Fachhochschule Köln
University of Applied Sciences Cologne

Köln, August `06

Kurzfassung (Paper) der Bachelor-Arbeit

Entwicklung eines Systems zur automatischen Extraktion von Merkmalsvektoren für die prosodische Steuerung von Sprachsynthese-Systemen

Andreas Lehmann
Lehmann.Andreas@web.de

ABSTRACT

Speech synthesis becomes more and more important with regards to make situations in everyday life easier to handle for handicaped people or to make informations accessible to them. But non-handicaped people also get in touch with speech synthesis. Science is engaged in speech synthesis for a long time, and is still far away from breaking up. If one deals with speech synthesis, one has to take the natural speech into account, especially what makes speech sound natural. That brings us to the aspect of *prosody*.

This paper gives a short overview of my thesis. In this work, I depict why prosody is important for synthetic speech, and what the elements of prosody are. The focus of the work is the design of a system, that extracts prosodic features from speech signals and makes them useable for the modification of synthetic speech.

KEYWORDS

Prosody, Prosodie, speech synthesis, Sprachsynthese, prosodic features, prosodische Merkmalsvektoren, speech signal processing, Sprachsignalverarbeitung, voicing detection, Anregungsartbestimmung, fundamental frequency extraction, Grundfrequenzbestimmung, speech signal segmentation, Sprachsignalsegmentierung, parametrisation of fundamental frequency, Grundfrequenz-Parametrisierung

1 Einleitung

Synthetische Sprache nimmt schon jetzt einen wichtigen Platz ein. An Bedeutung gewinnt sie noch im Hinblick auf Barrierefreiheit, da Sprachsynthese-Systeme ein mächtiges und wichtiges Werkzeug sind, um benachteiligten Menschen den Zugang zu Informationen zu ermöglichen oder den Umgang mit Computern zu erleichtern. Die möglichen Einsatzbereiche sind aber auch abgesehen davon sehr zahlreich, werden aber bisher kaum ausgeschöpft, da die *Natürlichkeit* der künstlichen Sprache noch größtenteils unbefriedigend ist. Das ist noch ein großes Manko. Somit ist auch die Motivation dieser Arbeit gegeben.

Synthetische Sprache soll natürlich klingen und verständlich sein. Um die Natürlichkeit zu erhöhen, muss also natürliche Sprache genauer untersucht werden. Was macht die Natürlichkeit der Sprache aus?

Prosodie ist bedeutend für die Natürlichkeit und die Verständlichkeit der Sprache. Was darunter zu

verstehen ist, wird in folgendem Abschnitt geklärt.

Hauptaugenmerk der Arbeit lag auf der Entwicklung eines Systems, das aus natürlicher Sprache prosodische Merkmale extrahiert und diese verarbeitet, sodass sie für die Sprachsynthese nutzbar sind. Neben den erforderlichen Grundlagen, wird in diesem Dokument das System beschrieben ohne zu sehr auf Details einzugehen.

2 Grundlagen

2.1 Prosodie

Prosodie beschreibt rhythmische und melodische Phänomene in der Sprache, die von Sprecher zu Sprecher und von Sprache zu Sprache unterschiedlich sind. Sie trägt also charakteristische Züge.

Es gibt genau drei Ebenen, auf denen Prosodie repräsentiert und untersucht werden kann. Bedeutend für diese Arbeit sind die *akustische* und die *linguistische* Ebene.

Auf akustischer Ebene wird Prosodie beschrieben durch *Dauer*, *Energie* und *Grundfrequenz*. Diese Beschreibung ist im Sinne der Signalverarbeitung, da alles direkt messbare Größen sind. Dauer beschreibt die Dauer einer akustischen Einheit, z.B. eines Lautes, Energie beschreibt die nötige Energie zur Erzeugung eines Signals. Die Grundfrequenz ist der wichtigste Parameter. Es ist die Frequenz, in der die Stimmbänder bei der Erzeugung des Sprachsignals schwingen (nur bei *stimmhaften* Lauten, daneben gibt es noch *stimmlose*, bei denen keine Schwingung vorliegt). Diese Merkmale bilden die Merkmalsvektoren, die es zu extrahieren gilt.

Auf linguistischer Ebene erfolgt die Beschreibung der Prosodie anhand von *Intensität, Quantität* und *Intonation*. Quantität beschreibt die zeitliche Dauer sprachlicher Einheiten. Die Vokaldauer ist im Deutschen teilweise distinktiv (z.B. Maße – Masse). Intensität bedeutet Betonung oder auch Akzentuierung. Durch Akzentuierung können bestimmte Informationen in den Fokus gestellt werden oder die Bedeutung eines Satzes geändert werden. Somit ist auch die Intensität distinktiv. Man unterscheidet zwischen *Wort-* und *Satzakzent*. Wortakzent ist die Betonung einer Silbe innerhalb eines Wortes (lexikalischer Akzent), der Satzakzent die Betonung eines Wortes in einem Satz (inhaltlicher Akzent). Die Intonation beschreibt die melodischen Phänomene, man redet auch von Sprachmelodie. In natürlicher Sprache wird anhand der Intonation unterschieden zwischen Frage- und Aussagesätzen.

Die Merkmale auf beiden Ebenen hängen zusammen. Es reicht also nicht aus, nur die akustischen Merkmale zu extrahieren, sondern muss diese auf die linguistischen Merkmale beziehen für eine komplette Beschreibung der Prosodie. Abbildung 1 veranschaulicht den Zusammenhang.

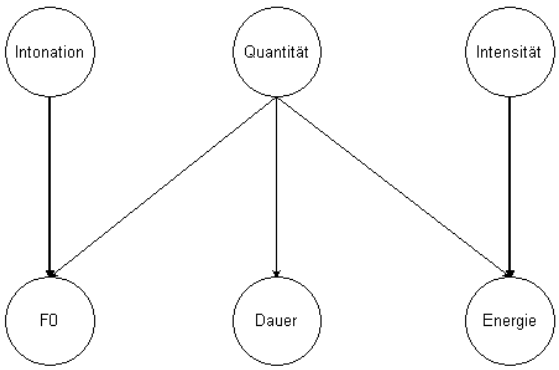

Abbildung 1: Zusammenhang der Merkmale

2.2 Sprachsynthese-Systeme

Sprachsynthese-Systeme dienen der computerbasierten Erzeugung künstlicher Sprache. Man kann sie kategorisieren in *text-to-speech* Systeme und *concept-to-speech* Systeme. Beim ersten wird aus einem beliebigen Text ein Sprachsignal erzeugt und dabei mit einer passenden Prosodie versehen. Dazu muss der Text vorher untersucht werden, und die Aussprache der Wörter durch Ausspracheregeln oder ~wörterbücher bestimmt werden. Diese Systeme sind unabhängig vom Kontext einsetzbar. Da diese Systeme kein Wissen über Inhalt und Zusammenhang dessen was sie als

Text einlesen besitzen, müssen sie eine möglichst neutrale, allgemeine Prosodie erzeugen. Bei concept-to-speech Systemen liegt kein zu synthetisierender Text vor, sondern nur eine interne Information, die es darzustellen gilt. Diese Systeme müssen also zuerst Text generieren. Da die Domäne eher beschränkt ist, liegen solchen Systemen komplette Informationen zu Pragmatik, Semantik und Syntax vor.

Bei den Techniken zur Sprachsynthese haben sich zwei Methoden etabliert, die nur kurz angerissen werden. Zum einen ist es die *regelbasierte* Methode. Bei diesem Ansatz wird keinerlei vorliegendes Sprachmaterial benötigt. Die künstliche Sprache wird anhand eines komplexen Regelsatzes erzeugt, der die Artikulationsgesten bei der Spracherzeugung nachbilden soll. Dieser Ansatz ist zwar flexibel, aber bietet keine akzeptable Natürlichkeit. Die Alternative dazu ist die *konkatenative* Synthese. Das Prinzip ist, dass einzelne Sprachsegmente miteinander verkettet werden. Es soll das bestmögliche Ergebnis mit einer möglichst kleinen Menge an Segmenten erreicht werden. In der Praxis wird daher meist auf *Diphone* zurückgegriffen. Ein Diphon besteht jeweils aus zwei halben Phonen. Daneben gibt es noch Ansätze mit variabler Segmentgröße (*automatic unit selection synthesis*).

Die Verkettung und Modifizierung dieser Segmente basiert auf Sprachmodellen. Hervorzuheben ist z.B. das PSOLA-Verfahren (*Pitch-Synchronous Overlap Add*), das die Modifizierung von Dauer, Energie und Grundfrequenz ermöglicht. Da die verschiedenen PSOLA-Verfahren komplex sind, kann im Rahmen dieses Dokuments nicht näher drauf eingegangen werden.

3 System zur Extraktion der Merkmale

Oberstes Ziel des Systems ist es, die *Verständlichkeit* und *Natürlichkeit* synthetischer Sprache zu verbessern, indem natürliche Sprache analysiert wird. Im Hinblick darauf haben sich folgende Anforderungen ergeben:

- Berücksichtigung von Wort- und Satzakzent
- Unterscheidung von Satzmodi
- Berücksichtigung natürlichsprachlicher Phänomene

Bezüglich eines Einsatzes am Computer ergeben sich noch weitere Anforderungen:

- Konfigurierbare Prosodie
- Allgemeine Schnittstelle des prosodischen Moduls
- Unterschiedliche akustische Hervorhebung schriftlicher Phänomene wie kursive oder fette Schrift

Allerdings konnten nicht alle Anforderungen komplett Berücksichtigt werden.

Abbildung 2 zeigt den Aufbau des entwickelten Systems.

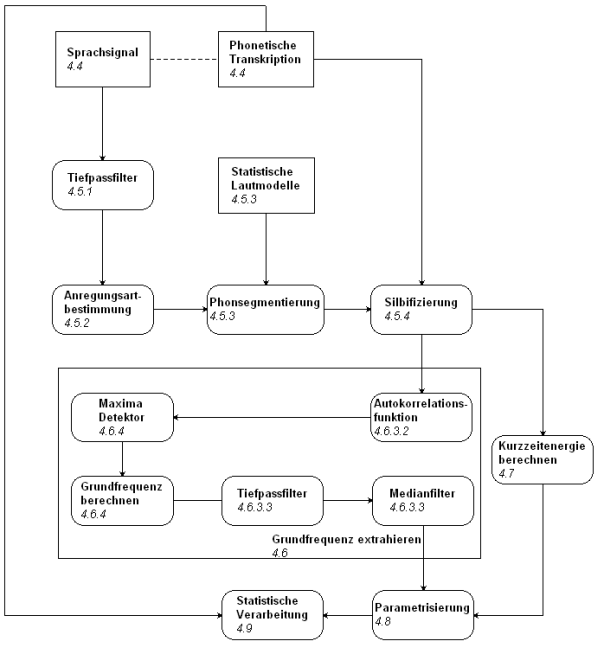

Abbildung 2: Architektur des Systems

3.1 Eingabedaten

Neben dem Sprachsignal als PCM-Datei mit 16Bit und 44.100Hz liegt eine phonetische Transkription vor in *SAMPA* (Speech Assessment Methods Phonetic Alphabet), die im weiteren

Verlauf benötigt wird. Eine solche Transkription ist z.B.:

Ich bin ein Ge nie!
[(I,C)] [(b,I,n)] [(a,I,n)] {[(Z,e)/(n,i:)]}!

SAMPA-Symbole werden durch „ , “ getrennt, Wörter durch „[]“ und Silben durch „()“ eingegrenzt. Die Silbe nach „/“ trägt den Wortakzent, das Wort in „{}“ den Satzakzent. Der Satzmodus wird durch „!“ oder „?“ markiert.

3.2 Vorverarbeitung

Filterung

Bevor die eigentlichen Merkmale extrahiert werden können, muss das Sprachsignal verarbeitet werden. Dazu wird das Signal zuerst *tiefpassgefiltert* mit einer Grenzfrequenz von 3.4Khz, und somit Frequenzen aus dem Signal entfernt, die nicht im Frequenzbereich von Sprache liegen. Dafür wurde ein *Typ I Chebyshev-Filter 9. Ordnung* gewählt.

Anregungsartbestimmung

Nach der Filterung wird eine Trennung von Sprache/Pause vorgenommen und gleichzeitig unterschieden zwischen *stimmhafter* und *stimmloser* Anregung. Dazu stehen verschiedene Techniken zur Verfügung. Ich habe mich für eine Schwellenwertanalyse entschieden, da diese am leichtesten zu realisieren ist, und bei guter Parameterwahl eine sichere Unterscheidung erlaubt. Als Unterscheidungskriterien ist die Wahl auf die *Zero-Crossing-Rate* (Anzahl an Nulldurchgängen des Signals), die Kurzzeitenergie und die *4Hz Modulationsenergie* gefallen. Die ZCR ist bei stimmlosen Lauten größer als bei stimmhaften, bei Energie ist es umgekehrt. Zur Abtrennung von Pausen ist vor allem die 4Hz Modulationsenergie wichtig. Diese ist bei Sprache deutlich größer. Die Begründung ist, dass die Silbenrate ungefähr 4Hz beträgt. Im ersten Schritt wird das Signal gefenstert (*Hammingfenster*, Länge 40ms, Verschiebung 10ms) und die drei Parameter werden berechnet.

Die ZCR berechnet sich durch:

$$ZCR_i = \sum_{n=0}^{1763} |sgn\,[x_i(n)] - sgn\,[x_i(n-1)]|$$

$$sgn\,[x(n)] = 1 \Rightarrow x(n) \geq 0; -1 \Rightarrow x(n) < 0$$

Für die Kurzzeitenergie wird folgende Formel benutzt:

$$E_i = \sum_{n=0}^{1763} |x_i(n)|^2$$

Für die 4Hz Modulationsenergie wird das Signal zuerst gefiltert mit einer Grenzfrequenz von 4Hz, die Energie des gefilterten Signals berechnet und durch die Energie des nicht gefilterten Signals geteilt. Nachdem das für das gesamte Signal geschehen ist, wird für jeden Parameter der Schwellenwert berechnet. Dazu wird der *Median* der Wertelisten bestimmt. Im nächsten Schritt werden die berechneten Werte mit den Schwellenwerten Verglichen und das Segment wird als stimmhaft/stimmlos oder Pause etikettiert. Dafür gilt:

- Stimmhaft: ZCR<=Schwelle & Energie>=Schwellenwert & Modulationsenergie>=Schwelle
- Stimmlos: ZCR>Schwelle & Energie<Schwellenwert & Modulationsenergie>=Schwelle
- Pause: ZCR<Schwelle & Energie<Schwellenwert & Modulationsenergie<Schwelle

Segmentierung

Nach der Anregungsartbestimmung wird das Signal in einzelne Phone und Silben segmentiert. Bei der Phonsegmentierung wird auf das frei verfügbare System *MAUS* (Münchener Automatisches Segmentationssystem) zurückgegriffen, das mit Hilfe von *Hidden-Markov-Modellen* und der *Viterbi-Suche* die Phongrenzen im Sprachsignal durch Alignment detektiert und in einer Liste speichert. Bei HMM handelt es sich um *stochastische Modelle*, die für jedes der SAMPA-Symbole gebildet werden. Die HMM werden anhand der phonetischen Transkription nacheinander geschaltet. Mehr zu MAUS ist unter

http://www.phonetik.uni-muenchen.de/Forschung/Verbmobil/VM14.7.html zu finden.

Die Silben werden anhand der phonetischen Transkription und den vorliegenden Phongrenzen gebildet.

3.3 Grundfrequenzbestimmung

Die Bestimmung der Grundfrequenz (F0) ist sehr wichtig für die prosodische Analyse, und gleichzeitig auch eine der am schwersten zu lösenden Aufgaben. Es gibt etliche mögliche Fehlerquellen, als Beispiel sei genannt, dass Sprache nichtstationär ist und daher rasch drastisch variieren kann. Zur Bestimmung der Grundfrequenz wurde ein Verfahren nach dem Prinzip der *Kurzzeitanalyse* gewählt. Die Bestimmung erfolgt mit der normierten *Autokorrelationsfunktion*, gegeben durch

$$NAKF(k) = \frac{\sum_{n=0}^{N-1-M} x(n) * x(n+k)}{\sum_{n=0}^{N-1-M} x(n+k) * x(n+k)}$$

N gibt die Länge des Signals an und *M* die Größe des Analysefensters. Diese erlaubt Aussagen über die Ähnlichkeit eines Signals zu seiner um *k* verschobenen Version. Das Sprachsignal wird silbenweise gefenstert (*Hammingfenster*, Länge 40ms damit 2-3 Grundperioden min. im Fenster, Verschiebung 10ms) und für dieses Fenster die Autokorrelationskoeffizienten berechnet. Die Verschiebung *k* beträgt dabei je 1 Sampel. Wurden für das gesamte Sprachsignal die Werte ermittelt, müssen lokale Maxima gesucht werden, da die F0 (bzw. T0 = Grundperiode) eines Segments anhand des größten Maximas für eine Verschiebung *k* ungleich 0 zu bestimmen ist. Durch die Position dieses Maximums in der Liste wird T0 berechnet für das ite gefensterte Signalelement.
*T0 = n * 40/1764ms* -> n=Position des Maximums Aus T0 wird dann F0 berechnet durch *F0 = 1/(T0/1000) Hz.*

Somit erhält man mehrere Schätzwerte für die F0 des Sprachsignals (ca. 21 pro Silbe). Diese Listen werden aufgrund zu erwartender Messfehler und Ungenauigkeiten gefiltert mit einem *Medianfilter* der Länge 5 und einem *Butterworth-Tiefpassfilter* 9. Ordnung und einer Grenzfrequenz von 800Hz.

3.4 Kurzzeitenergiebestimmung

Im Gegensatz zu der Bestimmung der Grundfrequenz, ist die Bestimmung der Kurzzeitenergie simpel zu realisieren. Die Kurzzeitenergie des Signals im Fenster (*Hammingfenster*, Länge 40ms, Verschiebung 10ms) berechnet sich wie schon erwähnt anhand der Formel:

$$E_i = \sum_{n=0}^{1763} |x_i(n)|^2$$

Dies wird für das gesamte Signal berechnet.

4.0 Parametrisierung

Da sowohl Grundfrequenz- als auch Energiewerte als absolute Werte vorliegen, werden diese Konturen parametrisiert. Dadurch erhält man relative Werte, was eine spätere Weiterverarbeitung erleichtert.

Bei der Wahl der Parameter ist darauf zu achten, dass Extrema bei der F0 gut abgebildet werden, da diese perzeptiv bedeutend sind. Außerdem darf nach einer Parametrisierung keine wichtige Information verloren gehen, wie z.B. Verlauf der F0 je nach Satzmodus etc. Folgende Parameter wurden daher zur Parametrisierung von F0 gewählt:

- Durchschnittliche F0 eines Phones
- Relative Position *dt* von Extrema zum Phonbeginn in %
- Relative Abweichung *dF0* der Extremaamplituden von der durchschn. F0 des Phones in %
- Relative Abweichung *dF0* der Amplituden an den Phongrenzen von der durchschn. F0 des Phones in %
- Verhältnis der durchschn. F0 des ersten Wortes zum letzten Wort in %

Abbildung 3 verdeutlicht die Parameter. Für die Parametrisierung der Energie genügen:

- Durchschnittliche Energie eines Phones/Silbe/Wortes

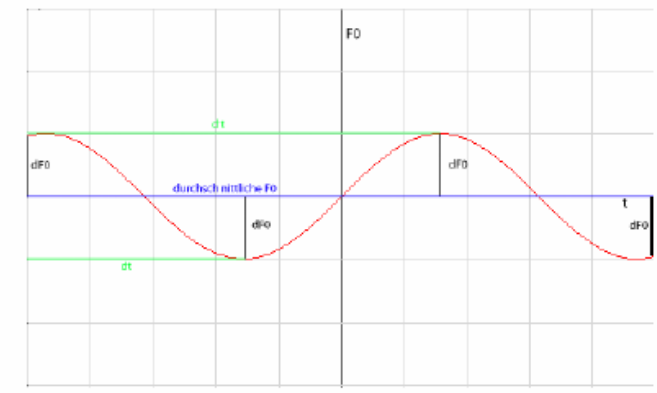

Abbildung 3: F0-Parameter auf Phonebene

Die Berechnung der einzelnen Parameter wird hier nicht weiter aufgeführt. Sind alle Parameter berechnet, werden die Parameter in Dateien abgespeichert nach folgendem Aufbau:

Beispiel.dat

phonetische Transkription;
Anzahl Wörter;
Anzahl Silben;
Anzahl Phone;
Satzmodus;

$F0_{Verhaeltnis}$;

 # Phone
1,SAMPA-Symbol,$\varnothing F0_{Phon}$,$\varnothing E_{Phon}$,Anzahl Extrema,[Δt,$\Delta F0$]...[Δt,$\Delta F0$],
[$\Delta F0_{Phongrenze}$,$\Delta F0_{Phongrenze}$];
.
.
.
 # Silben
1,[Nummern der Phone],Zeichen für Akzentuierung (0 - kein Akzent, 1 - Akzent), $\varnothing E_{Silbe}$;
.
.
.
 # Worte
1,[Nummern der Silben],Zeichen für Akzentuierung (0 - kein Akzent, 1 - Akzent), $\varnothing E_{Wort}$;

5.0 Weiterverarbeitung der Daten

Die parametrisierten Daten benötigen noch eine Weiterverarbeitung, um auch für die Sprachsynthese nutzbar zu sein. Aus diesem Grund werden die Daten unter linguistischen Gesichtspunkten statistisch ausgewertet. Es werden für *Satzmodus, Satz-* und *Wortakzent* signifikante Verhalten der Energie- oder Grundfrequenzkonturen betrachtet, und durchschnittliche Werte berechnet. Dazu muss ein Parser aus allen parametrisierten Daten die jeweils benötigten einlesen. Da je nach Satzmodus die F0-Werte innerhalb der letzten Silbe steigen oder fallen, muss so z.B. die durchschnittliche, relative Abweichung der F0 vom ersten zum letzten Phon einer Silbe berechnet werden. Bei der Untersuchung der Akzente ist auf Energieverhältnisse, Anstieg und Abfall der F0 bei Maxima und die relative Position der Maxima zu achten. Eine genauere Beschreibung ist der Arbeit zu entnehmen.

Die so errechneten Werte bilden eine Art Regelsatz, der in ein *XML-Dokument* abgebildet wird, damit Synthese-Systeme darauf zugreifen können. Ein Auszug aus dem XML-Dokument:

```
<prosodieRegelsatz>

  <satzmodi>
    <fragesatz>
      <abweichungF0>durchschn. delta f0</abweichungF0>
    </fragesatz>
    <aussagesatz>
      <abweichungF0>durchschn. delta f0</abweichungF0>
   </aussagesatz>
</satzmodus>
```

6.0 Fazit

Im Rahmen der Arbeit konnte ein System entwickelt werden, dass aus natürlicher Sprache prosodische Merkmalsvektoren extrahiert. Allerdings ist die Realisierung noch verbesserungswürdig. Ansatzpunkte sind vor allem die Grundfrequenzbestimmung und die Parametrisierung. Die Weiterverarbeitung der Daten ist im Hinblick auf die prosodische Steuerung von Sprachsynthese-Systemen auch zu überarbeiten.

Prosodie in synthetischer Sprache ist kein triviales Thema und auch noch nicht perfektioniert. Ansatzpunkte zur Verbesserung gibt es genug, und diese müssen auch angegangen werden. Dabei ist die Untersuchung der natürlichen Sprache unumgänglich. Diese Arbeit bietet da nur ein Grundgerüst, auf das aufgebaut werden kann.

7.0 Literatur (Auszug)

(1) Wolfgang Hess; *Digitale Signalverarbeitung*, 1. Auflage, Stuttgart, B.G. Teubner 1998

(2) Wolfgang Hess; *Grundlagen der Phonetik – Kapitel 4*, http://www.ikp.uni-bonn.de/dt/lehre/materialien/prosodie/gph_4f_4p.pdf

(3) Wolfgang Hess; *Prosodie – Kapitel 2*, http://www.ikp.uni-bonn.de/dt/lehre/materialien/prosodie/pros_2f.pdf

(4) Thierry Dutoit; *An Introduction to Text-to-Speech Synthesis*, 1. Auflage, Dordrecht, Kluwer Academic Publishers 1997

(5) Thierry Dutoit; *TTSBOX 1.0 Documentation*, http://tcts.fpms.ac.be/projects/ttsbox/ttsbox1.0.pdf

(6) Daniela Steinbrecher ; *SAMPA – Deutsch*, http://coral.lili.uni-bielefeld.de/Classes/Winter95/Grundkurs/grundkur/node22.html.

(7) Sassan Ahmadi; *Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm,* *h*ttp://www.ece.uvic.ca/499/2004b/group11/publications/00759042.pdf