Automatische Verständlichkeitsbewertung tracheoösophagealer Ersatzstimmen über das Telefon

Korbinian Riedhammer[1], Tino Haderlein[1], Elmar Nöth[2], Hikmet Toy[1], Ulrich Eysholdt[1], Fra nk Rosanowski[1]

[1]Abteilung für Phoniatrie und Pädaudiologie des Klinikums der Universität Erlangen-Nürnberg, Bohlenplatz 21, 91054 Erlangen
[2]Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen-Nürnberg,
Martensstraße 3, 91058 Erlangen
E-Mail: Tino.Haderlein@informatik.uni-erlangen.de

Einleitung

Die tracheoösophageale Ersatzstimme ist heute “state of the art” der Stimmrehabilitation nach einer Laryngektomie. Das zwischen Luft- und Speiseröhre eingebrachte Ventil lenkt den exspiratorischen Luftstrom in die Speiseröhre. Diese Luft versetzt die Schleimhaut im pharyngoösophagealen Segment in Schwingungen und erzeugt so das primäre Ersatzstimmensignal, das dann wie bei der normalen Stimmgebung in den Ansatzräumen gefiltert und vom Mund abgestrahlt wird.

Da eine maschinelle Bewertung der Verständlichkeit mithilfe eines automatischen Spracherkennungssystems prinzipiell möglich ist [1], wurde nun untersucht, inwieweit die maschinelle Bewertung auch über das Telefon durchführbar ist: Dazu müssten die Bewertungen der Maschine und einer Vergleichsgruppe von Experten korrelieren.

Material und Methode

Das auf Hidden-Markov-Modellen basierende Spracherkennungssystem war unabhängig vom gegenwärtigen Projekt am Lehrstuhl für Mustererkennung der Universität Erlangen-Nürnberg entwickelt und bereits in zahlreichen Forschungsprojekten erfolgreich eingesetzt worden. Von einer Ausgründung des Lehrstuhls (www.sympalog.de) wird es mit Erfolg zum Einsatz in Telefondialogsystemen vertrieben.

In dieser Studie wurden 41 Patienten mit einer tracheoösophagealen Ersatzstimme untersucht. Das Durchschnittsalter innerhalb der Gruppe betrug 62,0±7,7 Jahre, zwei der Patienten waren weiblich. Jede Testperson las den “Nordwind und Sonne”-Text vor und wurde dabei mit einem “dnt Call 4U Comfort”-Headset (Abtastfrequenz 16 kHz, Amplitudenauflösung 16 bit) aufgenommen. Durch Abspielen dieser Nahbesprechungsaufnahmen über ein Telefon entstanden Aufnahmen von simulierten Telefonanrufen (Abtastfrequenz 8 kHz, Amplitudenauflösung 16 bit).

Um das Spracherkennungssystem für Telefonaufnahmen verwenden zu können, wurde es mit Daten trainiert, deren akustische Qualität in etwa der von Telefonaufnahmen entsprach. Dazu wurde die ursprüngliche Trainingsmenge, die aus Nahbesprechungsdaten bestand (Abtastfrequenz 16 kHz, Amplitudenauflösung 16 bit), mithilfe eines Tiefpassfilters auf Telefonqualität (Abtastfrequenz 8 kHz, Amplitudenauflösung 16 bit) gebracht. Alle Frequenzanteile über 3400 Hertz wurden dadurch aus den Daten entfernt.

Die Nahbesprechungsaufnahmen der Patienten wurden von fünf Experten hinsichtlich ihrer Verständlichkeit auf einer Skala von 1 („sehr gut verständlich“) bis 5 („unverständlich“) bewertet. Als maschinelles Bewertungskriterium diente die Wortakkuratheit WA des Spracherkennungssystems.

Ergebnisse

Die Patienten mit einer tracheoösophagealen Ersatzstimme erreichten im Falle der Nahbesprechungsaufnahmen eine durchschnittliche WA von 35,3%±13,7%, bei den simulierten Telefonaufnahmen ergab sich ein Durchschnittswert von 28,4%±10,3%. Die entsprechenden Werte für die jeweiligen Sprecher sind in Abb. 1 dargestellt.

Der Vergleich der maschinellen Bewertungen zur durchschnittlichen Expertenbewertung für den jeweiligen Sprecher ergab eine Korrelation von r=-0,82 für die Nahbesprechungsaufnahmen und r=-0,69 für die simulierten Telefonaufnahmen. Zwischen den Expertenbewertungen und der maschinellen Auswertung ergab sich eine Übereinstimmung von =0,41 für die Originalaufnahmen und =0,42 für die

Telefonsprache (gewichtetes nach Cicchetti [2]). Die Inter-Rater-Korrelation innerhalb der Experten lag bei =0,45.

Diskussion

Nach diesen Ergebnissen ist eine maschinelle Bewertung der Ersatzstimme auch per Telefon prinzipiell möglich. Die Inter-Rater-Korrelation wird nur wenig reduziert, wenn das automatische Erkennungssystem der Expertengruppe hinzugefügt wird. Die niedrigere Wortakkuratheit bei der automatischen Erkennung ist auf den Qualitätsverlust durch die Telefonübertragung zurückzuführen. Da die Aufnahmen schon vorgefertigt waren, zeigen sie auch Einflüsse des ursprünglichen Aufnahmemikrofons. Die Trainingsdaten des Spracherkennungssystems hingegen sind nicht über ein Telefon gelaufen, sondern wurden nur mit der entsprechenden Frequenz abgetastet. Wird diese Diskrepanz in den akustischen Gegebenheiten von Trainings- und Testdaten bereinigt, sind Ergebnisse wie auf den Nahbesprechungsdaten zu erwarten. Weiterhin werden durch die Modifikation der aus dem Frequenzspektrum errechneten Sprachmerkmale verbesserte Erkennungsleistungen erwartet. Dies ist Gegenstand weiterer Arbeiten.

Danksagung

Diese Arbeit wird von der Deutschen Krebshilfe (Fördernr. 106266) gefördert.

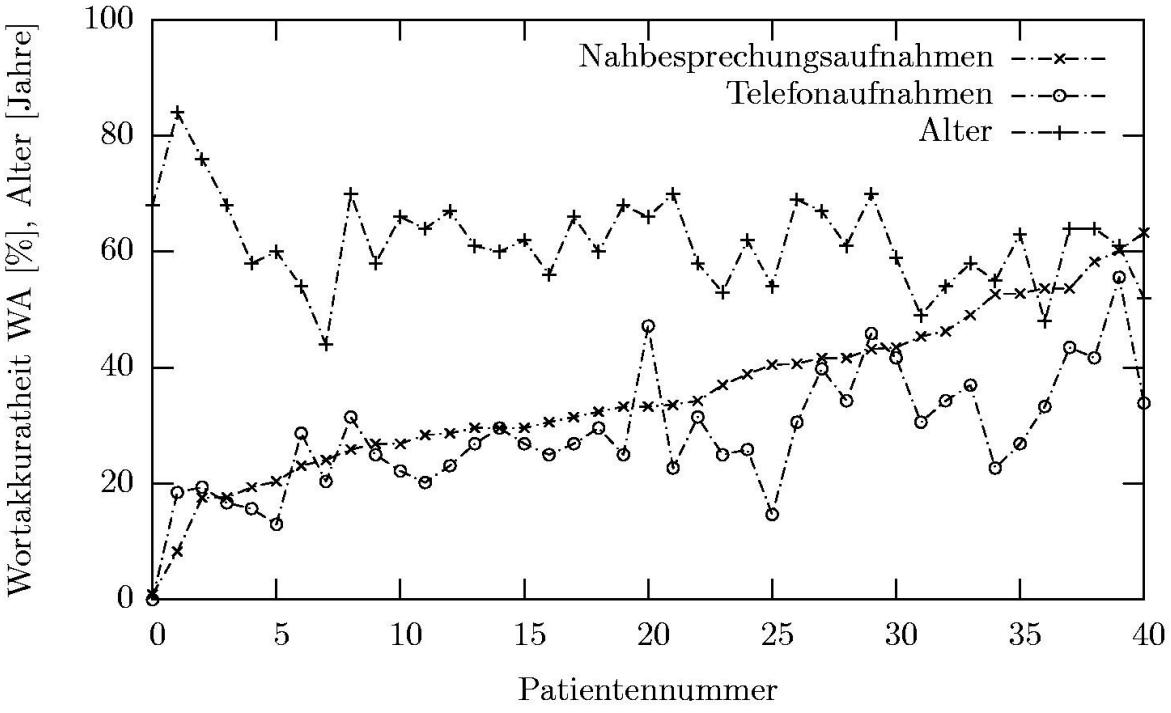
Literatur

[1] Schuster M, Haderlein T, Nöth E, Lohscheller J, Eysholdt U, Rosanowski F (2006)
Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. Eur Arch Otorhinolaryngol 263(2):188-193

[2] Cicchetti DV (1976)
Assessing inter-rater reliability for rating scales: Resolving some basic issues.
Br J Psychiatry 129(5):452-456

Abbildung

Abb. 1: Vergleich der Wortakkuratheiten (WA) des jeweiligen Spracherkennungssystems; Sprecher sortiert nach der WA auf Nahbesprechungsaufnahmen


Wortakkuratheit WA [%], Alter [Jahre]
Patientennummer
Nahbesprechungsaufnahmen
Telefonaufnahmen
Alter
0
20
40
60
80
100
0
5
10
15
20
25
30
35
40