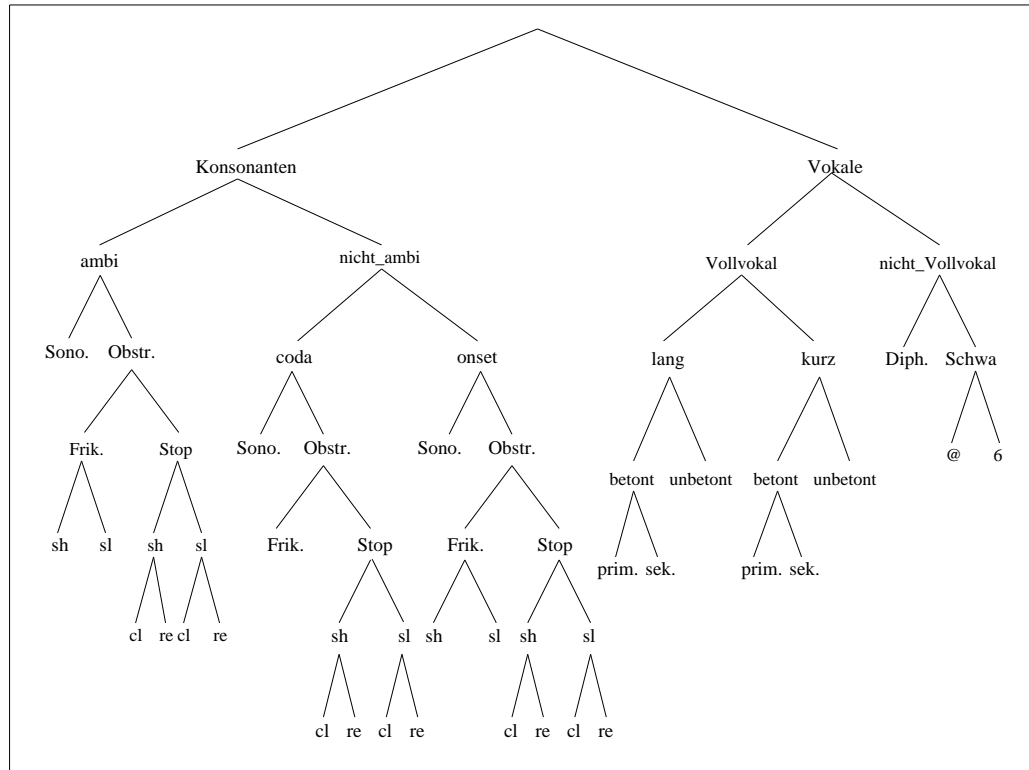


Abbildung 12: Kategorienbaum zum Binärbaum abgeändert. Folgende Abkürzungen wurden verwendet: ambi = ambisyllabisch, Sono. = Sonorant, Obstr. = Obstruent, Frik. = Frikativ, sh = stimmhaft, sl = stimmlos, cl = Verschlußphase, re = Öffnungsphase, Diph. = Diphthong, prim. = primäre Betonung, sek. = sekundäre Betonung, un. = unbetont

Zur Bestimmung der Faktorwerte mußten noch einige Funktionen geschrieben werden:

(segment_position SEGMENT)

bestimmt die Stellung des Segments in der Silbe (ambisyllabisch, Onset, Coda, Nukleus)

(word_position SEGMENT)

bestimmt die Stellung der Silbe im Wort (initial, medial, final, monosyllabisch)

(utt_position SEGMENT)

bestimmt die Stellung der Phrase im Satz (initial, medial, final, eine Phrase)

(phrase_position SEGMENT)

bestimmt die Stellung des Worts in der Phrase (initial, medial, final)

(word_length SEGMENT)

bestimmt die Anzahl der Silben im Wort (eine, mehr)

(phrase_length SEGMENT)

bestimmt die Anzahl der Wörter in der Phrase (eines, mehr)

(left_bound SEGMENT)

bestimmt die linke Grenze (keine, Silben-, Wort-, Phrasen-, Satzgrenze)

(right_bound SEGMENT)

bestimmt die rechte Grenze (keine, Silben-, Wort-, Phrasen-, Satzgrenze)

(word_class SEGMENT)

bestimmt die Wortklasse (Funktions- / Inhaltswort)

In der Datei *duration.scm* wird in der Funktion *Duration* die zu der gewählten Dauermodellierung gehörende Funktion aufgerufen.

7.2.1 Kompromisse

Da *Festival* nicht zwischen primärer und sekundärer Betonung unterscheidet, wird für alle betonten Vokale das Modell für primär betonte Vokale verwendet.

Die Plosive werden von *Festival* nicht in Verschuß- und Öffnungsphase aufgespalten. Es werden daher für jeden Plosiv die beiden Phasen modelliert und die beiden erhaltenen Dauerwerte zu einer Gesamtdauer addiert.

7.2.2 Einfügen einer neuen Modellierung

Wird eine neue Dauermodellierung erzeugt, so kann die vorliegende Implementierung relativ einfach abgeändert werden.

Bleibt die Baumstruktur bei der neuen Modellierung erhalten, so muß lediglich die Assoziationsliste, welche die zur Lautdauerberechnung benötigten Daten enthält, entsprechend der neuen Modellierung abgeändert werden.

Ändert sich auch die Baumstruktur, so muß zusätzlich noch der erzeugte CART angepaßt werden.

8 Schlußfolgerung

8.1 Zukünftige Arbeiten

Ein ausführlicher Test des vorliegenden Lautdauermodells konnte nicht durchgeführt werden. Folgende Arbeiten sollten die von der Modellierungsfunktion berechneten Lautdauern überprüfen und mit Lautdauern aus anderen Verfahren vergleichen. Da die Modellierung die natürliche Sprache möglichst gut nachahmen sollte, sind für eine derartige Analyse perzeptive Tests unerlässlich.

Im Rahmen dieser Auswertung müßte auch nach der Ursache für die relativ schlechten Korrelations- und rms-Werte der beiden Schwa gesucht werden. Mit den Ergebnissen aus diesen Tests kann eine Verbesserung der vorliegenden Lautdauermodellierung oder - bei grundlegend neuen Erkenntnissen - eine komplett neue Modellierung vorgenommen werden.

8.2 Verbesserungsvorschläge

Durch die Verwendung eines komplett handgelabelten Sprachkorpus könnte die Dauer der einzelnen Segmente exakter bestimmt und damit ein besseres Modell erzeugt werden.

Ein umfangreicheres Korpus, welches auch mehrere längere Texte enthält und nicht nur einzelne Sätze, könnte auch zu einer besseren Modellierung beitragen. Damit würde ein häufigeres Vorkommen der einzelnen Segmente mit den selben Merkmalen gewährleistet und es könnte somit auch eine größere Zahl an Dauerwerten für diese Segmente erhalten werden. Dies würde eine genauere Modellierung ermöglichen.

Um eine vom Sprecher unabhängige Lautdauermodellierung zu erreichen, sollten die selben Texte, von mehreren Personen gesprochen, vorliegen. Für jeden Sprecher sollte eine eigene Lautdauermodellierung erstellt und die erhaltenen Modelle miteinander verglichen werden. Dabei könnten eventuelle sprecherspezifische Eigenheiten erkannt und in dem, zur Synthese zu verwendenden, Modell korrigiert werden.

8.3 Ausblick

Für zukünftige Modellierungen könnte die Untersuchung subsegmentalen und asynchronen Timings eine weitere Möglichkeit bieten, eine natürlicher klingende Sprachsynthese zu erhalten.

Außerdem sollte jedoch der Untersuchung von Regelmäßigkeiten in natürlicher Sprache mehr Aufmerksamkeit zukommen. So können Pausen und eine verlangsamte Sprechgeschwindigkeit nicht nur an Satzzeichen festgestellt

werden. Daher wäre es interessant, zu erfahren, wie Sprecher ihre Sprachausgabe in Einheiten (“chunks”) aufteilen.

Schließlich bedeutet auch jede Verbesserung von Modulen, welche Eingaben für Dauermodule bereitstellen, eine Verbesserung der Lautdauermodellierung. Am wichtigsten ist hier die Arbeit an der Textanalyse, um Phrasengrenzen und phonologisch-intonationale Kategorien bestimmen zu können.

Literatur

- [Moe00] Möbius, Bernd: German and multilingual speech synthesis. Habilitationsschrift 2000
- [Bre99] Breitenbücher, Mark: Datenbasierte Methoden der Sprachsynthese. Diplomarbeit 1999
- [Cam91] Campbell, W. Nick; Isard, Stephen D.: Segment durations in a syllable frame. In: *Journal of Phonetics*, 19, 37 - 47 (1991)
- [Kla76] Klatt, Dennis H.: Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. In: *Journal of the Acoustical Society of America*, 59, 1209 - 1221 (1976)
- [San94] van Santen, Jan P. H.: Assignment of segmental duration in text-to-speech synthesis. In: *Computer Speech and Language*, 8, 95 - 128 (1994)
- [San93] van Santen, Jan P. H.: Exploring N-way Tables with Sums-of-Products Models. In: *Journal of Mathematical Psychology*, 37, 327 - 371 (1993)
- [San97] van Santen, Jan P. H.: Segmental Duration and Speech Timing. In: *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Sagisaka, Yoshinori; Campbell, Nick; Higuchi, Norio. New York: Springer 1997
- [Cam92] Campbell, W. Nick: Syllable-based segmental duration. In: *Theories, Models and Designs*. Bailly, Gérard; Benoît, Christian; Sawallis, Thomas R. Amsterdam: Elsevier 1992
- [San95] van Santen, Jan P. H.: Computation of Timing in Text-to-Speech Synthesis. In: *Speech Coding and Synthesis*. Kleijn, W.-Bastiaan; Paliwal, Kuldip-K. Amsterdam: Elsevier 1995
- [Koh94] Kohler, K. J.: Lexica of the Kiel PHONDAT Corpus: Read Speech Vol. I + II. In: *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel*, 27 u. 28
- [Koh92] Kohler, K. J.: Phonetisch-Akustische Datenbasis des Hochdeutschen. Kieler Arbeiten zu den PHONDAT-Projekten 1989-1992. In: *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel*, 26
- [Sch97] Shih, Chilin; Ao, Benjamin: Duration Study for the Bell Laboratories Mandarin Text-to-Speech System. In: *Progress in Speech Synthesis*. van Santen, Jan; Sproat, Richard W.; Olive, Joseph P.; Hirschberg, Julia. New York: Springer 1997