Institut für
Technische Informatik und
Kommunikationsnetze

# Erkennung von Silbenakzenten aus Sprachsignalen

Cécile Bucher

Semesterarbeit SA-2007-41

Sommersemester 2007

Institut für Technische Informatik
und Kommunikationsnetze

Betreuer: H. Romsdorfer

Verantwortlicher: Prof. Dr. L. Thiele

# Kurzfassung

Um die Prosodie aktueller Sprachsynthesesysteme zu erzeugen, werden zumeist statistische Modelle verwendet. Um solche Modelle trainieren zu können, braucht man korrekt annotierte Trainigssätze. Für die Prosodiesteuerung sind dabei vor allem die korrekte Position und Stärke der Satzakzente sowie der Phrasengrenzen wichtig.

Da das Annotieren von Hand sehr aufwändig ist, ist es Ziel dieser Arbeit aus dem Sprachsignal, der phonologischen Repräsentation und der gegebenen Lautsegmentierung die unbetonten von den betonten Silben automatisch zu unterscheiden und diese in 4 Akzentstärken abzustufen.

Dazu wurde ein Programm geschrieben das Features aus Signalfiles extrahiert. Es werden verschiedene Features zur Dauer, zur Signalenergie, zur Grundfrequenz und auch aus der Phonologischen Beschreibung verwendet. Zusätzlich können auch noch Informationen über Phrasen und Wortgrenzen verwendet werden. Mit den Features werden anschliessend verschiedene Modelle trainiert und getestet. Es wird das Featuresubset gesucht wo die Tests am besten abschneiden. Schlussendlich wird dann nur noch dieses verwendet.

Es wurden verschiedene Modelle untersucht. Verwendet wurde ein 1 stufiges Modell mit einem neuronalen Netz und 1 zwei stufiges Modell mit 2 neuronalen Netzen, welches zuerst in betont/unbetont unterteilt und anschliessend in die 4 Akzentklassen. Ein lineares Modell wurde ebenfalls getestet.

Zuerst wurden die Modelle auf dem gleichen Korpus getestet wie sie trainiert wurden. Hier kann eine zu knapp 80% erfolgreiche Klassierung über 5 Klassen erreicht werden. Anschliessend wurde ein Übertagung auf einen anderen Korpus getestet. Einmal von einer männlichen auf eine weibliche Stimme ein andermal von deutsch auf französisch. Bei der Übertragung auf eine andere Stimme verliert man etwa 6%, beim Übertragen auf eine andere Sprache verliert man etwas mehr. Die trainierten Modelle können auch adaptiert werden, indem dem neuronalen Netz einige Sätze des neuen Korpus dazugegeben werden und damit nochmals einige Trainingszyklen durchgeführt werden. Beim Übertragen in eine andere Sprache stellte sich diese Adaption schon bei 20 Sätzen als sehr nützlich heraus. Der stärkere Verlust bei der Übetragung kann hier mehr als ausgeglichen werden. Die Adaption auf eine andere Stimme scheint schwieriger und ist weniger wirksam.

# Inhaltsverzeichnis

# Figurenverzeichnis

# Tabellenverzeichnis

# 1 Einleitung

Das Ziel dieser Arbeit ist es aus dem Sprachsignal, der phonologischen Repräsentation und der gegebenen Lautsegmentierung die unbetonten von den betonten Silben automatisch zu unterscheiden und diese in 4 Akzentklassen abzustufen.

In diesem Kapitel wird die Definition der Akzentklassen sowie der Phonologischen Repräsentation angegeben.

## 1.1 Akzentklassen

Wie auch in [Tra95] werden in dieser Arbeit 4 Akzentklassen unterschieden.

**Phrasenhauptakzent** gibt es in jeder Phrase nur einen. Er beinhaltet die Hauptinformation der Phrase. Im Sprachsignal zeichnet sich der Phrasenhauptakzent hauptächlich durch eine starke Grundfrequenzbewegung und durch eine längere Silbendauer aus. Meist ist der Phrasenhauptakzent der "Pitch Accent" mit der stärksten Grundfrequenzbewegung.

Der **"Pitch Accent"** zeichnet sich ebenfalls durch eine starke Grundfrequenzbewegung und durch eine längere Silbendauer aus.

Als **"Non-Pitch Accent"** bezeichnet man eine Betonung auf Worhauptakzentposition welche sich lediglich durch die längere Silbendauer auszeichnet.

**Wortnebenakzente** gibt es meist nur in Komposita. Meist bekommt die betonte Silbe des zweiten Worts im Kompositum den Wortnebenakzent. Dieser zeichnet sich ebenfalls nur durch eine längere Silbendauer aus.

**Unbetonte Silben** haben weder eine starke Grundfrequenzbewegung noch eine längere Silbendauer.

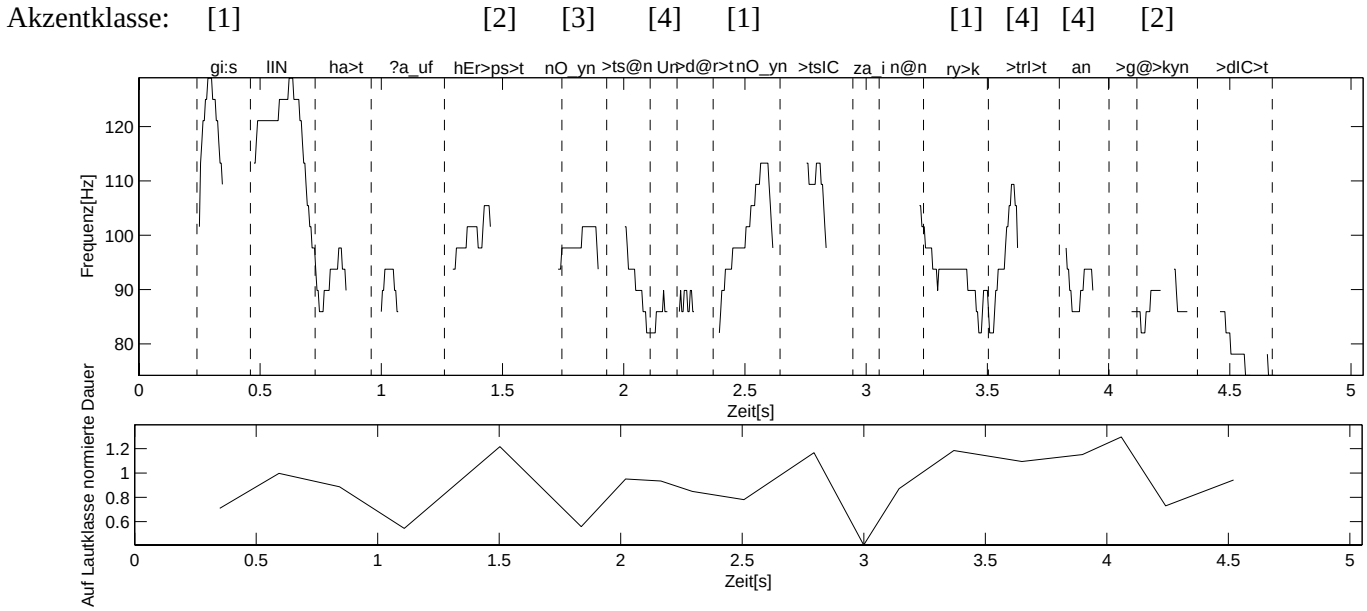
Fig. 1 zeigt die Grundfrequenz und die auf die Lautklassen normierten Nucleusdauern eines Satzes, welcher alle oben beschriebenen Akzentklassen enthält.

## 1.2 Phonologische Repräsentation

Die Annotation eines Trainigssatzes basiert auf einer abstrakten Beschreibung der phonologischen Eigenschaften dieses Satzes, der so genannten phonologischen Repräsentation. Die phonologische Repräsentation beinhaltet neben der phonetischen Transkription der Lautsequenz Informationen bezüglich Sprache auch Betonung, Phrasierung und Silbeneinteilung des Satzes. Als Beispiel hier die phonologische Repräsentation des Satzes "Die Lawinengefahr hat leicht abgenommen".

```
(P) di- la-[1]vi:-n@n-g@-[4]fa:r- #{2} (T) hat- [2]la_iCt-
  [1]?ap-g@-[4]nO-m@n.
```

In der phonologischen Repräsentation können folgende Spezialsymbole auftreten:




**Figur 1:** *Grundfrequenz und auf Lautklassen normierte Nucleusdauern des Satzes "Gisling hat auf Herbst neunzehnhundertneunzig seinen Rücktritt angekündigt"*

**#{n}** Steht für eine Phrasengrenze, wobei n=1 die Satzgrenze bezeichnet, n=1 eine satz-interne Phrasengrenze mit Pause und n>1 ein satz-interne Phrasengrenze ohne Pause kennzeichnet.

**#(X)** Ist eine Phrasentyp Markierung, welche am Anfang der Phrase platziert wird. Es gibt 5 Phrasentypen, bezeichnet mit (P) für *progredient*, (T) für *terminal*, (S) für *statement*, (Y) *y/n-question*, (E) für *exclamation*.

\ **X** \ kennzeichnet einen Sprachwechsel. So wechselt die Sprache beispielsweise mit '\E\' nach Englisch, mit '\F\' nach Französisch und mit '\G\' nach Deutsch.

- kennzeichnet eine Silbengrenze. Da eine Phrasengrenze zugleich eine Silbengrenze ist, ist die Markierung direkt vor einer Phrasengrenze optional.

**[ n ]** markiert eine Silbenzakzent. n=1 bezeichnet einen Phrasenhauptakzent, n=2 einen "Pitch Accent", n=3 einen "Non-Pitch Accent" und n=4 einen Wortnebenakzent. Mit n=0 können optional unbetonte Silben gekennzeichnet werden.

# 2 Beschreibung der Prosodie Korpora

In diesem Kapitel werden die verwendeten Sprachkorpora beschrieben.

## 2.1 Männlicher Sprecher, Deutsch

Dieser Korpus enthält 186 Sätze welche alle von demselben männlichen Sprecher gesprochen wurden. Bei den Sätzen handelt es sich ausschliesslich um deutsche Aussagesätze im Stil eines Nachrichtensprechers.

| | |
|---:|---:|
| Unbetonte Silben | 3930 |
| Akzentklasse [1] | 1002 |
| Akzentklasse [2] | 647 |
| Akzentklasse [3] | 329 |
| Akzentklasse [4] | 680 |
| Totale Anzahl Silben | 6588 |

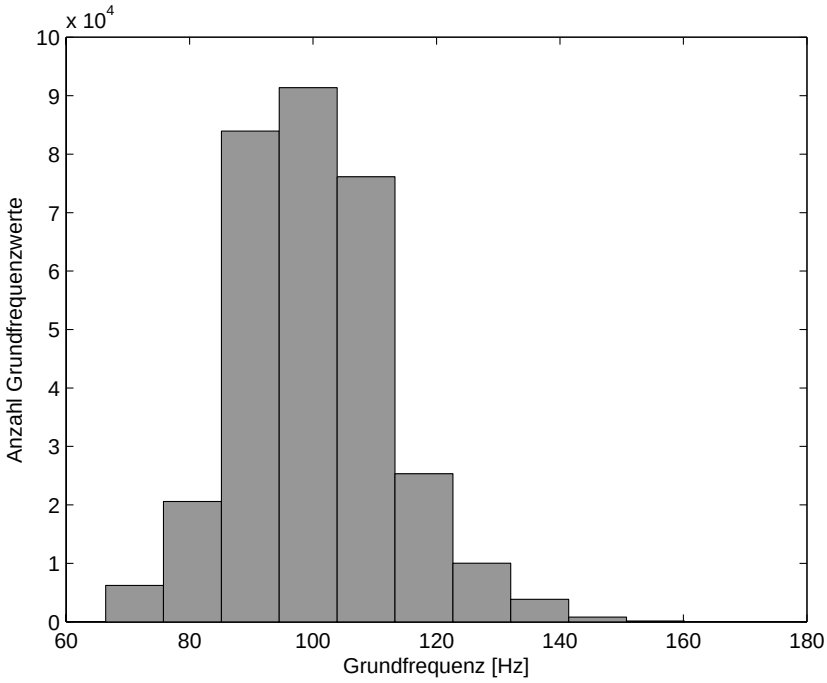
**Tabelle 1:** *Inventar der Silben im Korpus, mit Angabe der Akzentklassen Zugehörigkeit*

Tabelle 1 gibt an wieviele Silben im Korpus enthalten sind und zu welcher Akzentklasse sie gehören.

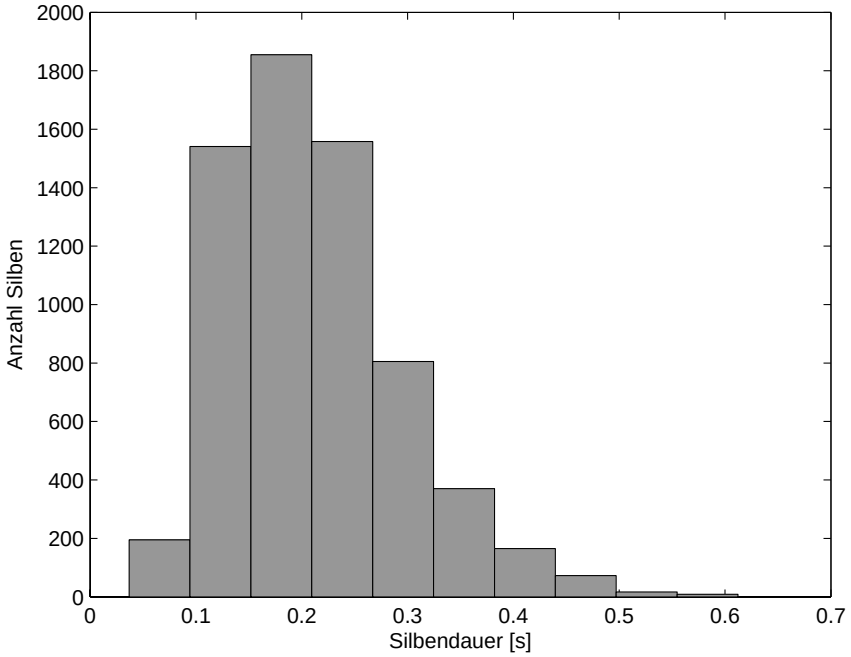
| Eigenschaft | Mittelwert | Varianz |
|---|---:|---:|
| Grundfrequenz [Hz] | 100.44 | 155.2 |
| Silbendauer [s] | 0.212 | 0.007 |
| Nucleusdauer [s] | 0.085 | 0.002 |
| Silbenzahl pro Phrase | 6.55 | 10.95 |

**Tabelle 2:** *Mittelwert und Varianz einiger Eigenschaften über den ganzen Korpus*

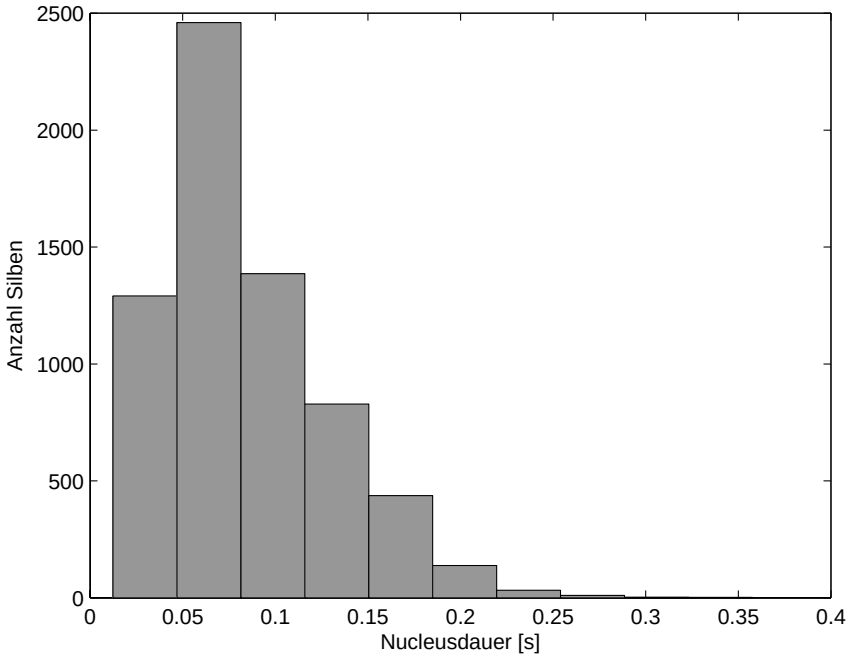
Tabelle 2 gibt die Mittelwerte einiger Eigenschaften im Korpus an. In Fig. 2 bis 5 sind die Histogramme zu den selben Eigenschaften dargestellt.



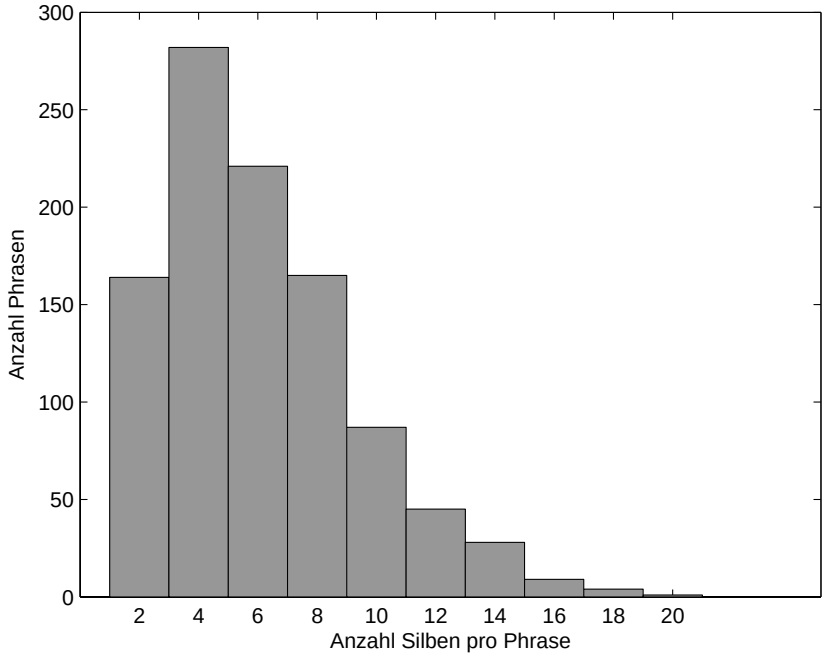

**Figur 2:** *Histogramm der Grundfrequenzwerte im männlichen Prosodie Korpus*




**Figur 3:** *Histogramm der Silbendauern im männlichen Prosodie Korpus*




**Figur 4:** *Histogramm der Nucleusdauern im männlichen Prosodie Korpus*




**Figur 5:** *Histogramm der Silbenanzahl pro Phrase im männlichen Prosodie Korpus*

## 2.2 Weibliche Sprecherin, Deutsch

Aus diesem Korpus wurden 400 Sätze verwendet, die wurden alle von derselben weiblichen Sprecherin gesprochen wurden. Die Sätze sind kürzer und unterschiedlicher als die im in 2.1 beschriebenen Korpus, das heisst es gibt nicht nur Aussagesätze sondern auch Fragesätze.

Tabelle 3 gibt an wieviele Silben in den 400 Sätzen des Korpus enthalten sind und zu welcher Akzentklasse sie gehören.

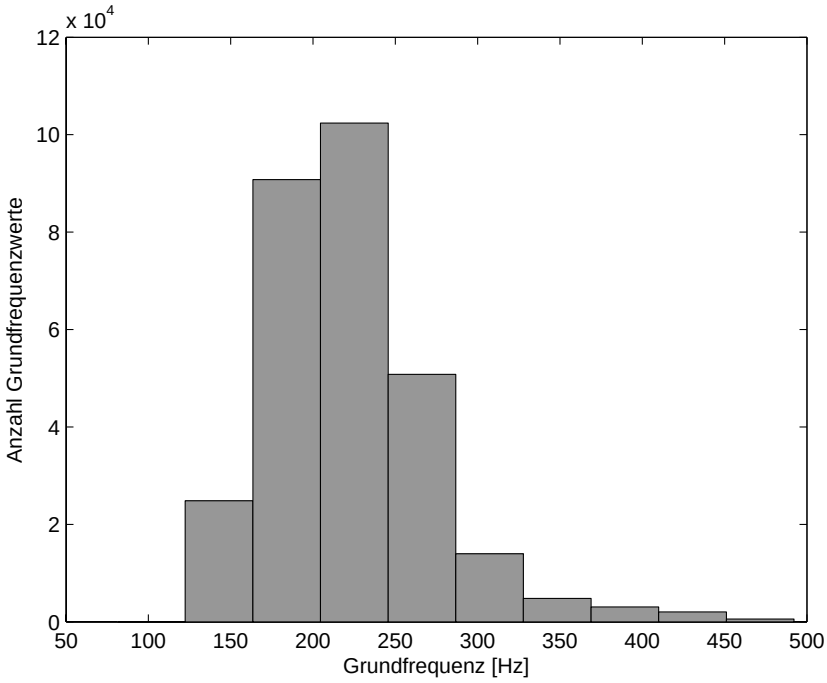
| | |
|---:|---:|
| Unbetonte Silben | 3471 |
| Akzentklasse [1] | 1102 |
| Akzentklasse [2] | 617 |
| Akzentklasse [3] | 207 |
| Akzentklasse [4] | 278 |
| Totale Anzahl Silben | 5675 |

**Tabelle 3:** *Inventar der Silben im Korpus, mit Angabe der Akzentklassen Zugehörigkeit*

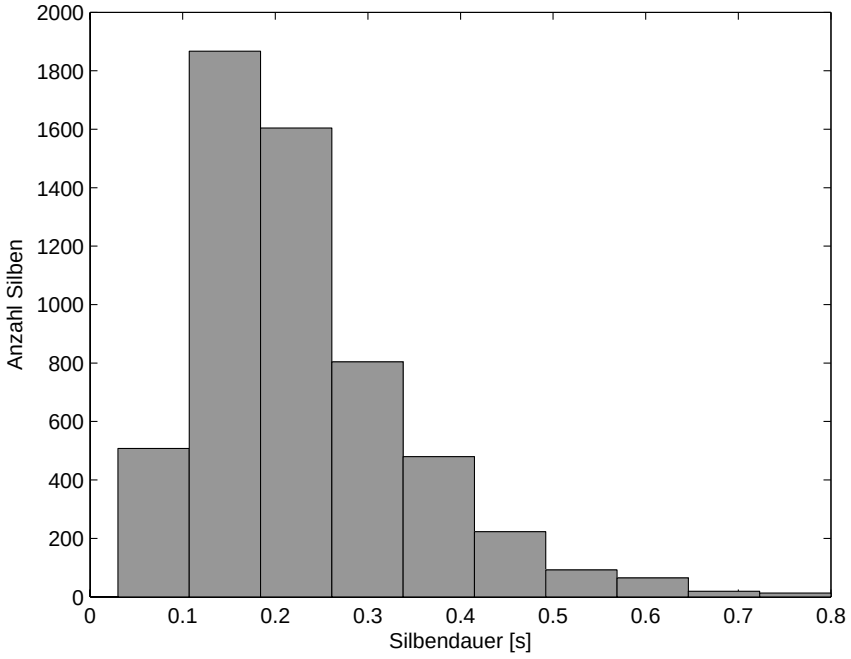
| Eigenschaft | Mittelwert | Varianz |
|---|---:|---:|
| Grundfrequenz [Hz] | 222.51 | 2479.8 |
| Silbendauer [s] | 0.229 | 0.013 |
| Nucleusdauer [s] | 0.085 | 0.003 |
| Silbenzahl pro Phrase | 5.54 | 7.55 |

**Tabelle 4:** *Mittelwert und Varianz einiger Eigenschaften über die verwendeten 400 Sätze im Korpus*

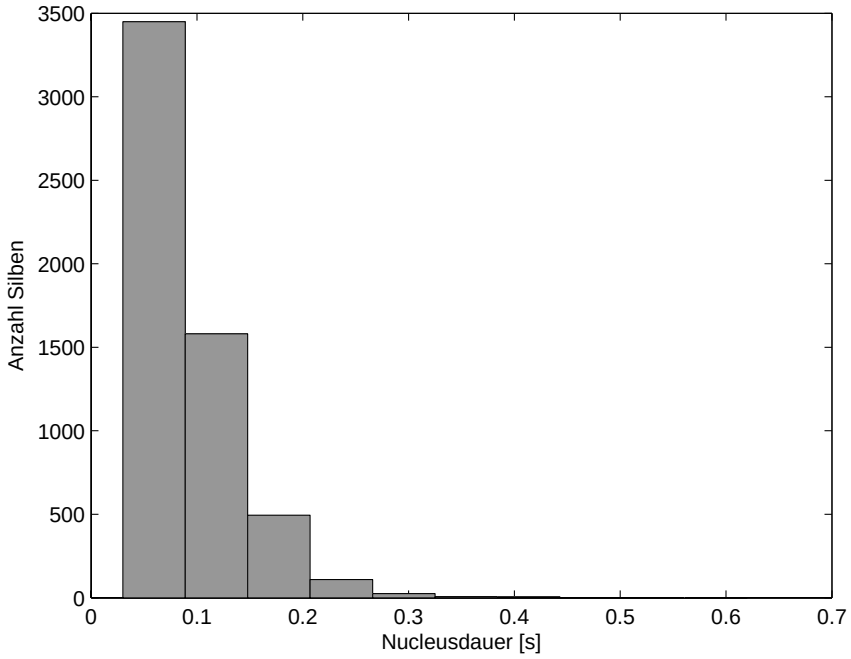
Tabelle 4 gibt die Mittelwerte einiger Eigenschaften des Korpus an. In Fig. 6 bis 9 sind die Histogramme zu den selben Eigenschaften dargestellt.



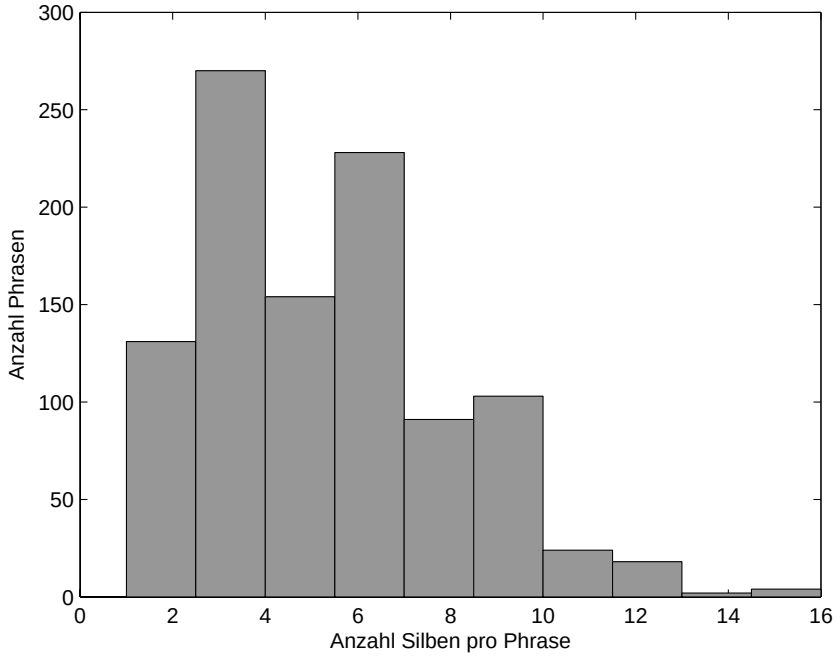

**Figur 6:** *Histogramm der Grundfrequenzwerte im weiblichen Prosodie Korpus*




**Figur 7:** *Histogramm der Silbendauern im weiblichen Prosodie Korpus*




**Figur 8:** *Histogramm der Nucleusdauern im weiblichen Prosodie Korpus*




**Figur 9:** *Histogramm der Silbenanzahl pro Phrase im weiblichen Prosodie Korpus*

## 2.3 Weibliche Sprecherin, Französisch

Aus diesem Korpus wurden 50 Sätze verwendet, die wurden alle von derselben weiblichen Sprecherin gesprochen wurden wie auch die im in 2.2 beschriebenen Korpus. Die Sätze sind kürzer und unterschiedlicher als die im in 2.1 beschriebenen Korpus, das heisst es gibt nicht nur Aussagesätze sondern auch Fragesätze.

Tabelle 5 gibt an wieviele Silben in den 50 Sätzen des Korpus enthalten sind und zu welcher Akzentklasse sie gehören.

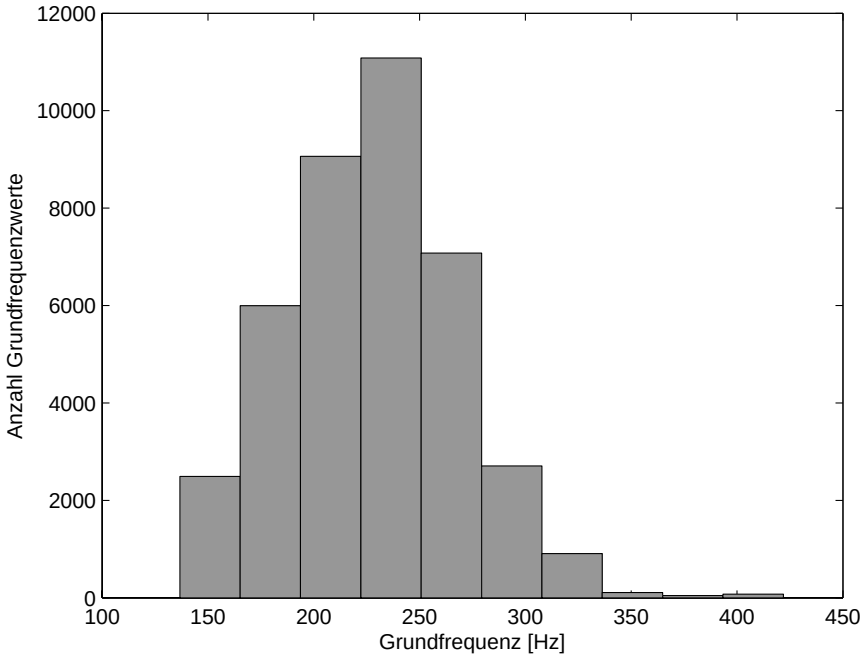
| | |
|---|---|
| Unbetonte Silben | 714 |
| Akzentklasse [1] | 363 |
| Akzentklasse [2] | 169 |
| Akzentklasse [3] | 102 |
| Akzentklasse [4] | 61 |
| Totale Anzahl Silben | 1307 |

**Tabelle 5:** *Inventar der Silben im Korpus, mit Angabe der Akzentklassen Zugehörigkeit*

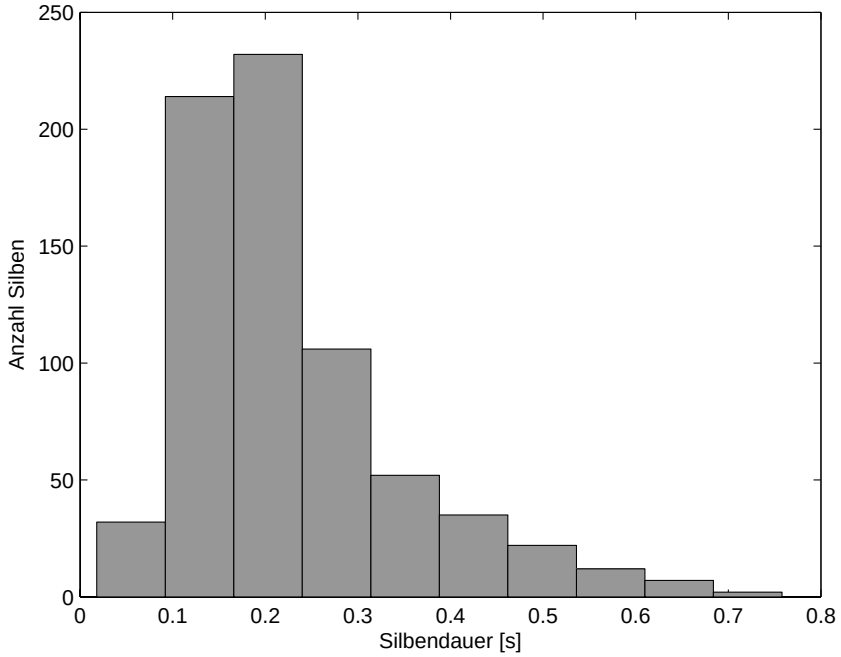
| Eigenschaft | Mittelwert | Varianz |
|---|---|---|
| Grundfrequenz [Hz] | 227.72 | 1637.6 |
| Silbendauer [s] | 0.228 | 0.014 |
| Nucleusdauer [s] | 0.096 | 0.003 |
| Silbenzahl pro Phrase | 4.38 | 3.78 |

**Tabelle 6:** *Mittelwert und Varianz einiger Eigenschaften über die verwendeten 50 Sätze im Korpus*

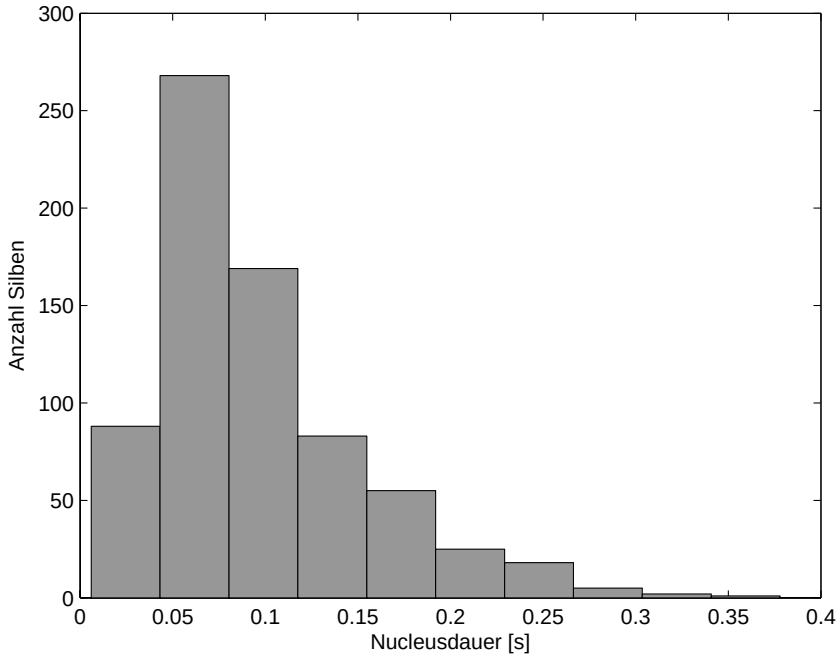
Tabelle 6 gibt die Mittelwerte einiger Eigenschaften des Korpus an. In Fig. 10 bis 13 sind die Histogramme zu den selben Eigenschaften dargestellt.



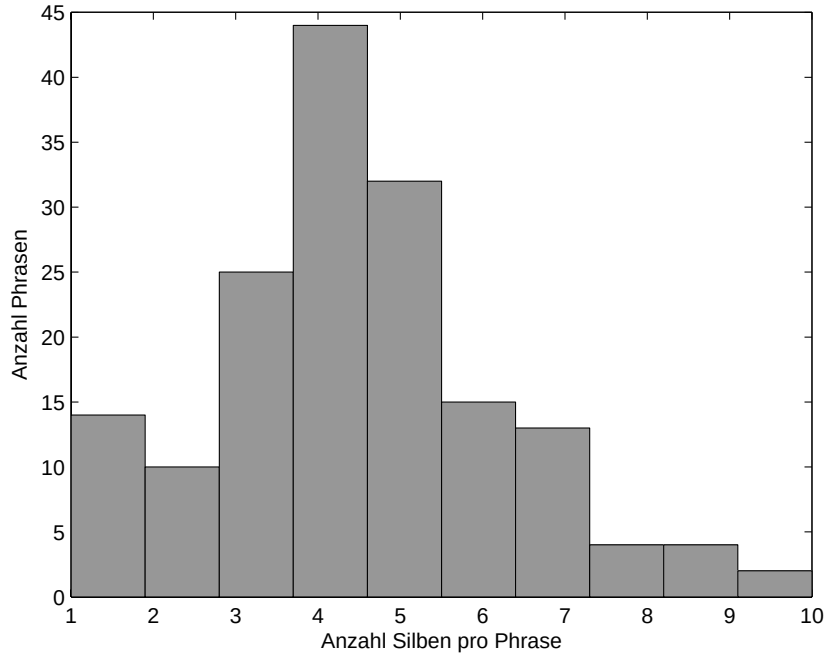

**Figur 10:** *Histogramm der Grundfrequenzwerte im französischen Prosodie Korpus*




**Figur 11:** *Histogramm der Silbendauern im französischen Prosodie Korpus*




**Figur 12:** *Histogramm der Nucleusdauern im französischen Prosodie Korpus*




**Figur 13:** *Histogramm der Silbenanzahl pro Phrase im französischen Prosodie Korpus*

# 3 Aufbau des Detektionsprogramms

In diesem Kapitel wird der Aufbau des Rahmenprogramms zur Akzentklassendetektion beschrieben.

## 3.1 Feature Extraktion

Zuerst wird eine grosse Anzahl von Features berechnet, welche versucht die vorhandenen Daten möglichst vollständig auszuwerten. Dafür müssen zuerst 4 verschiedene File Typen eingelesen werden. Es sind dies die Wavefiles, die F0files, die Labelfiles und die Ptrfiles. Aus den Wavefiles kann die Samplingfrequenz und die Signalenergie gelesen werden. Die F0files enthalten die Grundfrequenzwerte, die Labelfiles die Anfangs und Endpunkte der Laute und deren phonologische Bezeichnung. In den Ptrfiles befindet sich die komplette Phonologische Beschreibung jedes Satzes inklusive Phrasentypen und Akzentklassen.

Für die Features welche über den ganzen Korpus normiert werden, muss zuerst der ganze Korpus eingelesen werden um die benötigten Parameter zu ermitteln. Anschliessend werden die auf den Satz normierten Features für jeden Satz einzeln berechnet.

Eine ausführliche Beschreibung aller Features befindet sich im Kapitel 4. Beim Extrahieren der Features werden alle Featurewerte in sogenannte Featurefiles geschrieben. Somit muss die Extraktion jeweils nur einmal durchgeführt werden.

## 3.2 Einlesen der Featurefiles

Aus den Featurefiles soll nun eine Featurematrix entstehen. Nach dem Einlesen der Featurewerte für jeden Satz, werden zuerst noch einige Feature hinzugegeben welche direkt aus der Lautidentität des jeweiligen Lautes berechnet werden können. Diese Features sind im Abschnitt 4.4 beschrieben. Alle Features die über 3 Kontext Silben berechnet werden, werden einmal berechnet und dann jeweils der Silbe vorher und der Silbe nachher dazugegeben. Zu jeder Silbe gibt es dann also die Features der aktuellen Silbe plus die der vorherigen und die der nacherigen Silbe.

Die Akzentwerte aus der vorgegebenen Annotation werden in einer Targetmatrix abgespeichert, welche für jede Silbe eine Zeile und für jede Akzentklasse eine Spalte hat.

Anschliessend werden Feature- und Targetmatrix gleich in ein Trainings und Testset eingeteilt. Es wurden jeweils 60% als Trainingset und 40% als Testset verwendet.

## 3.3 Normierung

Damit die extrahierten Features Werte in vergleichbaren Grössen haben, müssen sie als erstes normiert werden. Je nach Modell wurde eine andere Normierung verwendet, mehr dazu im Abschnitt 5.1. Das Testset wird dabei immer auf das Trainingsset normiert.

## 3.4 Feature Auswahl

Nun muss heraus gefunden werden welche Features dem Modell bei seinem Entscheid tatsächlich helfen und welche es eher verwirren. Daher muss nun das optimale Featuresubset gefunden werden, dabei wird folgendermassen vorgegangen:

Zunächst wird das jeweilige Modell mit allen Features trainiert und der initiale Klassifizierungsfehler wird berechnet. Um nun das optimale Featuresubset zu finden, wird der Reihe nach immer ein Feature eliminiert, das heisst wenn man am Anfang N Features hat, werden nun alle Konfigurationen mit N-1 Features durchgerechnet. Anschliessend wird das Feature eliminiert, ohne welches der Klassifizierungsfehler am kleinsten war. Mit dem neuen Subset von N-1 Features wird das selbe wiederholt, bis am Schluss nur noch 1 Feature übrigbleibt. Das optimale Featuresubset wird nun an der Stelle gefunden wo der Fehler minimal war.

## 3.5 Auswahl der Netzkonfiguration

Bei den Modellen die neuronale Netze (siehe Abschnitte 5.2 und 5.3) verwenden, muss auch noch die optimale Anzahl der Hiddenknoten gefunden werden. Dafür wurden jeweils 3 Möglichkeiten ausprobiert: 2 Hiddenknoten, 4, und eine Variable Zahl welche nach [Bis95] eine Abschätzung für die ideale Hiddenknoten Anzahl berechnen soll. Diese Abschätzung wird folgendermassen berechnet:

$$\text{Anzahl Hiddenknoten} = \frac{\frac{\text{Anzahl Trainingsdaten}}{10} - \text{Outputs}}{\text{Inputs} + \text{Outputs} + 1} \tag{1}$$

Beim linearen Modell (Abschnitt 5.4) gibt es nur eine Netzkonfiguration.

Für jede Netzkonfigurationen wird nun ein Modell trainiert und es wird gespeichert welche Konfiguration am besten abschneidet.

## 3.6 Training

Für das Training wird vom Trainingset ein Evaluationsset abgespalten. Anschliessend wir das neuronale Netz zufällig initialisiert und mit dem verbleibenden Trainingset trainiert. Das Netz wird solange trainiert bis entweder 100 Trainingszyklen vorbei sind oder bis es auf dem Evaluationsset wieder schlechter wird. Damit soll ein Übertraining verhindert werden.

Um der Möglichkeit eines schlechten Abschneidens, durch schlechte Initialisierung, auszuweichen, können mehrere Kreuzvalidierungen berechnet werden und die gleiche Netzkonfiguration kann mehrmals trainiert werden.

Die Kreuzvalidierungen funktionieren folgendermassen: Die Unterteilung von Validierungsset und Trainingset kann verschieden gewählt werden. Bei der hier verwendeten sechsfachen Kreuzvalidierung werden zuerst die ersten 5/6 der Trainingsätze als Trainingset definiert und der letzte Sechstel ist nun das Validierungsset. Beim nächsten Lauf wird nun der zweitletzte Sechstel zum Validierungsset, und der Rest bleibt im Trainingsset. So geht es weiter bis alle 6 Validierungen durch sind.

Beim linearen Modell (siehe Abschnitt 5.4) ist kein Training nötig, die Gewichte können hier in einer einzigen Operation aus den Trainingsdaten berechnet werden.

## 3.7 Testen

Jedes dieser Modelle wird dann mit einem Testset geprüft. Im Testset werden die Konfusionsmatrix der 5 Klassen und die Fehlerrate berechnet.

Zurückgegeben wird am Schluss für jedes Featuresubset der Fehlerwert der Netzkonfiguration und der Validierung die im Testset am besten abschneidet.

# 4 Features

Falls nichts anderes erwähnt sind alle infolge beschriebenen Features jeweils über drei Silben berechnet.

## 4.1 Dauerfeatures

Es wurden 5 Features zur Dauer berechnet, die Lautspezifische Dauer, die Silbendauer und die Nucleusdauer.

**D1** Die Lautspezifische Dauer ist die Dauer des Silbenkerns (Nucleus) normiert auf den Mittelwert des spezifischen Lautes über der ganzen Korpus. Das heisst alle a's werden auf den Mittelwert der a's normiert alle e's auf den Mittelwert aller e's und so weiter.

**D2** Die Silbendauer berechnet die gesamt Dauer aller Laute die zur jeweiligen Silbe gehören, sie wird auf den Mittelwert aller Silben des jeweiligen Satzes normiert.

**D3** Die Nucleusdauer berechnet wie oben die Dauer des Silbenkerns, wird aber nun auf den Mittelwert aller Silbenkerne im Satz normiert.

**D4** Die Pausendauer beschreibt, die Länge der Pause nach der jeweiligen Silbe, sie wird auf die Pausendauern im Satz normiert.

**D5** Das 5.te Dauer Feature beschreibt den Abstand zwischen zwei Silbenkernen, zu einer Silbe gehört jeweils das Feature welches den Abstand zwischen dem Ende des aktuellen Silbenkern und dem Anfang des nächsten Silbenkerns, der Silbenkernabstand wird wiederum auf die Silbenkernabstände im Satz normiert.

## 4.2 Energiefeatures

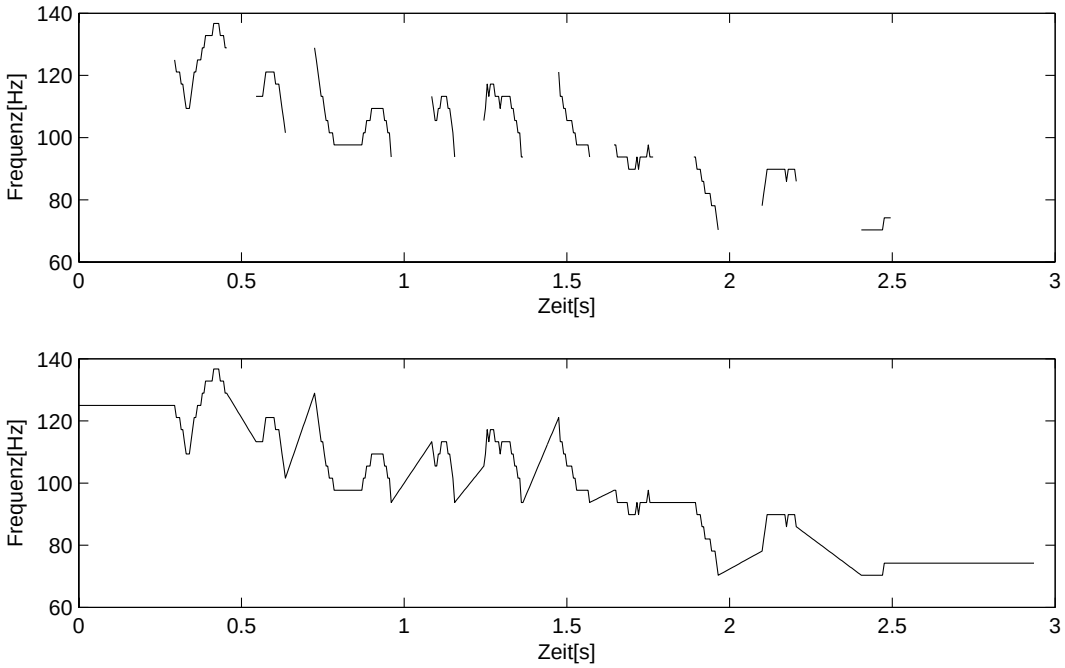
Es wurden 6 Features zur Nucleus Energie berechnet.

**E1** Das erste Feature zur Energie enthält die totale RMS Energie (über alle Frequenzen) über den Nucleus.

Fünf weitere Features enthalten die spektrale Emphasis, die in 5 Frequenzbänder - 0-500Hz (**E2**), 200-1000Hz (**E3**), 500-2000Hz (**E4**), 1000-4000Hz (**E5**) und über 2000Hz (**E6**) - aufgeteilt ist. Alle diese 6 Features sind auf den jeweiligen Satz normiert.

## 4.3 Grundfrequenzfeatures

Für jede Silbe wurden 9 Features zur Grundfrequenz berechnet. Die verwendeten Grundfrequenzwerte wurden an den stimmlosen Stellen jeweils interpoliert, in Fig. 14 sehen wir oben eine Grundfrequenz ohne Interpolation an den stimmlosen Stellen und unten eine mit. Für alle Features wurde die Grundfrequenz zuerst logarithmiert und anschliessend linear normiert. Diese Vorbehandlung hilft, dass die Grundfrequenzwerte für Frauen und Männer Stimmen ziemlich genau gleich verteilt sind.




**Figur 14:** *Grundfrequenz des Satzes "Friedliche Massenkundgebung in Peking" oben ohne die stimmlosen Stellen und unten mit Interpolation an diesen*

**F0_1** Der Grundfrequenzonset beschreibt eine 3 Punkt Abtastung der Grundfrequenz über den Anteil der Silbe vor dem Kern. Dies ergibt 2 Features aus den 2 Steigungen zwischen den 3 Punkten.

**F0_2** Die Grundfrequenz im Nucleus wird durch eine 5 Punkt Abtastung beschrieben. Dies ergibt 4 Features aus den 4 Steigungen zwischen den 5 Punkten.

**F0_3** Der Grundfrequenzcoda beschreibt eine 3 Punkt Abtastung der Grundfrequenz über den Anteil der Silbe nach dem Kern. Dies ergibt 2 Features aus den 2 Steigungen zwischen den 3 Punkten.

**F0_4** Ein weiteres Feature entsteht aus der maximalen Grundfrequenzschwankung des Silbenkerns, der Differenz zwischen maximaler und minimaler Grundfrequenz und innerhalb des Silbenkerns.

## 4.4 Features aus der Phonologischen Beschreibung

Es wurden 8 Features aus der Phonologischen Beschreibung verwendet.

**Ph1** Zwei Features beschreiben die horizontale Zungenposition, dabei handelt es sich um binäre Features. Das eine Feature ist 1 wenn die Zungenspitze hinten ist, das andere ist 1 wenn die Zungenspitze vorne ist, sind beide Null ist die Zungenspitze in der Mitte.

**Ph2** Die Funktion 'isHighF0' ist ein binäres Feature, sie beschreibt ob der Vokal eine höhere Grundfrequenz hat.

**Ph3** Diese Feature beschreibt die Lippenposition, 1 bedeutet dass die Lippen geschlossen sind. 2 heisst geschlossen bis halb offen. 3 heisst halboffen bis offen und 4 heisst offen.

**Ph4** Die Funktion 'F1' beschreibt die Lage des ersten Formaten, 1 bedeutet tiefe Frequenz, 2 mittlere, 3 hohe.

**Ph5** Die Funktion 'isLong' gibt an ob es sich bei dem Laut im Nucleus um einen Vokal mit Längezeichen handelt oder nicht.

**Ph6** Die Funktion 'isDiph' gibt an ob es sich bei dem Laut im Nucleus um einen dem Diphthong handelt oder nicht.

**Ph7** Dieses Feature gibt an ob es sich bei dem Laut im Nucleus um ein Schwa handelt, da ein solches immer unbetont ist, sollte dieses Feature beim Entscheid zwischen betonten und unbetonten Silben sehr hilfreich sein. Es gibt nur die zwei Zustände "Schwa" und "nicht-Schwa". Darin inbegriffen sind auch Schwa-r Laute, welche in der phonologischen Beschreibung als *'6'* bezeichnet werden.

## 4.5 Hinzugabe von Informationen über Phrasen und Wortgrenzen

Es stellte sich als sehr hilfreich heraus, zur Klassifizierung zusätzlich zu den obigen Features auch noch die Angaben zu verwenden wo sich Phrasen und Wortgrenzen befinden. Diese Informationen können auf mehrere Arten verwendet werden. Entweder können dem Modell zusätzliche Features hinzugegeben werden, oder man kann eine nachträgliche Korrektur auf den schon klassifizierten Daten durchführen.

### 4.5.1 Phrasen und Wortgrenzen als zusätzliche Features, Variante 1

Mit den Phrasengrenzen als Features wurden wiederum 2 Varianten ausprobiert.

Bei der ersten Methode werden dem Modell jeweils zwei zusätzliche Features dazu gegeben. Diese sind binär und haben folgende Bedeutung. 11 bedeutet Phrasengrenze, 10 Wortgrenze und 00 ist eine Silbengrenze. Diese Features wurden dem Modell dann ebenfalls über einen 3 silbigen Kontext mitgegeben.

### 4.5.2 Phrasen und Wortgrenzen als zusätzliche Features, Variante 2

Bei der zweiten Variante werden die Features etwas anders berechnet.

Für die Phrasengrenze werden 2 Features berechnet, welche jeweils die Anzahl Silben bis zur letzten Phrasengrenze und bis zur nächsten angeben.

Für die Wortgrenze wird ein binäres Feature verwendet, 1 bedeutet Wortgrenze am Ende dieser Silbe, 0 bedeutet keine Wortgrenze. Dieses Feature wurde jeweils über einen Kontext von 6 Silben berechnet und zwar jeweils die 3 Silben vorher, die aktuelle Silbe und 2 Silben nachher.

### 4.5.3 Nachträgliche Korrektur auf Grund von Phrasen und Wortgrenzen

Bei der nachträglichen Korrektur stellte es sich als vorteilhaft heraus, zuerst jeweils nur über 4 Klassen zu unterscheiden. Es sind dies die Akzentklassen [0], [1], [2], [3,4]. Folgende Korrekturregeln wurden angewandt:

**1.** Pro Phrase muss es genau einen Phrasenhauptakzent geben. Gibt es in einer Phrase keinen Phrasenhauptakzent, wird dieser auf [1] gesetzt wo die Summe der Wahrscheinlichkeit der Zugehörigkeit von [1] und [2] am grössten ist, in der jeweiligen Phrase. Gibt es in einer Phrase mehr als einen Phrasenhauptakzent wird nur der letzte, der als [1] klassifizierten auf [1] gelassen. Die übrigen werden auf [2] gesetzt. Dies kommt daher, dass in der Deutschen Sprache die meisten Sätze endbetont sind.

**2.** Pro Wort gibt es höchstens einen Akzent grösser [4]. Wenn es mehrere solche gibt, werden alle bis auf den ersten im Wort oder den als Phrasenhauptakzenten erkannten auf [4] gesetzt.

### 4.5.4 Feature zum Phrasentyp

Bei all diesen Varianten konnte zusätzlich auch noch vier Features zum Phrasentyp dazu gegeben werden. Die fünf Phrasentypen (siehe Abschnitt 1.2) wurden folgendermassen codiert: 1000 ist (P), 0100 ist (S), 0010 ist (T), 0001 ist (E) und 0000 ist (Y).

# 5 Modellansätze

Für die Einteilung der Silben in 5 verschiedene Akzentklassen, wurden 3 verschiedene Modelle implementiert, diese werden im folgenden beschrieben.

## 5.1 Normierung

Je nach Modell wurde eine andere Normierung der Daten gewählt. Bei der Verwendung von neuronalen Netzen (Abschnitte 5.2 und 5.3 ) hat sich eine lineare Normierung mit Mittelwert 0 und Varianz 1 bewährt.

Für das lineare Modell (Abschnitt 5.4) konnte diese nicht verwendet werden, weil sich sonst die negativen und die positiven Werte der verschiedenen Features gegenseitig aufheben würden. Daher wurden hier alle Features auf 0-1 normiert. Das heisst, zuerst wurden alle Daten so verschoben, dass der minimale Wert 0 wird und anschliessend werden sie durch den neuen Maximalwert geteilt.

## 5.2 1 stufiges Modell mit neuronalem Netz (1 stufiges MLP)

Dieses Modell verwendet ein neuronales Netz mit jeweils so vielen Inputs wie Features verwendet werden. Einen Hiddenlayer mit 2-7 Knoten und je nach Verwendung Variante der Phrasen und Wortgrenzen (siehe Abschnitt 4.5) 4 oder 5 Outputs.

## 5.3 2 stufiges Modell mit zwei neuronalen Netzen (2 stufiges MLP)

Beim 2 stufigen Modell entscheidet ein neuronales Netz mit nur einem Ausgang ob die Silben betont oder unbetont sind. Anschliessend wird nur mit den betonten Silben ein zweites Modell trainiert welches die 4 Akzentklassen unterscheidet. Für die zwei Stufen wurden zwei verschiedene Featuresubsets verwendet. In beiden Stufen werden neuronale Netze mit einem Hiddenlayer mit jeweils 2-7 Hiddenknoten verwendet.

## 5.4 Lineares Modell

Wie auch in [Pfi06] beschrieben kann auch ein einfaches lineares Modell verwendet werden. Dieses Modell verwendet folgenden Ansatz:

$$Y = Cq \tag{2}$$

Y ist dabei eine 5 spaltige Matrix mit den Wahrscheinlichkeiten, dass eine Silbe einer der 5 Klassen angehört. C ist die Featurematrix, welche für jedes y eine Zeile mit den zur jeweiligen Silbe gehörigen Features hat, q enthält die Gewichte. Die Gewichte können wie folgt berechnet werden.

$$q = (C^t C)^{-1} C^t Y \tag{3}$$

Wobei hier ein Trainingsset mit bekanntem Y verwendet wird. Y enthält hier keine Wahrscheinlichkeiten sondern binäre Werte ob zur Akzentklasse gehörig oder nicht. Da mehr Trainingsdaten als Features verwendet werden, berechnet Matlab hier eine Optimierung auf das Trainingsset.

# 6 Resultate

## 6.1 Featuresubsets zu den 3 Modellen

In diesen Abschnitt werden die für die folgenden Tests verwendeten Featuresubsets aufgelistet. Wenn man die Featureelimination mit verschiedenen Modellen durchführt, ergeben alle ein anderes Featuresubset. Daher wird hier für jede Modell das verwendete Featuresubset aufgelistet.

Dabei werden die in Kapitel 4 verwendeten Featurebezeichnungen verwendet plus eine Angabe zum Kontext. D1[-1] bedeutet beispielsweise das Feature D1 der vorhergenden Silbe, 0 bedeutet die aktuelle Silbe und +1 steht für die nachfolgende Silbe.

### 6.1.1 Verwendetes Featuresubset im 1 stufigen MLP

Dauerfeatures: D4[-1,0], D5[0]
Energiefeatures: E1[-1], E4[+1], E6[+1]
Grundfrequenzfeatures: F0_1[0,+1], F0_2[-1,0], F0_3[-1], F0_4[-1,0]
Phonologische Features: Ph1[0,+1], Ph2[0,+1], Ph3[-1], Ph4[-1,0], Ph5[0], Ph6[0], Ph7[+1]

Dieses Modell verwendet mit Abstand am wenigsten Features. Vor allem die Features zur Grundfrequenz und die zur Phonologischen Repräsentation sind sehr gut vertreten.

### 6.1.2 Verwendete Featuresubsets im 2 stufigen MLP

**Featuresubset der ersten Stufe:**
Dauerfeatures: D1[-1,0,+1], D2[-1,0,+1], D3[-1,0,+1], D4[-1,0,+1], D5[0,+1]
Energiefeatures: E1[-1,0,+1], E2[-1,0,+1], E3[-1,0,+1], E4[0,+1], E5[-1,+1], E6[-1,0,+1]
Grundfrequenzfeatures: F0_1[-1,0,+1], F0_2[-1,0,+1], F0_3[-1,0,+1], F0_4[-1,0,+1]
Phonologische Features: Ph1[-1,+1], Ph2[0,+1], Ph3[0], Ph5[0,+1], Ph6[0], Ph7[0,+1]

**Featuresubset der zweiten Stufe:**
Dauerfeatures: D1[0,+1], D2[-1,0,+1], D3[-1,0], D4[-1,0], D5[-1]
Energiefeatures: E1[-,+1], E2[0,+1], E3[+1], E4[-1,0], E5[-1]
Grundfrequenzfeatures: F0_1[-1,+1], F0_2[-1,0,+1], F0_3[-1,+1], F0_4[+1]
Phonologische Features: Ph1[-1,0,+1], Ph2[-1,0], Ph3[-1,0], Ph4[-1,0], Ph5[-1,0,+1], Ph6[-1]

Dieses Modell verwendet für die zwei Stufen verschiedene Features. Beide Stufen verwenden sehr viele Features zur Dauer und zur Energie. Die erste Stufe verwendet zudem alle Grundfrequenzfeatures. Bei den phonologischen Features fällt auf, dass für die erste Stufe tendenziell eher die phonologische Information zur aktuellen und zur nachfolgenden Silbe wichtig scheint bei der zweiten Stufe aber eher die phonologische Information zur vorhergenden Silbe und ebenfalls zur aktuellen.

### 6.1.3 Verwendetes Featuresubset des linearen Modells

Dauerfeatures: D1[0], D2[-1,0], D3[+1], D4[-1,0], D5[-1]
Energiefeatures: E1[-1,0,+1], E2[-1,+1], E3[0,+1], E4[+1], E5[-1,0], E6[-1,0,+1]
Grundfrequenzfeatures: F0_1[-1,0,+1], F0_2[-1,0,+1], F0_3[-1,0,-1], F0_4[0,+1]
Phonologische Features: Ph1[-1,0,+1], Ph2[0,+1], Ph3[0,+1], Ph4[-1,0,+1], Ph5[0], Ph6[-1,0,+1], Ph7[-1,0,+1]

Das lineare Modell verwendet sehr viele Features, es ist schwierig zu sagen welche hier wichtig sind.

Über alle Modelle hinweg gesehen, fallen einige Features als wichtiger auf als andere. Die Pausendauer (D4) zum Beispiel wird von allen Modellen verwendet und auch immer über 2-3 Silben. Auch die Features zur Grundfrequenz scheinen sehr wichtig, dabei wird bei der vorhergehenden Silbe vor allem Nucleus und Coda verwendet. Bei der aktuellen Silbe vor allem Onset und Nucleus und bei der nachfolgenden vor allem der Nucleus. Bei den Features zur spektralen Emphasis ist wenig System festzustellen, sie werden zwar oft verwendet aber erstaunlicherweise immer wieder andere Frequenzbänder. Bei den phonologischen Features sind vorallem die zur aktuellen Silbe bei allen Modellen sehr beliebt.

## 6.2 Vergleich der 3 Modelle innerhalb eines Korpus

Zuerst wurden alle 3 im Kapitel 5 beschriebenen Modelle mit dem gleichen Korpus getestet wie sie trainiert wurden. Alle 3 Modelle wurden jeweils ohne Verwendung von Phrasen und Wortgrenzen sowie mit deren Zugabe in allen 3 Varianten (Abschnitt 4.5) getestet.

Tab. 7 zeigt die Klassifizierungraten der 3 Modelle mit den 3 Verwendungen der Phrasen und Wortgrenzen sowie der Phrasentypen. Man sieht deutlich, dass beim linearen Netz die Zugabe von Wort und Phrasengrenze weniger bringt als bei den neuronalen Netzen. Dieser Zusammenhang scheint für das lineare Netz etwas zu kompliziert zu sein. Die Nachkorrektur bringt bei allen Modellen etwa gleichviel. Die Zugabe vom Phrasentypen scheint bei allen Modellen nur von kleinem Nutzen zu sein, falls überhaupt. Das lineare Modell schneidet hier vorallem bei mit der Nachkorrektur sehr gut ab. Das scheint aber nur so, weil die Features genau für dieses Testset und dieses Trainingsset ausgewählt wurden. Verwendet man ein etwas anderes Testset wird das Resultat schnell wieder schlechter.

| Modell | Lineares Netz | 1 stufiges MLP | 2 stufiges MLP |
|---|---:|---:|---:|
| Ohne Wort und Phrasengrenzen (OGR) | 73.70% | 73.81% | 73.78% |
| Phr. Features Variante 1 (PhV1) | 75.90% | 76.02% | 76.28% |
| Phr. Features Variante 2 (PhV2) | 77.19% | 78.60% | 78.63% |
| Mit Nachkorrektur (NK) | 81.02% | 80.11% | 79.81% |
| NK mit Phrasentyp (NK_PhT) | 80.83% | 79.85% | 79.58% |
| PhV2 mit Phrasentyp (PhV2_PhT) | 77.27% | 79.43% | 78.86% |

**Tabelle 7:** *Erfolgsrate über alle 5 Klassen im männlichen Prosodie Korpus*

Um diese Erfolgsraten etwas genauer aufzuschlüsseln, ist in Tab. 8 für zwei Modelle noch

die Erfolgsrate nach Akzentklassen aufgespalten. Die Prozentwerte geben hier an wieviele der vorhandenen Akzente in einer Akzentklasse richtig klassifiziert wurden. Es fällt auf, dass bei der Verwendung der Nachkorrektur die Erfolgsraten etwas gleichmässiger sind als wenn man die Phrasengrenzen als Features verwendet. Aber beide Varianten haben vor allem mit den Akzentklassen [2] und [3] grosse Mühe, diese sind aber auch am seltensten.

| Akzentklassen | 2 stufiges MLP mit NK | 2 stufiges MLP PhV2 |
|---|---:|---:|
| [0] | 88.97% | 89.22% |
| [1] | 80.00% | 78.52% |
| [2] | 53.96% | 57.74% |
| [3] | 36.59% | 18.69% |
| [4] | 69.63% | 62.53% |

**Tabelle 8:** *Erfolgreiche Klassifizierungsrate nach Akzentklassen aufgespaltet*

Tab. 9 zeigt die Resultate der gleichen Tests wie Tab.7 im weiblichen Prosodie Korpus. Es fällt auf, dass die Erfolgsrate deutlich schlechter ist. Das kann zum einen daran liegen, dass die Features mit dem männlichen Prosodie Korpus eliminiert wurden, zum anderen aber auch daran, dass es im weiblichen Prosodie Korpus mehr verschieden Phrasentypen gibt. Daher wird in Tab.10 für das 2 stufige Modell auch noch die erfolgreiche Klassifizierung in den verschiedenen Phrasentypen aufgelistet. Diese Tests wurden nur für ein Modell gemacht. Es ist aber anzunehmen, dass die Verteilungen in den anderen Modellen ähnlich sind. Es fällt deutlich auf, dass die Klassifizierung in progredienten Phrasen am besten funktioniert. Das erstaunt nicht sehr, da die progredienten erstens am häufigsten sind und zweitens die Merkmale zur Grundfrequenzbewegung hier deutlich ausgeprägt sind. Hier sieht man auch deutlich, dass die Nachkorrektur nur bei den progredienten Phrasen wirklich gut funktioniert bei den anderen aber nicht. Das ist auch nicht weiter verwunderlich, da die Nachkorrektur mit den Phrasenhauptakzenten am Phrasenende auch hauptsächlich auf diese abgestimmt ist. Wenn man die Phrasen und Wortgrenzen als Features verwendet, werden auch die terminalen Phrasen ziemlichen gut, diese treten am zweit häufigsten auf. Die zusätzliche Verwendung des Phrasentyps gibt ausser bei den (E)-Phrasen nur kleine Verbesserungen. Bei den (E)-Phrasen gibt es jedoch einen sehr starken Nutzen, diese kommen aber sehr selten vor und werden somit vermutlich erst durch die Verwendung des Phrasentyps angemessen gewichtet.

| Modell | Lineares Netz | 1 stufiges MLP | 2 stufiges MLP |
|---|---:|---:|---:|
| Ohne Wort und Phrasengrenzen (OGR) | 75.20% | 75.51% | 73.66% |
| Phr. Features Variante 1 (PhV1) | 75.42% | 77.18% | 75.42% |
| Phr. Features Variante 2 (PhV2) | 78.86% | 78.50% | 77.53% |
| Mit Nachkorrektur (NK) | 76.30% | 76.87% | 75.55% |
| NK und Phrasentyp (NK_PhT) | 76.48% | 77.05% | 75.99% |
| PhV2 Phrasentyp (PhV2_PhT) | 75.95% | 78.59% | 77.71% |

**Tabelle 9:** *Klassifizierungsrate über alle 5 Klassen im weiblichen Prosodie Korpus*

| Phrasentyp | OGR | NK | PhV2 | PhV2_PhT |
|---|---|---|---|---|
| P | 76.27% | 80.00% | 79.82% | 80.17% |
| T | 70.42% | 71.83% | 76.66% | 77.26% |
| S | 74.49% | 74.83% | 75.85% | 77.21% |
| Y | 69.58% | 68.28% | 71.19% | 72.11% |
| E | 66.66% | 60.00% | 64.44% | 71.11% |

**Tabelle 10:** *Erfolgsrate nach Phrasentyp im weiblichen Prosodie Korpus des 2 stufigen Modells*

## 6.3 Übertragung und Adaption vom männlichen auf den weiblichen Prosodie Korpus

Nun wurden die mit dem männlichen Prosodie Korpus trainierten Modelle auf dem weiblichen Prosodie Korpus getestet. Es ergab sich sich als vorteilhaft die Testdaten auch hier immer auf die Trainingsdaten zu normieren. Egal wie gross das Testset war, die Normierung auf das Trainigsset wird immer mindestens gleich gut, wie eine separate Normierung. Somit muss man dann auch nicht unterscheiden ob man das Klassifizierungsmodell auf einen ganzen Korpus oder bloss auf einen Satz anwendet.

Tab.11 zeigt die Resultate der Tests mit Training mit männlichen Prosodie Korpus und Testen im weiblichen.

| Modell | Lineares Netz | 1 stufiges MLP | 2 stufiges MLP |
|---|---|---|---|
| OGR | 66.87% | 68.41% | 66.43% |
| PhV2 | 69.07% | 67.75% | 70.39% |
| NK | 66.47% | 69.25% | 71.98% |

**Tabelle 11:** *Klassifizierungsrate über alle 5 Klassen der übertragenen Modelle*

Es wurde auch versucht die trainierten Modelle mit wenig Daten aus dem neuen Korpus zu adaptieren. Dazu gibt man dem trainierten Modell einige Sätze aus dem neuen Korpus hinzu. Diese werden mit dem Verhältnis 6:4 in ein Trainingsset und ein Evaluationsset eingeteilt. Nun wird das Modell mit den neuen Trainingsdaten nochmals solange trainiert bis es auf dem Evaluationsset nicht mehr besser wird. Es wurden Tests mit 20, 50 und 100 Sätzen gemacht.

Tab. 12 zeigt die Adaption des 2 stufigen Modells mit den unterschiedlichen Anzahlen an Adaptionsdaten.

## 6.4 Übertragung und Adaption vom deutschen auf den französischen Prosodie Korpus

Zum Schluss wurden die auf dem deutschen weiblichen Korpus trainierten Modelle noch einem französischen Testset der gleichen Sprecherin getestet und anschliessend wurde noch eine Adaption mit 20 Sätzen durchgeführt.

| Anzahl Adaptionsdaten | 2 stufiges MLP OGR | 2 stufiges MLP mit NK |
|---|---:|---:|
| 20 Sätze | 68.28% | 73.12% |
| 50 Sätze | 70.52% | 72.42% |
| 100 Sätze | 73.34% | 74.58% |

**Tabelle 12:** *Klassifizierungsrate über alle 5 Klassen der adaptierten von der männlichen auf die weibliche Stimme*

Die Resultate sind in Tab. 13 und 14 dargestellt. Man sieht deutlich, dass beim Übetragen auf eine andere Sprache die Adaption von viel stärkerem Nutzen ist als beim Übertragen auf eine andere Stimme. Ohne Adaption funktionieren die übertragenen Modelle hier deutlich am schlechtesten (zumindest ohne Nachkorrektur), aber mit einer Adaption mit nur 20 Sätzen können schon um die 6% gewonnen werden. Es fällt auf, dass sie Nachkorrektur nach Phrasengrenzen, welche eigentlich für die deutsche Sprache geschrieben auch für Französisch sehr nützlich ist.

| Modell | Lineares Netz | 1 stufiges MLP | 2 stufiges MLP |
|---|---:|---:|---:|
| Ohne Wort und Phrasengrenzen | 63.86% | 65.26% | 63.16% |
| Phr. Features Variante 2 | 63.50% | 62.80% | 63.16% |
| Mit Nachkorrektur | 70.82% | 69.47% | 70.53% |

**Tabelle 13:** *Klassifizierungsrate über alle 5 Klassen der übertragenen Modelle*

| Anzahl Adaptionsdaten | 2 stufiges MLP OGR | 2 stufiges MLP mit NK |
|---|---:|---:|
| 20 Sätze | 69.82% | 76.49% |

**Tabelle 14:** *Klassifizierungsrate über alle 5 Klassen der adaptierten Modelle*

# 7 Diskussion und Ausblick

Auf 3 Prosodie Korpora wurden 3 Modelle zur Detektion von unbetonten und betonten Silben sowie zu deren Klassifizierung in 4 Akzentklassen getestet. Für ihre Berechnungen verwenden die Modelle Features aus Dauer, Signalenergie, Grundfrequenz und aus der Phonologischen Repräsentation. Ein zusätzliche Verwendung von Informationen zu Phrasen und Wortgrenzen ergab sich als nützlich. Erreicht man ohne Verwendung dieser Information eine erfolgreiche Klassifizierungsrate zwischen 73% und 75% so erreicht man mit etwa 77-80%, je nachdem wie der Korpus aufgebaut ist.

Versuche zum Training auf einem deutschen männlichen Prosodie Korpus und Testen auf einem deutschen weiblichen, zeigen je nach Modell eine erfolgreiche Klassifierungsrate zwischen 66% und 72%. Durch eine Adaption des vortrainierten neuronalen Netzes kann auf dem einen Modell eine Steigerung von 72 auf 74% erreicht werden. Die Adaption funktioniert also, ist aber nicht sehr wirksam und braucht recht viele Adaptionsdaten. Beim Training auf dem selben Korpus erreicht, das selbe Modell knappe 76%.

Versuche zum Training auf dem deutschen weiblichen Prosodie Korpus und Testen auf dem französischen Korpus der gleichen Sprecherin, ergaben zuerst eine starke Verschlechterung von ca. 75% auf ca. 63%. Allerdings zeigte hier eine Adaption mit nur 20 Sätzen aus dem französischen Korpus eine starke Wirkung, somit konnte wieder eine erfolgreiche Klassifizierung von 70% erreicht werden, mit der Verwendung der Nachkorrektur erreicht man sogar 76%. Eine Adaption innerhalb der gleichen Stimme aber auf eine anderen Sprache scheint für das Modell also viel leichter zu lernen zu sein, als eine Adaption innerhalb der gleichen Sprache auf eine andere Stimme.

Ursprünglich wurde zwar eine etwas höhere Erfolgsrate bei der Klassifizierung erhofft. Wenn man berechnet wieviel richtig klassifiziert ist, wenn alles unbetont ist (je nach Korpus 54-61%), konnte aber immerhin gezeigt werden, dass die Modelle doch recht viele Zusammenhänge lernen und auch eine Übertragung vom einen auf den anderen Korpus funktioniert zu mindest teilweise. Wenn man in [Pic96] liest, dass die bei manueller Annotation lediglich eine Übereinstimmung von 80% beim betont/unbetont Entscheid erreicht haben, könnte man das Resultat sogar als erstaunlich betrachten, allerdings fragt es sich wie genau die sich abgesprochen haben wie eine betonte Silbe definiert ist.

Bei diesen Klassifizierungsraten kann ich wohl nicht empfehlen die Akzenklassen vollständig automatisch durchzuführen. Ein Menschlicher Annotierer kann sich aber viel Arbeit ersparen, indem er die automatische Annotation durchführt und anschliessend nur noch von Hand kontrolliert ob die Resultate Sinn ergeben. Wenn man zusätzlich zur gefundenen Klasse für jede Silbe auch noch die Wahrscheinlichkeit aller Klassen betrachtet, kann man die zweit wahrscheinlichste Klasse auch gerade als alternativ Vorschlag verwenden wodurch auch die Korrektur weniger aufwändig wird.

In zukünftigen Projekten könnte man noch untersuchen, ob die Modelle besser werden, wenn die verschiedenen Akzentklassen gleichmässiger über die Trainingsdaten verteilt werden. Auch eine gleichmässigere Verteilung der Phrasentypen könnte helfen.

# Literatur

[Bis95] CH. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[Pfi06] R. Pfister & B. Beutler. *Sprachverarbeitung I ,Vorlesungsskript für das Wintersemester 2006/2007*. Department ITET, ETH Zürich, 2006.

[Pic96] G. Pickering, B. Williams & B. Knowles. Analysis of transcriber differences in SEC". *In Knowles, G., Wichmann, A. & Alderson, P. (eds), Working with speech*, London: Longman 1996.

[Tra95] Ch. Traber. *SVOX: The Implementation of a Text-to-Speech-System for German*. vdf, 1995.

# Anhang A

Institut für
Technische Informatik und
Kommunikationsnetze

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo

Sommersemester 2007
(SA-2007-41)

Semesterarbeitsaufgabenstellung

für

Frau Cécile Bucher

Betreuer: H. Romsdorfer ETZ D97.5

Ausgabe: 19. März 2007
Abgabe: 22. Juni 2007

## Erkennung von Silbenakzenten aus Sprachsignalen

### Einleitung

Die Prosodie (Grundfrequenzkontur und Lautdauersequenz) aktueller Sprachsynthesesysteme wird zumeist mittels statistischer Modelle erzeugt. Um solche Modelle trainieren zu können, braucht man korrekt annotierte Trainingssätze. Für die Prosodiesteuerung sind dabei vor allem die korrekte Position und Stärke der Satzakzente und der Phrasengrenzen wichtig. Da die Prosodie der aufgenommenen Trainingssätze häufig von der vom Text abgeleiteten Standardakzentuierung und -phrasierung abweicht, wird deren Annotation in einem sehr zeitaufwändigen Schritt manuell korrigiert. Deshalb wäre es wünschenswert, diese Abweichungen von der Standardakzentuierung und -phrasierung automatisch erfassen und die Annotation entsprechend korrigieren zu können.

### Aufgabenstellung

In dieser Arbeit sollen unter Verwendung eines bestehenden, von Hand korrigierten deutschen Satzkorpus verschiedene, in der Literatur beschriebene Verfahren zur automatischen Silbenakzentbestimmung untersucht und miteinander verglichen werden (siehe zum Bei-

spiel [1, 2, 3, 4, 5, 6]).

Die Annotation eines Trainingssatzes basiert auf einer abstrakten Beschreibung der phonologischen Eigenschaften dieses Satzes, der so genannten phonologischen Repräsentation. Diese phonologische Repräsentation beinhaltet neben der phonetischen Transkription der Lautsequenz Informationen bezüglich Sprache, Betonung, Phrasierung und Silbeneinteilung des Satzes. Als Beispiel sei die phonologische Repräsentation des Satzes "Friedliche Massenkundgebung in Peking." angegeben:

```
#{P:0} fr[2]i:t-lI-C@- m[1]a-s@n-k[4]Un-ke:-bUN #{T:2} ?In- p[1]e:-kIN .
```

In der phonologischen Repräsentation können folgende Spezialsymbole auftreten:

`#{X:n}` markiert eine Phrasengrenze, wobei `n=0` eine Satzgrenze, `n=1` eine satz-interne Phrasengrenze mit Pause, und `n>1` eine satz-interne Phrasengrenze ohne Pause kennzeichnet. `X` gibt den Typ der auf die Phrasengrenze folgenden Phrase an, wobei `X=P` eine progrediente Phrase und `X=T` eine terminale Phrase bezeichnen. '`.`' markiert optional das Satzende.

`\X\` kennzeichnet einen Sprachwechsel. So wechselt die Sprache beispielsweise mit '`\E\`' nach Englisch, mit '`\F\`' nach Französisch und mit '`\G\`' nach Deutsch.

`-` kennzeichnet eine Silbengrenze. Da eine Phrasengrenze zugleich eine Silbengrenze ist, ist die Markierung '`-`' direkt vor einer Phrasengrenze optional.

`[n]` markiert einen Satzakzent. '`[1]`' kennzeichnet den Phrasenhauptakzent, '`[2]`' einen "Pitch Accent" (das ist ein Akzent mit einer starken Grundfrequenzbewegung), '`[3]`' markiert einen "Non-Pitch Accent" auf der Worthauptakzentposition, und '`[4]`' einen Wortnebenakzent. '`[E]`' markiert einen emphatischen Akzent. Unbetonte Silben können optional mit '`[0]`' gekennzeichnet werden.

Grundsätzlich kann in dieser Arbeit bei der Detektion der Satzakzente und der Phrasengrenzen davon ausgegangen werden, dass die Lautsegmentierung und die phonologische Repräsentation des Satzes bereits vorliegen. Dies erfordert aber im Gegensatz zu den meisten in der Literatur behandelten Verfahren einen etwas modifizierten Ansatz. Mit Hilfe dieser Zusatzinformationen sollte jedoch auch eine höhere Erkennungsleistung erzielbar sein.

Die folgenden Aufgaben stellen sich im Rahmen dieser Semesterarbeit:

1. Einarbeitung in die Literatur zu Detektion von Satzakzenten, z.B. [1, 2, 3, 4, 5, 6, 7].
2. Einarbeitung in die Literatur zu Mustererkennungsalgorithmen, z.B. [8, 9].
3. Ermittlung der optimalen Feature-Kombination für die Detektion der Satzakzente sowohl mit als auch ohne Berücksichtigung der Lautsegmentierung und der phonologischen Repräsentation.
4. Grundsätzliche Durchführbarkeitstests der Algorithmen anhand des manuell korrigierten, deutschen Prosodiekorpus eines männlichen Sprechers.

5. Tests zur Übertragbarkeit der Algorithmen auf eine Frauenstimme anhand eines zweiten deutschen Prosodiekorpus.

6. Tests zur Übertragbarkeit der Algorithmen auf andere Sprachen anhand eines französischen Prosodiekorpus von derselben Sprecherin.

Die ausgeführten Arbeiten und die erhaltenen Resultate sind in einem Bericht zu dokumentieren (siehe dazu [10]), der in gedruckter und in elektronischer Form abzugeben ist. Zusätzlich sind im Rahmen eines Kolloquiums zwei Präsentationen vorgesehen: etwa drei Wochen nach Beginn soll der Arbeitsplan und am Ende der Arbeit die Resultate vorgestellt werden. Die Termine werden später bekannt gegeben.

## Literaturverzeichnis


[1] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. Detection of phrase boundaries and accents. 1994.

[2] F. Tamburini. Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system. In *Proceedings of Eurospeech'03*, Geneva, Switzerland, September 2003.

[3] A. Batliner, A. Feldhaus, S. Geißler, A. Kießling, T. Kiss, R. Kompe, and E. Nöth. Integrating syntactic and prosodic information for the efficient detection of empty categories. volume 1, pages 71–76, 1996.

[4] V. Strom. Detection of accents, phrase boundaries and sentence modality in german with prosodic features. In *Proceedings of Eurospeech'95*, volume 3, pages 2039–2041, Madrid, 1995.

[5] V. Strom and C. Widera. What's in the "pure" prosody? In *Proceedings of ICSLP'96*, volume 3, pages 1497–1500, Philadelphia, PA, 1996.

[6] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji. Modeling and automatic detection of english sentence stress for computer-assisted english prosody learning system. In *Proceedings of ICSLP 2002*, pages 749–752, Denver, Colorado, USA, September 2002.

[7] B. Pfister und R. Beutler. *Sprachverarbeitung I.* Vorlesungsskript für das Wintersemester 2005/2006, Departement ITET, ETH Zürich, 2005.

[8] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

[9] I. T. Nabney. *Netlab. Algorithms for Pattern Recognition.* Springer, London, 2002.

[10] B. Pfister. *Richtlinien für das Verfassen des Berichtes zu einer Semester- oder Diplomarbeit.* Institut TIK, ETH Zürich, März 2004.
(http://www.tik.ee.ethz.ch/~spr/SADA/richtlinien_bericht.pdf).

[11] B. Pfister. *Hinweise für die Präsentation der Semester- oder Diplomarbeit*. Institut TIK, ETH Zürich, März 2004.
(http://www.tik.ee.ethz.ch/~spr/SADA/hinweise_praesentation.pdf).

Zürich, den 19. März 2007