

Sprachverarbeitung: Übung 9

Prosodiesteuerung Teil 2: Grundfrequenzsteuerung

Die Aufgabe der vorangegangenen Übung und dieser Übung ist es, eine Prosodiesteuerung für die Sprachsynthese zu realisieren. Eine Prosodiesteuerung setzt die abstrakten prosodischen Elemente der phonologischen Beschreibung eines Satzes in die folgenden physikalischen Grössen um:

- die Lautdauer
- den Tonhöhenverlauf (Grundfrequenzverlauf)
- den Intensitätsverlauf

Die Signalelemente werden bei der Synthese dann entsprechend dieser Grössen modifiziert. Ein Verfahren, diese Eigenschaften bei der Synthese von Sprachsignalen zu steuern, wurde in der Übungen 7 mit der LPC-Analyse-Synthese behandelt.

Nachdem in der vorangegangenen Übung eine Dauersteuerung implementiert wurde, folgt nun in dieser Übung die Implementation der Grundfrequenzsteuerung.

Allgemeine Hinweise

Bei der Grundfrequenzsteuerung geht es um den Zusammenhang zwischen der abstrakten phonologischen Darstellung von Lautsprache und den konkret zu realisierenden Grundfrequenzwerten.

In der Vorlesung wurde ein einfacher Ansatz zur Beschreibung dieses Zusammenhangs vorgestellt, nämlich die Grundfrequenzsteuerung mit einem linearen Ansatz. Dieser Ansatz geht von einer stückweise linearen Approximation des Grundfrequenzverlaufs zwischen den Grundfrequenzwerten in der Mitte der Silbenkerne aus. Der Grundfrequenzwert $\tilde{F}_0(j)$ der j -ten Silbe wird anhand der Deklinationsgeraden und einer Abweichung F_e von der Deklinationsgeraden geschätzt mit

$$\tilde{F}_0(j) = F_a + \frac{j-1}{J} F_d + F_e \quad (1)$$

Dabei ist J die Phrasenlänge in Anzahl Silben, F_a die Grundfrequenz am Anfang der Deklinationsgeraden und F_d die Steigung der Deklinationsgeraden pro Silbe. Die Grössen F_a , F_d und F_e hängen von linguistischen Faktoren ab.

Die in der Vorlesung behandelte Grundfrequenzsteuerung mit linearem Ansatz laut Formel (1) ist ein additives Modell und kann in folgender Form dargestellt werden:

$$\tilde{F}_0 = c_1 p_1 + c_2 p_2 + \dots + c_K p_K = \mathbf{c} \mathbf{p} \quad (2)$$

\tilde{F}_0 ist der Grundfrequenzschätzwert der Silbenkernmitte. \mathbf{c} ist ein Zeilenvektor und beinhaltet die entsprechenden Regelbedingungen. \mathbf{p} ist ein Spaltenvektor und enthält die Regelparameter.

In dieser Übung geht es darum, diesen Ansatz zu implementieren, die notwendigen linguistischen Faktoren zu ermitteln, und schliesslich die Grundfrequenzsteuerung anhand von Beispielsätzen zu testen.

Für das Bestimmen der p_i werden die gleichen Trainingsdaten wie in Übung 8 verwendet. Die Daten können wiederum mit der Funktion `load_training_data` geladen werden. Die Ausgaben dieser Funktion sind der Vektor `freq` mit den Grundfrequenzwerten und die Struktur `F` mit den folgenden Angaben zu jedem der insgesamt 21510 Laute der Trainingsdaten:

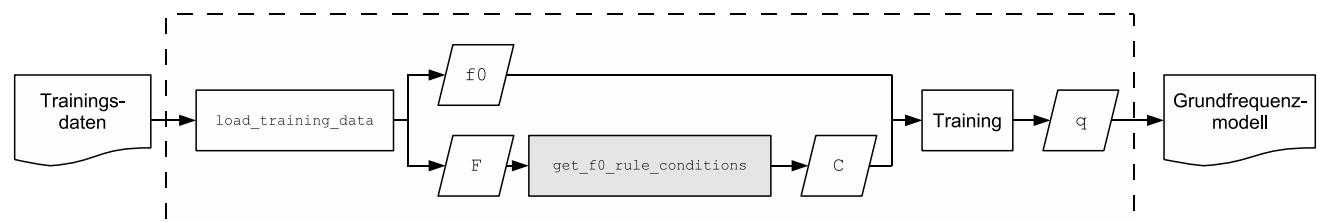
```
F =
    phone_id: [21510x3 char]      (Lautbezeichnung in ETHPA)
    phone_number: [21510x1 double] (Lautnummer innerhalb der Silbe)
    syllable_size: [21510x1 double] (Anzahl Laute in der Silbe)
    phone_position: [21510x1 char] (Lautposition: Ansatz, Nukleus, Koda)
    accent: [21510x1 double]      (Betonungsstärke der Silbe)
    syllable_number: [21510x1 double] (Silbennummer in der Phrase)
    phrase_size: [21510x1 double]  (Anzahl Silben in der Phrase)
    phrase_type: [21510x1 char]    (Phrasentyp)
    phrase_boundary: [21510x1 char] (Phrasengrenze nach der Silbe)
```

Jedes Feld der Struktur `F` ist ein Vektor mit 21510 Elementen. Elemente der Felder mit gleichem Index enthalten somit Angaben für denselben Laut der Trainingsdaten. Eine Beschreibung des Wertebereichs der Felder ist in der Hilfe der Funktion `load_training_data` zu finden. Zur selektiven Ausgabe der Struktur `F` kann die Funktion `dispF` benutzt werden.

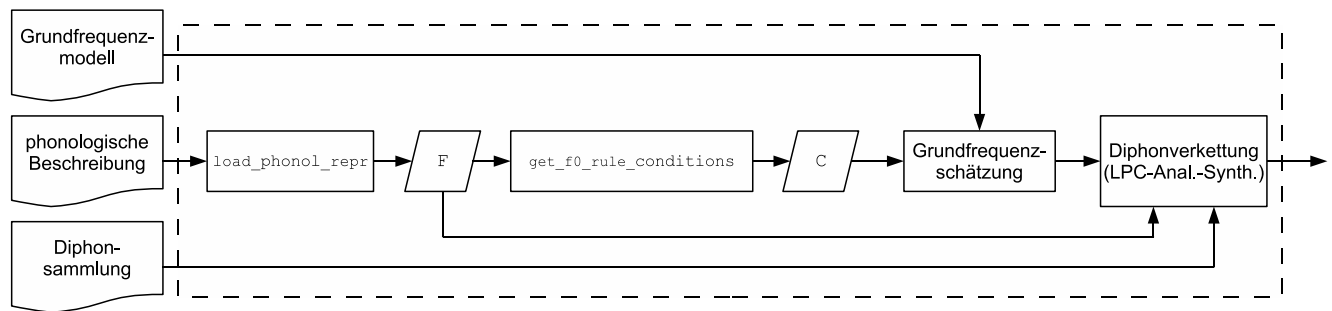
In den folgenden Teilaufgaben soll die Funktion `get_f0_rule_conditions` implementiert werden, mit der festgelegt wird, welche linguistischen Einflussfaktoren in der Grundfrequenzsteuerung berücksichtigt werden sollen. Sie soll pro Silbe den Vektor `c` ermitteln, der angibt, welche der linguistischen Einflussfaktoren auf die Silbe zutreffen. Durch Einsetzen von `c` in Formel (2) kann die Grundfrequenz der Silbe geschätzt werden.

Aufgabe 1: Einfache Grundfrequenzsteuerung

Gegeben ist eine sehr einfache Funktion `get_F0_rule_conditions_frame`, die für jeden in den Trainingsdaten enthaltenen Silbenkern nur berücksichtigt, dass es sich um einen Silbenkern handelt. Sie gibt also für die Trainingsdaten mit insgesamt 21510 Lauten eine `C`-Matrix der Grösse



Figur 1: Blockdiagramm für das Training des Grundfrequenzmodells (entspricht dem Matlab-Skript `create_f0_model.m`)



Figur 2: Blockdiagramm des Matlab-Skripts `synthesize_speech_sig`

$S \times 1$ aus. S ist dabei die Anzahl der in den Trainingsdaten enthaltenen Silben. Da alle Zeilen von C gleich sind, sind auch die nach Formel (1) ermittelten Grundfrequenzen alle gleich. Dies entspricht einer Grundfrequenzsteuerung unter Verwendung der mittleren Grundfrequenz.

Sie können mit dem Matlab-Skript `create_f0_model.m`, das dem Blockdiagramm von Figur 1 entspricht, für die Funktion `get_F0_rule_conditions_frame` das entsprechende Grundfrequenzmodell ermitteln. Das Grundfrequenzmodell wird mit dem Skript `synthesize_speech_sig` für die Sprachsynthese angewendet. Dieses Matlab-Skript ist im Blockdiagramm von Figur 2 dargestellt. Beim Anhören der synthetisierten Signale werden Sie feststellen, dass diese Grundfrequenzsteuerung ziemlich schlecht ist.

Machen Sie nun einen ersten Verbesserungsschritt, indem Sie eine Funktion `[C desc] = get_F0_rule_conditions(F)` schreiben, die als zweiten Einflussfaktor die Silbenposition innerhalb der Phrase berücksichtigt. Dieser Wert soll die Steigung der Deklinationsgeraden berücksichtigen, also den Faktor $(j-1)/J$ in Formel (1) beinhalten.

Sie können nun Ihre neue Funktion `get_F0_rule_conditions` einsetzen, um zuerst mit dem Skript `create_f0_model` das Grundfrequenzmodell zu trainieren und nachher mit dem Skript `synthesize_speech_sig` zu testen. In beiden Skripten werden sowohl die gegebene Funktion `get_f0_rule_conditions_frame` als auch Ihre Funktion `get_f0_rule_conditions` verwendet. Sie erhalten also zwei Grundfrequenzmodelle. Bei der Synthese wird für den gewählten Satz für jedes Grundfrequenzmodell ein separates Signal erzeugt. Zusätzlich wird noch als Vergleich ein Sprachsignal generiert, dessen Grundfrequenzen mit dem Sprachsynthesensystem SVOX geschätzt worden sind.

Hören und sehen Sie sich mit dem Skript `synthesize_speech_sig` einige der Beispielsätze an. Betrachten Sie ausserdem die ermittelten Regelparameter p . Was sagen sie aus? Welche Unterschiede zur Lautdauersteuerung erkennen Sie?

Wenn das Erzeugen und Anhören der Beispielsätze geklappt hat, können Sie zur nächsten Aufgabe gehen.

Aufgabe 2: Untersuchen der linguistischen Einflussfaktoren

Wie bei der Lautdauer gibt es auch bei der Grundfrequenz viele linguistische Einflussfaktoren (siehe Hilfe zur Funktion `syllable_has_properties`), welche die Grundfrequenz potentiell beeinflussen. Beachten Sie, dass es sich bei einigen dieser Einflussfaktoren um Eigenschaften der

Silbe selbst handelt (z.B. “Silbe hat Akzent 2. Grades”), bei andern um Eigenschaften der Phrase, in welcher die Silbe steht (z.B. “Silbe in einer zweisilbigen Phrase”).

Um herauszufinden, welche linguistischen Einflussfaktoren sich erheblich auf die Grundfrequenz auswirken, können Sie die Funktion `f0_histogram` verwenden. Mit dieser Funktion kann die Auswirkung eines Einflussfaktors auf die drei Grössen F_a , F_d und F_e beobachtet werden.

Zur besseren Beurteilung der Stärke des Einflussfaktors berechnet die Funktion auch die Fisher-Distanz.¹ Diese ist umso grösser, je stärker sich die statistischen Verteilungen der Lautdauern mit bzw. ohne einen betrachteten Einflussfaktor unterscheiden. Sie ist null, wenn der Einflussfaktor die Lautdauer gar nicht beeinflusst.

Ermitteln Sie durch Überlegen und Experimentieren (mit `f0_histogram`) die wichtigsten linguistischen Einflussfaktoren.

Aufgabe 3: Grundfrequenzsteuerung mit weiteren Einflussfaktoren

Erweitern Sie nun die Funktion `get_f0_rule_conditions` so, dass die in Aufgabe 2 gefundenen linguistischen Einflussfaktoren in der Grundfrequenzsteuerung berücksichtigt werden. Verwenden Sie auch jeweils einen eigenen Einflussfaktor für die unterschiedliche Deklination von progredienten Phrasen und Terminalphrasen. Testen Sie die verbesserte Grundfrequenzsteuerung wiederum mit den Skripten `create_f0_model` und `synthesize_speech_sig`.

Das Skript `create_f0_model` gibt den mittleren Fehler (RMS) zwischen den gemessenen Grundfrequenzen (der Trainingsdaten) und den mit dem Grundfrequenzmodell geschätzten Werten an. Wenn für die Grundfrequenzsteuerung die 10 bis 15 wichtigsten linguistischen Einflussfaktoren berücksichtigt werden, dann wird der mittlere Fehler weniger als 10 Hz betragen.

Aufgabe 4: Verwendung der eigenen Lautdauersteuerung

Bis jetzt wurde für die synthetisierten Signale die Lautdauersteuerung des Sprachsynthesystems SVOX verwendet. Sie können zur Erzeugung der Signale sowohl die durchschnittliche Lautdauer als auch Ihre eigene Lautdauersteuerung aus Übung 8 verwenden, indem Sie die entsprechenden Funktionen für die Regelbedingungen zusammen mit der entsprechenden `mat`-Datei mit den Modellparametern in das aktuelle Verzeichnis kopieren. Für die Dauersteuerung, die allen Lauten die durchschnittliche Dauer zuordnet, sind das die Dateien `get_dur_rule_conditions_frame.m` und `DurModel0.mat`, für Ihre eigene Lautdauersteuerung die Dateien `get_dur_rule_conditions.m` und `DurModel1.mat`. Beurteilen Sie das Ergebnis bei Verwendung Ihrer eigenen Lautdauersteuerung.

¹Die Fisher-Distanz ist ein Mass für die Verschiedenheit von zwei Wahrscheinlichkeitsverteilungen. Sie ist definiert als $d_f = (m_0 - m_1)^2 / (\sigma_0^2 + \sigma_1^2)$, wobei m_0 und m_1 die Mittelwerte und σ_0^2 und σ_1^2 die Varianzen der beiden Verteilungen sind.