Erkennung von Phrasengrenzen und Phrasentyp aus Sprachsignalen

Juri Baumberger und Jonas Sonnenmoser

Semesterarbeit SA-2007-49

Sommersemester 2007

Institut für Technische Informatik
und Kommunikationsnetze

Betreuer: H. Romsdorfer

Verantwortlicher: Prof. Dr. L. Thiele

Zusammenfassung

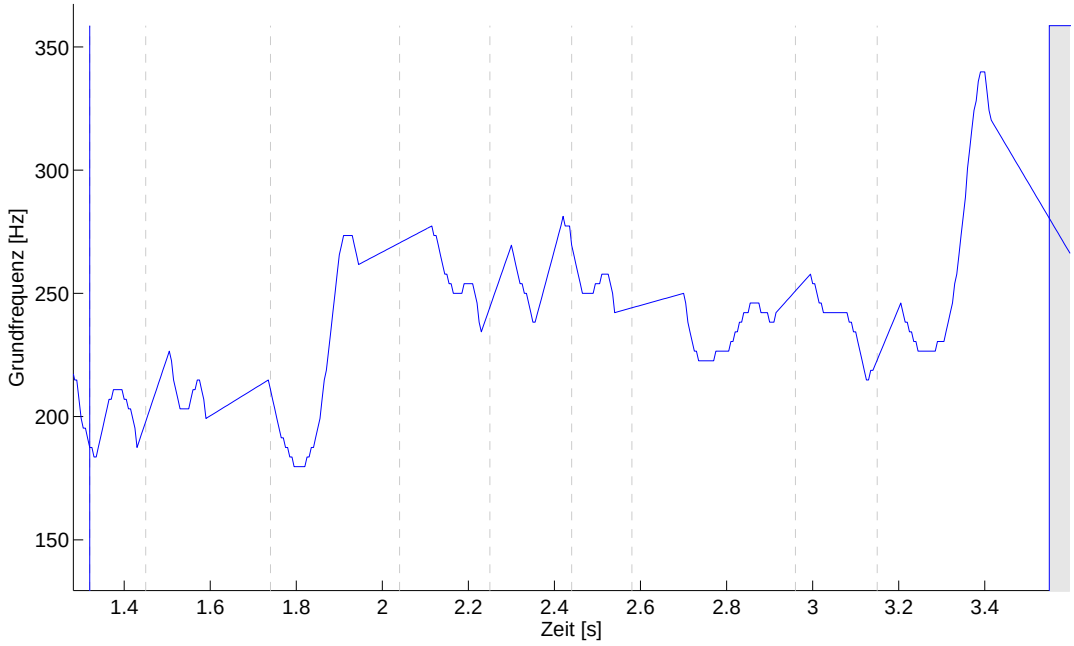
In aktuellen Sprachsynthese-Systemen ist die Steuerung der Prosodie eine zentrale Aufgabe. Die dafür verwendeten statistischen Modelle, erfordern eine grosse Menge an Trainingsdaten. In diesen müssen nun zuerst u.A. Phrasengrenzen und Phrasentypen korrekt eingezeichnet werden. Dies wird bislang meist von Hand gemacht, was eine mühsame und zeitraubende Angelegenheit ist.

In dieser Arbeit haben wir Modelle — insbesondere neuronale Netze — zur automatischen Detektion von prosodischen Phrasengrenzen und Phrasentypen auf ihre Klassifizierungsleistung getestet. Der Fokus der Arbeit lag darauf, geeignete Features aus dem Sprachsignal, dessen phonologischer Beschreibung und dem entsprechenden Grundfrequenzverlauf zu extrahieren und des weiteren die Adaption der Modelle auf unterschiedliche Sprecher und Sprachen zu testen. Die Detektion von Phrasengrenzen und Phrasentypen wurde dabei getrennt behandelt.

Inhaltsverzeichnis

Figurenverzeichnis

Tabellenverzeichnis



Figur 1: *Beispiel des Grundfrequenzverlaufs einer progredienten Phrase: ('...die fast gleichzeitig entstanden sind,...')*

1 Einleitung

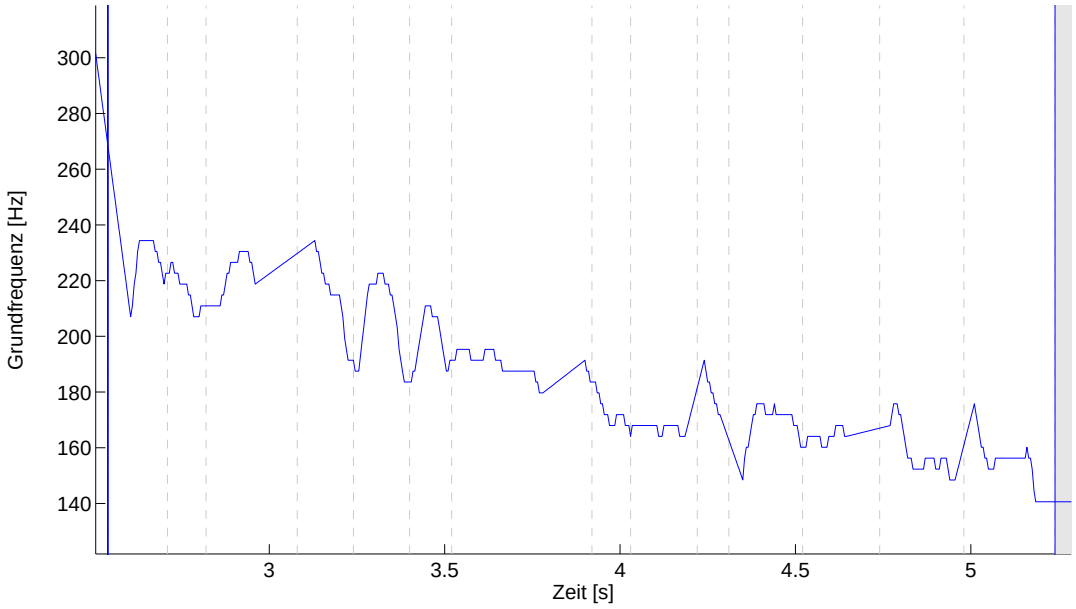
In diesem Kapitel werden wir die Klassifizierung der Phrasengrenzen und Phrasentypen, die phonologische Repräsentation und Ansätze zur Detektion erläutern.

1.1 Phrasentypen

Grundsätzlich lassen sich prosodische Phrasen in eine Vielzahl von Klassen unterteilen. Da wir nur eine relativ begrenzte Menge von Trainingsdaten zur Verfügung hatten, haben wir uns auf die folgenden 5 Phrasentypen beschränkt, die im folgenden grob beschrieben werden:
Progredient (P), Terminal (T), Ja-Nein-Frage (Y), Exclamation (E) und Statement (S).
Wir werden im Rest des Abschnitts grob erklären, wodurch sich diese auszeichnen und unterscheiden.

- **Progrediente Phrase (P)**: Progrediente Phrasen zeichnen sich durch einen generell steigenden Grundfrequenzverlauf aus. Sie sind Phrasen, welche auf eine weitere Aussage hinführen. Ein Beispiel eines typischen Grundfrequenzverlaufs ist gegeben in Fig. 1

- **Terminale Phrase (T)**: Terminale Phrasen zeichnen sich tendenziell durch einen linear fallenden Grundfrequenzverlauf aus. Sie sind Phrasen, welche eine zuvor angefangene Aussage abschliessen. Ein Beispiel eines typischen Grundfrequenzverlaufs ist gegeben in Fig. 2.



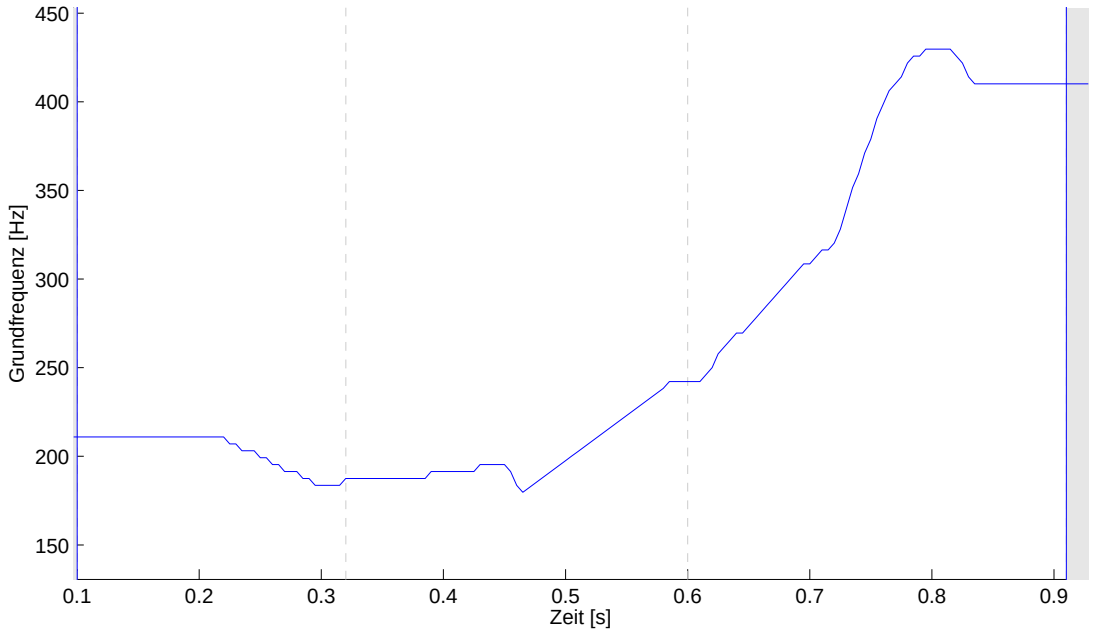
Figur 2: *Beispiel des Grundfrequenzverlaufs einer terminalen Phrase: (’...oder möchten sie den Nachnamen geändert haben.’)*

- **Ja-Nein-Frage (Y)**: Ja-Nein-Fragen zeichnen sich dadurch aus, dass sie gegen Ende einen steil ansteigenden Grundfrequenzverlauf aufweisen und in ihrer Aussage den Ja-Nein-Fragecharakter enthalten. Ein Beispiel eines typischen Grundfrequenzverlaufs ist gegeben in Fig. 3.
- **Exclamation (E)**: Eine Exclamation-Phrase zeichnet sich häufig durch Befehlscharakter (oft Imperativ), hohe Grundfrequenz und Lautstärke aus. Ein Beispiel eines typischen Grundfrequenzverlaufs ist gegeben in Fig. 4.
- **Statement (S)**: Eine Statement-Phrase zeichnet sich dadurch aus, dass ihre Aussage i.A. alleine stehen kann, Zusätzlich weist der Grundfrequenzverlauf eine erhöhte Varianz auf. Mit der T-Phrase gemeinsam ist der generell leicht abfallende Grundfrequenzverlauf. Ein Beispiel eines typischen Grundfrequenzverlaufs ist gegeben in Fig. 5.

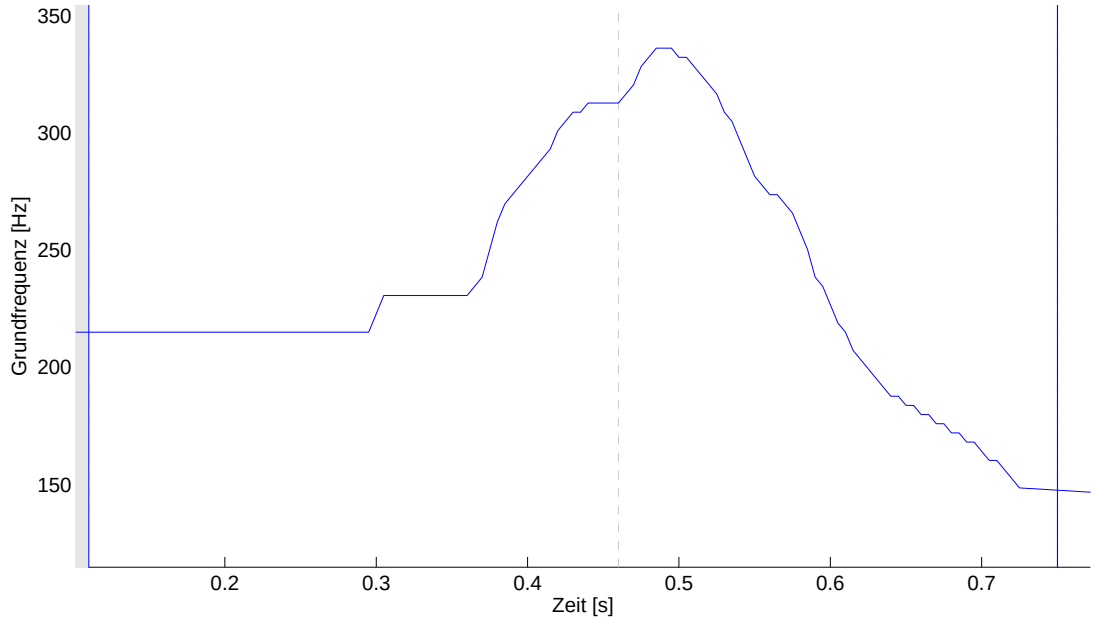
1.2 Phrasengrenzen

Um Phrasengrenzen klassifizieren zu können, haben wir uns für folgende 4 Typen entschieden: Keine Grenze, schwache Grenze (’/ ’), starke Grenze (’// ’), Satzgrenze (’///’). Es wird jede Silbengrenze untersucht und entschieden, um welche der 4 Klassen es sich handelt. Die Klassen im Detail:

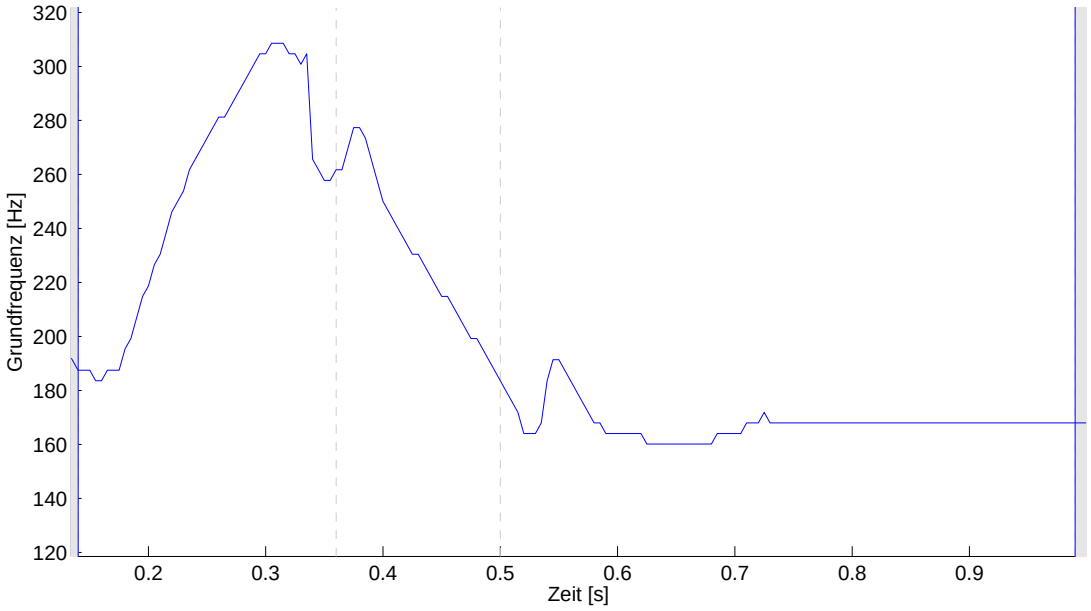
- **Satzgrenze (’///’)**: Eine Satzgrenze bezeichnet das Ende eines Satzes. Da in den verwendeten Sprachkorpora das Ende der Aufnahme immer auch das Satzende bedeutet, ist die



Figur 3: *Beispiel des Grundfrequenzverlaufs einer Ja-Nein-Frage: (’Vermittlungen?’)*



Figur 4: *Beispiel des Grundfrequenzverlaufs einer Exclamation-Phrase: (’Schlimmer!’)*



Figur 5: *Beispiel des Grundfrequenzverlaufs einer Statement-Phrase: (’Dübendorf.’)*

Detektion simpel.

- **Starke Grenze (’// ’)**: Eine starke Phrasengrenze zeichnet sich dadurch aus, dass der Sprecher an dieser Stelle eine Pause macht (meist 100-200 ms lang). Da in den verwendeten Sprachkorpora die Pausen bereits annotiert sind, ist die Detektion simpel.

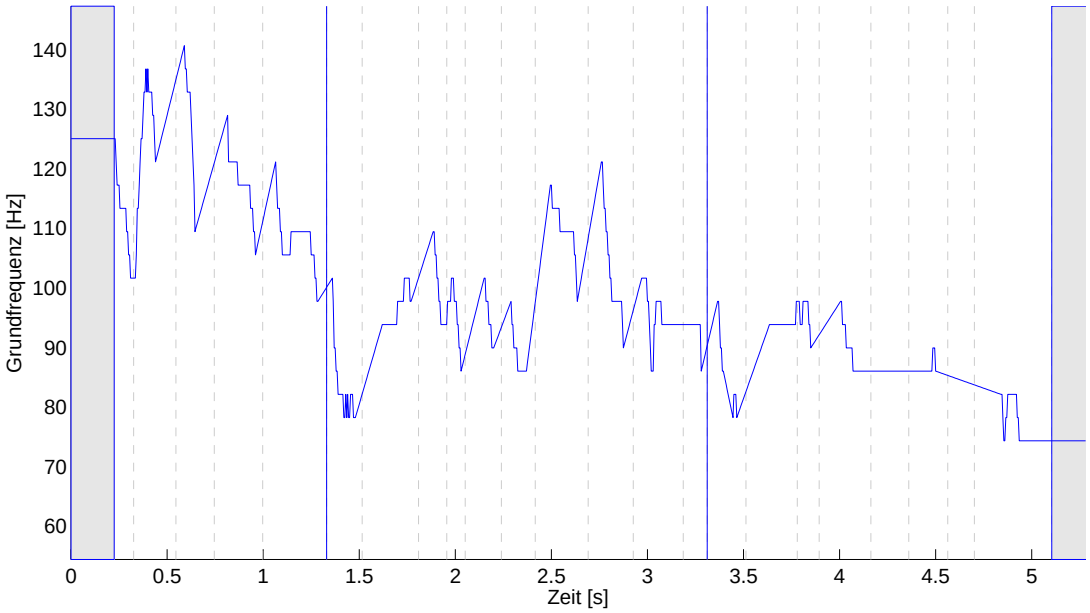
- **Schwache Grenze (’/ ’)**: Eine schwache Phrasengrenze zeichnet sich dadurch aus, dass der Sprecher an dieser Stelle mit der Grundfrequenz neu ansetzt (’Grundfrequenzreset’), aber keine Pause macht. Ein Beispiel ist gegeben in Abb 6. Die schwachen Grenzen sind als durchgezogene, senkrechte Linien eingezeichnet.

- **Keine Grenze**: Diese Klassifikation bezeichnet Silbengrenzen, die keine Phrasengrenzen sind und trifft dann zu, wenn keine der obigen 3 Klassifikationen zutrifft. In Fig. 6 sind sie als gestrichelte senkrechte Linien eingezeichnet.

1.3 Phonologische Repräsentation

Für alle Sätze eines Korpus liegen neben dem Audiosignal (.wav-File) und dem Grundfrequenzverlauf (.F0-File) auch die phonologische Beschreibung in einem sogenannten ptr-File vor. Darin ist hauptsächlich die phonologische Beschreibung im ETHPA enthalten, aber auch Informationen über Silbengrenzen, Phrasengrenzen, Phrasentyp und Silbenakzente. Der Beispielsatz

’Der amerikanische Präsident Reagan wird bei seinem bevorstehenden Staatsbesuch in der Bundesrepublik Deutschland auch ein ehemaliges Konzentrationslager besuchen.’



Figur 6: *Beispiel des Grundfrequenzverlaufs an schwachen Phrasengrenzen: (’Touristikbulletin / der schweizerischen Verkehrszentrale / vom 22. April.’)*

wird wie folgt notiert:

```
(P) der- [4]a-me-ri-[2]ka:-nI-S@- [4]prE-zi-[3]dEnt- [1]re:-g@n-
#{1} (P) [4]vIr- pa_i- za_i-n@m- b@-[2]fo:r-[4]Ste-@n-d@n-
[1]Sta:ts-b@-[4]zu:x- #{2} (P) ?In- der- [2]bUn-d@s-[4]re-pu-[4]bli:k-
[1]dO_ytS-[4]lant- #{1} (T) [3]a_ux- a_in- [2]?e:-@-[4]ma:-lI-g@s-
[4]kOn-tsEn-tra-[1]tsjo:ns-[4]la:-g@r- b@-[2]zu:-x@n.
```

Dabei bezeichnet

- `(P)` den Phrasentyp mit den in 1.1 definierten Typen, in diesem Fall progredient,
- `{1}` die Art einer Phrasengrenze, wobei `{1}` eine starke Grenze bezeichnet und `{2}` eine schwache,
- `'-'` eine Silbengrenze,
- `[1]...[4]` einen Silbenzakzent. `[1]` bezeichnet einen Phrasenhauptakzent, `[2]` einen ’Pitch Accent’, `[3]` einen ’Non-Pitch Accent’ und `[4]` einen Wortnebenakzent. Mit `[0]` können unbetonte Silben gekennzeichnet werden.

Zusätzlich ist in einem Label-File (.lab-File) für jeden Laut sein Anfangszeitpunkt in Sekunden notiert.

1.4 Ansätze zur Bestimmung der Phrasengrenzen und Phrasentypen

1.4.1 Detektion der Phrasengrenzen

Für den Anfang definieren wir, dass die zu einer Silbe gehörige Silben-/Phrasengrenze immer die Grenze **nach** der Silbe meint.
Für die Detektion der Phrasengrenzen werden anschliessend für jede Silbe eine Menge von Features (reelle Zahlen) aus Audiosignal, Grundfrequenzverlauf und gelabelter phonologischer Beschreibung extrahiert. Die genaue Beschreibung dieser Features folgt in Abschnitt 4. Die Werte dieser Features werden am Eingang des statistischen Modells angelegt. Der Ausgang des Modells soll nun eine Klassifikation des Phrasentyps am Ende der betreffenden Silbe liefern.
Da Detektion mit Information über nur eine Silbe nicht sehr gut funktioniert, werden auch Features aus je 1-2 Silben vor und nach der betrachteten Silbe mitgegeben (’context window’). Die Erfahrung hat gezeigt, dass ein 5 Silben breites Fenster (aktuelle Silbe + je 2 Silben vorher und nachher) die beste Detektionsperformance liefern.
Das statistische Modell wird anschliessend mit den Features von von Hand annotierten, korrekten Klassifikationsdaten gefüttert und so trainiert.

1.4.2 Detektion des Phrasentyp

Die Detektion des Phrasentyps läuft etwa analog zur Detektion der Phrasengrenzen. Es wird angenommen, dass die Phrasengrenzen bereits annotiert sind und somit die einzelnen Phrasen individuell behandelt werden können. Statt über einzelne Silben werden hier die Features über eine gesamte Phrase extrahiert. Auf ein context window wird allerdings verzichtet, da Phrasentypen wenig abhängig vom Kontext sind. Das Training verläuft des weiteren analog zur Phrasengrenzendetektion.

2 Beschreibung der Prosodiekorpora

Für Training und Test der Phrasengrenzen- und Typendetektion, haben wir Korpora von zwei Sprechern (männlich und weiblich) verwendet. Von der Männerstimme (’Schuster’) existiert ein deutschsprachiger Korpus. Die Sprecherin der Frauenstimme (’Heim’) ist bilingue und es liegt sowohl ein deutschsprachiger, als auch ein französischsprachiger Korpus vor.
Da die phonologische Repräsentation bei der Frauenstimme zu Beginn nur eingeschränkt brauchbar war, haben wir zusammen mit H. Romsdorfer von den eigentlich vorhandenen mehreren 1000 aufgenommenen Sätzen nur 400 (deutsch) und 50 (französisch) entsprechend korrigiert und verwendet.
In den folgenden Abschnitten sind Statistiken über die Korpora aufgelistet

2.1 Anmerkungen zur Grundfrequenzverteilung

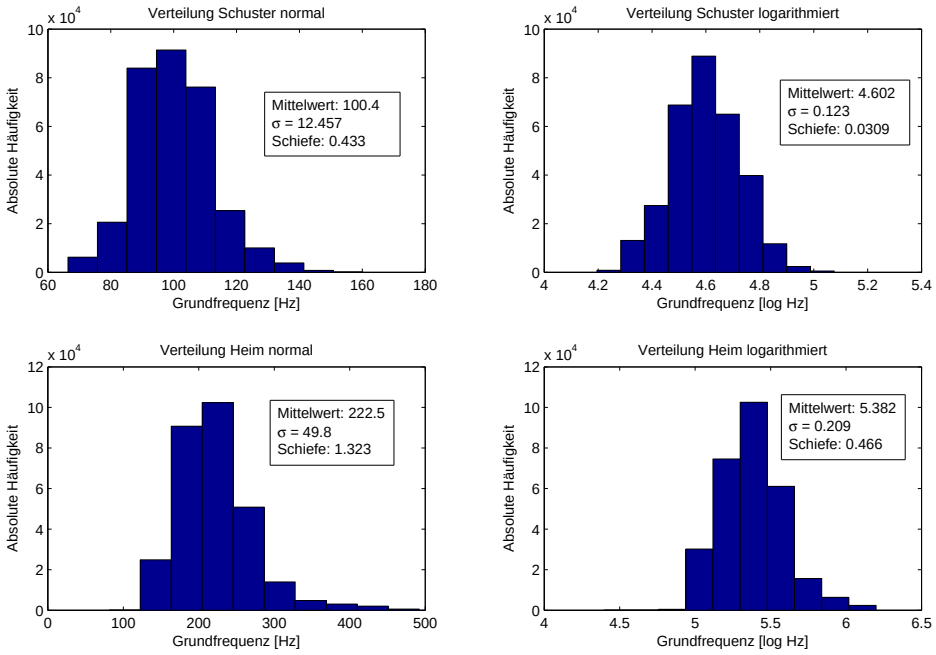
Die Verteilung der Grundfrequenzwerte zwischen der Männer- und der Frauenstimme ist wie anzunehmen sehr unterschiedlich und in Fig. 7 dargestellt.
Wie zu erwarten, liegt der Mittelwert der Grundfrequenz bei der Männerstimme deutlich tiefer als bei der Frauenstimme. Die Tatsache, dass die menschliche Wahrnehmung der Grundfrequenz etwa logarithmisch ist, findet sich wieder in einer leichten Rechtsschiefe der unlogarithmierten Verteilungen. Durch eine Logarithmierung, wird die Schiefe deutlich reduziert.
Trotz Logarithmierung ist die Standardabweichung bei der Frauenstimme immer noch höher als die der Männerstimme. Dies liegt an einigen Ausreissern gegen oben, während die Grundfrequenz der Männerstimme im logarithmierten Bereich schon sehr nahe an die Normalverteilung kommt.

2.2 Länge der Silben und Silbenkerne vor einer Phrasengrenze

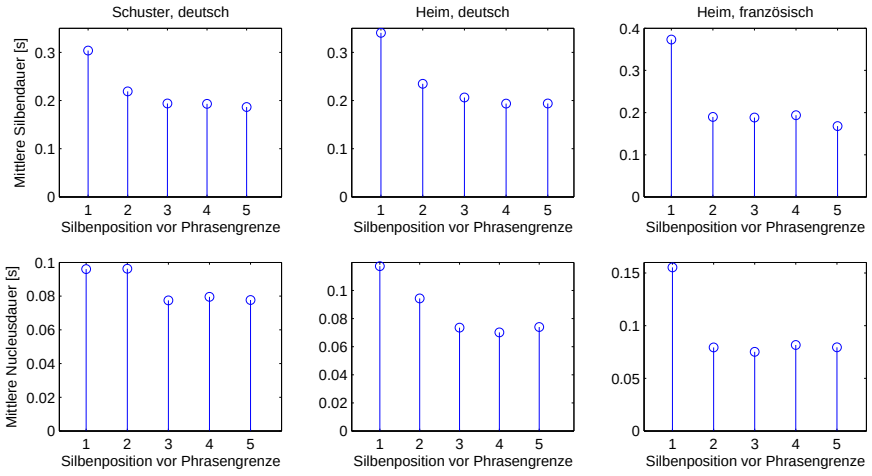
Weiter haben wir für die drei Korpora die mittlere Länge der Silben und Silbenkerne vor einer Phrasengrenze ausgewertet. Die Ergebnisse sind in Fig. 8 tabelliert.
Die Ergebnisse decken sich in etwa mit der Hypothese des ’final lengthening’, die besagt, dass die Länge der Silben und Silbenkerne vor einer Phrasengrenze zunimmt [4]. Es gilt zu berücksichtigen, dass es sich hierbei nur um Mittelwerte handelt und die Verteilungen durchaus überlappen.

2.3 Länge der Silbenkerne nach Lauten sortiert

Zusätzlich haben wir hier anhand des Schusterkorpus für alle Silbenkerne die mittlere Länge und Varianz der Länge nach Lauten getrennt analysiert und in Tabelle 1 aufgelistet.
Es ist erkennbar, dass die Länge der Silbenkerne stark vom Laut abhängt. Dies gibt Anlass, die Nucleusdauer bei der Verwendung als Feature auf den jeweiligen Laut zu normieren.



Figur 7: *Grundfrequenzverteilung der beiden deutschen Sprecher*



Figur 8: *Mittlere Silben- und Silbenkerndauer vor einer Phrasengrenze. Mit 1 wird die Silbe unmittelbar vor der Phrasengrenze bezeichnet, mit 2 die vorletzte Silbe vor der Grenze, etc.*

Laut	Mittlere Dauer [s]	Varianz der Dauer [s^2]	Absolute Häufigkeit
a͡u	0.161270	0.00157650	123
a͡i	0.150260	0.00151170	399
aː	0.147910	0.00112860	293
oː	0.130470	0.00063370	125
eː	0.130060	0.00073187	231
iː	0.118720	0.00077447	222
a	0.088784	0.00085727	725
ɔ	0.081027	0.00068958	246
ɛ	0.078602	0.00054880	513
i	0.072357	0.00054665	260
o	0.070228	0.00057441	106
e	0.067357	0.00065610	309
ʊ	0.061442	0.00037333	418
ɪ	0.060429	0.00030617	607
ə	0.049053	0.00054483	1502

Tabelle 1: *Verteilung der Nucleusdauer über alle Laute, die häufiger als 100 Mal auftreten im Schusterkorpus*

2.4 Männlicher Sprecher, deutsch

Der Schusterkorpus besteht aus Sätzen, die aus Nachrichtenberichten entnommen sind und von einem professionellen Sprecher gesprochen wurden. Da er keinerlei Frage- und Ausrufesätze enthält, ist er für das Training eines Typdetektors nicht geeignet, Deshalb ist der Phrasentyp nur sehr rudimentär annotiert. Meist sind alle Phrasen eines Satzes ausser der letzten als progredient und die letzte als terminal annotiert. Ausserdem ist das Laut-Labelling in Handarbeit nachkorrigiert.

- Name des Sprechers: Schuster
- 186 Sätze in deutscher Sprache
- Totale Aufnahmedauer $\approx$ 1600 Sekunden ($\approx$ 27 Minuten)
- 998 Prosodische Phrasen: 827 progredient + 171 terminal (mit Vorbehalt, s.o.)
- 6588 Silben(grenzen): 186 Satzgrenzen + 351 Starke Phrasengrenzen + 461 Schwache Phrasengrenzen + 5590 Silbengrenzen ohne Phrasengrenze
- Die Statistik über die Grundfrequenzverteilung ist in den oberen beiden Plots von Fig. 7 abgebildet.

2.5 Weibliche Sprecherin, deutsch

Der Heimkorpus besteht aus Sätzen, die in sehr unterschiedlichem Kontext stehen. Neben Aussagesätzen, enthält er auch Fragen und Befehle und ist somit fürs Training der Typendetektion

geeignet. Das Laut-Labelling ist hier rein maschinell entstanden und daher potentiell ungenau.

- Name der Sprecherin: Heim
- 400 Sätze in deutscher Sprache
- Totale Aufnahmedauer $\approx$ 1500 Sekunden ($\approx$ 25 Minuten)
- 1030 Prosodische Phrasen: 498 progredient + 221 terminal + 143 Ja-Nein-Fragen + 29 Exclamations + 139 Statements
- 5675 Silben(grenzen): 401 Satzgrenzen + 172 Starke Phrasengrenzen + 416 Schwache Phrasengrenzen + 4686 Silbengrenzen ohne Phrasengrenze
- Die Statistik über die Grundfrequenzverteilung ist in den unteren beiden Plots von Fig. 7 abgebildet.

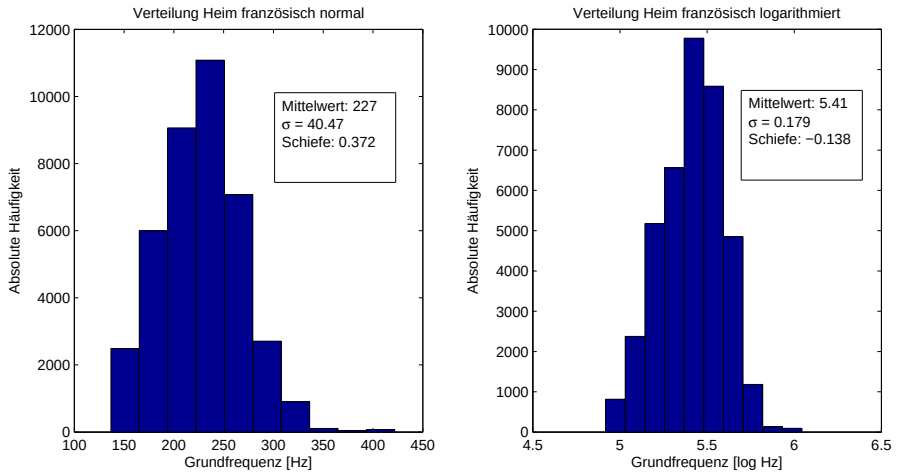
2.6 Weibliche Sprecherin, französisch

Der französische Heimkorpus besteht wie der deutsche ebenfalls aus sehr unterschiedlichen Sätzen.

- Name der Sprecherin: Heim
- 50 Sätze in französischer Sprache
- Totale Aufnahmedauer $\approx$ 200 Sekunden ($\approx$ 3 Minuten)
- 163 Prosodische Phrasen: 87 progredient + 25 terminal + 19 Ja-Nein-Fragen + 6 Exclamations + 26 Statements
- 714 Silben(grenzen): 50 Satzgrenzen + 36 Starke Phrasengrenzen + 77 Schwache Phrasengrenzen + 551 Silbengrenzen ohne Phrasengrenze
- Die Statistik über die Grundfrequenzverteilung ist in Fig. 9 abgebildet.

3 Aufbau des Rahmenprogramms

In den folgenden Abschnitten werden wir die Funktions- und Verwendungsweise des Rahmenprogramms zur Phrasengrenzen- und -Typdetektion erläutern.
Der Ablauf für Phrasengrenzen und Phrasentypen läuft getrennt, aber im wesentlichen analog ab, weshalb wir auf eine weitere Aufteilung verzichten.



Figur 9: *Grundfrequenzverteilung des französischen Korpus*

3.1 Übersicht

Der generelle Ablauf ist hier skizziert:

1. Generierung von Dateilisten
2. Profilgenerierung
3. Feature-Extraktion
4. Training
5. Test

3.2 Generierung von Dateilisten

Für den Ablauf der Feature-Extraktion und des Trainings der statistischen Modelle werden einige Dateilisten benötigt, die hier kurz beschrieben sind. Alle Dateilisten sind so aufgebaut, dass die Dateien für einen Satz auf einer Zeile stehen und durch Leerzeichen getrennt sind.

- **files.lst**
  wird benötigt, um die Features für die Phrasengrenzendetektion zu berechnen. Für jeden Satz sind in dieser Reihenfolge die folgenden Dateinamen eingetragen:
  Dateiname der **WAV**-Datei, Dateiname der **F0**-Datei, Dateiname der **lab**-Datei, Dateiname der **ptr**-Datei, Dateiname der **fea**-Datei.
  Die **fea**-Datei ist dabei die Datei, in die die Feature-Extraktion für Phrasengrenzen ihren Output schreibt.

- **files_typ.lst**
  wird benötigt, um die Features für die Phrasentypendetektion zu berechnen. Für jeden

Satz sind in dieser Reihenfolge die folgenden Dateinamen eingetragen:
Dateiname der **WAV**-Datei, Dateiname der **F0**-Datei, Dateiname der **lab**-Datei, Dateiname der **ptr**-Datei, Dateiname der **fea_phr**-Datei.
Die **fea_phr**-Datei ist dabei die Datei, in die die Feature-Extraktion für Phrasentypen ihren Output schreibt.

- **feafiles.lst**
  Wird beim Training der Phrasengrenzen benötigt und enthält für jeden Satz nur den Dateinamen der **fea**-Datei.

- **feafiles_typ.lst**
  Wird beim Training der Phrasentypen benötigt und enthält für jeden Satz nur den Dateinamen der **fea_phr**-Datei.

- **ptreditor.lst**
  Wird für das ptreditor-Programm benötigt.Für jeden Satz sind in dieser Reihenfolge die folgenden Dateinamen eingetragen:
  Name der **ptr**-Datei, Name der **new.ptr**-Datei, Name der **lab**-Datei, Name der **WAV**-Datei, Name der **F0**-Datei, Name der **fea**-Datei, Name der **fea_phr**-Datei.

3.3 Profilgenerierung

Für manche Features ist es für die Normierung nötig, Informationen aus dem ganzen Corpus zusammenzutragen. So zum Beispiel die Grundfrequenz- und Lautdauerverteilung. Das Profil kann dann zur späteren Verwendung als Datei gespeichert werden.
Die Profilgenerierung wird gestartet über die Funktion `createProfile`. Für alle Sätze des Corpus werden die wav-, F0-, ptr- und lab-Files ausgelesen.
Extrahiert werden:

- Mittelwert und Varianz der logarithmierten Grundfrequenzverteilung.
- Mittelwert und Varianz der Silbendauer
- Mittelwert und Varianz der Signalleistung in allen Silbennuclei
- Mittelwert und Varianz der Signalleistung in allen Silben
- Mittelwert und Varianz der Nucleusdauer für jeden als Nucleus auftretenden Laut separat (siehe Abschnitt 2.3 und insbesondere Tabelle 1)

3.4 Feature-Extraktion

Für jeden Satz werden Features extrahiert und als Datei (**.fea**-Datei für Phrasengrenzen, **.fea_phr**-Datei für Phrasentypen) gespeichert. Dies geschieht bei den Phrasengrenzen über die Funktion `calcPhrBndFeatures` und bei Phrasentypen über die Funktion `calcPhrTypeFeatures`.
Für Phrasengrenzen werden die Features silbenweise extrahiert, d.h. pro Feature existiert für

jede Silbe ein Wert. Für Phrasentypen werden die Features phrasenweise extrahiert, d.h. pro Feature existiert für jede Phrase ein Wert. Die Beschreibung der Features sowohl für Phrasengrenzen, als auch -Typen folgt in Abschnitt 4

3.5 Training

Ein statistisches Modell — wahlweise ein neuronales Netz ('multi layer perceptron', MLP) oder ein 'generalisiertes lineares Modell' (GLM) — wird nun mit den zuvor extrahierten Features trainiert. Das Verfahren für verschiedene statistische Modelle ist unterschiedlich und im folgenden kurz erklärt.

3.5.1 Neuronales Netz

Im Falle eines neuronalen Netzes werden die Daten satzweise unterteilt in Trainings-, Evaluations- und Testset. Das Trainingsset ist dabei die Datenmenge, die tatsächlich im verwendeten Trainingsalgorithmus verwendet wird. In unserem Fall ist der Algorithmus der 'scaled conjugate gradient algorithm' (SCG) [1]. Das Evaluationsset wird während der Trainingsiterationen ausgewertet, um festzustellen, ob der Fehler darauf ebenfalls gesunken ist.
Das Training wird dann abgebrochen, wenn der Fehler auf dem Evaluationsset angestiegen ist. Dies dient dazu, Overfitting-Effekte zu verhindern. Andernfalls würde das Netz die Trainingsdaten 'auswendig lernen' und auf unbekannten Daten nur sehr schlecht performen.
Es ist zudem notwendig, eine Anzahl von Hidden-Variables und Hidden-Layers zu definieren. Wir haben uns dabei auf Netze mit nur einem Hidden-Layer beschränkt und die Zahl der Hidden-Variables nach der folgenden Formel aus [1] abgeschätzt:

$$\text{nhidden} = \frac{\left(\frac{\text{ntrain}}{10}\right) - \text{nout}}{\text{nin} + \text{nout} + 1}, \tag{1}$$

wobei **nhidden** = Anzahl Hidden-Variables, **ntrain** = Anzahl der Trainingsdaten, **nin** = Anzahl Eingangsgrössen (Features) und **nout** = Anzahl Ausgänge (Klassen).
Da das resultierende Netz nun trotzdem noch in gewisser Weise auf das Evaluationsset trainiert ist, wird zur endgültigen Evaluation der Klassifizierungsleistung nochmal ein separates Set — das Testset — verwendet.

3.5.2 GLM

Im Falle eines GLM wird der 'iterative reweighted least squares algorithm' (IRLS) [3] verwendet. Dieser kommt ohne ein Evaluationsset aus. Der endgültige Fehler wird aber auch hier über ein vom Trainingsset getrenntes Testset evaluiert.
Zusätzlich besteht die Wahl zwischen unterschiedlichen Ausgangsaktivierungsfunktionen. Im linearen Fall reduziert sich das 'Training' auf das Lösen eines linearen Gleichungssystems. Im Fall 'softmax' kommt der oben erwähnte IRLS-Algorithmus [3] zum Einsatz.

3.5.3 Eingangsnormalisierung

Für die Eingangsnormalisierung stehen im wesentlichen zwei Methoden zur Auswahl: 'linear' und 'principle component analysis' (PCA).

Linear Bei der linearen Normalisierung wird die Verteilung jeder Eingansdimension für sich normiert, so dass Mittelwert 0 und Varianz 1 wird.

PCA Bei der PCA-basierten Normalisierung wird zuerst der Mittelwert der Datenverteilung abgezogen und dieser Raum anschliessend so orthogonal transformiert, dass die einzelnen Dimensionen unkorreliert sind. Anschliessend werden die nun unkorrelierten Dimensionen wieder einzeln linear normalisiert auf Varianz 1. Im weiteren Verlauf der Arbeit haben wir in der Regel diese Art der Normalisierung verwendet, da sich damit generell bessere Ergebnisse erzielen liessen als mit der linearen Normalisierung. Allerdings ist die PCA-Normalisierung durch ihre höhere Komplexität auch fehleranfälliger als die lineare.

3.6 Test

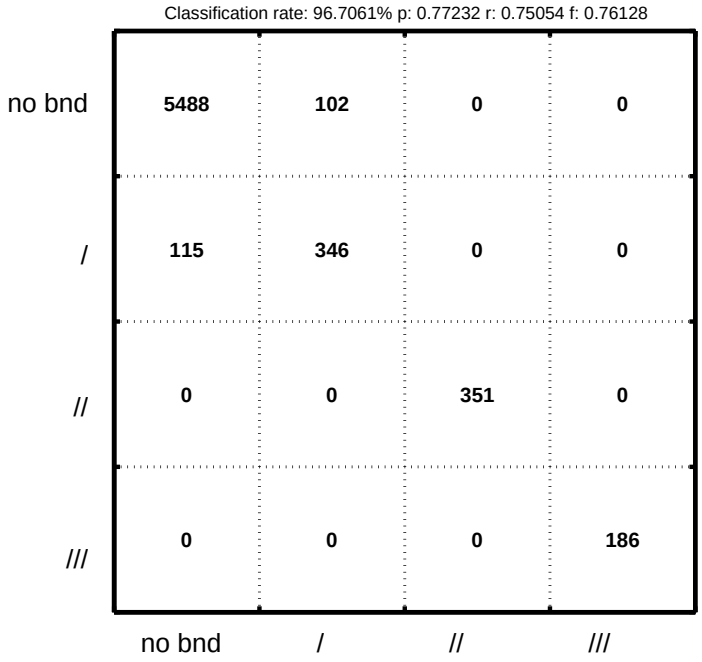
3.6.1 Confusion Matrix

Das Testen eines der obengenannten Modelle erfolgt wie schon erwähnt auf einem vom Trainingsset getrennten Testset. Die Auswertung der Klassifizierungsleistung erfolgt in einer sogenannten confusion matrix.

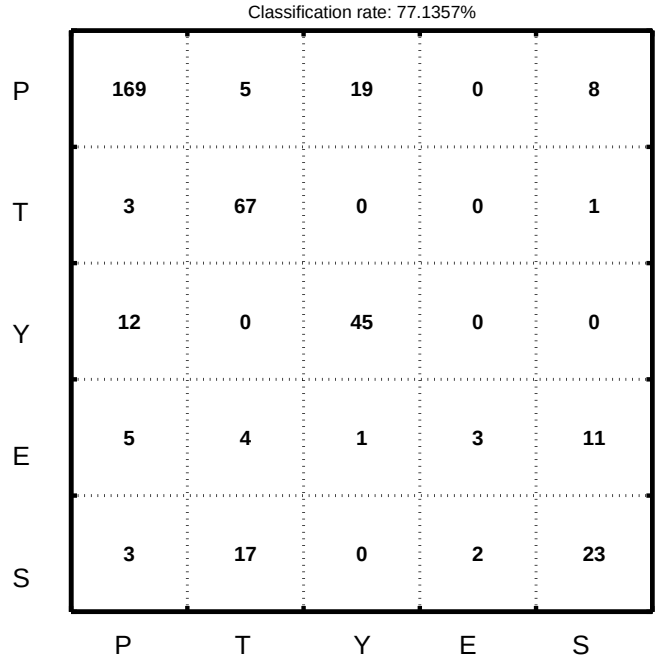
Eine typische Confusion-Matrix für Phrasengrenzen ist in Fig. 10 abgebildet. Dabei steht für jede Klasse von Phrasengrenze in der Vertikalen wie oft sie welcher Klasse in der Horizontalen zugeordnet wurde. Hier wurden z.B. 115 schwache Grenzen fälschlicherweise als keine Grenze detektiert, während an 102 Orten, wo keine Grenze ist, eine schwache Grenze detektiert wurde. Die korrekten Detektionen stehen also auf der Diagonalen.
Die classification rate berechnet sich aus der Summe der Diagonalelemente dividiert durch die Summe über alle Elemente und ist oben angemerkt. Zusätzlich wurde für die Klasse 'schwache Grenze' precision (p), recall (r) und f-score mit Gewichtungsfaktor 1 (f) berechnet.
Für die Phrasentypen sieht die Matrix ähnlich aus, nur dass die Klassen anders sind, wie aus Fig. 11 hervorgeht.

3.6.2 Auswertung im ptreditor

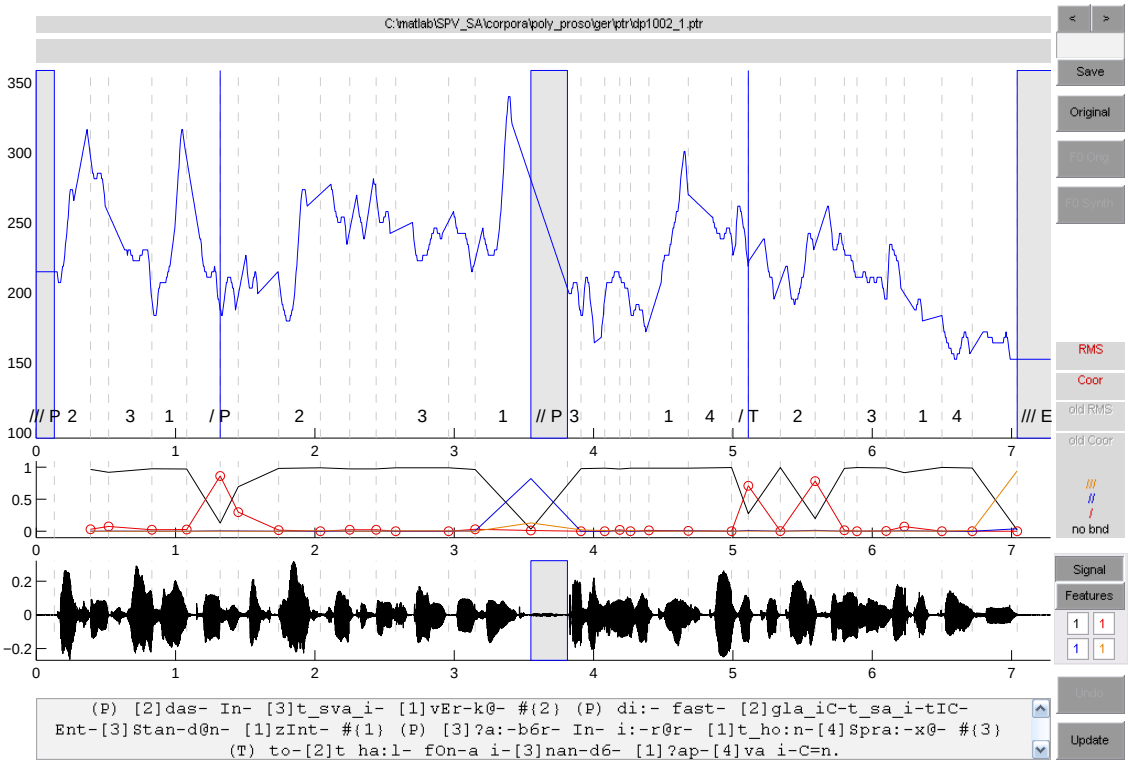
Phrasengrenzen Zusätzlich kann die Leistung eines Modells auch im ptreditor-Programm anhand von beliebigen Sätzen getestet werden. Die Benutzeroberfläche ist in Fig. 12 abgebildet und wird hier kurz beschrieben:
Der oberste Plot enthält den Grundfrequenzverlauf mit eingezeichneten Silben- und Phrasengrenzen, sowie Silbenakzenten.
Darunter folgt der Ausgang des Detektors, wobei die Farben der Linien in der Box rechts davon aufgeschlüsselt sind. In Fig. 12 ist z.B. eine falsch detektierte schwache Phrasengrenze bei etwa



Figur 10: *Beispiel einer Confusion Matrix bei der Phrasengrenzendetektion*



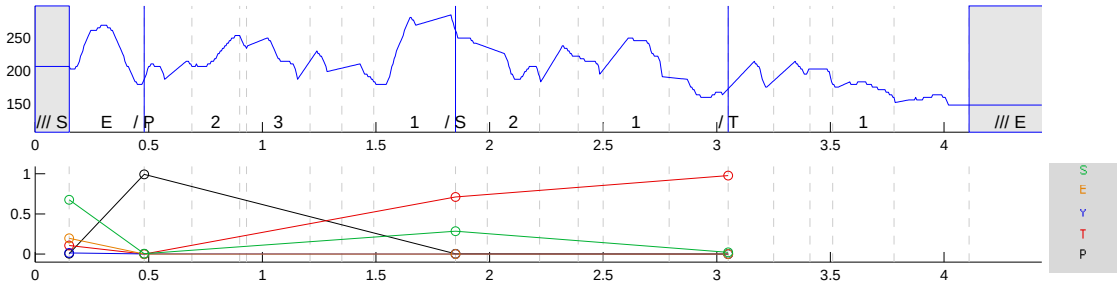
Figur 11: *Beispiel einer Confusion Matrix bei der Phrasentypendetektion*



Figur 12: *Ptreditor im Einsatz mit einem Phrasengrenzenmodell*

5.5 s zu sehen.
Im untersten Plot ist das Audiosignal eingetragen. Dies kann nützlich sein, um Pausen schnell zu erkennen.
Es besteht auserdem die Möglichkeit, durch einen Klick auf den 'Features'-Button sich den Verlauf von bis zu 4 Features während des Satzes anzusehen. Dabei wird der Plot des Audiosignals ausgetauscht.
Im Textfeld darunter ist ausserdem die phonologische Beschreibung aus der ptr-Datei enthalten, welche dort auch geändert werden kann, um Phrasengrenzen zu verschieben, neue einzuzeichnen und ihren Typ zu ändern.

Phrasentyp Der ptreditor wurde von uns so erweitert, dass er auch Modelle für Phrasentypendetektion verarbeitet. Der Ausgang des Detektors sieht dabei etwas anders aus, wie in Fig. 13 dargestellt. Die Klassifizierung bezieht sich dabei auf die Phrase, an deren Anfang sie eingezeichnet ist.



Figur 13: *Detektionsoutput für Phrasentypen*

4 Beschreibung der Features

Durch die Trennung der Phrasengrenzen- und Phrasentypendetektion resultieren auch zwei getrennte Featuresets: Die Phrasengrenzenfeatures werden silbenweise berechnet und mittels einem Kontextfenster von 5 Silben (die 2 vergangenen Silben, die aktuelle Silbe und nächsten 2 Silbe) in Verbindung gebracht.
Dagegen beziehen sich die Phrasentypenfeatures auf ganze Phrasen und besitzen kein Kontextfenster, da hierfür durchschnittlich zu wenige Phrasen pro Satz existieren.
Jede dieser Featuresets wird zusätzlich noch unterteilt in Features, welche lediglich aus dem Sprachsignal gewonnen werden können, und denjenigen, die zusätzlich noch auf die phonologische Beschreibung zugreifen.

4.1 Normierung und Profilinformationen

Da die Aufgabenstellung vorsieht, den Phrasengrenzen-und Phrasentypendetektor auf verschiedene Korpora mit jeweils verschiedenen Sprechern anzuwenden, ist die Normierung der Features ein zentrales Thema.
Durch die Segmentierung der Korpora in einzelne Sätze, ist der Trivialansatz die Normierung über die Features des jeweiligen Satzes. Da jedoch gewisse Features sehr selten auftreten (z.B. Pausen) oder aber der gesamte Satz nicht repräsentativ ist (Satz in Befehlsform weist viel höhere Energie und F0-Werte über den gesamten Satz), besteht die Möglichkeit sogenannte Sprecherprofile anzulegen, welche versuchen die Eigenheiten des Sprechers zu quantifizieren.
Dabei werden die Durchschnittswerte und Varianzen der zu normierenden Features über den ganzen Korpus berechnet.

4.2 Phrasengrenzenfeatures

Die Features, welche aus dem reinem Sprachsignal extrahiert werden, können in drei grundlegende Kategorien eingeteilt werden: in Dauerfeatures, welche die zeitliche Ausdehnung einzelner Grössen beschreiben, in Energiefeatures und in Merkmale der Grundfrequenz.

4.2.1 Dauerfeatures

Silbendauer (*syldur_norm*) Um das *'final lengthening'* (siehe Abschnitt 2.2 oder [4]) zu erfassen, wird hier die Silbendauer verwendet, wobei diese über alle Silben des Satzes normiert wird. Dies geschieht durch Division der Silbendauer über die mittlere Silbendauer des Satzes.
Alternativ wird bei vorliegenden Profilinformationen der Silbendauermittelwert des gesamten Korpus von der Silbendauer abgezogen und dieser Wert anschliessend durch die Standardabweichung der Silbendauer geteilt.

Nucleusdauer (*nucd_norm*) Die Silbenkerndauer wird berechnet und über den *Silbendauermittelwert* des Satzes normiert.
Alternativ wird mittels Profilinformation der ensprechende Nucleus über die jeweilige

durchschnittliche Lautdauer im Korpus normiert, da laut Tabelle 1 die Lautdauern sehr unterschiedlich sein können.

Pausendauer (*pausedur_norm*) Die Pausendauer ist laut [2] eine weiteres starkes Indiz für eine Phrasengrenze. Sie wird über die mittlere Silbendauer des Satzes normiert. Analog zur Silbenlänge wird mittels Profilinformation über den Silbenmittelwert des gesamten Korpus normiert.

Silbenkernabstand (*nucdist_norm*) Beschreibt den zeitlichen Abstand zweier aufeinanderfolgender Silbenkerne, d.h. es ist die Dauer zwischen dem Ende des aktuellen Silbenkerns und dem Anfang des folgenden Silbenkerns. Normiert wird über die mittlere Nucleusdauer im Satz oder über die mittlere Silbendauer des gesamten Korpus.

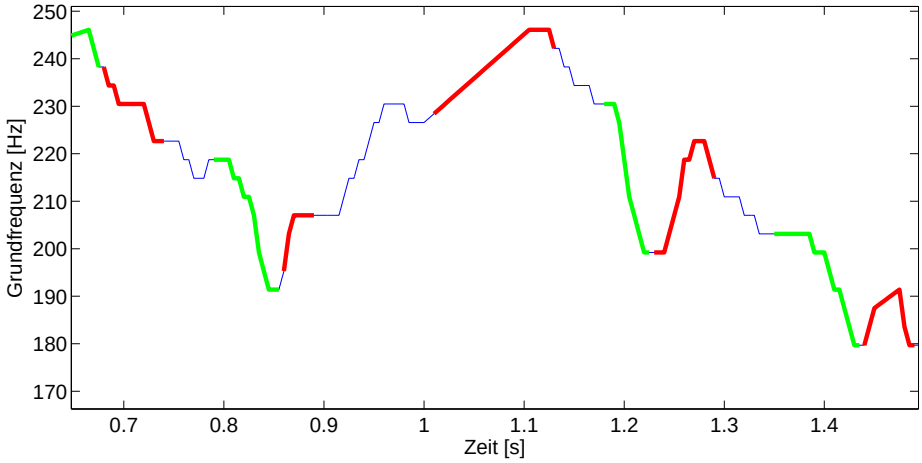
4.2.2 Grundfrequenzfeatures

Die Grundfrequenz wird mittels Grundfrequenzdetektor gewonnen. Dabei werden stimmlose Segmente linear interpoliert.
Die Grundfrequenz wird stets logarithmiert, da diese logarithmisch wahrgenommen wird und gemäss Abschnitt 2.1 dann annähernd normalverteilt ist. Falls nicht anders erwähnt, werden dann die Mittelwerte der logarithmierten Features abgezogen und durch die Standardabweichung geteilt, wobei sich diese Werte auf die einzelnen Sätze beziehen, bzw. bei Profilinformation auf den gesamten Korpus.
Somit sollten die Grundfrequenzverteilungen von Männer- und Frauenstimme vergleichbar sein.

Abtastung des Nucleus (*C(1)...C(4)*) Der Grundfrequenzverlauf des Nucleus wird mittels 5-Punkte Interpolation äquidistant abgetastet und anschliessend normiert, dass das Maximum der abgetasteten Werte auf 1 zu liegen kommt und das Minimum auf 0, wobei die zeitliche Ausdehnung auch noch auf die Dauer 1 normiert wird.
Von diesen 5 Punkte werden 4 Steigungen berechnet, welche als Features verwendet werden.

Onset (*minOnset, maxOnset*) Als Onset wird derjenige Anteil der Silbe bezeichnet, der zwischen Silbenanfang und Silbenkernanfang anzutreffen ist (rote Abschnitte in Fig. 14). Da diese generell einen etwa linearen Verlauf aufweist, werden nur 2 Punkte ermittelt, das Minimum und Maximum. Sollte kein Onset vorhanden sein, wird der Silbenkernanfang einfach konstant fortgesetzt.
Diese 2 Werte werden dann durch das Maximum bzw. Minimum der Nucleuswerte derselben Silbe dividiert.

Coda (*minCoda, maxCoda*) Als Coda bezeichnet man den Silbenanteil zwischen Ende des Nucleus und dem Silbenende (grüne Abschnitte in Fig. 14). Da diese generell einen etwa linearen Verlauf aufweist, werden nur 2 Punkte ermittelt, das Minimum und Maximum. Sollte keine Coda existieren, wird das Silbenkernende konstant fortgesetzt.
Diese 2 Werte werden dann durch das Maximum bzw. Minimum der Nucleuswerte derselben Silbe dividiert.



Figur 14: *Onset (rot) und Coda (grün) in den Grundfrequenzverlauf eingezeichnet*

Dynamik innerhalb Silbenkerns (*F0_ampl, F0_relampl*) Absolutes Maximum und Minimum des Silbenkerngrundfrequenzsverlaufs beschreiben als weitere Features die Dynamik innerhalb des Nucleus.

Polynomiales Fitting (*P Grad 0...P Grad 2*) Der logarithmierte, mittelwertfreie F0-Verlauf der aktuellen Silbe wird mit einer Grenzfrequenz von 5Hz gefiltert und anschliessend für ein Fenster von -0.1s vor dem Silbenende bis 0.1s nach dem Silbenende interpoliert.
Auf diesem gefilterten Grundfrequenzabschnitt wird gemäss Least Squares ein Polynom 2.Ordnung gefittet, wobei die Koeffizienten als Features verwendet werden.

4.2.3 Energiefeatures

Energie des Nucleus (*filtnucE_norm*) Der RMS-Wert der Signal-Amplitude wird über des Nucleus berechnet und ist proportional zur Wurzel der Energie (fortan wird RMS-Wert kurz als Energie bezeichnet). Normiert wird über alle Nucleusenergien im Satz.

Spektrale Emphasis (*filtnucE_norm*) Es wird die jeweilige Energie im nasalen Band (0-500Hz), sonoranten Band (500-2000Hz) und des frikativen Band (2-16KHz) ausgewertet, wobei die verwendeten Aufnahmen hier durch 16KHz Samplingrate bereits auf 8 KHz bandbegrenzt sind. Normiert wird über alle Nucleusenergien im Satz.

4.2.4 Features aus der phonologischen Beschreibung

Folgende Features sind aus den vorhandenen Labelfiles gewonnen. Da diese oft relativ aufwändig zu generieren sind und oft auch manuell nachkorrigiert werden müssen um vernünftige Ergebnisse zu erzielen, sind sie optional.

Wortgrenzen (*isWordBoundary*) Diese Informationen werden mitgegeben, da Phrasengrenzen nur bei Wortgrenzen auftauchen können.

Silbenakzente Bei vorhandener Akzentuierung können diese als weitere Features verwendet werden. Dabei werden die einzelnen Klassen wie folgt kodiert:

- 0000 für unbetonte Silben
- 1000 für Phrasenhauptakzente
- 0100 für Pitch Accent
- 0010 für Non-Pitch Accent
- 0001 für Wortnebenakzent

Für eine genauere Erläuterung der Klassen sei auf [5] verwiesen.

isHigh (*isHigh*) Aus der Lautbeschreibung wird ermittelt, ob der Silbenkernlaut ein Laut mit generell höherem F0 ist.

4.3 Phrasentypfeatures

Phrasenlänge (*phraselength*) Die Dauer der Phrase wird logarithmiert, da diese nach [5] logarithmisch wahrgenommen wird.

Polynomial Fitting (*P(1)...P(3)*) Der logarithmierte, mittelwertfreie, mit 1.35 Hz gefilterte Grundfrequenzverlauf der Phrase wird für ein polynomialles Fitting mit einem Polynom 2.ter Ordnung verwendet, woraus sich die Polynomkoeffizienten als Features ergeben.

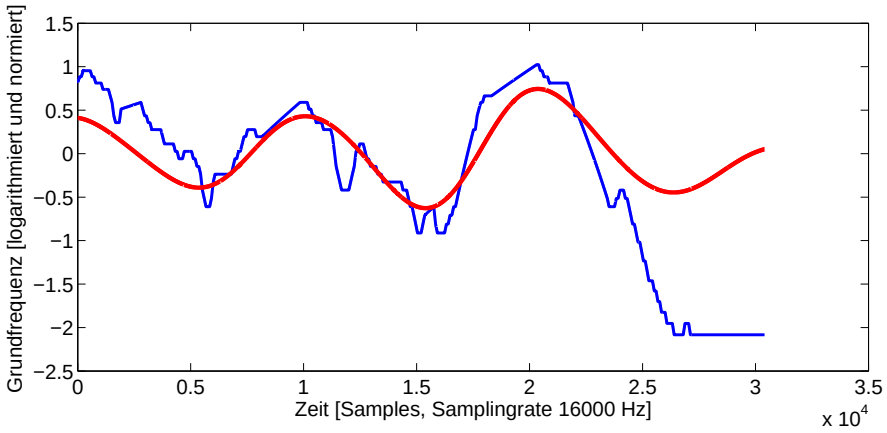
Energie (*E(1)...E(5)*) Die Phrase wird in zeitlich fünf gleich grosse Stücke aufgeteilt und jeweils der Effektivwert (RMS) des Zeitsignals berechnet. Die Energie ist proportional zum Quadrat dieser Grösse.

Rhythmus (*R*) Mit dieser Grösse versucht man die Regelmässigkeit im F0-Verlauf zu quantifizieren. Da diese Regelmässigkeit seine Periodizität im Bereich 1-2 Hz besitzt, wird das logarithmierte F0 in diesem Bereich mit einem Bandpassfilter gefiltert und als Feature das mittlere Amplitudenquadrat der Grundfrequenz in diesem Bereich gewählt. Ein Beispiel ist in Fig. 15 gegeben.

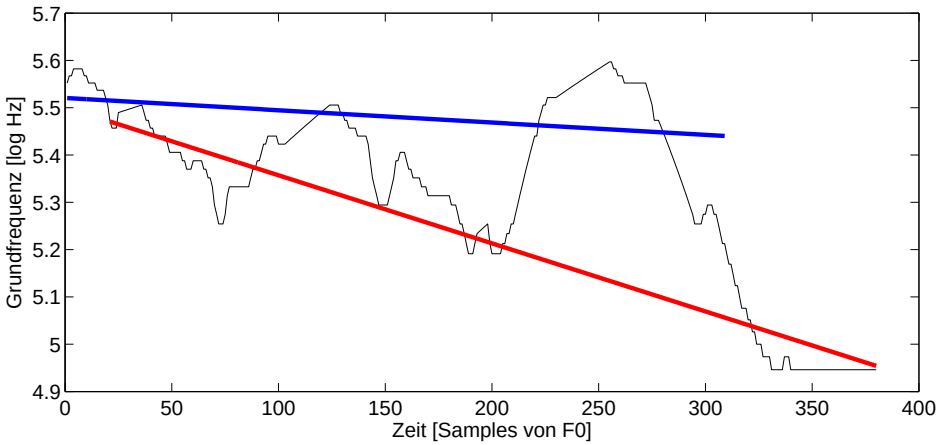
Dynamik der Grundfrequenz (*minF0, maxF0*) Maximum und Minimum des logarithmierten F0 über die gesamte Phrase.

Pause am Ende (*hasPauseAtEnd*) Beschreibt, ob direkt anschliessend an die Phrasengrenze eine Pause vorliegt.

Verlauf der Grundfrequenz (*meanF0 (1...5), minF0 (1...5), maxF0 (1...5), stdF0 (1...5)*) Zuerst werden die stimmlosen Abschnitte der logarithmierten Grundfrequenz entfernt, der Wertebereich auf [0,1], die Dauer auf 1 normiert und dann zeitlich in 5 gleich lange Segmente unterteilt. Für diese werden jeweils Minimum, Maximum, Mittelwert und Standardabweichung ermittelt und als Features verwendet.



Figur 15: *Beispiel für Rhythmusdetektion. Die rote Kurve bezeichnet den bandpassgefilterten F0-Verlauf, dessen mittleres Amplitudenquadrat als Feature 'Rhythmus' verwendet wird*



Figur 16: *Topline (blau) und Bottomline (rot) in einen Grundfrequenzverlauf gefittet*

Topline- und Bottomline (*topline, bottomline*) Der logarithmierte Grundfrequenzverlauf wird mit 0.5 Hz tiefpassgefiltert um den Grundverlauf zu charaktierisieren. Es wird nun eine Gerade nach Least Square gefittet für alle die F0-Werte, welche sich oberhalb diesem Grundverlaufs befinden. Analoges gilt für die die Werte unterhalb des Grundverlaufs. Die jeweiligen Steigungen der beides Geraden werden als Feature verwendet. Ein grafisches Beispiel ist in Fig. 16 gegeben.

4.3.1 Features der phonologischen Beschreibung für Phrasentypen

Satzenden (*isSentenceEnd*) Die Satzenden werden über die Phrasengrenzen des Typs '///' mitgeteilt. Bei Satzenden kann keine progrediente Phrase auftreten.

4.4 Auswahl der Features per Feature-Elemination

Es gilt festzustellen, welche der ermittelte Features die Detektionsleistung erhöhen und welche keinen oder gar einen negativen Einfluss auf die Detektion besitzen — indem z.B. zufällige Zusammenhänge auftreten, welche generell nicht gelten.
Hierzu wird zuerst das Modell mit allen N Features trainiert und die Klassifizierungsrate gespeichert.

In jeder Iteration wird die Klassifikationsleistung aller Featuresubsets mit N-1 Features berechnet, d.h. es wird jeweils berechnet wie gut die Performance ist, falls man auf ein Feature verzichtet. Die Leistung dieser Subsets wird mit Hilfe der sogenannten *Kreuzvalidierung* errechnet, bei der jedes Datum aus Trainings-und Evaluationsset $N_{cross} = 6$ mal verwendet, indem zuerst mittels des ersten Sechstels der Daten evaluiert wird und mit dem Rest trainiert, anschliessend auf dem zweiten Sechstel evaluiert wird und mit dem Komplement dazu trainiert, etc. bis alle Daten einmal zur Evaluation verwendet worden sind.

Dadurch stellt man sicher, dass die Featureelimination nicht lediglich auf einem einzelnen, fixes Set durchgeführt wird, welches möglicherweise gewisse Phänomene und Zusammenhänge nicht aufweist.

In jedem dieser Kreuzevaluationsdurchläufen wird dann das trainierte Netz mit den separaten Testdaten gefüttert und die Klassifikationsrate errechnet. Als Klassifikationsleistung der N-1 Features wird dabei die maximal erzielte Klassifizierungsrate in den Kreuzevaluationsergebnissen gewertet. Anschliessend wird dasjenige Feature entfernt, welches bei seiner Entfernung die beste Rate liefert.

Durch die Maximumsbildung werden jedoch unserer Meinung nach Ausreisser gewisser Konfigurationen, welche nur zufällig eine hohe Rate erreichen, zu stark gewichtet.
Erstrebenswerter wäre der Mittelwert als Mass, wobei Aussreisser durch ein Vorsortierung der Werte ausserhalb einem Band um den Mittelwerts nicht berücksichtigt werden.

5 Ergebnisse Phrasengrenzendetektion

5.1 Einleitung

Für Phrasengrenzen existierte zu Beginn unserer Arbeit bereits ein Detektor auf Basis eines neuronalen Netzes. Dieser erreichte auf dem Schusterkorpus eine Klassifizierungsrate von 94.91%. Diese Zahl mag auf den ersten Blick nach viel aussehen, allerdings ist zu bedenken, dass man auf dem gewählten Set mit einer trivialen Zuordnung auf die am häufigsten vorkommende Klasse (keine Grenze) auch schon eine Rate von etwa 86% erreicht. Wenn man ausserdem bedenkt, dass die Zuordnung von starken Phrasengrenzen und Satzgrenzen aus der phonologischen Beschreibung 1:1 abgeleitet werden kann und diese aus der Rechnung lässt, kommt man auf eine triviale Rate von 94.03%, womit die 94.91% nicht mehr sehr beeindruckend erscheinen. Diese triviale Rate hängt natürlich wiederum vom gewählten Testset ab und ist somit nicht konstant, sollte aber trotzdem als Grössenordnungen im Hinterkopf behalten werden bei den folgenden Ergebnissen unserer Optimierungsarbeit.
Um die Leistung der Netze besser einschätzen zu können, haben wir neben der Klassifizierungsrate auch den mit 1 gewichteten f-score für die schwache Grenze berechnet. Im ursprünglichen Zustand lag dieser bei 0.463 und wird fortan der Einfachheit halber mit 'f-score' bezeichnet.
Es muss auch berücksichtigt werden, dass die Leistung der Netze von der Wahl des Trainings- und Testsets abhängt, sowie von der zufälligen Initialisierung der Gewichte innerhalb des Netzes. Dadurch kann der f-score gut um $\pm$ 5% schwanken, grobe Ausreisser nach unten nicht mitberücksichtigt. Wir haben uns für diesen Bericht deshalb auf eine einzige Auswahl von Sets beschränkt, um eine gewisse Vergleichbarkeit der Resultate zu gewährleisten.

5.2 Wortgrenzen als Features

Eine deutliche Verbesserung konnte erzielt werden, indem wir dem Detektor die Information mitgaben, ob die betreffende Silbengrenze eine Wortgrenze ist. Falls es sich um keine Wortgrenze handelt, kann auch keine Phrasengrenze vorliegen. Damit können Fehlklassifizierungen reduziert werden, womit wir auf eine Rate von etwa 95.36% (f-score 0.548) kamen. Da bei dem verwendeten Netz die Klassen 'starke Grenze' und 'Satzgrenze' immer noch manchmal Fehler aufwiesen, haben wir dem Netz die Information einfachheitshalber explizit mitgegeben. Die Rate stieg dadurch erwartungsgemäss leicht auf 95.44% (f-score 0.586).

5.3 Korrektur des Korpus

Beim Analysieren der auftretenden Fehler fiel uns auf, dass die Phrasengrenzen im Korpus selbst oft inkonsistent gesetzt waren. Da inkonsistentes Trainingsmaterial die Detektionsperformance stark limitieren kann, befanden wir es für notwendig, den Korpus als nächstes konsistenter zu annotieren. Dies geschah in mehreren Durchgängen gemeinsam mit H. Romsdorfer. Durch diese Korrektur konnte die Klassifizierungsleistung abermals deutlich gesteigert werden auf 96.6% (f-score 0.714).

5.4 Versuch mit nur 2 Klassen

Da wir sowieso nur die Wortgrenzen betrachten, haben wir zusätzlich ein Modell getestet, welches nur an den Wortgrenzen trainiert wird und dort nur zwischen 'keine Grenze' und 'schwache Grenze' entscheidet. Dabei haben wir eine Rate von 87.7% erzielt, wobei die triviale Rate hier bei etwa 83% wesentlich tiefer liegt, da weniger Silbengrenzen keine Phrasengrenze sind. Der f-score lag aber nur bei 0.658, womit die Klassifizierung als schlechter eingeschätzt werden muss als beim Modell im vorherigen Abschnitt. Es hat den Anschein, dass diese Verschlechterung daher kommt, dass weniger Trainingsdaten für die Klasse 'keine Grenze' vorhanden sind. Wenn man nun die Daten für die wortinternen Silbengrenzen mitgibt, verbessert dies die Erkennung für die Klasse 'keine Grenze'. Dies führt unter anderem zu der Schlussfolgerung, dass sich wortinterne Silbengrenzen sich nicht wesentlich von Wortgrenzen unterscheiden, die keine Phrasengrenzen sind.
Aus diesen Gründen haben wir im weiteren Verlauf unser Modell wieder mit sämtlichen Silbengrenzen trainiert.

5.5 Feature-Elimination

Da wir nach dem Extrahieren einiger neuer Features eine sehr grosse Anzahl Eingänge (108) für das Netz hatten, wollten wir herausfinden, welche dieser Features sich eliminieren lassen, ohne die Leistung des Netzes zu beeinträchtigen. Dies mit dem Hintergedanken, dass eine kleinere Anzahl Eingänge tendenziell zu einem robusteren Modell führt. Daher haben wir die in 4.4 beschriebene Feature-Elimination laufen gelassen. Allerdings waren die Ergebnisse in 2 Durchläufen sehr unterschiedlich. Es zeichnete sich einzig der Trend ab, dass die Informationen aus der zweiten Silbe nach der Phrasengrenze eher unwichtig sind. Wir haben uns daher entschlossen, die Features mit etwas Intuition manuell auszusortieren und die Performance eigenhändig zu testen.
Das Featureset, welches dabei herauskam, ist in Tabelle 2 tabelliert.

Mit diesem Featureset (plus Wortgrenzen, starke und schwache Phrasengrenzen) aus 33 Features werden auf dem Schusterkorpus 96.9% Klassifizierungsleistung bei einem f-score von 0.76. Der grösste Vorteil dieser Reduktion macht sich allerdings erst bemerkbar, wenn man ein so trainiertes Netz auf einem anderen Korpus auswertet.

5.6 Übertragung und Adaption auf die deutsche und französische Frauenstimme

5.6.1 Direkte Übertragung

Ohne Featureselektion, d.h. mit den vollen 108 Features funktioniert ein auf dem Schusterkorpus trainiertes Netz zwar fast gleich gut, liefert auf dem deutschen Heimkorpus allerdings vernichtend schlechte Ergebnisse mit einem maximalen f-score von 0.23.
Also haben wir das Netz aus dem vorigen Abschnitt getestet, welches nur die dort bezeichneten Features verwendet. Bei der direkten Auswertung auf dem ganzen deutschen Heimkorpus kamen wir dabei auf eine Klassifizierungsrate von 90.3% und einen f-score von maximal 0.34.

Featurename	2 Sil. v. Gr.	1 Sil. v. Gr.	0 Sil. v. Gr.	1 Sil. n. Gr.	2 Sil. n. Gr.
nucd_norm	×	×	×	×	×
nucdist_norm	×	×	×	×	×
nucE_norm 0-500	-	×	×	×	×
nucE_norm 500-2000	-	-	-	-	-
nucE_norm 2000-max	-	-	-	-	-
filtnucE_norm	-	-	-	-	-
syldur_norm	×	×	×	×	×
pausedur_norm	-	-	-	-	-
F0_ampl	-	-	×	-	-
F0_relampl	-	-	-	-	-
C(1)	-	-	-	-	-
C(2)	-	-	-	-	-
C(3)	-	-	-	-	-
C(4)	-	-	-	-	-
P Grad 2	×	-	-	-	-
P Grad 1	×	-	×	×	-
P Grad 0	-	-	-	-	-
maxOnset	-	×	×	×	-
minOnset	-	-	-	-	-
maxCoda	-	-	-	-	-
minCoda	-	×	×	×	-

Tabelle 2: *Ausgewählte Features für Phrasengrenzendetektion. '×' steht für ein verwendetes Feature, '-' für ein nicht verwendetes Feature*

Bei der direkten Auswertung auf dem französischen Korpus kamen wir allerdings auf einen wesentlich höheren f-score von 0.56.

5.6.2 Adaption

Mit Adaption ist gemeint, dass ein mit dem Schusterkorpus trainiertes Netz mit wenigen Sätzen aus einem anderen Korpus (hier 20 Sätze aus dem deutschen Heimkorpus) mit dem SCG-Algorithmus weitertrainiert wird. Dabei kamen wir auf dem deutschen Heimkorpus auf einen f-score von 0.42. Der f-score auf dem französischen Korpus hat sich dabei auf 0.59 verbessert.
Die Adaption auf dem französischen Korpus direkt ist zwar machbar, lässt sich allerdings schlechter evaluieren, da zu wenig Sätze für Trainings-, Evaluations- und Testset vorhanden sind.
Es sei erwähnt, dass in Einzelfällen eine Adaption alleine auf dem deutschen Heimkorpus eine Verbesserung auf einen f-score von bis zu knapp 79% beim französischen Heimkorpus ergab. Dieses Ergebnis stellte sich allerdings als schwer reproduzierbar heraus, weshalb es hier nicht als Hauptbeispiel angeführt wird.

5.6.3 Gründe für die Diskrepanz zwischen deutscher und französischer Stimme

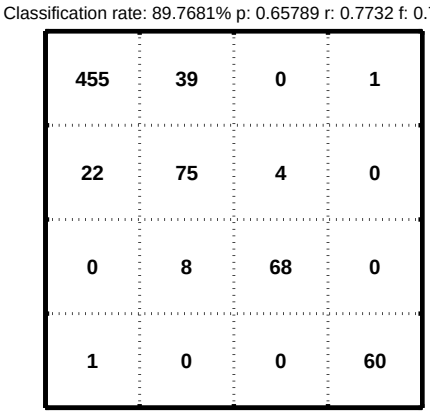
Die Tatsache, dass ein mit dem deutschen Heimkorpus adaptiertes Netz auf dem französischen Korpus besser abschneidet, als auf dem Korpus, wo es trainiert wurde sorgt für Verwunderung. Die Vermutung liegt nahe, dass es sich hier eher um ein Problem des deutschen Heimkorpus handelt denn um einen Mangel im Modell. Beim Durchhören des deutschen Heimkorpus fällt einerseits auf, dass die Grenzen auch subjektiv eher schlecht zu hören sind. Ausserdem haben wir in einigen Fällen festgestellt, dass das Laut-Labelling — welches automatisch geschieht — stark daneben liegt. Während für die Phrasentypendetektion eine Verschiebung um 100 ms kaum bedeutsam ist, kann dies die Phrasengrenzendetektion durchaus durcheinander bringen.

5.7 Test mit GLM

Zusätzlich haben wir testweise GLMs mit verschiedenen Aktivierungsfunktionen verwendet. Im Fall ’tanh/softmax’ kamen wir dabei bei Training und Übertragung auf etwa gleichwertige Resultate wie mit dem MLP. Im rein linearen Fall versagt das GLM allerdings bereits beim Training gründlich mit f-scores um 0.2, welche bei der Übertragung auf andere Korpora nochmals deutlich unterboten werden. Dies deutet darauf hin, dass das rein lineare Modell der Problemkomplexität nicht gewachsen ist. Andererseits ist die Komplexität bei den vorhandenen Daten auch nicht so hoch, dass ein neuronales Netz aus mehreren Layers benötigt wird.

5.8 Zusammenfassung und Ausblick

Ein beim Interpretieren der oben gefundenen Raten und f-scores wichtiger Punkt ist der, dass das Setzen von prosodischen Phrasengrenzen generell eine sehr subjektive Angelegenheit ist. Eine sehr hohe Klassifizierungsrate ist daher kaum erreichbar, da das Netz kaum besser sein

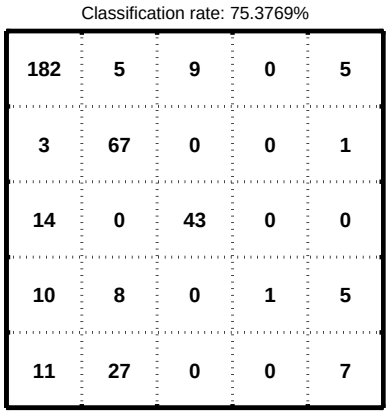


Figur 17: *Confusion Matrix für unser Modell ohne als Features mitgegebene starke Phrasengrenzen und Satzgrenzen, trainiert über alle Silbengrenzen, ausgewertet nur auf den Wortgrenzen*

kann als die Person, die die Trainingsdaten annotiert hat. Und je grösser ein Korpus ist, desto wahrscheinlicher ist auch, dass sich darin Inkonsistenzen befinden.
Eine Problemursache liegt darin, dass die annotierende Person die Phrasengrenzen oft nicht nur nach rein prosodischen, sondern auch syntaktischen Kriterien setzt — und somit auf einem anderen Abstraktionsniveau als das neuronale Netz arbeitet. Viele der 'Fehlklassifizierungen' des Netzes sind auch durchaus nicht abwegig. Dies ist allerdings schwierig zu fassen, da die Annotation binär geschieht, d.h. eine Silbengrenze ist eine Phrasengrenze oder eben nicht. In einer weiteren Arbeit könnte man daher beispielsweise versuchen, eine differenziertere Klassifizierung für schwache Phrasengrenzen zu finden und zu testen.

Vergleich mit anderen Arbeiten Im Vergleich mit anderen Arbeiten wie z.B. [2] schneidet unser Modell eher besser ab. In [2] wurden nur die Klassen 'keine Grenze', 'schwache Grenze' und 'starke Grenze' betrachtet. Zur besseren Vergleichbarkeit haben wir zusätzlich auf die Mitgabe der Targets für starke Grenzen verzichtet. Damit kamen wir über die ersten 3 Klassen auf einen mittleren f-score von 84% bei der in Fig. 17 abgebildeten Confusion Matrix.

In [2] wird eine 'maximum recognition rate' von 75.7% angegeben. Allerdings war das Setting dort ein etwas anderes. So wurden 10'000 Sätze von 100 nicht-professionellen Sprechern verwendet, welche automatisch annotiert worden waren. Durch die viel grössere Zahl von Trainingsdaten ist in jener Arbeit nach Formel (1) ein MLP mit wesentlich mehr Hidden-Variables (2 Layer mit 60 Variablen im ersten und 30 im zweiten Layer) möglich. Gleichzeitig ist nicht klar, wie konsistent der dort verwendete Satzkorpus in Bezug auf Phrasengrenzen war, was durchaus einen grossen Einfluss auf die Erkennungsrate hat.



Figur 18: *Anfängliche Phrasentypdetektion mit 11 Features*

6 Resultate der Phrasentypendetektion

6.1 Einleitung

Im Gegensatz zu der Phrasengrenzendetektion existierte kein Prototyp eines Detektors. Da sich die Ansätze für die Problemstellungen gleichen, konnten einige Teile übernommen werden und mussten lediglich angepasst werden. Die erste Konfiguration der Phrasendetektion bestand aus 400 grob korrigierten Sätzen aus dem Heim-Korpus. Diese wurden mittels folgendem Featureset detektiert:

- F0_relamp als Mass für die relative Schwankung der Phrase
- Die 5-Punkte-Interpolation C, wobei die 4 Steigungen dazwischen ausgewertet wurden
- und ein polynomiales Fitting P der Ordnung 4 über die gesamte Phrase.

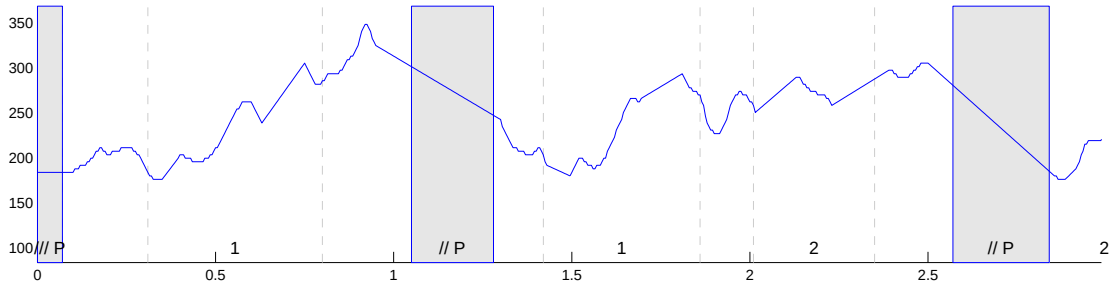
Es wurde damit eine Rate von über 75% erreicht.

Als nächstes wurde der Korpus gemeinsam mit H.Romsdorfer konsistenter notiert, wobei vor allem der Typ Statement (S) nach klareren Regeln bewertet wurde und versucht wurde, die Typen aufgrund der akustischen Wahrnehmung und nicht nach der Syntax zu bewerten.
Die Unterschiede in den Mengen der einzelnen Typen sind dadurch zu begründen, dass in der ersten Version kein separates Evaluationsset bestand und somit das Testset grösser war. Fortan wird für alle folgenden Plots dieses Korpus ein Trainset von fixen 250 Sätzen verwendet, die zufällig aus den 400 Sätzen gewählt wurden, d.h. für jede Konfiguration werden jeweils die gleichen 250 Sätze für das Training verwendet. Somit stellt man die Vergleichbarkeit sicher und hat gleichzeitig eine repräsentative Auswahl, da einige Phrasentypen und Phänomene in der ersten Hälfte gehäuft oder nur vereinzelt vorkommen.
Von den restlichen 150 Sätze werden zufällig 50 dem Evaluationsset zugewiesen und der Rest dem Testset.

Classification rate: 78.903%

96	1	9	0	4
0	46	0	0	5
12	0	23	0	0
1	0	0	0	5
7	6	0	0	22

Figur 19: *Phrasentypdetektion des korrigierten Korpus mit 11 Features*

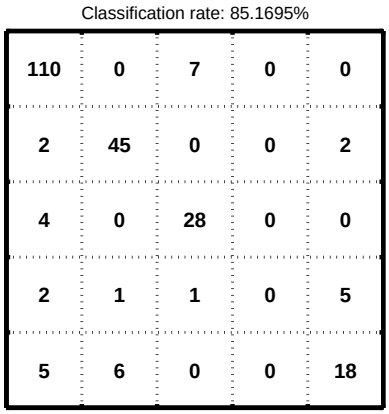


Figur 20: *Grundfrequenzverlauf einer P-Phrase mit Nebenbemerkung*

Man stellt hierbei fest, dass bei bestehendem Featureset Probleme bei der Klassifizierung zwischen den Typen Progredient (P) und Ja/Nein-Fragen (Y), Statement (S) und Terminal (T), S und P und Exclamation (E) insgesamt auftreten.
Die Verwechslungen von P gegenüber Y sind grösstenteils durch das Auftreten von Sätzen, welche Nebenbemerkungen beinhalten, zu begründen, welche sich akustisch nicht von den Ja/Nein-Fragen unterscheiden. Als Beispiel sei hier folgenden Satz erwähnt:

```
    Ein Söldner -- der Held der Geschichte -- wandert zwei
Jahre lang durch das atomare Inferno.
```

Dessen Grundfrequenzverlauf über die ersten 2 Phrasen ist in Fig. 20 abgebildet.
Es galt also noch Features zu finden, welche die generell höheren Schwankungen im Grundfrequenzverlauf von S gegenüber T und P-Phrasen erfassen und die generell höhere Energie und absolute Grundfrequenz des Typs E festhalten.



Figur 21: *Phrasentypdetektion mit den 19 besten Features*

6.2 Auswirkung neuer Features

Um die Performance zu steigern, wurden die in Kap. 4.3 erwähnten Features gewählt. Analog zur Phrasengrenzendetektion wurde die Featureelimination durchgeführt, wobei die Ergebnisse ein wenig konsistenter sind. Dennoch musste nach Plausibilitätsüberlegungen und manuellem Testen, ein geringfügig anderes Featureset als das vorgeschlagene gewählt werden.
Eine Auflistung der 20 gewählten Features ist in der folgenden Zusammenstellung aufgelistet:

Grundfrequenzfeatures : Das Minimum und Maximum des Grundfrequenzverlaufs (*F0_relamp* und *F0_amp*), die Koffizienten des polynomialen Fittings (*P1-P3*), Rhythmus (*R*), der durchschnittliche F0-Wert in vorletzten Segment (*meanF04*), minimaler F0-Wert im ersten, vorletzten und letzten Segment (*minF01, minF04, minF05*), maximaler Wert im ersten und letzten Segment (*maxF01, maxF05*), Standardabweichung der letzten 3 Segmente (*stdF03-stdF05*)

Energiefeatures : Energie im zweiten, dritten und letzten Segment (*E2, E3, E5*)

Dauerfeatures : Phrasenlänge (*phraselength*)

Phonologische Feature : Satzende (*isSentenceEnd*)

Auf Fig. 21 ist die daraus resultierende Confusion-Matrix illustriert. Es stellte sich eine Verbesserung um gut 3.5% auf 78.9% Erkennungsleistung.
Insgesamt sieht man eine geringfügige Verbesserung der Typen Y und S, welche durch die Neuinformation des absoluten F0-Werts, der Energie und Schwankungsgrössen zu erklären ist. Erstaunlich nützlich erweist sich ausserdem unser Rhythmus-Feature auf Phrasen des Typs S. Für eine genauere Beschreibung der Features sei hier auf die Featuresektion 4.3 verwiesen.

Erstaunlich schlecht bleibt der Typ E, welcher sich eigentlich durch die höhere mittlere Grundfrequenz und Energie leicht klassifizieren lassen müsste. Eine nähere Untersuchung

	P	T	Y	E	S
Mean F0	226.6007	194.1999	240.2831	214.1109	227.5499
RMS-Energy	0.05266	0.043747	0.05104	0.052408	0.065231

Tabelle 3: *Vergleich der Grundfrequenz- und Energiemittelwerte der verschiedenen Phrasentypen*

ergab, dass die Sprecherin die Ausrufe relativ kraftlos und unmotiviert ausspricht, und sich der Typ dadurch kaum von den anderen Typen hervorhebt.
Zu diesem Zwecke wurden in Tabelle 3 die Durchschnittswerte der Energie und Grundfrequenz für die einzelnen Phrasentypen aufgelistet.
Der mittlere Grundfrequenzwert ist in etwa identisch mit derjenigen einer progredienten Phrase und auch vergleichbar mit einigen Statement-Phrasen. Lediglich die mittlere Energie hebt sich ab, jedoch würde man einen signifikanteren Anstieg erwarten, bedenkt man, dass die wahrgenommene Lautheit logarithmisch von der Signalleistung abhängt.

Es sei hier festgehalten, dass die Sätze und Merkmale nicht sehr uniform über den Korpus verteilt sind und je nach Auswahl von Trainings- und Testset Unterschiede bis zu 5% in der Klassifikationsrate auftreten können.

6.2.1 Hinzufügen von phonologischen Features

Zuletzt wurde noch versucht die Auswirkung durch das Hinzugeben von Satzenden und Pausen zu analysieren. Bei einem Satzende kann keine progrediente Phrase auftreten unter der Annahme, dass der Satz korrekt betont wurde.
Die Auswertung mit diesen neuen Features ergab keine sichtliche Verbesserung der Erkennungsleistung, sie blieb weiterhin bei etwa 83-85% im Durchschnitt. Dies kann man darauf zurückführen, dass die Satzenden vom Detektor ohnehin kaum als progrediente Phrasen gewertet werden und die Verwechslungsraten von P und T sehr tief sind.

6.3 Übertragung des Netzes auf einen anderssprachigen Korpus

Als nächstes wird untersucht, inwiefern die gesammelten Ergebnisse eine generelle Gültigkeit besitzen und ob diese auch sprachübergreifend ist. Da der Schuster-Korpus nur aus 2 Phrasentypen besteht (P und T) und diese trivial annotiert sind (letzte Phrase immer T, Rest P) macht eine Übertragung auf diesen keinen Sinn, obwohl die Ergebnisse für die Übertragung von Frauen- auf Männerstimme sicherlich interessant wären.
Der deutsche Heimkorpus wird nun lediglich in ein Trainingsset- und ein Evaluationsset unterteilt, da die Auswertung ausschliesslich auf dem französischen Heimkorpus geschieht. Somit können mehr Trainingsdaten für das Netz verwendet werden, als bei der Auswertung im selben Netz.
Es wird lediglich noch eine Klassifizierungsrate von knapp über 75% erreichen. Das Netz scheint vor allem die Phrasentypen Y und S falsch zu erkennen. Dies kann unter Umständen mit der andersartigen Prosodie dieser Satztypen im Französischen zusammenhängen.

Classification rate: 75.4601%

87	0	0	0	0
2	12	0	0	11
6	0	11	1	1
3	0	0	0	3
9	4	0	0	13

Figur 22: *Übertragung aufs Französische*

Classification rate: 78.7234%

57	0	0	0	0
2	4	0	0	7
2	0	7	0	1
2	0	0	0	3
1	2	0	0	6

Figur 23: *Adaption auf den französischen Korpus*

6.3.1 Zusätzliche Adaption auf dem französischen Korpus

Um das Netz auf die Eigenheiten des neuen Korpus anzupassen, wird das bestehende Netz mit weiteren 25 Sätzen aus dem französischen Korpus weitertrainiert. Da insgesamt nur 50 Sätze auf dem französische Korpus vorhanden sind, haben wir für das Test- und Evalulationsset jeweils die identischen 25 restlichen Sätze verwendet, obwohl dadurch das Training stärker auf das Evaluationsset optimiert als gewöhnlich.
Generell lässt sich eine kleine Verbesserung der Rate von 1-2% erreichen jedoch bleibt die Rate in einigen Fällen konstant oder wird gar schlechter. Wir führen dies auf den sehr beschränkten Datensatz zurück, wobei möglicherweise gewisse Phrasentypen und Phänomene gar nicht auftreten und so auch nicht verbessert werden können.

Konfiguration	Neuronales Netz	Softmax	Linear
nur die ersten 11 Features	78.903%	78.193%	74.043%
19 besten Features nach Elim.	85.17%	83.81%	83.404%
inkl. phonologische Features	85.15%	83.96%	82.565%
Übertragung auf franz. Heim-Korpus	75.46%	73.616%	74.233%
mit Adaption von 25 Sätzen	78.723%	-	-

Tabelle 4: *Vergleich der Modelle*

6.3.2 Vergleich der verschiedenen Modelle für die Typendetektion

Es wird die Erkennungsrate des linearen Modells mit linearer Aktivierungsfunktion, dem linearen Modell mit Softmax-Aktivierungsfunktion und die Rate des neuronalen Netzes verglichen. Für eine Beschreibung der Modelle sei hier auf Kapitel 3.5 verwiesen. Aus dieser Zusammenstellung lässt sich erahnen, dass neuronale Netze bei vielen Features geringfügig besser abschneiden und vor allem bei der Übertragung auf eine andere Sprache durch ihre komplexere Struktur besser generalisieren können.

6.4 Zusammenfassung und Ausblick

Eine finale Rate von 80-85% auf dem deutschen Heimkorpus mag auf den ersten Blick nicht überwältigend sein, stellt jedoch eine starke Verbesserung gegenüber einem trivialen Schätzer, der lediglich auf den häufigsten Phrasentyp (meist progredient) schätzt, welcher je nach Testset auf etwa 50% kommen würde.
Zu berücksichtigen gilt auch, dass viele Sätze des Heimkorpus in ihrer Aussprache nicht den syntaktischen Erwartungen gerecht werden. So existieren viele Grenzfälle, in welcher der Typ dann mehr oder weniger zufällig zugewiesen wurde. Durch die absolute Festlegung auf einen einzelnen Typ, wird man der Grenzentscheidssituation nicht gerecht, da oftmals jeder der gemäss Detektion wahrscheinlichsten Typen vertretbar wäre. Ein differenzierteres Klassifikationsverfahren wäre wie bereits bei der Phrasengrenzendetektion erstrebenswert.

Bei der Übertragung und anschliessenden Adaption auf eine andere Sprache wurden in gewissen Fällen wenig schmeichelhafte Ergebnisse erzielt. Es wäre nötig, weitere Sätze des französischen Korpus zu annotieren und anschliessend zu testen, um sicherzustellen, dass die Verschlechterung der Rate tatsächlich auf auf unseren Ansatz und nicht auf den Mangel an repräsentativen Trainingsdaten zurückzuführen ist.

A Aufgabenstellung

Institut für Technische Informatik und Kommunikationsnetze
ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo

Sommersemester 2007
(SA-2007-49)

Semesterarbeitsaufgabenstellung

für

Herr Juri Baumberger
Herr Jonas Sonnenmoser

Betreuer: H. Romsdorfer ETZ D97.5

Ausgabe: 11. April 2007
Abgabe: 6. Juli 2007

Erkennung von Phrasengrenzen und Phrasentyp in 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, die Phrasengrenzen und der Phrasentyp 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

Phrasentyp- und Phrasengrenzenbestimmung untersucht und miteinander verglichen werden (siehe zum Beispiel [1, 2, 3, 4]).

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 Erkennung des Phrasentyps 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 Phrasentyp und Phrasengrenzen, z.B. [1, 2, 3, 4, 5].
2. Einarbeitung in die Literatur zu Mustererkennungsalgorithmen, z.B. [6, 7].
3. Ermittlung der optimalen Feature-Kombination für die Detektion der Phrasengrenzen sowohl mit als auch ohne Berücksichtigung der Lautsegmentierung und der phonologischen Repräsentation.

4. Aufstellung eines Sets von Phrasentypen (in Zusammenarbeit mit dem Betreuer), welches als Grundlage für die Detektion der Phrasentypen dienen soll. Für die Erstellung dieses Sets kann auf entsprechende Literatur zurückgegriffen werden [8].

5. Annotierung des Prosodiekorpus entsprechend dem neuen Phrasentypen-Sets.

6. Ermittlung der optimalen Feature-Kombination für die Detektion von verschiedenen Phrasentypen sowohl mit als auch ohne Berücksichtigung der Lautsegmentierung und der phonologischen Repräsentation.

7. Grundsätzliche Durchführbarkeitstests der Algorithmen anhand des manuell korrigierten, deutschen Prosodiekorpus eines männlichen Sprechers.

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

9. 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 [9]), 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] 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.

[3] 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.

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

[5] B. Pfister und R. Beutler. *Sprachverarbeitung 1.* Vorlesungsskript für das Wintersemester 2005/2006, Departement ITET, ETH Zürich, 2005.

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

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

[8] *Duden "Die Grammatik",* 7. Auflage. Bibliographisches Institut. Mannheim, Wien, Zürich, 2005.

[9] 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).

[10] 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 10. April 2007

Literatur

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

[2] A. Kiessling. Detection of phrase boundaries and accents, 1994.

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

[4] Benno Peters und Klaus J. Kohler und Thomas Wesener. Phonetische merkmale prosodischer phrasierung in deutscher spontansprache.

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