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

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

Automatische prosodische Stimmqualitätsbewertung aus Textaufnahmen bei großem Mikrofonabstand

Einleitung

Mit prosodischen Analyseverfahren können Stimm- und Sprecheigenschaften automatisch aus Textaufnahmen bewertet werden [1]. In der hier vorgestellten Studie wurde getestet, ob sich diese Methode auch zur Stimmqualitätsmessung eignet, wenn die untersuchten Personen kein Nahbesprechungsmikrofon (Headset) tragen, da die Aufnahmesituation gelegentlich als belastend empfunden wird. In der Therapiesitzung ist der Abstand zwischen Testperson und Logopädin üblicherweise klein. Es wurde ermittelt, wie sich die Übereinstimmung von perzeptiver und automatischer Bewertung ändert, wenn das Mikrofon für die Aufnahme einen größeren Abstand zur Testperson aufweist.

Material und Methode

68 Männer und 14 Frauen nach einer Larynxteilresektion mit einem Durchschnittsalter von 62,3±8,8 Jahren (min. 41,1, max. 86,1 Jahre) lasen den „Nordwind und Sonne“-Text vor und wurden mit einem Headset (Abtastfrequenz 16 kHz, Amplitudenauflösung 16 bit) aufgenommen. Fünf erfahrene Logopädinnen und Ärzte bewerteten in jeder Aufnahme die Gesamtqualität der Stimme auf einer visuellen Analogskala zwischen 0,0 („sehr gut“) und 10,0 („extrem schlecht“). Aus den Bewertungen für jede Person wurde jeweils eine Durchschnittsnote als Referenz für alle automatischen Evaluierungsszenarien gebildet.
Um Sprachproben mit anderem Mikrofonabstand zu erhalten, dabei aber sicherzustellen, dass diese sich nur in der Raumakustik und nicht in Stimmqualität, Störgeräuschen oder Vokabular unterscheiden, wurden die vorhandenen Headset-Aufnahmen künstlich

verhallt. Dazu wurden Raumimpulsantworten in einem Raum der Größe 580×590×310 cm gemessen. Durch schallschluckende Teppiche und Vorhänge wurde die Nachhallzeit T_{60}, in welcher der Nachhall um 60 dB abklingt, zwischen 250 und 400 ms variiert. Die angenommenen Sprecherpositionen lagen auf drei Halbkreisen mit 60, 120 und 240 cm Mikrofonabstand (s. Abb.). Für alle Positionen wurde die Mensch-Maschine-Korrelation der Stimmqualitätsbewertung für die entsprechend verhallte Aufnahme berechnet.

Die apparative Diagnostik verwendet sog. prosodische Merkmale. Basierend auf Wort- und Pausendauern, der Sprachgrundfrequenz F_0 und der Energie im Signal [2] wurden 33 prosodische Merkmale pro Wort bzw. pro Wort-Pause-Wort-Intervall erfasst. Die größte Gruppe umfasste die F_0-Merkmale, die u.a. Mittelwert, Minimum, Maximum, den Wert bei Stimmeinsatz und -ausklang sowie ihre jeweiligen Positionen im betrachteten Intervall enthielten. 15 weitere Merkmale, auf Abschnitten von jeweils 15 Wörtern Länge berechnet, umfassten Mittelwert und Standardabweichung von Jitter und Shimmer, weiterhin 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 subjektiv-auditive Bewertung für den gesamten Text erfolgte, wurden auch für jedes prosodische Merkmal alle pro Wort bzw. Aufnahmeabschnitt berechneten Werte über die gesamte Aufnahme gemittelt.

Mithilfe der Support-Vektor-Regression (SVR) wurde schließlich aus allen Messwerten die aussagekräftigste Kombination bestimmt und ein Vorhersagewert für die durchschnittliche perzeptive Bewertung der jeweiligen Testperson berechnet. Diese Optimierung wurde für die Headset-Aufnahmen und auch für die Sprecherposition durchgeführt, die akustisch von der Nahbesprechungssituation am stärksten abweicht (Impulsantwort h423165, T_{60}: 400 ms, 240 cm Mikrofonabstand, 165° Winkel zum Mikrofon; vgl. Abb.). Dann wurden mit den ermittelten Merkmalsmengen auch alle übrigen akustischen Szenarien untersucht.

Ergebnisse

Die durchschnittliche Stimmqualitätsnote der fünf bewertenden Personen lag für die 82 Sprecher bei 5,59±2,49 (min. 1,46; max. 9,52). Die Inter-Rater-Korrelation (ein Bewerter

gegen den Mittelwert der übrigen) war r=0,89. Die Mensch-Maschine-Korrelationen sind im Folgenden für alle Raumimpulsantworten zusammengestellt. Angegeben sind jeweils die Nachhallzeit T_{60}, der Mikrofonabstand („Abst."), der Sprechwinkel zum Mikrofon (α, s. Abb.) sowie die Korrelation unter Verwendung der besten Merkmalsmengen für Nahbesprechung (r_{nah}) bzw. verhallte Aufnahmen (r_{hall}).

Impuls-antwort	T_{60} (ms)	Abst. (cm)	α (°)	r_{nah}	r_{hall}	Impuls-antwort	T_{60} (ms)	Abst. (cm)	α (°)	r_{nah}	r_{hall}
Headset	—	3–5	90	**0,74**	0,68						
h411000	250	60	0	0,64	0,67	h421045	400	60	45	0,66	0,70
h411090	250	60	90	0,67	0,68	h421135	400	60	135	0,65	0,69
h412060	250	120	60	**0,74**	0,68	h422015	400	120	15	0,72	0,67
h412150	250	120	150	0,71	0,69	h422105	400	120	105	0,67	0,69
h413030	250	240	30	0,72	0,69	h423075	400	240	75	0,72	0,71
h413120	250	240	120	0,70	0,71	h423165	400	240	165	0,72	**0,73**

Die beste Merkmalsmenge für stark verhallte Aufnahmen enthielt die durchschnittliche Länge eines Wort-Pause-Wort-Intervalls und damit einen Hinweis auf Sprechanstrengung bzw. -tempo, den mittleren normierten Wert der F_0 sowie deren Standardabweichung. Für Nahbesprechungsaufnahmen kommen noch der F_0-Wert beim Stimmeinsatz sowie der prozentuale zeitliche Anteil der als stimmhaft erkannten Aufnahmeabschnitte hinzu.

Diskussion und Fazit

Im Gegensatz zur Verständlichkeitsbewertung [3] liegt die Stimmqualitätsbewertung aus verhallten Textaufnahmen mit den verfügbaren Messwerten im Vergleich mit der perzeptiven Bewertung noch nicht auf dem Niveau eines durchschnittlichen menschlichen Bewerters. Mensch-Maschine-Korrelationen bis r=0,74 zeigen jedoch die grundsätzliche Eignung des Verfahrens. Mikrofonabstand und Sprecherposition haben zum Teil nennenswerten Einfluss auf die Ergebnisse. Durch eine andere Zusammensetzung der Merkmalsmenge können diese jedoch ausgeglichen werden.

Danksagung: Wir danken Wolfgang Herbordt für Software und Daten zur Verhallung. Michael Döllinger wird von der Deutschen Krebshilfe, Fördernr. 111332, unterstützt.

Literatur

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

[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.139-152.

[3] Haderlein T, Döllinger M, Schützenberger A, Nöth E. Influence of Reverberation on Automatic Evaluation of Intelligibility. In Sojka P, Horák A, Kopeček I, Pala K (Hrsg.): Text, Speech and Dialogue (TSD 2016). LNAI. Springer. Im Druck.

Abbildung: Angenommene Sprecherpositionen zur Messung von Raumimpulsantworten für die künstliche Verhallung bei verschiedenen Nachhallzeiten T_{60}; das Mikrofon wird durch den schwarzen Punkt symbolisiert, der Sprecherwinkel α zum Mikrofon beginnt rechts oben mit 0° und wächst im Uhrzeigersinn bis 165°.