Automatische Bestimmung der Verständlichkeit expressiver Sprache

Tino Haderlein[1], Elmar Nöth[2], Michael Döllinger[1]

[1]Klinikum der Universität Erlangen-Nürnberg, Phoniatrische und pädaudiologische Abteilung
[2]Universität Erlangen-Nürnberg, Department Informatik, Lehrstuhl für Mustererkennung

Einleitung

Zur umfassenden Evaluation der chronisch heiseren Stimme fehlen bisher validierte objektive Verfahren, die gleichzeitig klinisch praktikabel sind. Etablierte apparative Methoden bewerten die Stimme lediglich anhand von Aufnahmen gehaltener Vokale [1,2,3]. Das wichtigste Sprechkriterium, die Verständlichkeit, kann durch einzelne Vokale jedoch nicht abgebildet werden. Möglich wird dies durch Methoden der automatischen Sprachanalyse. In dieser Studie wird exemplarisch an einer Gruppe von Larynxteilresezierten gezeigt, wie zu diesem Zweck automatische Spracherkennungsverfahren und prosodische Analyse eingesetzt werden können. Der Fokus lag außerdem auf der Verständlichkeitsmessung am Telefon, da das Telefon in der heutigen Zeit eines der wichtigsten Kommunikationsmittel ist.

Material und Methoden

Als Testsprecher dienten 82 Personen, davon 14 Frauen, nach einer Larynxteilresektion. Diese wurde aufgrund einer malignen Erkrankung im Kehlkopfbereich durchgeführt. Das Durchschnittsalter betrug 62,3 mit einer Standardabweichung von 8,8 Jahren (min. 41,1, max. 86,1 Jahre). Die Stichprobe ist damit repräsentativ für die Gesamtbevölkerung hinsichtlich des Auftretens von Kehlkopfkrebs. Jede Person las den "Nordwind und Sonne"-Text [4] vor und wurde synchron mit einem Nahbesprechungsmikrofon (Logitech Premium Stereo Headset 980369-0914; Abtastfrequenz 16 kHz, Amplitudenauflösung 16 bit) und über das Telefon (8 kHz, 16 bit) aufgenommen.

Als Vergleichsbasis für die automatische Evaluierung bewerteten fünf erfahrene Logopädinnen und Ärzte 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.

Zur objektiven Messung der Verständlichkeit kam ein Spracherkennungssystem zur Anwendung, das am Lehrstuhl für Mustererkennung der Universität Erlangen-Nürnberg entwickelt [5] und bereits in zahlreichen Forschungsprojekten erfolgreich eingesetzt worden war. Bei der Spracherkennung werden die Sprachdaten zuerst in kleine zeitliche Einheiten unterteilt (16 ms), dann werden deren temporale und spektrale Merkmale analysiert und jeder Zeiteinheit ein Lautsymbol zugewiesen. Die so erkannten Lautfolgen werden mithilfe einer vorgegebenen Vokabularliste auf Wörter abgebildet. In diesem Fall bestand die Liste aus den Wörtern des vorgelesenen Textes. Das Spracherkennungssystem, kurz „Erkenner" genannt, wertet zwei unterschiedliche Informationskanäle aus: Das Wissen über das akustische Signal wird mit stochastischen Lautmodellen (Hidden-Markov-Modelle) modelliert [6], das Wissen über die Abfolge von Wörtern einer Sprache wird durch stochastische Sprachmodelle repräsentiert. Für den Zweck der vorliegenden Studie wurde lediglich ein Unigramm-Sprachmodell verwendet, das die Auftretenswahrscheinlichkeiten der einzelnen Wörter berücksichtigt. Auf diese Weise werden bei der Erkennung die akustischen Eigenschaften des Gesprochenen stärker bewertet. Das Erkennungssystem arbeitet auf akustischer Ebene polyphonbasiert, d.h. koartikulatorische Effekte werden ausgenutzt, um Informationen über die Nachbarn des jeweiligen Phonems zu gewinnen und dadurch die Erkennungssicherheit für die Lautkette zu erhöhen [7].

Als automatisches Verständlichkeitsmaß wurde zunächst die Wortkorrektheit (WR) des Erkenners verwendet. Sie beschreibt den Anteil der korrekt erkannten Wörter und 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 (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 [8, S. 54f], die aus Nahbesprechungsdaten bestand (16 kHz, 16 bit), durch Anwendung 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 prosodische Merkmale berechnet, um die automatische Verständlichkeitsmessung zu verbessern. Ein Prosodiemodul [9] verarbeitet die Ausgabe des Worterkennungsmoduls und das Sprachsignal selbst, d.h. die Sprachaufnahme. Es hat also Zugriff auf die Zeitzuordnung des Erkenners, die angibt, an welcher Stelle im Signal welcher Laut erkannt wurde. Für jede zu untersuchende Sprachuntereinheit (hier: Wörter) wird ein Referenzpunkt festgelegt, von dem aus die prosodischen Merkmale berechnet werden. Dieser Punkt wurde an das Wortende gelegt (Abbildung 1): Das Wort ist eine wohldefinierte Einheit in der Spracherkennung, es kann von jedem Standarderkenner bereitgestellt werden, und das Wortende kann einfacher bestimmt bzw. definiert werden als zum Beispiel die Mitte des Kerns einer betonten Silbe. Für jeden Referenzpunkt wurden 28 prosodische Merkmale auf Einzelwörtern oder der Pause zwischen zwei Wörtern berechnet. Zehn dieser Merkmale, z.B. die Dauer oder die Energie, wurden zusätzlich auf Wort-Pause-Wort-Intervallen bestimmt. Abb. 1 zeigt eine Auswahl der berechneten Merkmale. Weitere Details sind in [9,10] zu finden.

Zusätzlich zu den 38 für jede Wortposition berechneten Merkmalen wurden 16 globale Merkmale auf Abschnitten von je 15 Wörtern Länge berechnet. Sie umfassen Mittelwert und Standardabweichung von Jitter und Shimmer, Anzahl, Dauer und maximale Dauer von stimmhaften und stimmlosen Abschnitten, das Verhältnis der Anzahl bzw. Dauer von stimmhaften zu stimmlosen Bereichen sowie das Verhältnis der Dauer von stimmhaften bzw. stimmlosen Abschnitten zur Gesamtdauer des Signals. Die Standardabweichung der Sprachgrundfrequenz F_0 wurde einerseits nur auf den stimmhaften Abschnitten berechnet, andererseits wurde die Anzahl der stimmlosen Bereiche dadurch berücksichtigt, dass jeder stimmlose Bereich mit dem F_0-Pseudowert 0 in die Summe einging.

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.

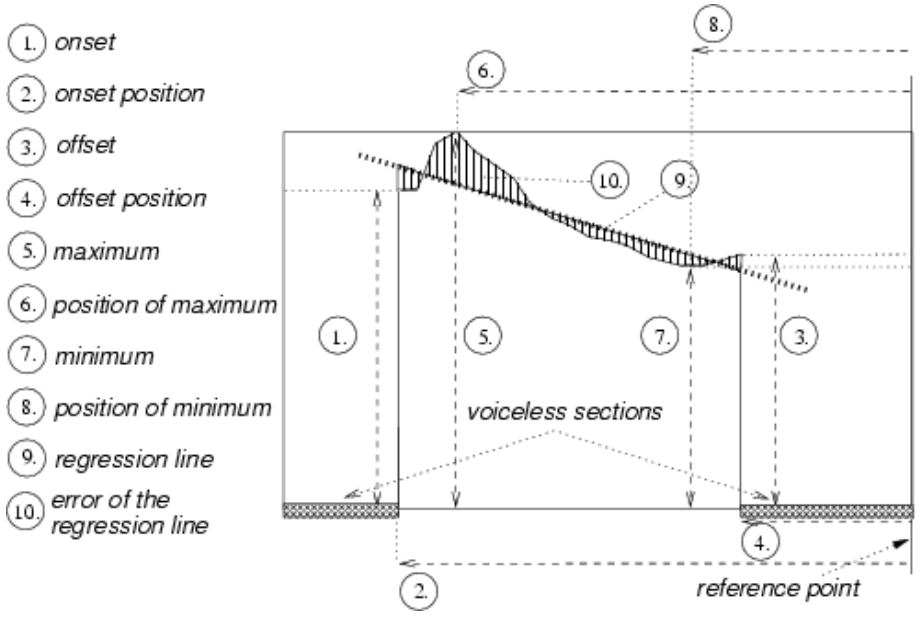

Abbildung 1: Verlauf der Sprachgrundfrequenz F_0 innerhalb eines Wortes mit je einem stimmlosen Bereich am Anfang und Ende und die berechneten prosodischen Merkmale (Auswahl; nach [11])

Mithilfe der Support-Vektor-Regression (SVR) [12] wurde schließlich aus der WR und den prosodischen Merkmalen die aussagekräftigste Kombination bestimmt und ein Vorhersagewert für die menschliche Verständlichkeitsbewertung des jeweiligen Patienten berechnet. Dieser Schritt wurde für die Headset- und die Telefonaufnahmen getrennt 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 [13]) sind in Tabelle 1 zusammengefasst:

	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

Tabelle 1: Korrelation nach Pearson für menschliche und automatische Verständlichkeitsbewertung

Die beste Merkmalsmenge aus WR und prosodischen Merkmalen war für beide Aufnahmequalitäten dieselbe (Tabelle 2). 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.

Merkmal	Headset	Telefon
Dauer der stillen Pause vor einem Wort	0,291	0,221
Standardabweichung der F_0 pro Wort (über alle Bereiche)	0,616	0,350
Standardabweichung des Jitter	0,243	0,332
Verhältnis Dauer stimmhafte Bereiche/Dauer der Aufnahme	-0,916	-0,775
Wortkorrektheit WR	-0,476	-0,641

Tabelle 2: SVR-Gewichtungsfaktoren der besten Merkmalsmenge für die Mensch-Maschine-Korrelation

Diskussion

Bei der perzeptiven Bewertung wurde die Verständlichkeit der Telefonaufnahmen erwartungsgemäß 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 perzeptiven Bewertung deshalb deutlich. Für die Telefonaufnahmen wird sogar der Referenzwert der menschlichen Inter-Rater-Korrelation übertroffen.

Da die Bewerter bei beiden Aufnahmetypen gleich gut übereinstimmen, sind die unterschiedlichen Gewichtungen in der Regressionsformel (Tabelle 2) offenbar nur durch das Fehlen der hohen Frequenzbereiche in der Telefonaufnahme begründet. Dadurch wird die Detektion stimmhafter Anteile und vor allem der F_0 schwieriger, was sich deutlich in den entsprechenden Gewichten niederschlägt. Das Fehlen der Information kann aber über die Wortkorrektheit wieder ausgeglichen werden. Die Übereinstimmung mit der perzeptiven Bewertung wird bei den Telefonaufnahmen sogar besser als bei den Headset-Aufnahmen (Tabelle 1). Offensichtlich extrahieren Mensch und Maschine aus den hohen Frequenzen nicht dieselbe Information bezüglich der Verständlichkeit, so dass das Fehlen dieser Bereiche sogar einen Vorteil für die Übereinstimmung der Bewertungen hat.

Im Hinblick auf die breite klinische Anwendung der Messmethode kann folgendes geschlossen werden: Die maschinelle Bewertung der Verständlichkeit ist auch per Telefon prinzipiell möglich, wie hier beispielhaft an der pathologischen Stimme nach Larynxteilresektion gezeigt wird. Die automatischen Methoden können im klinischen Einsatz als objektive „zweite Meinung“ bei der Stimm- und Sprechbewertung dienen. Sie eröffnen auch die Möglichkeit für neue Anwendungsfelder, wie z.B. die Bewertung von Restzuständen nach Aphasien oder von emotionalen Elementen bei Depressionen. Sie wurden außerdem so entworfen, dass sie für automatische Stimmbelastungstests und telemedizinische Anwendungen geeignet sind. Patient, Therapeut und Auswertungsrechner müssen sich dann nicht mehr am selben Ort befinden. Der Patient kann selbständig eine Aufnahme erstellen, die automatisch ausgewertet wird. Die Ergebnisse sind dann in der Klinik oder Arztpraxis abrufbar.

Danksagung

Diese Arbeit wurde von der Deutschen Krebshilfe (Fördernr. 107873) und von der Else Kröner-Fresenius-Stiftung (Fördernr. 2011_A167) gefördert.

Literatur

[1] Friedrich G, Dejonckere PH. The voice evaluation protocol of the European Laryngological Society (ELS) – first results of a multicenter study. Laryngorhinootol. 2005;84(10):744-752.

[2] Fröhlich M, Michaelis D, Strube HW, Kruse E. Acoustic voice analysis by means of the hoarseness diagram. J Speech Lang Hear Res. 2000;43(3):706-720.

[3] Wuyts F, de Bodt M, Molenberghs G, Remacle M, Heylen L, Millet B, van Lierde K, Raes J, van de Heyning P. The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach. J Speech Lang Hear Res. 2000;43(3):796-809.

[4] International Phonetic Association. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press, Cambridge, 1999.

[5] Stemmer G. Modeling Variability in Speech Recognition. Band 19 von Studien zur Mustererkennung. Logos Verlag, Berlin, 2005.

[6] Jelinek F. Statistical Methods for Speech Recognition. The MIT Press, Cambridge, 1997.

[7] Schukat-Talamazzini EG, Niemann H, Eckert W, Kuhn T, Rieck S. Automatic Speech Recognition without Phonemes. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech). European Speech Communication Association (ESCA), Berlin, 1993, 129-132.

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

[9] Zeissler V, Adelhardt J, Batliner A, Frank C, Nöth E, Shi RP, Niemann H. The prosody module. In Wahlster W (Hrsg.): SmartKom: Foundations of Multimodal Dialogue Systems. Springer, New York, 2006, 139-152.

[10] Zeissler V. Robuste Erkennung der prosodischen Phänomene und der emotionalen Benutzerzustände in einem multimodalen Dialogsystem. Band 35 von Studien zur Mustererkennung. Logos Verlag, Berlin, 2012.

[11] Buckow J, Warnke V, Huber R, Batliner A, Nöth E, Niemann H. Fast and Robust Features for Prosodic Classification. In Matoušek V, Mautner P, Ocelíková J, Sojka P (Hrsg.): Proc. Text, Speech and Dialogue (TSD’99). Band 1692 von Lecture Notes for Artificial Intelligence. Springer, Berlin, 1999, 193-198.

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

[13] Haderlein T, Nöth E, Batliner A, Eysholdt U, Rosanowski F. Automatic Intelligibility Assessment of Pathologic Speech over the Telephone. Logop Phoniatr Vocol. 2011;36(4):175-181.