Tino Haderlein[1,2], Catherine Middag[3], Jean-Pierre Martens[3], Michael Döllinger[1], Elmar Nöth[2]

[1]Phoniatrische und pädaudiologische Abteilung, Klinikum der Universität Erlangen-Nürnberg

[2]Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen-Nürnberg

[3]Vakgroep voor Elektronica en Informatiesystemen (ELIS), Universiteit Gent, Gent, Belgien

Untersuchung von Sprachaufnahmen heiserer Stimmen mittels phonologischer und phonemischer Merkmale

Einleitung

Gängige Methoden der automatischen Stimmanalyse stützen sich meist auf Aufnahmen gehaltener Vokale. Sprachaufnahmen hingegen enthalten Stimmeinsätze, Schwankungen der Grundfrequenz F_0 und auch Sprechpausen und erlauben so eine wesentlich umfangreichere Analyse. Neuere Entwicklungen verwenden akustische Modelle, die die phonologischen und phonemischen Eigenschaften einer Äußerung über die Zeit beschreiben. Diese Verfahren können auch zur sprachenunabhängigen Analyse verwendet werden, d.h. die Testsprecher müssen nicht dieselbe Sprache sprechen wie die Personen, mit denen die Modelle trainiert wurden [1]. In der vorgestellten Studie wurde untersucht, ob sich mittels solcher akustischer Modelle verschiedene Arten chronischer Heiserkeit differenzieren lassen.

Material

Als Testsprecher dienten 69 repräsentativ ausgewählte Personen deutscher Muttersprache mit chronischer Heiserkeit nichtmaligner Ursache. 13 Männer und 32 Frauen wurden zur Gruppe „funktionelle Dysphonie“, 9 Männer und 15 Frauen zur Gruppe „organische Dysphonie“ zusammengefasst (Tab. 1).

Tab. 1: Alter in Jahren und RBH-Bewertung für alle Sprecher (n=69) sowie für funktionelle (n=45) und organische (n=24) Dysphonien; Maximalwert für alle RBH-Kriterien war 3,00.

Gruppe	Alter				R			B			H		
	μ	σ	min	max	μ	σ	min	μ	σ	min	μ	σ	min
gesamt	48,9	16,2	20	85	1,57	0,84	0,00	1,21	0,82	0,00	1,86	0,86	0,00
funktionell	47,1	16,3	20	85	1,35	0,85	0,00	0,97	0,75	0,00	1,61	0,88	0,00
organisch	52,2	15,6	25	79	1,98	0,65	0,80	1,64	0,76	0,60	2,33	0,58	1,20

Jede Person las den „Nordwind und Sonne“-Text vor und wurde mit einem Nahbesprechungsmikrofon (Abtastfrequenz 16 kHz, Amplitudenauflösung 16 bit) aufgenommen. Zum Vergleich mit der automatischen Analyse wurde für jede Aufnahme aus RBH-Bewertungen von fünf Experten eine Durchschnittsnote gebildet.

Methode

Alle Aufnahmen wurden in 25 ms lange Abschnitte (Frames) mit einer Fortschaltzeit von 10 ms unterteilt. In jedem Frame wurden zwölf Mel-Frequenz-Cepstrumkoeffizienten (MFCC) und eine Energiekomponente berechnet. Daraus wurden mittels neuronaler Netzwerke 14 phonologische Eigenschaften pro Frame berechnet [2, 3]:

- Stimmquelle: Stimmgebung
- Artikulationsart: Stille, Konsonantnasalität, Vokalnasalität, Turbulenz (bei Frikativen und Plosiven)
- Artikulationsort (Konsonanten): labial, labiodental, alveolar, velar, glottal, palatal
- Vokalmerkmale: Vokalhöhe, Ort, Rundung

Die Netze waren mit flämischen Sprachdaten trainiert. Jede phonologische Eigenschaft wurde durch zwei Subnetzwerke analysiert. Eines davon ermittelte die Relevanz der Eigenschaft zum gegebenen Zeitpunkt (z.B. sind bei Konsonanten Vokalmerkmale nicht relevant), und das zweite, ob die jeweilige Eigenschaft (z.B. „labial“) aktuell vorlag oder nicht. Die zeitliche Analyse der Ausgaben der Netzwerke lieferte jeweils Mittelwert und Standardabweichung, den Prozentsatz, wie oft die jeweilige Ausgabe gemäß spezifischer Schwellwerte hoch, mittel und niedrig war, die mittlere Höhe der Maximalwerte und die Dauer einer Transition von einem niedrigen zum hohen Wert. Insgesamt wurden 504 phonologische Merkmale errechnet, wovon jedoch viele eine ähnliche Information trugen.

Phonemische Merkmale [2] basieren auf den Produktionswahrscheinlichkeiten für einzelne Laute, wie sie bei der automatischen Spracherkennung berechnet werden. Sie werden wiederum aus den phonologischen Eigenschaften mittels neuronaler Netze bestimmt. Tritt eine maximale A-posteriori-Wahrscheinlichkeit für einen konkreten Laut auf, werden über den zeitlichen Verlauf der Wahrscheinlichkeit Mittelwert, Standardabweichung und die Extremwerte bestimmt. Zusätzlich werden die Dauer der Frames, die diesem Laut zugewiesen wurden, und die mittlere Wahrscheinlichkeit dieses Lautes über alle Frames berechnet. Insgesamt erhält man auf diese Weise 495 phonemische Merkmale.

Zur Ermittlung der relevantesten Merkmale wurde das Prinzip der Ensemble Linear

Regression angewandt. Für jeweils zwei Merkmale wurde mittels linearer Diskriminanzanalyse (LDA) die Klassifikation aller Sprecher in die Klassen „funktionell“ und „organisch“ durchgeführt und die Erfolgsrate bestimmt. Jedes Experiment erfolgte dabei mit fünffacher Kreuzvalidierung auf allen Daten.

Ergebnisse

Für die Unterscheidung von funktioneller und organischer Dysphonie wurden zwei wichtigste Merkmale ermittelt, mit deren Hilfe sich eine Klassifikationsrate von 83% ergab:

- *consonantnasality_presence_meanmin:* Minima der Präsenz von Nasalität
- *h_meanneg:* mittlere minimale Wahrscheinlichkeit, dass ein /h/ gesprochen wurde (phonemisches Merkmal)

Abb. 1 zeigt die graphische Darstellung der beiden Merkmale.

Diskussion und Fazit

Patienten mit organischer Dysphonie wiesen in den beiden relevantesten Merkmalen tendenziell höhere Werte auf. Sie deuten auf eine höhere Behauchtheit in dieser Gruppe hin, was durch die RBH-Bewertung bestätigt wurde. Bereits zwei Merkmale erlauben eine effektive Visualisierung und Klassifikation. Sprachenunabhängige phonologische und phonemische Merkmale aus Sprachaufnahmen bei repräsentativ ausgewählten Patienten bilden somit eine geeignete Basis für weitere Untersuchungen.

Danksagung

Förderer dieser Arbeit waren die Else Kröner-Fresenius-Stiftung (Nr. 2011_A167), die Kampagne „Kom op tegen Kanker“ der Vlaamse Liga tegen Kanker VZW in Brüssel und The Netherlands Cancer Institute / Antoni van Leeuwenhoek Hospital (Amsterdam).

Literatur

[1] Middag C, Bocklet T, Martens J-P, Nöth E: Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment; in Proc. Interspeech. International Speech Communication Association, 2011, S. 3005-3008.

[2] Middag C. Automatic Analysis of Pathological Speech. Universiteit Gent, Gent, Belgien, 2012.

[3] Middag C, Saeys Y, Martens J-P: Towards an ASR-free objective analysis of pathological speech; in Proc. Interspeech. International Speech Communication Association, 2010, S. 294-297.

Abbildung

Abb. 1: Relevanteste Merkmale zur Unterscheidung funktioneller (F) und organischer (O) Dysphonie