Tino Haderlein, Andreas Maier, Elmar Nöth, Frank Rosanowski, Ulrich Eysholdt

Automatische Verständlichkeitsbewertung von Telefonaufnahmen Larynxteilresezierter mittels prosodischer Analyse

Einleitung

Objektiv-apparative Stimmbewertungen werden derzeit meist auf der Basis gehaltener Vokale durchgeführt. Jedoch reflektiert ein isolierter Vokal keine reale Kommunikationssituation. In früheren Arbeiten wurde gezeigt, dass prosodische Analyseverfahren verwendet werden können, um die Verständlichkeit von pathologischen Sprechern automatisch zu bewerten [1,2,3].

Das Telefon ist in der heutigen Zeit eines der wichtigsten Kommunikationsmittel. Dies trifft vor allem auf ältere Menschen zu, deren Mobilität eingeschränkt ist. Der Fokus dieser Studie lag deshalb auf der Verständlichkeit am Telefon und auf dem Einfluss der Signalqualität auf die Mensch-Maschine-Korrelation.

Material

Als Testsprecher dienten 82 Personen nach einer Larynxteilresektion. Das Durchschnittsalter innerhalb der Gruppe betrug 62,3±8,8 Jahre (min. 41,1, max. 86,1 Jahre), 14 der Patienten waren weiblich. Jede Testperson las den “Nordwind und Sonne”-Text vor und wurde dabei mit einem Nahbesprechungsmikrofon (Abtastfrequenz 16 kHz, Amplitudenauflösung 16 bit) und synchron über das Telefon (8 kHz, 16 bit) aufgenommen. Der Text enthält 71 verschiedene Wörter. Die Gesamtzahl der Wörter beträgt 108.

Als Vergleichsbasis für die automatische Evaluierung bewerteten fünf Experten das Kriterium „Gesamtverständlichkeit“ bei jedem Sprecher mit Noten von 1 („sehr gut verständlich“) bis 5 („extrem schlecht verständlich“). Aus den fünf Bewertungen für jede Aufnahme wurde jeweils eine Durchschnittsnote gebildet.

Methode

Basierend auf Wort- und Pausendauern, der Sprachgrundfrequenz F_0 und der Energie im Signal werden 95 prosodische Merkmale pro Wort bzw. pro Wort-Pause-Wort-Intervall und 15 Merkmale auf Abschnitten mit jeweils 15 Wörtern berechnet. Die Energie- und F_0-Werte liegen sowohl in Absolutwerten als auch normiert vor. Die Normierung erfolgt in Bezug auf die gesamte Aufnahme bzw. auf das betrachtete Intervall [1]. Die menschlichen Bewertungen erfolgten pro Aufnahme, d.h. jeder Bewerter gab einen einzigen Verständlichkeitswert für einen Sprecher ab. Deshalb wurden auch für jedes der prosodischen Merkmale alle pro Wort bzw. Aufnahmeabschnitt berechneten Einzelwerte über die gesamte Aufnahme gemittelt.

Ergebnisse

Die durchschnittliche Verständlichkeitsnote der fünf Bewerter für die 82 Sprecher lag im Falle der Headset-Aufnahmen bei 2,9, für die Telefonaufnahmen bei 3,3. Die durchschnittliche Inter-Rater-Korrelation, jeweils gemessen zwischen einem Bewerter und dem Durchschnitt der übrigen vier, betrug für beide Aufnahmetypen r=0,84. Die Berechnung der Intra-Rater-Korrelation, gemessen an 44 Aufnahmen, ergab r=0,83.

Als beste Korrelationen zwischen den menschlichen Verständlichkeitsbewertungen und den berechneten prosodischen Merkmalen wurden folgende Werte ermittelt:

	Headset	Telefon
Verhältnis der Dauer: stimmlose Bereiche/Aufnahme	0,56	0,65
Dauer der stillen Pause vor dem aktuellen Wort	0,57	0,57
Dauer eines Wort-Pause-Wort-Intervalls	0,66	0,64
normierte Energie eines Wort-Pause-Wort-Intervalls	0,66	0,51

Diskussion

Die perzeptive Bewertung hat ergeben, dass die Verständlichkeit der Telefonaufnahmen etwas schlechter ist als die der synchron erstellten Aufnahmen des Nahbesprechungsmikrofons. Die Inter- und Intra-Rater-Korrelation der Hörer wird dadurch jedoch nicht negativ beeinflusst.

Der Vergleich der menschlichen und der automatischen Bewertung zeigt den Einfluss der Signalqualität. Die Korrelationen der dauerbasierten Merkmale und der normierten Signalenergie zu den perzeptiv ermittelten Werten weisen darauf hin, dass die Sprechrate und die Stimmqualität in direktem Zusammenhang zur Verständlichkeit stehen. Die Detektion von stimmlosen Bereichen und der normierten Energie werden durch das Fehlen der Frequenzbereiche über 3400 Hertz beeinflusst. Diese werden während der Telefonübertragung aus der Aufnahme entfernt. Die dauerbasierten Merkmale unterliegen diesem Einfluss kaum. Deshalb unterscheiden sich bei ihnen die Korrelationen für die beiden Aufnahmetypen nur unwesentlich.

Frühere Untersuchungen mit anderen Stimmpathologien [1,2,3] zeigten bessere Korrelationen. Der Grund hierfür ist der relativ schwache Pathologiegrad der Stimme nach Larynxteilresektion. Einzelne Merkmale können diesen nicht trennscharf abbilden. Zukünftige Arbeiten werden deshalb die Kombination mehrerer Merkmale und die Bestimmung einer optimalen Merkmalsmenge mithilfe von Regressionsverfahren umfassen. Auch die Merkmalsmenge selbst wird um Merkmale erweitert werden, die den zeitlichen Verlauf der Werte der bisher betrachteten Merkmale berücksichtigen. Diese Information fehlt zur Zeit aufgrund der Mittelung der Merkmalswerte über die gesamte Aufnahmedauer.

Im Hinblick auf die breite klinische Anwendung der Messmethode kann folgendes geschlossen werden: Die maschinelle Bewertung der pathologischen Stimme nach Larynxteilresektion ist auch per Telefon prinzipiell möglich.

Danksagung

Diese Arbeit wurde von der Deutschen Krebshilfe (Fördernr. 107873) gefördert.

Literatur

[1] Haderlein T. Automatic Evaluation of Tracheoesophageal Substitute Voices. Band 25 von Studien zur Mustererkennung. Berlin: Logos Verlag; 2007.

[2] Bocklet T, Toy H, Nöth E, Schuster M, Eysholdt U, Rosanowski F, Gottwald F, Haderlein T. Automatic Evaluation of Tracheoesophageal Substitute Voice: Sustained Vowel versus Standard Text. Folia Phoniatr Logop 2009;61(2):112-6.
[3] Maier Andreas, Hönig F, Bocklet T, Nöth E, Stelzle F, Nkenke E, Schuster M. Automatic detection of articulation disorders in children with cleft lip and palate. J Acoust Soc Am 2009;126(5):2589-602.