"Irregularitäten" spontaner Sprache und ihre Verarbeitung mit automatischen Grundfrequenzverfahren

A. Batliner[1], A. Kießling[2], R. Kompe[2], E. Nöth[2]

[1]Institut für deutsche Philologie, Ludwig-Maximilians-Universität München
[2]Lehrstuhl für Informatik 5 (Mustererkennung), Friedrich-Alexander-Universität Erlangen

1 Problemstellung und Untersuchungsmaterial

Die Extraktion von Grundfrequenz-(F0-)Verläufen – und damit die Bestimmung ihres perzeptiven Korrelats, der Tonverläufe – ist für die automatische Spracherkennung zumindest relevant bei der Bestimmung des Satzmodus (wie Frage vs. Nicht-Frage), des Satzfokus ("wichtigster" Teil der Äußerung) und der Gliederung – und damit potentiell auch der Disambiguierung – von Äußerungen, sowie bei der Bestimmung von betonten und unbetonten Silben. Einschlägige Erfahrungen basieren zum größten Teil auf kontrolliertem, "elizitiertem", d.h. gelesenem Material, das von kooperativen, oft in irgendeinem Sinn auch "eingeweihten" Sprechern produziert wurde. Auch wenn diese Ergebnisse sicher zum großen Teil auf nicht-elizitiertes, spontanes Material übertragen werden können, so bleibt doch die Frage, wo und in welchem Ausmaß Unterschiede bestehen, und insbesondere, ob man bei zwar kooperativen, aber nicht geschulten Sprechern nicht doch mit zusätzlichen Schwierigkeiten rechnen muß, etwa mit Versprechern/Neuansätzen, Häsitationen oder Irregularitäten bei der Phonation.

Wir wollen uns in diesem Beitrag mit einer Form dieser Irregularitäten beschäftigen, den Laryngalisierungen [Hed90], d.h. mit speziellen Stimmphänomenen, die sich oft entweder durch ein völlig aperiodisches oder sehr langperiodisches Signal auszeichnen. Auditiv werden sie häufig als eine Art "Knarren" der Stimme wahrgenommen, beeinflussen aber den perzipierten Tonverlauf nicht, d.h. der Hörer verarbeitet sie ebenso wie stimmlose Passagen, über die er den Tonverlauf unbewußt extrapoliert. Laryngalisierungen können ebenfalls für die o.a. funktionalen Aspekte nutzbar gemacht werden – am bekanntesten ist wohl die Koexistenz von Laryngalisierung und final abfallendem Tonverlauf als Indikator einer Nicht-Frage. Auf mögliche weitere Funktionen wollen wir hier nicht eingehen. Eine Bestimmung und gegebenenfalls separate Verarbeitung von Laryngalisierungen ist aber schon deshalb erforderlich, weil F0-Algorithmen an diesen Stellen oft aufgrund der erwähnten Signaleigenschaften Schwierigkeiten haben.

Die unserem Beitrag zugrundegelegte Stichprobe von 181 Äußerungen einer Sprecherin (ca. 4,3 Minuten Sprachmaterial) ist Teil eines größeren Korpus, das in nächster Zeit bearbeitet wird. Aufnahmesituation war die folgende: Zwei Versuchspersonen (miteinander befreundete Studentinnen der Psychologie) saßen sich ohne Blickkontakt in einem Versuchsraum der LMU München gegenüber und gaben sich gegenseitig Anweisungen, was die Partnerin mit vor ihr aufgebauten Holzklötzchen machen sollte (sog. "Problemlösungsregister"). Das Experiment war so angelegt, daß eine "echt" spontane, lebhafte Unterhaltung zustandekam, ohne daß den Versuchspersonen bewußt war, daß ihre Sprache und nicht ihr kooperatives Verhalten untersucht wurde. Aus den Redebeiträgen wurden (nach in unserem Zusammenhang nicht interessierenden Gesichtspunkten) "vollständige" Äußerungen ohne Häsitationen und Abbrüche ausgewählt. (Äußerungen mit Häsitationen und Abbrüchen bilden eine andere Stichprobe, die gesondert bearbeitet wird.)

Die Aufnahmebedingungen entsprachen denen einer ruhigen Büroumgebung. Die Äußerungen wurden mit 12 Bit und einer Abtastfrequenz von 10 kHz digitalisiert.

2 Untersuchte Verfahren zur F0-Bestimmung

Für jede Äußerung der Stichprobe wurden mit drei Algorithmen automatisch F0-Konturen berechnet und dabei jedem Frame (12,8 ms) ein Wert zugewiesen: Das in Erlangen entwickelte

DPGF-Verfahren (**E**) definiert als F0-Kontur den optimalen Pfad in einer Menge von alternativen F0-Kandidaten [Not91]. Der Pfad wird mittels Dynamischer Programmierung gesucht und basiert auf den Abständen zum zeitlichen Vorgänger und einem mit einem Mehrkanalverfahren bestimmten Zielwert. Von W. Hess stand ein Cepstrum-Verfahren (**H**) [Hes91] und von H. Reetz ein regelbasiertes periodensynchrones Verfahren (**R**) [Ree89] zur Verfügung[1]. Für unsere Untersuchungen wurde der zulässige F0-Bereich bei E und H auf das sprecherspezifische Intervall 167-550 Hz, bei R auf das Standardintervall von 50-500 Hz beschränkt; ansonsten wurden die Verfahren nicht optimiert. Bei R ist die stimmhaft(*Sh*)/stimmlos(*Sl*)-Entscheidung implizit, bei E und H wurde sie durch verschiedene Präprozessoren durchgeführt.

3 Referenzdaten

Bei einer Klassifikation von *Sh*-Bereichen in Laryngalisierungen und Nichtlaryngalisierungen ist eine eindeutige Grenzziehung zwischen diesen beiden Klassen kaum möglich. Aus diesem Grund wurden Laryngalisierungen per Hand durch Anschauen und Anhören des Signals eher restriktiv segmentiert, d.h. nur deutlich ausgeprägte Laryngalisierungen wurden als solche markiert. Unsere Stichprobe weist unter 800 *Sh*-Bereichen (11240 Frames) 153 (861 Frames) solcher Laryngalisierungen auf (4,3 % der Gesamtstichprobe, 7,7 % aller *Sh*-Frames), die zwischen einem und 18 Frames (im Mittel 6 Frames) lang sind.

Die F0-Referenzkonturen wurden aus den automatisch berechneten Werten erstellt, wobei diese Werte auditiv und am Zeitsignal überprüft wurden. In erster Linie handelt es sich dabei um eine Handkorrektur der von H berechneten F0-Werte. Korrigiert wurden dabei lediglich *grob konturverfälschende* F0-Werte (s. [Not91], S. 132 ff), indem sie durch einen Wert von R, E oder einer von H ermittelten zweiten Alternative ersetzt wurden. War keiner dieser Werte passend, so wurde ein adäquater Wert aus dem Signal ermittelt. Besonders problematisch erwies sich die Erstellung von Referenzwerten an Laryngalisierungen: Zum einen lassen sich aufgrund der starken Irregularitäten an diesen Stellen oft nur schwer eindeutige Perioden im Signal bzw. deutlich ausgeprägte Harmonische im Spektrum feststellen; zum anderen entsprechen die tatsächlich akustisch vorhandenen sehr niedrigen F0-Werte nicht dem perzeptiven Eindruck. Deshalb wurde bei Laryngalisierungen die Referenzkontur auch aus den benachbarten F0-Werten extrapoliert; sie repräsentiert keine signalnahe F0-Kontur, sondern den Tonverlauf.

4 Problematik der automatischen F0-Bestimmung bei Laryngalisierungen

Laryngalisierte Bereiche bereiten auch bei der automatischen Extraktion der F0 Probleme. Die hier untersuchten drei Verfahren zeigen bei Laryngalisierungen häufig ein sehr unterschiedliches Verhalten und weichen zumeist stark von der Referenzkontur ab. Im wesentlichen ließen sich drei häufige Fehlertypen, die jedoch nicht unbedingt verfahrenstypisch sind, beobachten. Bei der in Abb. 1 dargestellten Laryngalisierung tritt z.B. jeder dieser Fehlertypen in der Kontur eines der drei Verfahren auf: R findet die tatsächlich vorliegende tiefe – aber den Tonverlauf nicht adäquat repräsentierende – F0 der Laryngalisierung "fehlerfrei", denn als periodensynchrones Verfahren ist es (ohne starke F0-Bereichsbegrenzung) besonders gut dafür geeignet. Hier setzt R allerdings die niedrige F0 von der Laryngalisierung ausgehend zu weit nach links fort. Mit H wurde die erste Harmonische dieser tiefen F0 berechnet, was zu einem konturverfälschenden Sprung nach oben führt. Die Kontur von E weist die geringste Abweichung zur Referenzkontur auf, zeigt aber viel zu starke Sprünge, die durch die Irregularitäten im Signal ausgelöst wurden. Häufig werden bei E für Laryngalisierungen auch zu niedrige F0-Werte angenommen, was aufgrund der Dynamischen Programmierung dazu führen kann, daß die Kontur auch in den regulären Abschnitten des betreffenden *Sh*-Bereichs zu tief berechnet wird. Vor der

[1]An dieser Stelle sei Herrn Hess und Herrn Reetz herzlich für die Bereitstellung der Programme gedankt.

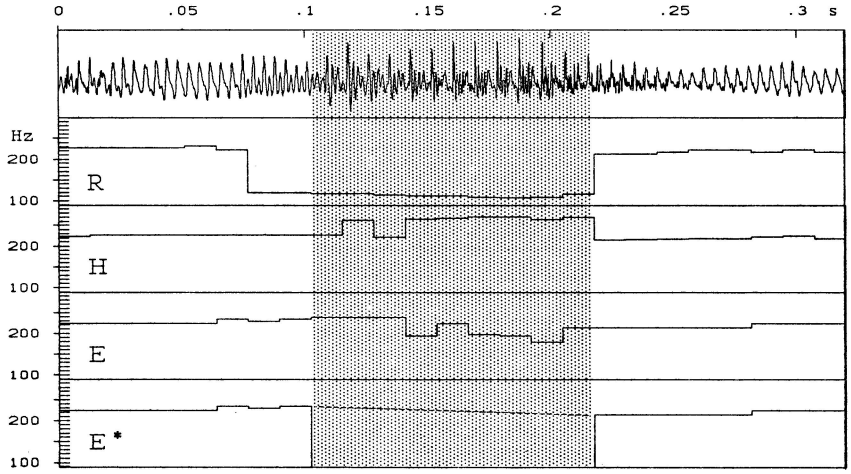

Abb. 1: *Mit vier verschiedenen Verfahren berechnete Konturen zum dargestellten Sprachsignal der Äußerung "oda wean die au' nich' gefressn" (Oder werden die auch nicht gefressen?). Der Diphthong /au/ wurde laryngalisiert und ist schraffiert unterlegt.*

Berechnung der Kontur E* mit dem Verfahren E wurde der laryngale Bereich manuell als *Sl* markiert. An der Laryngalisierung kann man durch Extrapolation (gerade Linie im schraffierten Bereich) einen dem auditiven Eindruck entsprechenden F0-Verlauf erhalten.

5 Quantitative Auswertungen

Die in Tab. 1 zusammengestellten Ergebnisse sollen die Problematik der F0-Analyse bei Laryngalisierungen verdeutlichen. Dargestellt sind die **Abweichung** der einzelnen Verfahren von der Referenzkontur, die **Änderung** der F0 aufeinanderfolgender Frames und die **Divergenz** der Verfahren untereinander. Die Spalten **RB** (**R***egulär* **B***ereiche*) beziehen sich auf zusammenhängende *Sh*-Bereiche ohne Laryngalisierung, **LF** bezeichnet ausschließlich **L***aryngalisierte* **F***rames*. Die nicht laryngalisierten *Sh*-Frames in den Bereichen, in denen eine Laryngalisierung auftritt, bleiben unberücksichtigt, weil Laryngalisierungen die F0-Berechnung in benachbarten regulären Frames negativ beeinflußen können (s.o.); somit entspricht hier das Verhalten der Verfahren nicht dem in rein regulären Bereichen. Als Maße für die Abweichungen wurden verwendet:

- die Wurzel des mittleren quadratischen Abstands **Q** der F0-Werte in Hz sowie
- der prozentuale Anteil der innerhalb eines Verfahrens *grob* voneinander abweichenden Frames **G**, operational definiert als F0-Differenz von mindestens 30 Hz. Die entsprechenden Werte sind in der Tabelle kursiv wiedergegeben.

In der Tabelle gibt es also für jede Konstellation zwei Maßzahlen: So ist z.B. der Wert 3 (rechts oben) so zu verstehen, daß in 3% der RB-Frames die Differenz zwischen den F0-Werten von H und E mindestens 30 Hz beträgt. Der entsprechende Wert für den mittleren quadratischen Abstand zwischen H und E beträgt 26 Hz. Die großen Werte für LF im Vergleich zu RB in der Spalte *Abweichung* bedeuten, daß bei Laryngalisierungen häufig nicht das Äquivalent der

	Abweichung				*Änderung*			
	Q (Hz)		G (%)		Q (Hz)		G (%)	
	LF	RB	LF	RB	LF	RB	LF	RB
E	63	17	*26*	*2*	20	18	*12*	*4*
H	43	20	*15*	*1*	36	18	*19*	*4*
R	124	63	*61*	*20*	51	48	*24*	*15*

	Divergenz			
	Q (Hz)		G (%)	
	LF	RB	LF	RB
H/E	62	26	*29*	*3*
H/R	122	63	*64*	*21*
E/R	109	62	*63*	*21*

Tab. 1: *Abweichung und Änderung der F0 von 3 Verfahren, bzw. ihre gegenseitigen F0-Divergenz in laryngalisierten Frames (LF) vs. regulären Bereichen (RB).*

perzipierten Tonhöhe berechnet wird. Die ebenfalls hohen Unterschiede zwischen LF und RB in der Spalte *Divergenz* zeigen weiterhin, daß die Verfahren nicht konsistent die *tatsächliche F0* berechnen. In den größeren *Änderung*en bei Laryngalisierungen (LF vs. RB) spiegeln sich die irregulären Signaleigenschaften in den F0-Berechnungen wider. Das uneinheitliche Verhalten der drei Verfahren kann als ein Indikator für Laryngalisierungen angesehen werden[2].

6 Zusammenfassung und Ausblick

Die F0-Bestimmung bei Laryngalisierungen erweist sich sowohl manuell als auch automatisch als problematisch. Zudem können in regulären Signalabschnitten Fehler entstehen, die durch benachbarte Laryngalisierungen hervorgerufen werden, wenn die Verfahren nicht rein lokal (frameweise) arbeiten. Es wäre wünschenswert, ein Verfahren zu besitzen, das eine dem perzipierten Tonverlauf entsprechende, bei Laryngalisierungen nicht signalnahe Kontur berechnet, und dabei beispielsweise in den folgenden Schritten vorgeht:

1. *Sh*/*Sl*-Entscheidung,
2. Detektion von Laryngalisierungen,
3. F0-Berechnung für alle nicht laryngalisierten *Sh*-Frames,
4. Extrapolation der F0-Kontur bei Laryngalisierungen.

Diese Vorgehensweise würde auch die mittlere Abweichung bei nichtlaryngalen Stellen senken: So beträgt die Abweichung Q von E* (*Sl*-Setzung und damit Nicht-Berücksichtigung der laryngalen Frames bei der Berechnung durch E) 16,5 % gegenüber 22,5 % bei E, der Anteil der Grobfehler G senkt sich von 2,5 % bei E auf 2,1 % bei E*.

Literatur

[Hed90] P. Hedelein, D. Huber: *Pitch Period Determination of Aperiodic Speech Signals.* In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, S. 361–364, 1990.

[Hes91] W. Hess: *Persönliche Mitteilung.* 1991. Institut für Kommunikationsforschung und Phonetik, Universität Bonn.

[Not91] E. Nöth: *Prosodische Information in der automatischen Spracherkennung — Berechnung und Anwendung.* Niemeyer, Tübingen, 1991.

[Ree89] H. Reetz: *A Fast Expert Program for Pitch Extraction.* In *Proc. European Conf. on Speech Technology*, S. 476–479, 1989.

[2]In unserem Zusammenhang geht es nicht um einen qualitativen Vergleich der Verfahren (vertikaler Zahlenvergleich in der Tabelle). Deshalb wurde auch nicht versucht, die Verfahren zu optimieren. Es ist nur ein horizontaler Vergleich jeweils zwischen LF und RB sinnvoll.