Akustische Analyse pathologischer Stimmen in fortlaufender Sprache

Hans Werner Strube[1], Dirk Michaelis[2], Jan Lessing[1,2], Sven Anderson[1,2]
[1]*Drittes Physikalisches Institut* bzw. [2]*Abteilung Phoniatrie und Pädaudiologie, Universität Göttingen*

Einleitung

Für die Erkennung und Behandlung von Stimmstörungen ist die Bewertung der Stimmqualität von zentraler Bedeutung. Hierzu wurden in Kooperation zwischen Physik (Strube) und Phoniatrie (Prof. Kruse) akustische Analyseverfahren entwickelt. Als besonders fruchtbar erwies sich hierbei die Stimmcharakterisierung durch das „Göttinger Heiserkeits-Diagramm“ (GHD, Abbildung 1) [2], das ein kombiniertes Jitter-Shimmer-Irregularitätsmaß als Abszisse und ein neues Maß für den relativen Rauschanteil (GNE, Glottal-to-Noise Excitation ratio) [3] als Ordinate benutzt, als grobe Entsprechungen der subjektiven Rauigkeit und Behauchtheit. Normalstimmen liegen links unten, aphone Stimmen rechts oben im Diagramm.

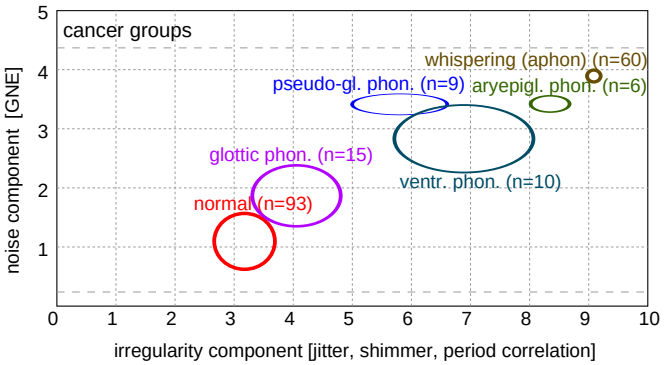

Abbildung 1: Göttinger Heiserkeits-Diagramm, Beispiel: Verteilungen verschiedener Krebs-Gruppen.

Die Stimmanalyse basierte zunächst auf gehaltenen Vokalen. In der klinischen Diagnostik ist aber eine Stimmanalyse aus fortlaufender Sprache erforderlich, um die stimmliche Erkrankung unter normaler Belastung objektiv beurteilen und optimal behandeln zu können. Die gehaltene Phonation entspricht ja eher der Singstimme im Vergleich zur natürlicheren fortlaufenden Sprache. Zur umfassenden Beschreibung der Stimmqualität stellt demzufolge die Analyse fortlaufender Sprache eine wesentliche Erweiterung zur Analyse gehaltener Phonation dar.

Die Methoden der Vokalanalyse sollten auf stimmhafte Abschnitte in fortlaufender Sprache z. T. übertragbar sein. Hierzu wurde ein Verfahren entwickelt, solche Abschnitte automatisch zu erkennen. Die Hauptschwierigkeit bestand darin, dass die linguistisch stimmhaften Laute bei starken Stimmstörungen keineswegs stimmhaft realisiert sein müssen. Es wurde ein lineares Modell der Sprachproduktion angenommen, das sowohl für Normalstimmen als auch für den gesamten pathologischen Bereich bis hin zur Aphonie anwendbar ist (z-Transformations-Notation):

$$X(z) = E(z)G(z)V(z)L(z).$$

$E(z)$ ist die Anregungsfunktion, $G(z)$ Glottispulsmodell, $V(z)$ Vokaltraktmodell und $L(z)$ Abstrahlung an den Lippen. Bei festem Glottispulsmodell und fester Abstrahlung stellt eine Parametrisierung des Vokaltraktes eine Größe zur Differenzierung zwischen stimmlosen und stimmhaften Phonemen dar, unabhängig von ihrer tatsächlichen phonatorischen Realisierung, die von der Stimmstörung abhängt.

Neben den Messgrößen, die dem GHD zugrunde liegen, werden auch neue Größen untersucht, die z. T. nur für laufende Sprache Sinn haben. Die akustischen Analysen werden durch Auswertung von stroboskopischen und Hochgeschwindigkeits-Videoaufnahmen der Glottis ergänzt.

Ausrüstung und Datenbank

Die Aufnahmen und Verarbeitungen geschahen auf Standard-PCs mit guten Soundkarten unter Linux in einem Ethernet-LAN. Die Sprachaufnahmen wurden in einem akustisch gedämmten Aufnahmeraum vorgenommen, der keine lärmenden Geräte enthielt. Eine spezielle grafische Oberfläche zur Aufnahme, zum Schneiden und Markieren der Stimmaufnahmen wurde programmiert.

Es liegen zzt. rund 77000 Sprachaufnahmen als WAV-Dateien mit 48 kHz Abtastfrequenz vor: Vokale ä a e i o u in Tonhöhe normal, tief, hoch; fortlaufende Sprache (phonetisch ausgewogener Standardtext „Nordwind und Sonne“); Spontansprache. Zur Archivierung und automatischen Verwaltung der Stimm- und Videoaufnahmen sowie der medizinischen Befunde (ca. 70 verschiedene phoniatrische Diagnosen, mit Patientenbezug) wurde eine umfangreiche MySQL-Datenbank aufgebaut. Sie läuft mit einem PHP-Web-Frontend auf einem Linux-PC, ist ans Patienteninformationssystem SAP der Klinik zum automatischen Personal-Stammdatenabgleich angeschlossen und hat eine Schnittstelle zu den Videostroboskopie-Arbeitsplätzen. Für jede patientenbezogene Stimmanalyse kann eine PDF-Datei mit Farbausdruck erstellt werden.

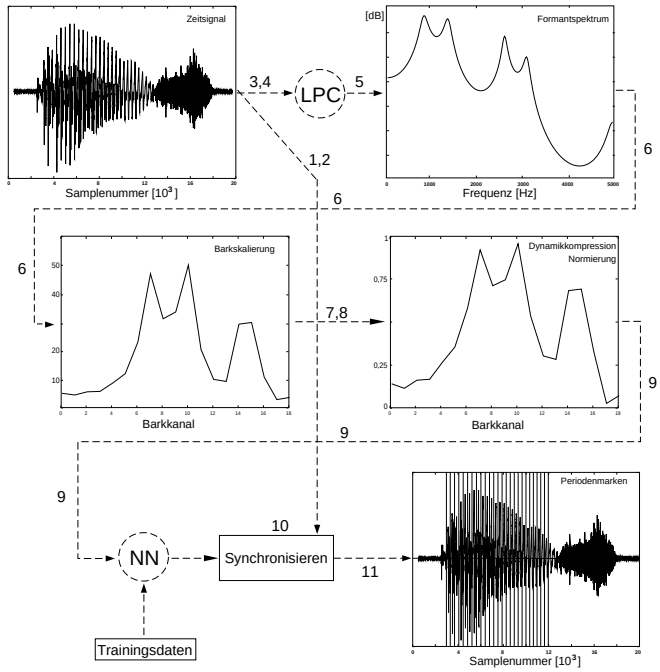

Abbildung 2: Stimmhaft/stimmlos-Unterscheidung, s. Text

Bestimmung stimmhafter und stimmloser Intervalle

Anfänglich wurde eine Stimmhaftigkeitsdetektion mit Nulldurchgangs- und Korrelationstechniken direkt auf dem Sprachsignal getestet, die aber bei stark gestörten Stimmen zu wenige stimmhafte Intervalle erkennt. Daher ist eine Betrachtung der spektralen Einhüllenden (Formantstruktur) vorzuziehen. Die Methode benutzt ein 3-Schicht-Perzeptron (19/6/1 Zellen) mit sigmoider Aktivierungsfunktion (Werte 0 bis 1) als Klassifikator. Die Mustervektoren für seine Eingangsschicht werden wie folgt gebildet *(Zahlen nach Abb. 2)*:
Die mit 48 kHz digitalisierten Sprachsignale werden mit 12 kHz herunter-abgetastet und in überlappende, Hann-gefensterte 40-ms-Abschnitte mit 10 ms Vorschub zerlegt *(3, 4)*. Anhand einer empirischen Energieschranke werden Pausen eliminiert. Eine LPC-Analyse 16-ter Ordnung (Autokorrelationsmethode; Präemphase 0,9735) liefert ein Modellspektrum *(5)*, das durch Summierung mit überlappen-

den Trapezfenstern in 19 Frequenzgruppen (Bark-Skala) umgerechnet wird *(6)*. Es wird mit Exponent 0,25 dynamikkomprimiert und auf sein Maximum über Zeit und Frequenzgruppen normiert *(7, 8)*.

Als Lernverfahren des Perzeptrons *(9)* dient eine beschleunigte Backpropagation [4] mit Lernrate 0,01 und Trägheitsterm (Momentum) 0,8. Die Klassifikationsschwelle am Ausgang liegt bei 0,45. Die Gewichte werden mit Zufallszahlen im Bereich 0 bis 1 initialisiert.

Die Trainingsmenge bestand zunächst aus 32 phonetisch segmentierten Texten (16mal „Nordwind und Sonne“, 16mal „Berlin Butter“) aus der Phondat-Datenbasis von 16 verschiedenen Sprechern mit Normalstimme. Das Training erfolgte in 10000 Iterationen für Soll-Netzausgabe 0,1 (stimmlos) und 0,9 (stimmhaft). Zur Prüfung wurde mit 31 Sprechern trainiert und mit dem 32. getestet. Die Fehlerquote betrug 8 %. Mit obiger Schwelle (0,45) werden davon nur 25 % falsch stimmhaft klassifiziert; falsche stimmlos-Erkennung ist weniger schädlich.

Danach musste ein Nachtraining mit pathologischer Sprache vorgenommen werden. Für die stimmhaften Abschnitte wurden hierbei nur acht gehaltene Vokale verwendet, für die stimmlosen aber manuell ausgewählte Konsonanten aus laufender Sprache („Nordwind und Sonne“).

Es ging bisher nicht um Erkennung der Laute, sondern nur ihrer (linguistischen) Stimmhaftigkeit. In Zukunft ist auch der Einsatz von HMMs zur Lauterkennung oder zum zeitlichen Alignment vorgesehen. Ebenso werden zzt. Perzeptrons zur automatischen Erkennung der sechs gehaltenen Vokale entwickelt.

Analyse fortlaufender Sprache

Die Analyse wurde nur auf zusammenhängenden als stimmhaft klassifizierten Intervallen von mindestens 70 ms Länge durchgeführt. Geplant ist eine Wichtung mit der Länge, da lange Intervalle aussagekräftiger sind. In diesen Intervallen werden Periodenmarken gesetzt (Abbildung 2 *(11)*) und akustische Parameter bestimmt, z. B.:

Periodenlängen mit Waveform-Matching-Algorithmus;

Jitter (3), Shimmer (3), Periodenkorrelationskoeffizient, GNE.

Aus diesen kann wieder ein GHD erstellt werden. Die Lage der Stimmen im GHD ist bei laufender Sprache eine andere als bei gehaltenen Vokalen, so dass eine Neueichung nötig wird. Die Achsendefinition, die auf einer Hauptachsenanalyse im hochdimensionalen Raum verschiedener Messgrößen basiert, ist neu durchzuführen. Die Varianzen der Messpunkte im GHD sind aufgrund der Lautabhängigkeit natürlich größer als bei gehaltenen Vokalen, aber die Mittelwerte behalten ihre Aussagekraft.

Neben dem GHD werden weitere Größen untersucht, deren Korrelation zu den medizinischen Befunden z. T. noch offen ist. Dies sind:

Langzeitspektrum: (a) über alles, (b) nur über linguistisch stimmhafte Intervalle;

Signal/Rausch-Verhältnis nach Qi et al. [5];

Pitch-Amplitude (1. Maximum der AKF des Prädiktionsfehlersignals);

Spectral Flatness Ratio (des Prädiktionsfehlersignals);

Zeitverhältnisse Pausen / stimmhafte Abschnitte / stimmlose Abschnitte.

Auf der Basis der akustischen Größen sollen Gruppenanalysen verschiedener Phonationsmechanismen und Krebsgruppen (signifikante Trennung der Gruppen) vorgenommen werden. Geplant sind ferner evtl. Untersuchungen zur Verteilung der Voice Onset Time sowie zur Beziehung des Stimmfeldes zum emotionalen Ausdruck mithilfe von Theatersprechern.

Videoauswertungen

Die akustischen Messgrößen sollen zu Eigenschaften der Glottisschwingung in Beziehung gesetzt werden. Für die automatische Auswertung der in der phoniatrischen Diagnostik üblichen stroboskopischen Farbaufnahmen der Glottis wurde ein automatisches Bildsegmentierungsverfahren entwickelt, das mittels eines Perzeptrons Objekte wie Stimmlippen, Glottisöffnung usw. erkennt.

Die Einzelheiten der Schwingungen wurden mittels Hochgeschwindigkeitsaufnahmen (schwarzweiß, 4000 Bilder/s, 256×256 Pixel) untersucht. Die Glottisöffnung wurde durch einen dreidimensionalen (Fläche×Zeit) Algorithmus automatisch erkannt, dessen ältere Version wir auf der DAGA'02 beschrieben haben [1]. Zur Untersuchung des Zusammenhangs glottaler Größen mit akustischen Größen (z. B. GNE) wurden aus 15 Aufnahmen 277 Datensätze aus je 100 ms (nicht überlappend) erstellt, die jeweils die vier Werte GNE, mittlere Maximalfläche der Glottis, mittlere Minimalfläche (Restöffnungsfläche) der Glottis und Closed-Quotient (CQ, Zeitanteil der geschlossenen Glottis) enthielten. Spearmansche Rangkorrelation zeigte eine deutliche Beziehung zwischen CQ und GNE (Korrelation 0,53, $\neq 0$ mit Signifikanzniveau $1{,}6 \cdot 10^{-21}$), ebenso zwischen der minimalen Öffnungsfläche und dem GNE (Korrelation $-0{,}51$, $\neq 0$ mit Signifikanzniveau $1{,}5 \cdot 10^{-19}$). [Zahlen sind vorläufig!] Dies entspricht den Erwartungen: Je länger und dichter die Glottis geschlossen ist, desto kleiner sollte der Rauschanteil sein. Abbildung 3 zeigt die Ränge der Einzel-Messpunkte (die *n*-fach vorkommenden Nullwerte der Ordinaten wurden alle auf Rang $(n+1)/2$ gesetzt).

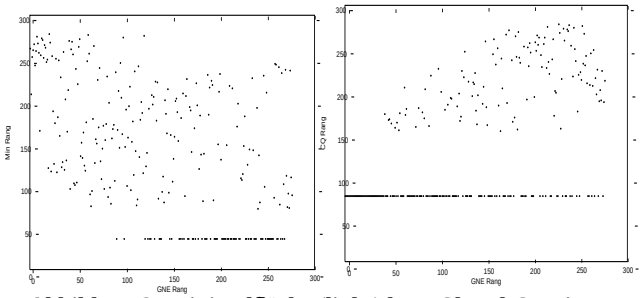

Abbildung 3: Minimalfläche (links) bzw. Closed-Quotient (rechts) über dem GNE (jeweils Rang-Werte).

Danksagung

Wir danken der DFG für die Förderung unter Kr 1469/2 und 1469/5.

[1] S. Anderson, D. Michaelis, H.W. Strube: Vollautomatische Glottisdetektion bei Hochgeschwindigkeitsaufnahmen. In: U. Jekosch (Ed.): *Fortschritte der Akustik – DAGA'02*, Oldenburg: DEGA 2002, 624–625 (CD-ROM: anderson.pdf)

[2] D. Michaelis, M. Fröhlich, H.W. Strube: Selection and combination of acoustic features for the description of pathologic voices. *J. Acoust. Soc. Am.* **103**, 1628–1639 (1998)

[3] D. Michaelis, T. Gramss, H.W. Strube: Glottal-to-Noise Excitation Ratio – a New Measure for Describing Pathological Voices. *Acustica / acta acustica* **83**, 700–706 (1997)

[4] A. van Ooyen, B. Nienhuis: Improving the Convergence of the Backpropagation Algorithm. *Neural Networks* **5**, 465–471 (1992)

[5] Y. Qi, R.E. Hillman, C. Milstein: The estimation of signal-to-noise ratio in continuous speech for disordered voices. *J. Acoust. Soc. Am.* **105**, 2532–2535 (1999)