Tino Haderlein, Elmar Nöth, Ulrich Eysholdt, Frank Rosanowski

Verständlichkeitsbewertung von Telefonaufnahmen Larynxteilresezierter mittels der Kombination von automatischer Spracherkennung und prosodischer Analyse

Einleitung

In früheren Arbeiten wurde gezeigt, dass prosodische Analyseverfahren verwendet werden können, um die Sprachverständlichkeit von pathologischen Sprechern automatisch zu bewerten [1,2]. Der Fokus dieser Studie lag auf der Verständlichkeit am Telefon und auf dem Einfluss der Signalqualität auf die Mensch-Maschine-Korrelation, da das Telefon in der heutigen Zeit eines der wichtigsten Kommunikationsmittel ist.

Material

Als Testsprecher dienten 82 Personen, davon 14 Frauen, nach einer Larynxteilresektion. Ihr Durchschnittsalter betrug 62,3±8,8 Jahre (min. 41,1, max. 86,1 Jahre). Jede Person las den “Nordwind und Sonne”-Text vor und wurde mit einem Nahbesprechungsmikrofon (Abtastfrequenz 16 kHz, Amplitudenauflösung 16 bit) und synchron über das Telefon (8 kHz, 16 bit) aufgenommen. 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

Ein bewährtes automatisches Verständlichkeitsmaß ist die Wortkorrektheit (WR) eines Spracherkennungssystems. Sie wird mittels $WR\,[\%] = 100 * [1-(N_{sub}+N_{del})/N_{ges}]$ berechnet, wobei N_{ges} die Anzahl aller gesprochenen Wörter, N_{sub} die Anzahl der vom System durch andere Wörter ersetzten Wörter (Substitutionen) und N_{del} die Anzahl der nicht erkannten Wörter (Deletionen) bezeichnet. Um das System für Telefonaufnahmen verwenden zu können, wurde es mit Daten trainiert, deren akustische Qualität der von Telefonaufnahmen entsprach. Dazu wurde die ursprüngliche Trainingsmenge, die aus

Nahbesprechungsdaten bestand (16 kHz, 16 bit), mithilfe eines Tiefpassfilters auf Telefonqualität (8 kHz, 16 bit) gebracht.

Basierend auf Wort- und Pausendauern, der Sprachgrundfrequenz F_0 und der Energie im Signal wurden außerdem 95 prosodische Merkmale pro Wort bzw. pro Wort-Pause-Wort-Intervall und 16 Merkmale auf Abschnitten von jeweils 15 Wörtern Länge berechnet. Da die menschlichen Bewertungen für den gesamten Text erfolgten, wurden auch für jedes prosodische Merkmal alle pro Wort bzw. Aufnahmeabschnitt berechneten Werte über die gesamte Aufnahme gemittelt.

Mithilfe der Support-Vektor-Regression (SVR) [3] wurde schließlich aus der WR und den prosodischen Merkmalen die aussagekräftigste Kombination bestimmt und ein Vorhersagewert für die menschliche Bewertung des jeweiligen Patienten berechnet. Dieser Schritt wurde jeweils für die Headset- und die Telefonaufnahmen durchgeführt.

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 berechneten Korrelationswerte (vgl. auch [2]) lauten:

	Headset	Telefon
Inter-Rater-Korrelation (ein Bewerter gegen Mittelwert der übrigen)	0,84	0,84
Mensch-Maschine-Korrelation (nur mit WR)	-0,62	-0,75
Mensch-Maschine-Korrelation (WR und prosodische Merkmale)	0,79	0,86

Die beste Merkmalsmenge war für beide Aufnahmequalitäten dieselbe. Sie enthielt neben der WR die Dauer der stillen Pause vor dem aktuellen Wort, die Standardabweichung des Jitter, das Verhältnis der Dauer von stimmhaften Bereichen und der gesamten Aufnahme sowie die Standardabweichung der F_0, in die jedoch auch die Dauer der stimmlosen Bereiche einbezogen wurde. Somit enthielt sie Information über die Aufnahmedauer. Wurden nur die stimmhaften Bereiche gezählt, war das Merkmal nicht erfolgreich.

Diskussion

Bei der perzeptiven Bewertung wurde die Verständlichkeit der Telefonaufnahmen etwas schlechter bewertet als die der synchron erstellten Nahbesprechungsaufnahmen. Bei der Mensch-Maschine-Korrelation zeigen die automatisch ausgewählten Merkmale, dass die Sprechrate und die Stimmqualität bzw. die Irregularität des Stimmsignals in direktem Zusammenhang zur Verständlichkeit stehen. Die Hinzunahme der prosodischen Merkmale zur WR als bisheriges alleiniges Maß für Verständlichkeit verbessert die Nachbildung der menschlichen Bewertung deshalb deutlich. Für die Telefonaufnahmen wird sogar der Referenzwert der menschlichen Inter-Rater-Korrelation übertroffen. 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] 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.

[2] Haderlein T, Maier A, Nöth E, Rosanowski F, Eysholdt U. Automatische Verständlichkeitsbewertung von Telefonaufnahmen Larynxteilresezierter mittels prosodischer Analyse In: Gross M, am Zehnhoff-Dinnesen A (Hrsg.): Aktuelle phoniatrisch-pädaudiologische Aspekte 2010, Warendorf: Darpe Industriedruck 2010, 165-167.

[3] Smola AJ, Schölkopf B. A Tutorial on Support Vector Regression. Statistics and Computing 2004;14(3):199-222.