Automatisierung des Postlaryngektomie-Telefontests (PLTT)

Tino Haderlein[1], Ulrich Eysholdt[1], Korbinian Riedhammer[2], Elmar Nöth[2], Frank 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

In früheren Arbeiten wurde gezeigt, dass automatische Spracherkennungsverfahren verwendet werden können, um die Verständlichkeit von Sprechern mit tracheoösophagealer Ersatzstimme (TE-Stimme) automatisch zu bewerten [1,2]. In diesem Beitrag wird eine automatische Version des Postlaryngektomie-Telefontests (PLTT, [3]) vorgestellt, der einen eingeführten Standardtest für die Verständlichkeit über das Telefon darstellt.

Material

PLTT-Aufnahmen von 31 Laryngektomierten (25 Männer und 6 Frauen, im Durchschnitt 63,4±8,7 Jahre alt) mit TE-Stimme waren vorhanden. Die Daten wurden mit einem Dialogsystem der Firma Sympalog Voice Solutions (www.sympalog.com) aufgenommen. Die naiven Hörer waren 8 männliche und 3 weibliche Studenten (22,5±1,2 Jahre), von denen niemand Erfahrung mit Stimm- und Sprachanalyse besaß.

Methode

Ein naiver Hörer, der das Textmaterial nicht kennt, schreibt am Telefon auf, was ein Patient am anderen Ende vorliest. Der PLTT-Wortschatz besteht aus 400 einsilbigen Wörtern und 100 Sätzen, von denen jeweils 22 Wörter und 6 Sätze zufällig ausgewählt werden. Die ersten beiden Wörter und der erste Satz dienen

zum Einhören. Der Sprecher liest nur die vorgedruckten Texte vor. Sonstige Äußerungen sind nicht erlaubt. Der Test beginnt mit dem Lesen der Wörter. Versteht der Zuhörer ein Wort nicht, sagt er genau einmal: „Bitte wiederholen Sie das Wort.“ Sätze dürfen nicht wiederholt werden. Die Zahl der auf Anhieb richtig verstandenen Wörter, mit 5 multipliziert, ergibt die Wortverständlichkeit i_{Wort} in Prozent. Wörter, die wiederholt wurden, werden nicht gewertet. Wird ein Satz vollständig korrekt verstanden, werden zwei Punkte vergeben. Ein Punkt wird gegeben, wenn ein Wort fehlt oder nicht richtig verstanden wurde. In allen anderen Fällen erhält der Leser keinen Punkt. Die Satzverständlichkeit i_{Satz} in Prozent ist die mit 10 multiplizierte Summe aller Punkte für die Sätze. Die prozentuale Gesamtverständlichkeit i_{total} wird dann durch $i_{total}=(i_{Wort}+i_{Satz})/2$ berechnet. Die Hörer in dieser Studie hatten keinen direkten Telefonkontakt zu den Sprechern, sondern spielten die gespeicherten Aufnahmen ab, die sie jederzeit anhalten konnten, um die verstandene Äußerung zu notieren.

Das auf Hidden-Markov-Modellen basierende Spracherkennungssystem war am Lehrstuhl für Mustererkennung der Universität Erlangen-Nürnberg entwickelt und bereits in zahlreichen Forschungsprojekten erfolgreich eingesetzt worden. Kommerziellen Erfolg beim Vertrieb des Systems mit Telefondialogsystemen erzielt wiederum die Firma Sympalog. Ein Spracherkennungssystem kann nur diejenigen Wörter erkennen, die in seiner Vokabularliste gespeichert sind. Eine solche Liste wurde von allen Wörtern im PLTT erstellt. Dies ist jedoch nicht genug, um einen menschlichen Zuhörer zu simulieren. Ein Mensch kennt mehr Wörter als die, die im Test auftreten, was zu Fehlern beim Verstehen führen kann. Um dies im automatischen Test zu simulieren, wurde die Wortschatzliste des Erkennungssystems auch um Wörter ergänzt, die zu denen des tatsächlichen Wortschatzes phonetisch ähnlich sind. Auf diese Weise wurde das PLTT-Vokabular, das aus 738 Wörtern bestand („PLTT-klein“), auf 1017 Wörter erweitert („PLTT-groß“). Dann wurde die gesamte Aufnahme jeweils eines Patienten von dem System verarbeitet und die Worterkennungsrate berechnet. Sie gibt Auskunft darüber, wie viele Wörter prozentual korrekt bzw. gar nicht erkannt wurden und wie viele durch andere Wörter ersetzt, also „falsch

verstanden“ wurden. Ihr Maximalwert beträgt somit 100%, ihr Minimalwert 0%.

Ergebnisse

Tabelle 1 zeigt die PLTT-Resultate der einzelnen Hörer. Obwohl sie nie zuvor TE-Stimmen gehört hatten, ist die Inter-Rater-Korrelation für die Gesamtverständlichkeit i_{total} für alle Personen größer als 0,8. Jedoch schwanken die perzeptiven Resultate stark innerhalb der Hörergruppe. Tabelle 2 zeigt die Durchschnittswerte von Worterkennungsrate und PLTT-Ergebnissen für die menschlichen Bewerter und die Spracherkennungssysteme. Die Korrelation zwischen den PLTT-Messwerten und der automatisch erhobenen Worterkennungsrate ist in Tabelle 3 zu finden.

Diskussion

Der Unterschied von i_{total} über die ganze Sprechergruppe für den „besten“ und den „schlechtesten“ Bewerter beträgt mehr als 20 Punkte, was zeigt, wie stark der Test vom jeweiligen Zuhörer abhängt. Die Standardabweichung von i_{total} ist jedoch für alle Bewerter sehr ähnlich. Die Worterkennungsrate des Spracherkennungssystems ist vor allem deshalb sehr niedrig, weil das System mit Normalstimmen trainiert wurde. Dies simuliert einen naiven Hörer, der nie zuvor TE-Stimmen gehört hat, also genau die Art von Hörer, die für den PLTT gefordert wird. Kein Satz wurde entsprechend den PLTT-Richtlinien vollständig korrekt erkannt (i_{Satz}=0). Die Worterkennungsrate für die menschlichen Bewerter wurde aus deren Niederschrift der Aufnahmen berechnet.

Obwohl die automatische Erkennung so schlechte Resultate erzielte, war die Korrelation zu den menschlichen Bewertungen hoch. Der Grund dafür ist, dass das entscheidende automatische Maß nicht der Durchschnitt der Erkennungsrate ist, sondern ihre Standardabweichung und damit die Breite des angenommenen Wertebereiches. Da zwischen der Worterkennungsrate und dem menschlichen Maß i_{total} Korrelationen bis über 0,9 erzielt wurden, kann festgestellt werden, dass der PLTT durch eine objektive, automatische Version ersetzbar ist.

Danksagung

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

Tabellen

Tabelle 1

Menschliche Bewertungsergebnisse für 31 PLTT-Aufnahmen; angegeben sind jeweils Mittelwert (μ) und Standardabweichung (σ) sowie die Korrelation nach Pearson (r) und Spearman (ρ) für den jeweiligen Hörer zum Durchschnitt der anderen zehn.

	i_{Wort}				i_{Satz}				i_{total}			
Hörer	μ	σ	r	ρ	μ	σ	r	ρ	μ	σ	r	ρ
BM	31,4	21,5	0,91	0,90	48,1	30,4	0,88	0,88	39,8	22,6	0,92	0,90
BT	29,2	20,7	0,83	0,80	52,2	29,5	0,88	0,87	40,7	21,8	0,90	0,90
CV	43,9	20,1	0,90	0,90	45,9	28,6	0,83	0,81	44,9	22,5	0,89	0,89
GM	39,4	21,4	0,87	0,83	48,1	28,9	0,88	0,86	43,8	23,3	0,93	0,93
HT	43,0	21,3	0,89	0,84	57,2	28,6	0,77	0,76	50,1	22,4	0,86	0,86
KC	41,7	20,8	0,93	0,91	46,9	28,8	0,65	0,60	44,3	20,8	0,85	0,83
MM	47,2	23,5	0,91	0,90	50,0	28,7	0,79	0,78	48,6	23,9	0,92	0,92
PC	34,1	21,9	0,87	0,83	48,4	29,5	0,87	0,85	41,3	22,4	0,92	0,92
SM	51,1	21,3	0,82	0,82	59,4	24,0	0,82	0,75	55,2	19,2	0,90	0,86
ST	53,6	23,0	0,92	0,92	67,5	29,6	0,72	0,64	60,6	23,4	0,86	0,85
WW	40,3	19,2	0,89	0,90	56,6	25,2	0,87	0,89	48,4	19,7	0,94	0,94

Tabelle 2

Durchschnittliche PLTT-Maße und Worterkennungsrate (WER) für elf naive Hörer und die beiden Spracherkennungssysteme mit großem und kleinem Erkennungsvokabular; angegeben sind jeweils Mittelwert (μ) und Standardabweichung (σ).

	i_{Wort}		i_{Satz}		i_{total}		WER	
	μ	σ	μ	σ	μ	σ	μ	σ
Erkenner PLTT-klein	16,6	12,6	13,1	13,0	0,0	0,0	6,6	6,5
Erkenner PLTT-groß	13,7	11,2	10,9	10,9	0,0	0,0	5,5	5,8
11 naive Hörer	55,3	21,4	41,4	21,3	52,8	28,3	47,1	22,0

Tabelle 3

Korrelation zwischen durchschnittlichen PLTT-Maßen von elf naiven Hörern und der Worterkennungsrate (WER) der beiden Spracherkennungssysteme mit großem und kleinem Erkennungsvokabular; angegeben ist jeweils die Korrelation nach Pearson (r) und Spearman (ρ).

	r	ρ
Erkenner PLTT-klein	0,88	0,93
Erkenner PLTT-groß	0,85	0,92

Literatur

[1] Schuster M, Nöth E, Haderlein T, Steidl S, Batliner A, Rosanowski F. Can You Understand Him? Let's Look at His Word Accuracy - Automatic Evaluation of Tracheoesophageal Speech. Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 2005; 61-4

[2] Schuster M, Haderlein T, Nöth E, Lohscheller J, Eysholdt U, Rosanowski F. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. Eur Arch Otorhinolaryngol. 2006. 263(2):188-93

[3] Zenner HP. The postlaryngectomy telephone intelligibility test (PLTT). In Herrmann IF (ed.): Speech Restoration via Voice Prosthesis. Berlin: Springer; 1986. 148-52.