Tino Haderlein, Tobias Bocklet, Florian Hönig, Elmar Nöth, Frank Rosanowski

Automatische Stimmanalyse nach Larynxteilresektion mithilfe akustischer Sprechermodellierung

Einleitung

Objektiv-apparative Stimmbewertungen werden derzeit meist auf der Basis gehaltener Vokale durchgeführt. Jedoch reflektiert ein isolierter Vokal keine reale Kommunikationssituation. In früheren Arbeiten wurde gezeigt, dass automatische Spracherkennungsverfahren verwendet werden können, um pathologische Sprecher automatisch zu bewerten [1]. Grundlage der Methode war die Annahme, dass das Spracherkennungssystem umso weniger Wörter eines vorgegebenen Standardtextes „versteht“, je schlechter die Stimmqualität des Sprechers ist. In dieser Studie wird nun gezeigt, dass eine hohe Korrelation zwischen menschlicher und maschineller Qualitäts- und Verständlichkeitsbewertung auch dann erzielt werden kann, wenn aus allen gesprochenen Lauten lediglich ein akustisches Modell des jeweiligen Sprechers aufgebaut wird, ohne eine Spracherkennung als solche durchzuführen.

Material und Methoden

Als Testsprecher für diese Querschnittsuntersuchung dienten 85 Personen mit Krebserkrankungen des Kehlkopfes, davon 65 nach einer Larynxteilresektion. Das Durchschnittsalter innerhalb der Gruppe betrug 60,7±9,2 Jahre (min. 34,0, max. 83,0 Jahre), zehn der Patienten waren weiblich. Jede Testperson las den „Nordwind und Sonne“-Text vor und wurde dabei mit einer Abtastfrequenz von 16 kHz und einer Amplitudenauflösung von 16 bit aufgenommen.

Als Vergleichsbasis für die automatische Evaluierung bewerteten fünf Experten die Gesamtqualität der Stimme auf einer visuellen Analogskala mit einer Breite von 10 Zentimetern, wobei der linke Rand (Wert 0,0) dem Prädikat „sehr gut“ und der rechte Rand (Wert 10,0) dem Prädikat „sehr schlecht“ entsprach. Weiterhin wurde das Kriterium „Gesamtverständlichkeit“ bei jedem Sprecher mit Noten auf einer Skala von 1 („sehr gut verständlich“) bis 5 („extrem schlecht verständlich“) bewertet. Aus den fünf Bewertungen

pro Kriterium für jede Aufnahme wurde jeweils eine Durchschnittsnote gebildet.

Die aus der Sprachaufnahme gewonnenen automatischen Messwerte („Merkmale“) heißen Mel-Frequenz-Cepstrum-Koeffizienten (MFCC). Die Sprachaufnahme wird hierfür in Abschnitte von je 16 ms Dauer unterteilt, wobei sich aneinandergrenzende Abschnitte um 6 ms überlappen. Die Ränder dieser „Frames“ werden mit einem sog. Hamming-Fenster geglättet. Das Frequenzspektrum der einzelnen Frames wird mit 25 sich überschneidenden, dreieckigen Masken gefiltert, die die Frequenzbandverarbeitung des menschlichen Gehörs nachbilden. Durch Rücktransformation der Summenwerte der Einzelfilter in den Zeitbereich erhält man das Cepstrum, dessen Koeffizienten als Merkmale für die akustischen Sprechermodelle dienen. Der Merkmalsvektor eines Frames besteht dabei aus der Signalenergie, den MFCCs Nr. 2 bis 12 und jeweils der 1. Ableitung dieser Messwerte.

Alle Merkmalsvektoren von allen Sprechern werden zunächst zusammengefasst, um ein statistisches, universelles Hintergrundmodell (Universal Background Model, UBM) zu erzeugen. Dieses stellt sich als gewichtete Summe aus Normalverteilungen, d.h. als Gaußsches Mischverteilungsmodell (GMM), dar. Von dieser Grundlage eines allgemeinen Modells für pathologische Stimmen werden dann mittels Maximum-A-Posteriori-Adaption (MAP) die akustischen Modelle für die einzelnen Sprecher abgeleitet.

Aus den einzelnen Sprechermodellen werden nun wiederum verschiedene Größen als charakteristische Merkmale des jeweiligen Sprechers herangezogen, und zwar die Mittelwerte, Gewichte und Kovarianzen der 128 Gaußdichten des GMM. Mithilfe der Support-Vektor-Regression (SVR) wird dann eine Funktion ermittelt, die aus der Kombination der Merkmale einen Vorhersagewert für die menschliche Bewertung des jeweiligen Patienten liefert. Eine detaillierte Beschreibung des Verfahrens ist in [2] zu finden.

Ergebnisse

Die Korrelation zwischen der menschlichen Verständlichkeitsbewertung und der Bewertung der Gesamtqualität betrug r=0,96. Die Inter-Rater-Korrelation kann in Tab. 1 abgelesen werden. Die Korrelation zwischen dem maschinell berechneten SVR-Vorhersagewert und der tatsächlichen menschlichen Durchschnittsbewertung lag bei r=0,79 für die Gesamtqualitätsbewertung und r=0,73 für die Verständlichkeit.

Tab. 1: Korrelation zwischen einem jeweils Bewerter und dem Durchschnitt der übrigen

Bewerter	MS	RO	SA	SU	VD
Verständlichkeit	0,82	0,76	0,80	0,86	0,83
Gesamtqualität	0,85	0,81	0,88	0,93	0,87

Diskussion

Die Korrelation zwischen menschlicher und automatischer Bewertung zeigt, dass die verwendeten Methoden zur objektiven Stimmevaluierung geeignet sind. Die Daten dieser Studie zeigten in Übereinstimmung mit früheren Untersuchungen der Arbeitsgruppe [1] eine hohe Korrelation zwischen menschlicher Verständlichkeits- und Gesamtbewertung. Daraus kann geschlossen werden, dass aus einer der beiden Bewertungen verlässlich auf die jeweils andere geschlossen werden kann. Somit ist die automatische Evaluierung der Sprachverständlichkeit nicht auf eine detaillierte Analyse der gesprochenen Wörter angewiesen, sondern ist auch mithilfe eines rein akustischen Modells möglich. Die hohe Übereinstimmung zwischen den beiden Bewertungskriterien wird jedoch in der Literatur nicht bestätigt [3]. Zur Bestätigung der aktuellen Ergebnisse sind deshalb weitere Untersuchungen an einer größeren Stichprobe geplant.

Danksagung

Diese Arbeit wird von der Deutschen Krebshilfe (Fördernr. 107873) gefördert.

Literatur

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

[2] Bocklet T, Haderlein T, Hönig F, Rosanowski F, Nöth E. Evaluation and Assessment of Speech Intelligibility on Pathologic Voices Based upon Acoustic Speaker Models. 3rd Advanced Voice Function Assessment International Workshop (AVFA2009), Madrid, 2009; 89-92.

[3] Preminger JE, Van Tasell DJ. Quantifying the relation between speech quality and speech intelligibility. J Speech Hear Res 1995; 38(3):714-25.