Tino Haderlein

Automatische Bewertung von Stimmstörungen aus Textaufnahmen

Einleitung und Hintergrund

Etablierte apparative Methoden zur Evaluierung der chronisch heiseren Stimme bewerten die Stimme lediglich anhand von Aufnahmen gehaltener Vokale. Das wichtigste Sprechkriterium, die Verständlichkeit, kann durch einzelne Vokale jedoch nicht abgebildet werden. Im Rahmen dieser Habilitation wurden Methoden der automatischen Sprachanalyse zur Untersuchung von Stimmstörungen adaptiert und weiterentwickelt. Dafür kamen automatische Spracherkennungsverfahren, prosodische Analyse und akustische Sprechermodelle zum Einsatz. Die untersuchten Evaluierungskriterien umfassten die Stimmqualität [1,2], die Rauigkeit, Behauchtheit und Heiserkeit (RBH-Schema) [1] sowie die Sprachverständlichkeit [2,3].

Material

Sprachaufnahmen des „Nordwind und Sonne“-Textes von drei Patientengruppen mit verschiedenen Ursachen von Heiserkeit wurden ausgewertet. 82 Personen (68 Männer und 14 Frauen, Durchschnittsalter 62,3±8,8 Jahre) wurden nach einer krebsbedingten Larynxteilresektion aufgenommen [2,3]. Weiterhin standen 54 vollständig laryngektomierte Männer (62,2±10,1 Jahre) [2] mit tracheoösophagealer Ersatzstimme und 73 chronisch Heisere (24 Männer, 49 Frauen, 48,3±16,8 Jahre) [1] ohne maligne Erkrankung zur Verfügung. Da kein objektiver Bewertungsstandard für Pathologien von Stimme und Sprache existiert, wurden die Durchschnittswerte einer Expertengruppe als Referenz für die automatische Messung definiert. Der Schwerpunkt der perzeptiven Evaluierung waren die Stimmqualität und die Sprachverständlichkeit. Bei den Heiseren mit benigner Ursache wurden auch Rauigkeit, Behauchtheit und Heiserkeit (RBH-Schema) beurteilt.

Methode

Die Verständlichkeit wurde einerseits durch die Verarbeitung der Textaufnahmen mit einem automatischen Spracherkennungssystem gemessen, das einen Zuhörer simuliert.

Je mehr Erkennungsfehler es macht, desto geringer ist die Verständlichkeit des Patienten. Bei der Verständlichkeitsanalyse durch prosodische Analyse werden Laut-, Wort- und Pausendauern kontinuierlicher Sprache gemessen, die Stimmqualität durch Messung der Grundfrequenz F_0 sowie der Lautheit und deren Schwankungen bestimmt. Die Rauigkeit wurde durch die Zählung von stimmhaften und stimmlosen Bereichen ermittelt. Mithilfe eines Regressionsverfahrens wurde eine Untermenge von Merkmalen bestimmt, die gemeinsam die menschliche Bewertung nachbilden. Die Studie wurde mit Nahbesprechungs- und Telefonaufnahmen Larynxteilresezierter durchgeführt.

Die in einer Stimm- oder Sprachaufnahme enthaltenen Frequenzanteile geben Aufschluss über die Stimmqualität. Das Cepstrum stellt eine kompakte Repräsentation des geglätteten Frequenzspektrums dar. Es wird durch gewichtete Summen von Gaußschen Verteilungen beschrieben. Mit diesen Mischverteilungen wurde aus allen gesprochenen Lauten ein akustisches Modell des jeweiligen Sprechers aufgebaut. Mittelwert und Varianz der Gaußdichten der Sprechermodelle dienten als charakteristische Merkmale dieses Sprechers. Mit einer nichtlinearen Regression wurde daraus die menschliche Bewertung des Patienten nachgebildet. Die Stimmqualität und auch die Verständlichkeit von teilweise und vollständig laryngektomierten Personen wurden auf diese Weise beurteilt.

Cepstrale Parameter wurden nicht nur bei Spracherkennung und Sprechermodellierung benutzt, sondern auch als Maße für die Stimmqualität. Ihr Vorteil ist, dass sie nicht abhängig von einer korrekten Bestimmung der Grundfrequenz sind. Die cepstralen Maße wurden mit gängigen perturbationsbasierten Qualitätsmaßen, wie Jitter, Shimmer oder dem Signal-Rausch-Abstand (HNR), verglichen. In einer Kooperation mit der Universität Bonn wurde ihre Eignung zur Messung von Rauigkeit, Behauchtheit und Heiserkeit gemäß dem klinischen RBH-Schema bei Heiserkeit benigner Ursache untersucht.

Ergebnisse

Bei der Verständlichkeitsbewertung mittels automatischer Spracherkennung wurden für Totallaryngektomierte Mensch-Maschine-Korrelationen bis zu $|r|=0,87$ ermittelt, für Teilresezierte nur $|r|=0,62$. Der Grad der Stimmstörung weist bei dieser Personengruppe

keine so große Variabilität auf wie bei den Kehlkopflosen. Deshalb kann ein einzelner Messwert die kleineren Unterschiede zwischen den Sprechern nicht zufriedenstellend auflösen. Die Hinzunahme der prosodischen Merkmale verbesserte die Nachbildung der menschlichen Verständlichkeitsbewertung so, dass r=0,79 erreicht wurde. Automatische Spracherkennung und prosodische Analyse können somit für verschiedene Stimmstörungen zur Verständlichkeitsbewertung eingesetzt werden.

Bei der akustischen Sprechermodellierung lag die Korrelation zwischen dem maschinell berechneten Wert und der menschlichen Durchschnittsbewertung bei r=0,79 für die Stimmqualität und r=0,73 für die Verständlichkeit. Somit kann auch ein lautbasiertes akustisches Modell zur automatischen Evaluierung der Verständlichkeit beitragen.

Die automatische Nachbildung der RBH- und Stimmqualitätsbewertung mit cepstralen Parametern (bis zu $|r|=0,73$) ist den bisherigen Perturbationsmaßen überlegen ($|r|\leq 0,63$). Sie weisen auch bei schwacher Heiserkeit noch eine zumindest moderate Mensch-Maschine-Korrelation auf (bis zu $|r|=0,49$), während eingeführte Maße (Jitter, Shimmer, HNR etc.; $|r|\leq 0,37$) versagen. Außerdem erzielte die textbasierte Bewertung bessere Ergebnisse als die vokalbasierte.

Diskussion und Fazit

Die Verwendung von Sprach- statt der bisher üblichen Vokalaufnahmen erlaubt auch die objektive Bewertung der Verständlichkeit. Für einzelne Messwerte war die Mensch-Maschine-Korrelation bei Textaufnahmen höher als bei gehaltenen Vokalen. Cepstrale Parameter sind nicht wie Jitter auf eine Detektion der F_0 angewiesen und sind deshalb auch bei starken Störungen anwendbar. Die neue Entwicklung stellt eine Ergänzung zur subjektiven und herkömmlichen objektiven Stimmbewertung dar und kann im klinischen Einsatz als objektive „zweite Meinung“ dienen.

Danksagung

Diese Arbeit wurde von der Deutschen Krebshilfe (Fördernr. 107873) und der Else Kröner-Fresenius-Stiftung (Fördernr. 2011_A167) gefördert.

Literatur

[1] Moers C, Möbius B, Rosanowski F, Nöth E, Eysholdt U, Haderlein T. Vowel- and Text-based Cepstral Analysis of Chronic Hoarseness. J Voice. 2011 Sep 21 [Epub ahead of print]. PMID: 21940144

[2] Bocklet T, Riedhammer K, Nöth E, Eysholdt U, Haderlein T. Automatic Intelligibility Assessment of Speakers After Laryngeal Cancer by Means of Acoustic Modeling. J Voice 2012;26(3):390-7. PMID: 21820272

[3] Haderlein T, Nöth E, Batliner A, Eysholdt U, Rosanowski F. Automatic Intelligibility Assessment of Pathologic Speech over the Telephone. Logoped Phoniatr Vocol 2011;36(4):175-81. PMID: 21875389