

Irregularitäten im Sprachsignal – störend oder informativ ?

A. Kießling, R. Kompe, E. Nöth

Universität Erlangen-Nürnberg, Informatik 5, Martensstr. 3, 8520 Erlangen, FRG

e-mail: kiessl@informatik.uni-erlangen.de

A. Batliner

Institut für deutsche Philologie, L.M.-Universität München, Schellingstr. 3, 8000 München 40, FRG

1 Einleitung

Bei der Produktion menschlicher Sprache wirken im wesentlichen drei funktionelle Organgruppen zusammen: die *Atmungsorgane* erzeugen einen Luftstrom, der entweder die *Glottis* in quasiperiodische Vibration versetzt (bei stimmhafter Anregung) oder diese weitgehend ungehindert passiert (bei stimmloser Anregung). Das dabei entstehende glottale Signal regt die Luftsäule im *Vokaltrakt* zum Resonieren an und die akustische Schallwelle wird über Mund und/oder Nase freigesetzt. Auch bei eindeutig stimmhaften Signalen lassen sich häufig mehr oder weniger irreguläre Übergangsbereiche beobachten, in denen eine eindeutige Festlegung der Grundperioden schwierig oder sogar unmöglich ist. Typische Beispiele für solche problematischen Übergangsbereiche sind das Einschwingen bzw. Ausklingen der Stimme am Anfang bzw. am Ende eines stimmhaften Bereichs, koartikulatorische Erscheinungen im Transitionsbereich aufeinanderfolgender Laute sowie spezielle Phonationsphänomene wie Laryngalisierungen. In der automatischen Sprachverarbeitung werden Irregularitäten der letzten Art oft nur als störend empfunden, insbesondere für Grundfrequenz-(F0)-Analyseverfahren stellen sie eine der Hauptfehlerquellen dar (vgl. [Hes83, S. 49], [Kom89]).

Wir wollen uns im vorliegenden Beitrag mit diesem speziellen Stimmphänomen – der **Laryngalisierung** – beschäftigen. Nach einer kurzen Charakterisierung wird erläutert, daß Laryngalisierungen für automatische Spracherkennungssysteme nicht ausschließlich negative Eigenschaften besitzen müssen, sondern durchaus für funktionale Aspekte nutzbar gemacht werden können, was durch die Ergebnisse der durchgeführten Untersuchungen über Auftreten und Vorkommen von Laryngalisierungen in der Stichprobe belegt wird. Anschließend werden Untersuchungen bezüglich der starken Periodenschwankungen in laryngalisierten Bereichen und die dabei erzielten Resultate beschrieben. Im Ausblick wird über einen ersten Ansatz zur automatischen Laryngalisierungsdetektion berichtet und ein Anwendungsbeispiel für automatisch detektierte Laryngalisierungen vorgestellt.

2 Laryngalisierungen

Eine eindeutige Definition des Begriffs *Laryngalisierung* ist schwierig, da der für ihre Produktion verantwortliche Mechanismus noch immer nicht exakt geklärt ist. Hinzu kommt, daß in der Literatur unterschiedliche Bezeichnungen mehr oder weniger synonym verwendet werden¹. In Anlehnung an

¹Eine Ausnahme stellen [Hub88], [Hed90] dar, wo Laryngalisierungen anhand spezifischer Signalgestaltmerkmale in vier Mustertypen aperiodischer glottaler Anregung unterteilt werden: **Glottalisierung**, **Creak**, **Creaky Voice** und **Diplophonische Phonation**.

[Hub88], [Hes83] wollen wir *Laryngalisierung* als Oberbegriff für Signal- oder Lautphänomene verwenden, die auf irgendeine Art und Weise von “*irregulären, biphasigen oder außergewöhnlich niederfrequenten* (typischer F0-Bereich: 30-60 Hz, mitunter sogar noch tiefer) *Stimmbandvibrationen herrühren*”. Der Mechanismus ihrer Produktion läßt sich im allgemeinen dadurch charakterisieren, daß die Stimmbänder relativ kurz, dick und schlaff sind und an ihrem hinteren Ende zusammengehalten werden, so daß sie nur an ihrem vorderen Ende langsam und mit relativ geringer Amplitude schwingen können. Im Sprachsignal ist eine Laryngalisierung in der Regel an der nahezu kompletten Dämpfung zwischen zwei aufeinanderfolgenden Anregungsimpulsen sowie häufig durch relativ große Variationen der Leitamplitudenhöhe und/oder der Grundperiodenlänge von einer Periode zur nächsten zu erkennen. Auditiv werden Laryngalisierungen häufig als eine Art “Knarren” der Stimme wahrgenommen, ohne daß sie jedoch den perzipierten Tonhöhenverlauf entscheidend beeinflussen, d.h. “Knarren” und Tonhöhenverlauf überlagern sich unabhängig. [Lad82, S. 129] schlägt folgende Anleitung zur Produktion von Laryngalisierungen vor: “...*You can probably learn to produce it by singing the lowest note that you can - and then trying to go even lower.*”.

Man könnte annehmen, daß Laryngalisierungen in normaler Alltagssprache ein ausgesprochen seltenes Phänomen sind. Daß dem nicht so ist, zeigen die Untersuchungen unserer Stichprobe (s.u.). Sie werden allerdings sowohl sprach- als auch sprecherspezifisch unterschiedlich häufig eingesetzt (nicht in jeder Sprache, nicht bei jedem Sprecher treten sie mit der gleichen Häufigkeit auf) und erscheinen vor allem deshalb interessant, weil sie für unterschiedliche funktionale Aspekte nutzbar gemacht werden können. So werden Laryngalisierungen u.a. für folgende Funktionen verwendet:

- In manchen Sprachen (v.a. afrikanische, bei denen man unter anderem auch spezielle laryngalisierte Verschußlaute kennt) werden sie zur **Laut-** oder **Wortdifferenzierung** eingesetzt (vgl. [Lad82], [Lav80]).
- Zur **Satzmodusdifferenzierung** – am bekanntesten ist wohl die häufig beobachtbare Koexistenz von Laryngalisierung und final abfallendem Tonverlauf als Indikator einer Nicht-Frage.
- Beim Ausdruck von “gelangweilter Resignation”, Autorität, Gereiztheit, Mitleid, Klage oder Beschwerde (vgl. [Lav80, S. 126], [Hub88]) erfüllt die Laryngalisierung eine **paralinguistische Funktion**.
- Zur **Signalisierung morphosyntaktischer Grenzen**: Laryngalisierungen treten häufig am Phonationsbeginn oder -ausklang sowie an syntaktischen Grenzen wie Satz-, Satzteil-, Wort- oder Morphemgrenzen auf. Weiterhin werden sie zur Markierung eines (laryngalen) Hiat (z.B. in “*da achte*” vs. “*dachte*”) eingesetzt, insbesondere wenn die Verbindungsstelle nicht durch andere prosodische Mittel wie Pausensetzung, phrasenfinale Dehnung oder F0-Fall-Rise-Muster gekennzeichnet ist (vgl. [Hub88]).

In [Hub88] wird von interessanten Untersuchungen in diesem Bereich für das Schwedische berichtet; für das Deutsche sind diesbezüglich, vor allem im Hinblick auf automatische Spracherkennungssysteme, noch keine Untersuchungen durchgeführt worden. In einem ersten Ansatz soll deshalb im folgenden qualitativ und quantitativ untersucht werden, inwieweit die Aussagen von [Hub88] auf das Deutsche übertragbar sind. Wie die Ergebnisse zeigen, scheint für die automatische Spracherkennung insbesondere die Signalisierung morphosyntaktischer Grenzen durch Laryngalisierungen von Bedeutung und sollte in Zukunft größere Beachtung geschenkt werden.

3 Untersuchungsmaterial

Für die Untersuchung laryngalisierter Sprachbereiche wurde eine Stichprobe von 16 Äußerungen (Gesamtdauer: ca. 100 Sekunden) von vier (2 weiblichen, 2 männlichen) ungeschulten (‘naiven’) Sprechern

verwendet. Es handelte sich dabei um frei gesprochene, telefonische Auskunftsdialoge aus dem Anwendungsbereich *'Intercity-Zugauskunft'*, weshalb ein natürlicher Gebrauch von Laryngalisierungen vorausgesetzt werden kann. Die Sprachsignale waren mit 100-6400 Hz bandpassgefiltert und mit 16 kHz abgetastet. Die Laute wurden von einer erfahrenen Phonetikerin punktgenau hand-transkribiert, wobei auch Besonderheiten, wie das Auftreten von Laryngalisierungen und glottalen Verschlüssen, markiert waren.

4 Ergebnisse

Die in der Handklassifikation als laryngalisiert markierten Signaltbereiche wurden zunächst einer visuellen und einer auditiven Überprüfung auf typische Merkmale und Charakteristika von Laryngalisierungen unterzogen. Zum einen wurde untersucht, ob und wie gut die markierten Laryngalisierungen anhand von Strukturmerkmalen im Sprachsignal visuell zu erkennen waren. Unabhängig davon wurde an den markierten Stellen durch mehrmaliges Anhören subjektiv ermittelt, wie stark ausgeprägt sich ein laryngalisierungs-typisches Knarren, Krächzen oder Klopfen wahrnehmen ließ. Desweiteren wurde untersucht, welche Laute in der Stichprobe laryngalisiert realisiert waren bzw. an welchen morphosyntaktischen Stellen (d.h. in welchen Silben, an welchen Wortpositionen) die Laryngalisierungen aufgetreten sind.

Obwohl die Stichprobe aufgrund ihrer geringen Größe nicht als repräsentativ angesehen werden kann, lassen sich doch gewisse Tendenzen ableiten, die noch an einer umfangreicheren Stichprobe verifiziert werden sollen. Folgende Eigenschaften konnten festgestellt werden:

- Insgesamt traten 63 Laryngalisierungen auf. Im Mittel wurde alle zwei Sekunden eine Laryngalisierung produziert, nahezu jedes fünfte Wort enthielt eine Laryngalisierung bzw. 4,4 % der Gesamtdauer der Stichprobe waren laryngalisiert. Lediglich eine Äußerung enthielt keine Laryngalisierung. Zwischen weiblichen und männlichen Sprechern bestanden in der Auftretenshäufigkeit kaum Unterschiede.
- Es besteht ein enger Zusammenhang zwischen der visuellen Erkennbarkeit einer Laryngalisierung im Zeitsignal und ihrer auditiven Wahrnehmbarkeit. Insbesondere waren alle als deutlich hörbar markierten Laryngalisierungen gut bis sehr gut visuell erkennbar; alle nur schlecht visuell erkennbaren lassen kaum oder gar kein laryngalisierungstypisches Knarren vernehmen.
- Nahezu alle (ca. 95 %) als laryngalisiert gekennzeichneten Laute waren Sonoranten.
- Die Untersuchung der Frage, in welcher Silbe eines (ein- oder mehrsilbigen Wortes) die Laryngalisierungen auftreten ergab, daß alle Silben, die eine Laryngalisierung enthielten, direkt vor oder nach einer Morphemgrenze lokalisiert waren.
- Innerhalb der Silben, die eine Laryngalisierung aufwiesen, waren am häufigsten die Silbenkerne laryngalisiert (81 %), wesentlich seltener dagegen Silbenkopf und Koda (je 9,5 %).
- 23 % aller laryngalisierten Silben waren betont, was dem Verhältnis aller betonten Silben zur Gesamtanzahl der Silben in der Stichprobe entspricht (vgl. hierzu. [Not91, Kap. 3.2.3]). Laryngalisierte Silben sind somit ungefähr gleichhäufig betont wie normal phonierte Silben.
- 54 % aller Laryngalisierungen traten im ersten Laut eines Wortes auf, 17 % im letzten Laut eines Wortes; insgesamt markierten Laryngalisierungen also zu 71 % eine Wortgrenze. Auffällig war, daß sich nahezu alle laryngalisierten Wortanfänge durch eine gute bis sehr gute visuelle Erkennbarkeit auszeichneten, und daß etwa 66 % dieser Wortgrenzen ohne (subjektiv) deutlich hörbare Pause realisiert waren.
- bei 26 % [bzw. 67 %] aller in der Stichprobe aufgetretenen Einsilber [bzw. Mehrsilber], die mit einem Vokal begannen, war dieser initiale Vokal laryngalisiert.

Fazit: Der relativ hohe Anteil an durch Laryngalisierungen gekennzeichneten Wortgrenzen, insbesondere wenn diese nicht durch deutlich hörbare Pausen hervorgehoben sind, zeigt, daß eine algorithmengestützte Detektion von laryngalisierten Signalstellen in der automatischen Spracherkennung durchaus sinnvoll eingesetzt werden kann.

Für alle Sätze der Stichprobe standen handkorrigierte periodensynchrone Referenzwerte für die in den stimmhaften Bereichen auftretenden Grundperioden zur Verfügung. Als Periodenanfang wurde jeweils der positive Nulldurchgang unmittelbar vor der Leitamplitude markiert. Mit Hilfe dieser Referenzperioden konnte untersucht werden, inwiefern sich Unterschiede in der Periodenschwankung zwischen normal phonierten und laryngalisierten Sonoranten beobachten lassen. In der Regel sind diese Schwankungen bei Laryngalisierungen viel stärker ausgeprägt als die in stimmhafter Sprache immer auftretenden (kleinen) stochastischen Schwankungen der Periodenlänge (der sog. Jitter).

Als Maß für die Periodenschwankung zweier aufeinanderfolgender Grundperioden p_i und p_{i+1} wurde dabei der Betrag der Halbtondifferenz (relativ zum Basiswert 1 Hertz) gewählt:

$$\delta Ht(p_{i+1}) = \left| \frac{12}{\ln(2)} \cdot (\ln(F_0(p_{i+1})) - \ln(F_0(p_i))) \right|$$

Gemessen wurde zum einen der Mittelwert der maximalen Periodenschwankung (in der Tabelle: $\max \delta Ht$) sowie der Mittelwert der durchschnittlichen Periodenschwankung (in der Tabelle: $\text{mitt} \delta Ht$) innerhalb der laryngalisierten bzw. normal phonierten Laute. In beiden Fällen konnte dabei für Laryngalisierungen ein deutlich höherer Wert (zwischen zwei bis dreimal höher) festgestellt werden, als bei den Vokalen oder den nichtvokalischen Sonoranten. Aus diesem Ergebnis läßt sich ableiten, daß eine automatische Lokalisierung von Laryngalisierungen mit Hilfe eines zuverlässigen Grundperioden-Detektors möglich sein müßte.

Lautgruppen	$\max \delta Ht$	$\text{mitt} \delta Ht$
Vokale	2.72	0.79
Nichtvokalische Sonoranten	2.29	0.66
Alle Sonoranten	2.74	1.00
Laryngalisierte Sonoranten	5.82	2.22

5 Zusammenfassung und Ausblick

Wie im vorliegenden Beitrag beschrieben wurde, könnten Laryngalisierungen für automatische Spracherkennungssysteme nicht ausschließlich negative Eigenschaften besitzen, sondern durchaus für funktionale Aspekte nutzbar gemacht werden. Aus ihrer Anwesenheit an bestimmten Stellen im Sprachsignal läßt sich durchaus nützliche Information ableiten, die z.B. zur Unterstützung der Wortgrenzendetektion, bei der Satzmodusbestimmung oder bei der Ermittlung der syntaktischen Struktur einer Äußerung eingesetzt werden kann.

In einem ersten Anwendungsbeispiel haben wir deswegen versucht, Laryngalisierungen auch automatisch zu detektieren. Die dabei gefundenen Bereiche wurden von der anschließenden Grundfrequenzberechnung ausgeschlossen. Auf diese Weise erhofften wir, eine Verbesserung des in Erlangen verwendeten Grundfrequenz-Algorithmus ([Not91], [Kom89]), dem laryngalisierte Signalbereiche erfahrungsgemäß Probleme bereiten, zu erreichen. Als Merkmale für die Detektion von Laryngalisierungen wurden zum einen die handkorrigierten Referenz-Grundperioden, zum anderen die Ergebnisse des in Erlangen eingesetzten Lautkomponentenklassifikators ([Reg88]) verwendet, denn eine Untersuchung der Klassifikatorergebnisse ergab, daß laryngalisierte Signalabschnitte im Mittel mit einer höheren Bewertung als Plosiv oder Frikativ klassifiziert werden als normal phonierte sonorante Abschnitte (vgl. [Kie90]). Ein Signalabschnitt wurde dann automatisch als *potentiell laryngalisiert* klassifiziert,

wenn der Lautkomponentenklassifikator nicht eindeutig auf einen Sonoranten entschieden hat, und die aus den Grundperioden ermittelten Periodenschwankungen über einem gewissen Schwellwert lagen. Auf diese Weise wurden etwa 65 % aller Laryngalisierungen vollständig oder teilweise gefunden. Erste vorläufige Untersuchungen, die allerdings noch an einer größeren Stichprobe verifiziert werden müssen, haben gezeigt, daß sich durch den Ausschluß der lokalisierten Bereiche von der anschließenden Grundfrequenzberechnung die Anzahl der fehlerhaft berechneten Grundfrequenzkonturen reduzieren läßt.

Durch die bisher betrachteten Ansätze und die dabei gewonnenen Ergebnisse ermutigt, werden sich unsere zukünftigen Arbeiten unter anderem weiter mit der automatischen Erkennung von Laryngalisierungen und dem Einsatz des dabei gewonnen Wissens in der automatischen Spracherkennung beschäftigen.

Die Arbeiten, die diesem Beitrag zugrundeliegen, wurden mit Mitteln des Bundesministeriums für Forschung und Technologie sowie der Deutschen Forschungsgemeinschaft gefördert.

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.
- [Hes83] W. Hess: *Pitch Determination of Speech Signals*. Vol. 3 von Springer Series of Information Sciences, Springer Verlag, Berlin, Heidelberg, New York, 1983.
- [Hub88] D. Huber: *Aspects of the Communicative Function of Voice in Text Intonation*. Dissertation, Göteborg/Lund, 1988.
- [Kie90] A. Kießling: *Optimierung des DPGF-Grundfrequenzverfahrens durch besondere Berücksichtigung irregulärer Signaltbereiche, Diplomarbeit*. Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, 1990.
- [Kom89] R. Kompe: *Ein Mehrkanalverfahren zur Berechnung der Grundfrequenzkontur unter Einsatz der Dynamischen Programmierung, Diplomarbeit*. Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, 1989.
- [Lad82] P. Ladefoged: *A Course in Phonetics, Second Edition*. Hartcourt Brace Jovanovich, New York, 1982.
- [Lav80] J. Laver: *The Phonetic Description of Voice Quality*. Cambridge University Press, Cambridge, 1980.
- [Not91] E. Nöth: *Prosodische Information in der automatischen Spracherkennung — Berechnung und Anwendung*. Niemeyer, Tübingen, 1991.
- [Reg88] P. Regel: *Akustisch-Phonetische Transkription für die automatische Spracherkennung*. VDI Verlag, Düsseldorf, 1988.