Ökonomische Repräsentation natürlicher Intonationskonturen

Joachim Mersdorf, Hans Kruschke, Oliver Jokisch, Diane Hirschfeld***
**RU Bochum, Lehrstuhl für Allgemeine Elektrotechnik und Akustik*
email: mersdorf@ika.ruhr-uni-bochum.de
***TU Dresden, Institut für Akustik und Sprachkommunikation*
email: diane@eakss2.et.tu-dresden.de

Für verschiedene Anwendungsfälle (Intonationsforschung, Prominenzdetektion in der Spracherkennung, Datenbeschreibung und Kategorisierung für die Sprachsynthese), ist es wünschenswert, natürliche prosodische Parameterverläufe ökonomischer darzustellen bzw. nur die linguistisch relevanten Merkmale zu extrahieren und aufzubewahren.
Im vorliegenden Artikel wird ein Verfahren zur Grundfrequenzglättung und -interpolation vorgestellt. Die Nützlichkeit des Verfahrens wird für drei Anwendungen demonstriert.

0 Problemstellung

Für die Untersuchung und statistische Beschreibung natürlicher Grundfrequenzkonturen ist es wünschenswert, ein stetiges und mehrfach differenzierbares Signal als Ausgangspunkt zur Verfügung zu haben.
Die Messung der Grundfrequenz mit gängigen Verfahren liefert jedoch Konturen mit Meßfehlern (typisch sind Grundfrequenzdoppler oder -halbierungen). Weiterhin ist der Grundfrequenzverlauf überlagert von Mikrointonationseffekten und kurzzeitigen Ausreißern. Teile der f0-Kontur sind durch Wechsel von stimmhafter zu stimmloser Anregung maskiert, indem f0-Werte zu 0 gesetzt werden. Übergangsbereiche sind Orte starker Mikrointonation, was eine einfache Interpolation solcher Bereiche erschwert.
Für Untersuchungen im Bereich der Grundfrequenzmodellierung bzw. zur Datenkategorisierung ist jedoch eine Analyse der langsamen Grundfrequenzänderungen (bis 20 Hz) von Interesse, die syntaktische oder semantische Ursachen haben. Da mikroprosodische Effekte lauttypisch sind, und somit in allen makroprosodischen Varianten in gleicher Weise auftreten, enthalten sie diesbezüglich keine zusätzlichen Informationen.

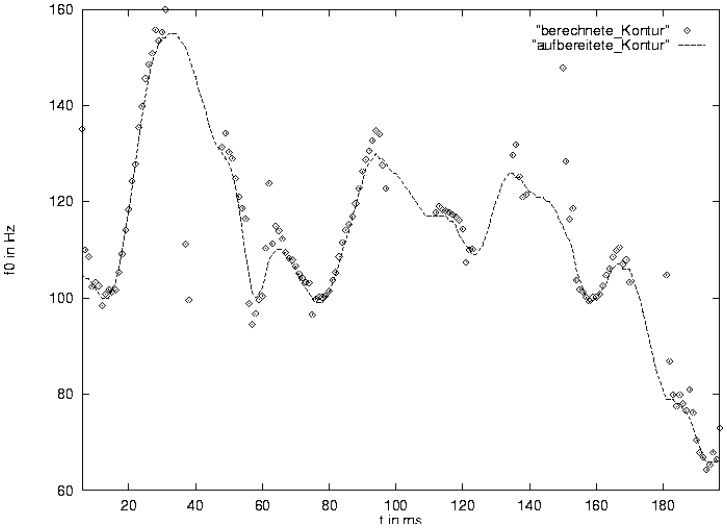

Bild 1 - Vergleich der geglätteten und originalen f0-Kontur

1 Aufbereitung der f0-Kontur

Der Weg zu geglätteten und stetigen Grundfrequenzverläufen führt über die Berechnung und Glättung der Rohkontur und die Interpolation stimmloser Abschnitte.
Zur Berechnung der Grundfrequenz können Standardverfahren verwendet werden. Die nachfolgenden Applikationen nutzen die Funktion *get_f0* des Programmpakets *esps*.

1.1 Glättung des Grundfrequenzverlaufs

Die wissensbasierte Glättung der Rohkontur korrigiert die für die Grundfrequenzberechung typischen Meßfehler mit Verfahren der Signalverarbeitung. Dabei werden folgende Eigenschaften des Grundfrequenzverlaufs vorausgesetzt:

1) Die Grundfrequenz gesprochener Sprache ändert sich innerhalb stimmhafter Signalabschnitte weitgehend stetig (d.h. keine sprunghaften Änderungen).
2) Der Grundfrequenzverlauf ist im wesentlichen immer streng monoton steigend oder fallend. Nahezu konstante Werte ergeben sich lediglich kurzzeitig ($t < 30$ ms) und nur in Bereichen lokaler Extrema der f0-Kontur.
3) Die Grundfrequenz verläuft innerhalb eines begrenzten Bereiches, z.B. zwischen 50 - 500 Hz (vgl. Hess 1983).

Der Glättungsalgorithmus umfaßt eine Sammlung von Algorithmen, die in ihrer Reihenfolge beliebig variiert und auch wiederholt werden können. Grundsätzlich werden zunächst Ausreißer beseitigt und anschließend wird eine Mittelwertfilterung durchgeführt.
Zur Korrektur von Ausreißern werden mittels eines Minimum- und eines Maximum-Filters f0-Werte gelöscht, die außerhalb eines vorgegebbaren Bereiches liegen. Eine weitere Funktion detektiert und beseitigt Sprünge zu Harmonischen und Subharmonischen der Grundfrequenz. Stimmhafte Bereiche, die aus weniger als einer vorgebbaren Anzahl von Werten bestehen, werden gelöscht, d.h. als stimmlos betrachtet. Stimmhafte Bereiche, deren Gesamtmittelwert unterhalb eines Schwellwertes gegenüber dem Konturmittelwert liegt, können ebenfalls gelöscht werden.
Besondere Aufmerksamkeit erfahren stimmhaft-stimmlos-Übergänge, da bekannt ist, daß die Algorithmen zur Berechnung der Grundfrequenz in diesem Bereich besonders fehleranfällig sind. Wird für aufeinanderfolgende Meßwerte in diesem Bereich ein unzulässiger Anstieg festgestellt, wird der zum Rand gelegene Wert gelöscht.
Innerhalb stimmhafter Bereiche erfolgt eine Korrektur von Ausreißern, indem einzelne Werte, die stark vom Umgebungsmittelwert abweichen, durch diesen ersetzt werden.
Das eigentliche Glätten der f0-Kontur erfogt durch ein adaptives gleitendes Mittelwertfilter. Da der Mittelwert durch Ausreißer verfälscht wird, ist deren vorherige Bereinigung dringend notwendig.

1.2 Interpolation

Bei der Interpolation wird von der gestaltpsychologisch motivierten Annahme ausgegangen, daß die intonatorische Geste eines Sprechers auf kognitiver Ebene kontinuierlich und über einen längeren Zeitraum geplant ist. Ihr Vorhandensein und ihre Meßbarkeit im Sprachsignal wird nur durch die zeitweilige Umschaltung in einen stimmlosen Anregungsmechanismus unterbrochen.
Zur Interpolation kürzerer stimmloser Abschnitte (bis 100 ms) wird die kubische Spline-Funktion verwendet (Sedgewick, 1992). Längere

stimmlose Abschnitte (bis 200 ms) werden linear interpoliert. Ab einer konfigurierbaren Dauer werden stimmlose Abschnitte nicht mehr interpoliert, da davon ausgegangen wird, daß es sich um Phrasengrenzen oder Sprechpausen handelt.
Bedeutsam für die Spline-Interpolation ist die Wahl der Stützstellen. Dafür werden aus dem bereits geglätteten Kurvenverlauf jeweils zu beiden Seiten der Randpunkt des stimmhaften Bereichs, der vorhergehende bzw. nachfolgende Extremwert und der von der diese beiden Punkte verbindenden Geraden am weitesten entfernte Punkt der f0-Kontur verwendet. So werden Interpolationswerte berechnet, die gut dem allgemeinen Verlauf der f0-Kontur angepaßt sind. In Bild 1 ist ein beispielhaftes Glättungs- und Interpolationsergebnis dargestellt.

2 Anwendungsbeispiele

2.1 Datengetriebene Grundfrequenzmodelle

Die ökomische und korrekte Repräsentation von f0-Konturen spielt eine fundamentale Rolle bei einer automatischen Generierung von Intonationsverläufen, z.B. in der Sprachsynthese. Das Grundkonzept der Überlagerung akzentbasierter und phrasentypischer Komponenten, ist in der Regel einheitlich. Ausgehend von der Güte der Ergebniskonturen, lassen sich verschiedene Gruppen von Intonationsmodellen klassifizieren:

(1) humaner Intonation ähnelnde f0-Konturen (z.B. nichtlineare, regelbasierte Ansätze wie bei *Fujisaki[1]*: Die funktionale Abbildung fungiert als artikulatorisch motiviertes *Stützkorsett*. Eine nachträgliche f0-Glättung ist unnötig.

(2) f0-Konturen, die perzeptiv identisch zu human produzierten Intonationsverläufen sind (z.B. *Adriaens, 1993*). Keine nachträgliche f0-Glättung.

(3) streuende f0-Werte sowie artikulatorisch unmotivierte *Ausreißer* (typisch für statistische oder datengetriebene Modelle wie *ANN nach Jokisch, 1998*). Die f0-Kontur muß nachträglich geglättet werden.

(4) *geglättete* f0-Ergebniskonturen, da bereits die zugrundeliegenden f0-Trainingsdaten entsprechende Eigenschaften besitzen.

Ansätze nach (3) erreichen zwar nominell eine variable Intonation, bedingen aber eine komplexe Modell-Topologie (z.B. für Stimmhaft-stimmlos-Phasen) und umfangreiche Trainingsstichproben (Größenordnung mehrere hundert bis tausend Phrasen). Eine nachträgliche Glättung zerstört u.U. beabsichtigte, mikroprosodische Effekte.

	Pit*Net*2.0 (Jokisch , 1998)	Pit*Net*3.0 (geglättete Konturen)
Netztyp	Elman (rekurrentes Netzwerk)	Elman
f0-Ausgabe-Neuronen	1 oder 3 pro Silbe	6 pro Silbe
Abtastinter-vall (je nach Silbenlänge)	ca. 30..300 ms; Mittelwert: ≈ 70 ms	ca. 15..50 ms; Mittelwert: ≈ 30 ms
Trainings-daten	≈ 300..400 Sätze	≥ 200 Sätze
Urteile	unzureichende Modellierung großer f0-Anstiege (interrogative Phrasengrenze) zeitliche Verzerrung der f0-Kontur keine mikroprosodische Abbildung	Trainingsdaten konsistenter, Umfang reduziert zeitliche Verzerrung der f0-Kontur reduziert mikroprosodische Effekte berücksichtigt

Tabelle 1: *Einfluß der Verwendung geglätteter f0-Konturen auf Intonations-ANN (Pit*Net*)*

Der Ansatz (4) stellt gewissermaßen eine Brücke zwischen regelbasierten (1 und 2) und datengetriebenen Intonationsmodellen dar. Avisierte Eigenschaften der f0-Kontur werden bereits in den Trainingsdaten modelliert. Der Aufwand des eigentlichen Intonationsmodells (z.B. *Artificial Neural Network..ANN*) reduziert sich. Die Untersuchungen zum Übergang von Ansatz (3) nach (4) sind in Tabelle 1 zusammengefaßt.

2.2 Transplantation von Grundfrequenzkonturen

Mit der geglätteten und interpolierten f0-Kontur ist es leicht möglich, eine natürliche Intonation auf natürliche oder synthetische Sprachsignale zu projizieren. Mit den in äquidistanten Abständen der aufbereiteten Kontur entnommenen f0-Werten wird direkt der Synthetisator angesteuert. Wird hingegen die Rohkontur ohne nachfolgende Glättung verwendet, besteht die Gefahr, daß Ausreißer als Stützwerte verwendet werden. Da die akustische Synthese linear zwischen den angegebenen Stützstellen interpoliert, führt dies zu wahrnehmbaren Verzerrungen des Intonationsverlaufes und zu einer verfälschten Abbildung der originalen f0-Kontur.

2.3 Etikettierung für die Korpusbasierte Sprachsynthese

Starke Dauer- und Grundfrequenzmanipulationen bei der Sprachsynthese im Zeitbereich sind die Hauptursachen für einen maschinellen, metallischen Klang synthetischer Sprache.
Bei der korpusbasierten Sprachsynthese wird daher der Ansatz verfolgt, prosodische Alternativvarianten der Inventarbausteine zu verwenden. Die Abweichung prosodischer Merkmale ist zum Beispiel ein wesentliches Kriterium zur Berechnung von Bausteinkosten. Die Stetigkeit des Grundfrequenzverlaufes an der Verkettungsstelle kann in die Berechnung der Verkettungskosten eingehen.
In jedem Falle ist eine prosodische Manipulation der Sprachbausteine bei großer Merkmalsähnlichkeit zur Zieläußerung nur in geringem Maße oder gar nicht mehr notwendig.
Für die prosodische Etikettierung kann prinzipiell der im Signal gemessene Grundfrequenzverlauf verwendet werden. Für die Verkettung von Bausteinen ist die Lage im Grundfrequenzbereich (z.B. über den Laut gemittelte f0-Meßwerte) sowie die Zugehörigkeit zu prosodischen Konturen (steigend, monoton bzw. fallend) von Interesse. Gewisse Abweichungen von den Sollwerten sind durchaus zu tolerieren bzw. durch nachträgliche leichte prosodische Manipulation korrigierbar.
Die Mittelung bzw. Approximation von Konturstücken ist jedoch sehr sensitiv gegenüber Ausreißern in der Ausgangskontur. Bevor entsprechende Meßwerte berechnet werden, muß die gemessene Grundfrequenzkontur geglättet und in stimmlosen Abschnitten interpoliert werden. Die so erhaltenen lautbasierten f0-Mittelwerte zeichnen sich durch erhöhte Konsistenz und Aussagekraft aus, als die aus den Rohkonturen zu ermittelnden Parameter.

3 Schlußfolgerungen

Wie gezeigt wurde, ist für die verschiedensten Anwendungen in Intonationsforschung und Datenklassifikation das in Abschnitt 1 beschriebene Verfahren nicht nur mit großem Vorteil einsetzbar, sondern geradezu die Voraussetzung für minimalen Aufwand und verläßliche und aussagekräftige Ergebnisse. Die Erschließung der vollen Funktionalität des vorgestellten Algorithmus verlangt jedoch einige Übung sowie tieferes Verständnis seiner Wirkungsweise.

4 Literatur

Adriaens, L.M.H.(1993): Ein Modell deutscher Intonation. Dissertation, TU Eindhoven.
Hess, W. (1983): "Pitch Determination of Speech Signals", Heidelberg
Hess, W. et.al.(1997): "Prosodic Modules for Speech Recognition and Understanding in VERBMOBIL" in: Y. Sagisaka et. all. "Computing Prosody", Springer, New York.
Jokisch, O.; Pescheck, M.(1998): "Neuronale Prosodiegenerierung - Einfluß der Trainingsdaten", in: Fortschritte der Akustik - DAGA 98, S.352-353, Zürich.
Sedgewick,R. (1992): "Algorithmen in C", Addison Wesley, Bonn