Tino Haderlein[1], Florian Hönig[1], Frank Jassens[2], Lea Mahlberg[2], Elmar Nöth[1], Alexander Wolff von Gudenberg[2]
[1]Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen-Nürnberg, Erlangen
[2]PARLO, Institut für Forschung und Lehre in der Sprachtherapie, Calden

Robustes Echtzeit-Feedback für die gebundene, weiche Sprechtechnik in der Stottertherapie

Einleitung

Die Kasseler Stottertherapie [1] vermittelt eine Sprechtechnik mit weichem Stimmeinsatz. Weitere Merkmale sind Lautdehnungen, sanfte Übergänge zwischen stimmhaften Lauten, reduzierte stimmlose Konsonanten sowie gehaltene Phonation innerhalb von Phrasen. Besonderer Wert wird auf selbständige Übungssitzungen gelegt, bei denen Patienten am Rechner auch außerhalb der Therapieeinrichtung den Stimmeinsatz trainieren können. Die bisher bei Sprechübungen eingesetzte Software [2] erkennt den Stimmeinsatz nur anhand der Lautstärke und ist deshalb wegen möglicher Störgeräusche fehleranfällig. Weiterhin müssen mehrere Parameter (z.B. die Ruhepegelschwelle) des Programms von den Benutzern kalibriert werden, welche damit jedoch oft überfordert sind. Dies schränkt die Einsatzmöglichkeiten deutlich ein und hat auch negativen Einfluss auf die Verlässlichkeit der Ergebnisse. In der vorgestellten Studie wurde ein neues Programm mit einer automatischen Selbstkalibrierung sowie einer verbesserten grafischen Darstellung des gewünschten Pegelverlaufs getestet.

Material

Für die ersten Auswertungen wurden 66 Aufnahmen mit weichem Stimmeinsatz von neun Männern und zwei Frauen ausgewählt. Diese wurden mit 66 Aufnahmen mit zu hartem Stimmeinsatz (sechs Männer, zwei Frauen) verglichen. Bis auf einen Mann waren alle Personen der zweiten Gruppe auch in der ersten enthalten. Verwendet wurden zunächst nur Aufnahmen einsilbiger Wörter, die mit einem Nahbesprechungsmikrofon erstellt worden waren (Abtastfrequenz 44,1 kHz, Amplitudenauflösung 16 bit). Zum Vergleich mit der automatischen Analyse wurden die Audiodaten von drei Therapeutinnen hinsichtlich des Stimmeinsatzes beurteilt. Von den vier möglichen Kategorien wurden in dieser Studie lediglich Daten untersucht, die hinsichtlich des Stimmeinsatzes als „gut“ oder „zu schnell laut“ bewertet worden waren. Der Fokus lag somit auf dem Anstieg des Sprechpegels.

Methode

Die neue Software arbeitet mit einem Verfahren zur automatischen Segmentierung der Aufnahme in stimmhafte und stimmlose Bereiche, um zu entscheiden, wann der Stimmeinsatz erfolgt. Die dort gemessene Lautstärke wird grafisch nicht mehr als reiner Energiewert, sondern logarithmiert über die Zeit angezeigt (Abb. 1). Gleichzeitig wird der maximal erlaubte Anstieg des Laustärkepegels pro Zeit als Linie („Limbo“) visualisiert bzw. eine Überschreitung unmittelbar zurückgemeldet. Basierend auf den Erkenntnissen aus rund 1000 Aufnahmen der Sprechübungen von 20 Patientinnen und Patienten der Kasseler Stottertherapie wurde der erlaubte Anstieg auf 15 dB pro Sekunde festgelegt. Der einzige Parameter, der noch zu Beginn des Programmlaufs zu bestimmen ist, ist der Pegel der normalen Sprechlautstärke, d.h. der initiale Abstand des Limbo von der Grundlinie. Die Automatisierung dieser Kalibrierung war Teil der vorgelegten Studie.

Zur Bestimmung der stimmhaften Abschnitte wurde ein bestehender Algorithmus zur Detektion der Grundfrequenz F_0 modifiziert. Der RAPT-Algorithmus (Robust Algorithm for Pitch Tracking) [3] ist im Wesentlichen unabhängig von der Lautstärke und Aussteuerung und erstellt Stimmhaft/Stimmlos-Segmentierungen sowie die F_0-Kontur gleichzeitig. Statt eines spektralen Abstandsmaßes auf Basis von linearer Vorhersage wurde jedoch ein Maß aus dem geglätteten Spektrum der schnellen Fourier-Transformation angewendet. Aufgrund häufig auftretenden Netzbrummens (50 Hz) in den Testdaten wurde der minimale F_0-Wert auf 55 Hz festgelegt. Der Anfangswert der Lautstärke wurde jeweils aus den initialen stimmhaften Segmenten bestimmt. Dafür wurden vom Pegel des Stimmsignals 3 dB abgezogen. Die Grundlinie des durchschnittlichen Sprechpegels wird dann in der grafischen Benutzeroberfläche angezeigt (s. Abb. 1).

Da bei der Therapie die durchgehende stimmhafte Phonation trainiert werden soll, beginnt das Programm beim Erkennen eines stimmlosen Lautes erneut mit dem Zeichnen der Pegelkurve. Stimmlose Abschnitte bis zu 50 ms Dauer werden jedoch toleriert.

Das gesamte Verfahren erfordert keine spezielle Hardware, lediglich die Verwendung eines Nahbesprechungsmikrofons, z.B. an einem Headset, wird empfohlen.

Anhand zweier Kriterien wurde in dieser Studie bewertet, ob ein Stimmeinsatz anhand der detektierten Energiekurve „zu schnell laut“ erfolgte. Einerseits wurde geprüft, ob die Pegelkurve der Testperson die Energiewerte der Limbo-Kurve überschritt. Andererseits wurde ermittelt, ob die Energiekurve des Stimmsignals in Teilbereichen stärker und somit schneller ansteigt als der Limbo. War eines oder beide Kriterien erfüllt, wurde auf einen zu schnellen Anstieg der Lautstärke entschieden.

Ergebnisse

46 von 66 „guten“ Stimmeinsätzen (69,7%) wurden als gut bewertet, 56 von 66 „zu schnell lauten“ Stimmeinsätzen (84,8%) wurden ebenfalls richtig dargestellt. Daraus ergibt sich eine Gesamterkennungsrate von 77,2%. Die Ergebnisse sind vor dem Hintergrund zu bewerten, dass auch bei 40,5% aller verfügbaren paarweisen Annotationen keine einstimmige Entscheidung zwischen „normal“ und „zu schnell laut“ erzielt wurde.

Abb. 1 zeigt ein Beispiel für einen Stimmeinsatz mit zu schnellem Pegelanstieg, bei dem außerdem der Limbo überschritten wird. Die Erkennung des Stimmeinsatzes ist sehr robust hinsichtlich Störgeräuschen. Die Verarbeitung läuft in 0,6-facher Echtzeit auf einem Prozessor mit 1,4 GHz Taktfrequenz; bei 3,16 GHz liegt der Echtzeitfaktor unter 0,1. Dabei wird der Energieverlauf zu Beginn der Aufnahme, während die Selbstkalibrierung läuft, unmittelbar nachgetragen. Für die automatische Berechnung der Kreuzkorrelation und Rückverfolgung des F_0-Wertes im Sprachsignal fällt lediglich eine Verzögerung von 85 Millisekunden an (bei 1,4 GHz Taktfrequenz).

Diskussion und Fazit

Der Stimmeinsatz in verrauschten Daten ist beim grundfrequenzbasierten Ansatz besser zu detektieren als bei früher verwendeten lautstärkebasierten Verfahren. Es müssen vor der Benutzung keine Parameter per Hand kalibriert werden. Das F_0-basierte Verfahren erlaubt genauere Rückmeldungen an den Benutzer in Echtzeit.

Danksagung

Förderer dieser Arbeit war die Hessen Agentur/Hessen ModellProjekte (Wiesbaden, Fördernr. 463/15-05).

Literatur

[1] Euler HA, Wolff von Gudenberg A: Die Kasseler Stottertherapie (KST). Ergebnisse einer computergestützten Biofeedbacktherapie für Erwachsene. Sprache - Stimme - Gehör 2000;24(2):71-79.

[2] Euler HA, Wolff von Gudenberg A, Jung K, Neumann K: Computergestützte Therapie bei Redeflussstörungen: Die langfristige Wirksamkeit der Kasseler Stottertherapie (KST). Sprache - Stimme - Gehör 2009;33(4):193-201.

[3] Talkin D: A Robust Algorithm for Pitch Tracking; in Kleijn WB, Paliwal KK (Hrsg.): Speech Coding and Synthesis, Amsterdam, Elsevier, 1995, S. 495-518.

Abbildung

Abb. 1: „Zu schnell lauter“ Stimmeinsatz mit Überschreitung der Limbo-Kurve