Menschliche und automatische Verständlichkeitsbewertung bei tracheoösophagealen Ersatzstimmen*

Tino Haderlein[1], Stefan Steidl[1], Elmar Nöth[1], Maria Schuster[2]

[1] *Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung, Martensstr. 3, 91058 Erlangen, Deutschland*

E-Mail: Tino.Haderlein@informatik.uni-erlangen.de

[2] *Universität Erlangen-Nürnberg, Abt. für Phoniatrie und Pädaudiologie, Bohlenplatz 21, 91054 Erlangen, Deutschland*

Einleitung

Die Anbahnung einer tracheoösophagealen Ersatzstimme (TE-Stimme) ist eine Möglichkeit, Patienten nach einer totalen Laryngektomie, d.h. Kehlkopfentfernung, die Fähigkeit zu sprechen zurück zu geben. Ein Ventil zwischen Luft- und Speiseröhre erlaubt es, den Luftstrom aus der Lunge umzuleiten und Gewebeschwingungen in der Speiseröhre zur Ersatzstimmgebung zu nutzen. Die Betroffenen durchlaufen eine Therapie, in der wiederholt evaluiert werden muss, ob und wie sich ihre Ersatzstimme hinsichtlich Kriterien wie Lautstärke, Verständlichkeit oder Prosodiefähigkeit entwickelt hat [1, 2]. Da die Beurteilung subjektiv erfolgt und das Verfahren für Arzt und Patienten aufwändig ist, erscheint eine Automatisierung und Objektivierung in diesem Bereich sinnvoll.

In unserer Arbeit untersuchen wir, wie gut tracheoösophageale Sprache von einem automatischen Spracherkennungssystem erkannt wird und ob die Ermittlung der Qualität einer Ersatzstimme zumindest teilweise automatisiert erfolgen kann. Dazu müssen die Bewertungen der Maschine und einer Vergleichsgruppe von Experten korrelieren. Im folgenden werden wir unsere ersten Ergebnisse zu diesen Arbeitsgebieten vorstellen.

Das Spracherkennungssystem

Das verwendete, HMM-basierte Spracherkennungssystem wurde am Lehrstuhl für Mustererkennung an der Universität Erlangen-Nürnberg entwickelt. Es wurde bereits in mehreren Forschungsprojekten erfolgreich eingesetzt und kann spontane, kontinuierliche Sprache mit einem Vokabular von bis zu 10000 Wörtern verarbeiten.

Eine detaillierte Beschreibung des für die Experimente eingesetzten Systems findet sich in [3]. Ausgangspunkt war ein Erkenner, der im Rahmen des Forschungsprojektes VERBMOBIL [4] entstanden war. Dieser war mit Aufnahmen von laryngealen Sprechern („Normalsprechern“) trainiert worden. Etwa 80% der 578 Trainingssprecher (304 Männer, 274 Frauen) waren zwischen 20 und 29 Jahre alt, weniger als 10% waren älter als 40. Das Durchschnittsalter der laryngektomierten Testsprecher hingegen lag bei über 60 Jahren. Daher wurde auch eine Vergleichsgruppe von 18 gesunden Männern aufgenommen, die hinsichtlich ihres Alters den TE-Patienten entsprachen. Jeder Sprecher las den „Nordwind und Sonne“-Text, der aus 108 Wörtern besteht, wovon 71 verschieden sind. Die Aufnahmen erfolgten mit einem „dnt Call 4U Comfort“-Headset sowie 16 kHz Abtastfrequenz und 16 bit Amplitudenauflösung. Die Gesamtdauer der 18 TE-Aufnahmen betrug 21 min, die Patienten sprachen darin 1980 Wörter. Darunter waren 32 verschiedene Wörter, die nicht im Text vorkamen und durch Lesefehler entstanden waren. Für die Experimente wurde das Vokabular des Erkenners auf die 103 (71+32) Wörter aus den Aufnahmen beschränkt. Genauere Ausführungen zu den Daten sind in [5] zu finden.

Spracherkennung auf TE-Sprache

Grundlegende Experimente

Bei grundlegenden Experimenten zur Verbesserung der Spracherkennung bei Laryngektomierten [5] wurde u.a. ein monophonbasierter VERBMOBIL-Erkenner eingesetzt. Seine Resultate zeigten Verbesserungen gegenüber polyphonbasierten Ansätzen bei den Sprechern, deren Stimmqualität als besonders schlecht eingestuft worden war. Da der Schwerpunkt der Experimente auf der akustischen Analyse lag, wurde nur ein Unigramm-Sprachmodell verwendet. Deshalb sind die Erkennungsraten niedriger als bei einem stärker differenzierten linguistischen Modell. Auf dem beschriebenen Erkenner erzielte eine Vergleichsgruppe von 16 laryngealen Sprechern (9 Männer, 7 Frauen) eine durchschnittliche Wortakkuratheit von 69,0% (min. 54,6%, max. 88,0%). Die Kontrollgruppe aus 18 älteren Männern erreichte eine durchschnittliche Wortakkuratheit von 57,6% (min. 46,8%, max. 71,6%). Die TE-Sprecher zeigten auf dem für sie nicht angepassten System immerhin Werte bis 50,0% (min. 10,0%) bei einem Mittelwert von 28,7%, obwohl die Testsprecher große Unterschiede hinsichtlich Verständlichkeit, Lautstärke, Rauheit und Atmungsgeräuschen aufwiesen. Außerdem sprachen einige der Patienten mit dialektaler Färbung.

Bewertung der Verständlichkeit durch Experten und Spracherkenner

Da sich eine schwer verständliche Stimme auch beim Menschen in der Zahl der „erkannten“ Wörter niederschlägt, so schlossen wir daraus, dass die Bewertung des Kriteriums „Verständlichkeit“ durch einen Menschen und die von einem Spracherkenner berechnete Wortakkuratheit korrelieren müssten. Die von fünf erfahrenen Bewertern für die 18 Testsprecher vergebenen ganzzahligen

*Diese Arbeit wurde teilweise durch das BMBF im Rahmen des SmartWeb-Projekts (Fördernr. 01 IMD 01 F) und die Deutsche Krebshilfe (Nr. 106266) finanziert. Die Verantwortlichkeit für den Inhalt der Studie liegt bei den Autoren.

Werte reichten von „1“ (sehr gut verständlich) bis „5“ (unverständlich). Die Inter-Rater-Korrelation ist in Tabelle 1 dargestellt. Zwischen jeweils einem Bewerter und dem Durchschnitt der vier anderen lagen die beiden niedrigsten Korrelationswerte bei 0,68 und 0,77, die übrigen zwischen 0,82 und 0,85. Der gewichtete Multi-Rater-κ-Wert über alle Sprecher betrug 0,44.

Danach wurde die Korrelation zwischen Mensch und Maschine bestimmt, wobei die Wortakkuratheit, gemessen über den gesamten Text des jeweiligen Sprechers, als automatische Bewertung herangezogen wurde. In Tabelle 2 sind die Ergebnisse jeweils für die einzelnen Zuhörer und für die Gesamtkorrelation, also bzgl. des Durchschnitts über alle fünf Bewerter, zusammengefasst. Letztere beträgt für den verwendeten monophonbasierten Erkenner –0,84. Der Koeffizient ist negativ, da hohe Erkennungsraten von „guten“ Sprechern mit einer kleinen Bewertungszahl stammten und umgekehrt. Abbildung 1 zeigt die durchschnittliche Bewertung jedes Sprechers und die entsprechende Wortakkuratheit. Die Wortakkuratheit wurde dann wie folgt auf die Noten der Likert-Skala abgebildet: Werte kleiner als 0 bekamen den Likert-Wert 5, dieser Fall trat jedoch in den Daten nicht auf. Ergebnisse unter 15% bekamen die Bewertung 4, die nächsten Intervallgrenzen waren 25 und 40%. Für eine 1 auf der Likert-Skala waren also mindestens 40% Wortakkuratheit nötig. Dann war eine Multi-Rater-κ-Berechnung zwischen Mensch und Maschine möglich. Sie erreichte einen Wert von 0,43, lag also im selben Bereich wie die Gruppe der Experten unter sich.

Die bisherigen Ergebnisse zeigen, dass ein deutlich sichtbarer Zusammenhang zwischen den Ergebnissen der menschlichen und der maschinellen Analysemethode besteht. Wir schließen daraus, dass die Wortakkuratheit ein wichtiger Bestandteil eines möglichen, automatisch gewonnenen Verständlichkeitsmaßes sein könnte und für weitere Untersuchungen im Bereich der pathologischen Stimm- und Sprachanalyse herangezogen werden sollte.

In den aktuellen Versuchen lasen die Patienten einen Standardtext. Um durch die Wortakkuratheit nur die Erkennungsfehler und nicht etwaige Lesefehler zu messen, wurde die erkannte Wortkette nicht mit der Textvorlage, sondern mit der von Hand erstellten, tatsächlich gesprochenen Wortkette verglichen. Für eine künftige klinische Anwendung müssen die beiden Fehlerquellen jedoch strikt getrennt werden, zumal dann die von Hand erstellte Transliteration nicht mehr vorliegen wird. In der automatischen Sprachverarbeitung bedient man sich zu diesem Zweck u.a. der Konfidenzmaße, die bewerten, ob ein vorgegebenes Wort wirklich gesprochen wurde. Damit können aus den Sprachaufnahmen nur die Abschnitte zur Bewertung ausgewählt werden, in denen der vorgegebene Text fehlerfrei gelesen wurde.

Bewerter	K	L	R	S	U
alle	+0,83	+0,82	+0,77	+0,85	+0,68

Tabelle 1: Inter-Rater-Korrelation für das Kriterium „Verständlichkeit“ zwischen je einem Bewerter und dem Durchschnitt aus den vier übrigen

Bew.	K	L	R	S	U	alle
Korr.	–0,81	–0,65	–0,81	–0,79	–0,55	–0,84

Tabelle 2: Korrelation zwischen menschlichen Bewertern und dem Spracherkenner für das Kriterium „Verständlichkeit“

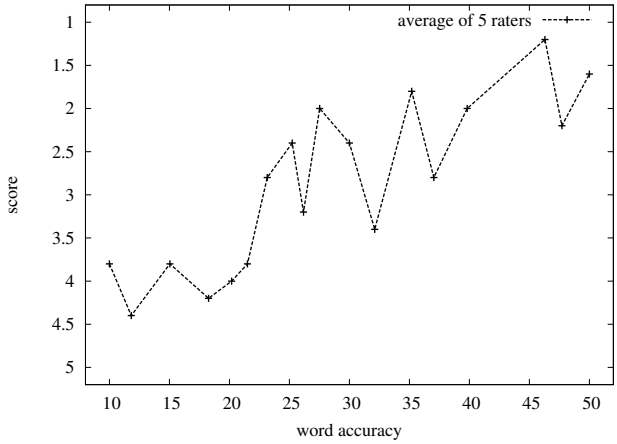

Abbildung 1: Wortakkuratheit und Durchschnittsbewertung von fünf Experten für 18 Patienten mit TE-Stimme

Zusammenfassung und Ausblick

Ein Spracherkennungssystem, das mit Aufnahmen laryngealer Sprecher trainiert worden war, wurde auf tracheoösophagealer Sprache eingesetzt. Die Ergebnisse dienten zur automatischen Bewertung der Verständlichkeit der Ersatzstimmen. Die Bewertungen von fünf Experten zeigten eine Korrelation von –0,84 zur vom Spracherkennungssystem berechneten Wortakkuratheit. Eine automatische Bewertung der Verständlichkeit und evtl. die Erweiterung auf weitere Stimmbewertungskriterien ist somit denkbar. Versuche zur prosodischen Analyse der TE-Stimmen laufen bereits.

Literatur

[1] Robbins J., Fisher H.B., Blom E.C., Singer M.I.: A Comparative Acoustic Study of Normal, Esophageal, and Tracheoesophageal Speech Production. Journal of Speech and Hearing Disorders, **49** (1984) 202–210

[2] Qi Y., Weinberg B.: Characteristics of Voicing Source Waveforms Produced by Esophageal and Tracheoesophageal Speakers. Journal of Speech and Hearing Research, **38** (1995) 536–548

[3] Stemmer G.: Modeling Variability in Speech Recognition. Dissertation am Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg (2005)

[4] Wahlster W. (Hrsg.): Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin (2000)

[5] Haderlein T., Steidl S., Nöth E., Rosanowski F., Schuster M.: Automatic Recognition and Evaluation of Tracheoesophageal Speech. Proc. 7th Int. Conf. Text, Speech & Dialogue (TSD), Brno, Tschechische Republik (2004) 331–338