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

[1]Lehrstuhl für Informatik 5 (Mustererkennung), Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen
[2]Phoniatrische und pädaudiologische Abteilung in der HNO-Klinik, Klinikum der Universität Erlangen-Nürnberg, Erlangen

Stimmqualitätsmessung mittels prosodischer Analyse verschiedener gelesener Textabschnitte

Einleitung

In früheren Arbeiten wurde gezeigt, dass prosodische Analyseverfahren verwendet werden können, um die Stimm- und Sprecheigenschaften von pathologischen Sprechern automatisch zu bewerten [1]. Bisher wurden jedoch die prosodischen Messwerte für alle Wörter des jeweils gelesenen Textes zusammengefasst. Durch die Mittelwertbildung über lange und kurze Wörter und alle Wortpositionen im Satz hinweg kann jedoch Information verlorengehen. In dieser Studie wurde deshalb untersucht, welchen Einfluss die Position und grammatikalische Wortklasse auf die automatische Stimmqualitätsbewertung haben.

Material

Als Testsprecher dienten 73 repräsentativ ausgewählte Personen (24 Männer, 49 Frauen) mit chronischer Heiserkeit nichtmaligner Ursache. 45 Personen litten an funktioneller Dysphonie, 24 an organischer Dysphonie und vier an Laryngitis. Das Durchschnittsalter betrug 48,3±16,8 Jahre (min. 19, max. 85 Jahre). Jede Person las den „Nordwind und Sonne“-Text vor und wurde dabei mit einem Nahbesprechungsmikrofon (AKG 420 C, AKG Acoustics, Wien; Abtastfrequenz 16 kHz, Amplitudenauflösung 16 bit) aufzeichnet. Als Referenz für die objektive Analyse bewerteten fünf erfahrene Personen, darunter eine HNO-Ärztin, ein HNO-Arzt und drei Logopädinnen, die Gesamtqualität der Stimme auf einer visuellen Analogskala der Breite 10 cm (0,0: „sehr gut“; 10,0: „sehr schlecht“).

Methode

Die automatisch erhobenen Messwerte, die Informationen über die jeweils zu ermittelnde Eigenschaft einer Stimme, eines Sprechers, oder für die automatische Spracherkennung enthalten sollen, werden in der Informatik als Merkmale bezeichnet. In der hier beschriebenen Studie zur Analyse der Stimmqualität wurden, basierend auf Wort- und Pausendauern, der Sprachgrundfrequenz F_0 und der Energie im Signal [2], 33 prosodische Merkmale pro Wort bzw. pro Wort-Pause-Wort-Intervall berechnet. Die größte Gruppe umfasst die F_0-Merkmale, die u.a. Mittelwert, Minimum, Maximum, den Wert bei Stimmeinsatz und -ausklang sowie ihre jeweiligen Positionen im betrachteten Intervall enthalten. 15 weitere Merkmale, auf Abschnitten von jeweils 15 Wörtern Länge berechnet, umfassen Mittelwert und Standardabweichung von Jitter und Shimmer, die 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 hier ebenfalls textbasiert ausgewertet.

Da die menschlichen Bewertungen für den gesamten Text erfolgen, wurde bisher auch jedes pro Wort bzw. Aufnahmeabschnitt berechnete prosodische Merkmal über die ganze Aufnahme gemittelt. Für diese Studie wurden deshalb folgende Szenarien betrachtet:

- Mittelwertbildung für jedes Merkmal über alle 108 Wörter (Referenzexperiment)
- Mittelung nur über Substantive (24 Wörter)
- Mittelung nur über Substantive und Verben (44 Wörter)
- Mittelung nur über Satzanfänge (erste drei Wörter eines jeden Satzes; 18 Wörter)
- Mittelung nur über Anfänge der sechs Haupt- und zehn Nebensätze (jeweils erste drei Wörter; insgesamt 48 Wörter)

Substantive und Verben wurden ausgewählt, da sie hinsichtlich der Verständlichkeit bedeutender sind als Funktionswörter, also Artikel, Präpositionen und Konjunktionen [3]. Die Anfänge von Sätzen und Nebensätzen, ohne Berücksichtigung der grammatikalischen Wortklasse, wurden aufgrund der klinischen Anwendung ausgewählt. Durch die große

Sprechanstrengung und kürzere Phonationszeit sind viele Stimmpatienten ohnehin gezwungen, ihre Äußerungen in kürzere Abschnitte zu zerlegen.

Ergebnisse

Die durchschnittliche menschliche Stimmqualitätsbewertung lag bei 4,74±2,51 (min. 0,32, max. 9,50). Die Inter-Rater-Korrelation (jeweils eines Bewerters mit dem Mittelwert der übrigen) betrug r=0,86. Tabelle 1 zeigt die Korrelation der menschlichen und maschinellen Bewertung für ausgewählte Merkmale (r≥0,50 für mindestens eine Konstellation).

Tab. 1: Mensch-Maschine-Korrelation r, abhängig vom gewählten Textausschnitt (ges.: gesamter Text, S: Substantive, S+V: Substantive und Verben, Anf: Satzanfänge, An+: Anfänge von Haupt- und Nebensätzen; der beste Wert pro Zeile ist jeweils fett gedruckt.)

prosodisches Merkmal	ges.	S	S+V	Anf	An+
Dauer der stillen Pause vor dem aktuellen Wort	**0.56**	0.55	0.55	0.33	0.46
Energieregressionskoeffizient (Wort-Pause-Wort-Intervalle)	0.37	0.24	0.37	0.40	**0.50**
normierte Energie eines Wort-Pause-Wort-Intervalls	**0.60**	0.56	0.49	0.54	**0.60**
normierte Dauer eines Wort-Pause-Wort-Intervalls	0.55	**0.57**	0.53	0.39	0.49
absolute Dauer eines Wort-Pause-Wort-Intervalls	0.47	**0.57**	0.49	0.35	0.43
Mittelwert des Jitter	0.67	0.69	***0.72***	0.66	0.67
Standardabweichung des Jitter	0.58	0.63	**0.64**	0.52	0.56

Diskussion und Fazit

Die hohen Korrelationswerte zur Dauer von Pausen oder Wort-Pause-Wort-Intervallen deuten auf eine Verbindung von Sprechanstrengung und Stimmqualität bei chronisch Heiseren hin. Die Stimmqualität selbst wird jedoch durch Merkmale auf Basis der Energie, d.h. der Amplitudenwerte in der Aufnahme, und besonders durch den Jitter am besten dargestellt. Bezüglich der Frage, welche Ausschnitte des Textes zur Berechnung herangezogen werden sollen, ergibt sich kein einheitliches Bild. Bis auf die Anfänge der Hauptsätze erweist sich jedes Berechnungsszenario bei bestimmten Merkmalen als

vorteilhaft. Für die jitterbasierten Merkmale empfiehlt sich die Analyse von Substantiven und Verben. Die menschliche Stimmqualitätsbewertung lässt sich hier allein durch den Mittelwert des Jitter mit einer Korrelation von r=0,72 annähern. Dies zeigt die grundsätzliche Eignung des Verfahrens. Durch die Kombination verschiedener Merkmale mittels Regressionsverfahren sind noch deutlich bessere Ergebnisse zu erwarten [1].

Danksagung

Die Arbeit von Herrn Döllinger wurde von der DFG (Fördernr. DO1247/8-1) gefördert.

Literatur

[1] Haderlein T, Döllinger M, Matoušek V, Nöth E. Objective Voice and Speech Analysis of Persons with Chronic Hoarseness by Prosodic Analysis of Speech Samples. Logoped Phoniatr Vocol 2016;41(3):106-16.

[2] 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. New York: Springer, 2006. S.139-52.

[3] Rubenstein H, Pickett J. Intelligibility of Words in Sentences. J Acoust Soc Am 1958;30(7),670.