Hörgeräte-basierte Parkinson-Sprachanalyse

Finn Spitz[1], Christin Baasch[1], Gerhard Schmidt[1], Ulrich Heute[1], Adelheid Nebel[2], Günther Deuschl[2]

[1] *Digitale Signalverarbeitung und Systemtheorie, Christian-Albrechts-Universität zu Kiel, E-Mail: {fisp, chrb, gus, uh}@tf.uni-kiel.de*

[2] *Neurologie, Christian-Albrechts-Universität zu Kiel, E-Mail: {a.nebel, g.deuschl}@neurologie.uni-kiel.de*

Einleitung

Von der Parkinson-Krankheit, einer der häufigsten Erkrankungen des zentralen Nervensystems, sind allein in Deutschland ca. 250.000 Menschen betroffen. Durch Medikation und Therapien können die motorischen Symptome behandelt werden. Gegen die oftmals auftretenden Sprachstörungen, wie eine monotone, nuschelnde Stimme, leises Sprechen und verstärkte Atemgeräusche, hilft jedoch in der Regel nur eine logopädische Therapie [7].
Um diese zu unterstützen und die Sprachqualität der Patienten in Echtzeit im Alltag messen zu können, wird die Möglichkeit untersucht, eine Sprachqualitätsschätzung mithilfe eines Hörgerätes durchzuführen. Die zugehörige Signalverarbeitung wird parallel, zusätzlich zu den im Hörgerät verwendeten Algorithmen, implementiert. Als Ergebnis wird eine Wertung errechnet, aus der erkennbar ist, ob der Sprachpegel ausreichend hoch und die Sprache verständlich ist. Dazu wird die nach Ramig modifizierte NTID-Skala verwendet, um Sprachabschnitte zu klassifizieren.
Zur korrekten Ausführung der Sprachanalyse wird das System durch eine Eigenspracherkennung (welche nicht Gegenstand dieses Beitrags ist) gesteuert, und es werden Umgebungsgeräusche reduziert. Eine neue Qualitätswertung wird in einem festen Zeitintervall bestimmt, in dem alle in Echtzeit extrahierten Merkmale gesammelt und mithilfe eines Klassifikators einer NTID-Stufe zugeordnet werden.
Eine solche Sprachqualitätsanalyse könnte auf einem Handy, welches mit dem Hörgerät verbunden ist, umgesetzt werden, um den Parkinson-Patienten auch in ihrem Alltag eine Rückmeldung über die Verständlichkeit ihrer Sprache geben zu können (siehe Abb. 1).

Abbildung 1: Motivation

Bewertung der Sprache

Zur Bewertung der Verständlichkeit der analysierten Sprachabschnitte wird die NTID-Skala modifiziert nach Ramig (siehe Tabelle 2) verwendet [2]. Diese Wertung wird genutzt, um durch Logopäden oder andere Testpersonen Sprachaufnahmen als Referenz klassifizieren zu lassen und die NTID-Stufe durch das vorgestellte System zu schätzen.

Tabelle 1: NTID-Skala modifiziert nach Ramig [2]

NTID-Stufe	Bedeutung
1	Die sprachlichen Äußerungen sind unverständlich.
2	Die sprachlichen Äußerungen sind mit Ausnahme einiger Wörter oder Phrasen unverständlich.
3	Die sprachlichen Äußerungen sind schwer zu verstehen, doch der Inhalt ist im wesentlichen verständlich. (Die Verständlichkeit kann sich bei längerem Zuhören erhöhen.)
4	Die sprachlichen Äußerungen sind mit Ausnahme einiger Wörter oder Phrasen verständlich.
5	Die sprachlichen Äußerungen sind, bei aufmerksamen Hinhören oder in leiser Umgebung, verständlich.
6	Die sprachlichen Äußerungen sind völlig verständlich.

Echtzeitsystem

Das in Abbildung 3 dargestellte System wird parallel zu den im Hörgerät verwendeten Algorithmen implementiert, wie in Abbildung 2 dargestellt. Dieses wird mithilfe eines Hörgerät-Dummies als ein echtzeitfähiges MATLAB-Programm umgesetzt. Somit soll abgeschätzt werden, ob eine Implementierung auf einem mit dem Hörgerät gekoppeltem Smartphone erfolgversprechend wäre. Hauptteil des Systems ist eine Sprachqualitätsschätzung, welche die 4 Audiosignale $u_i(t)$ eines Hörgeräte-Dummy-Pärchens als Eingang nutzt und daraus eine Wertung zur Verständlichkeit der analysierten Sprache errechnet. Abhängig von dieser Wertung könnte ein Feedback über einen Soundgenerator im Hörgerät an den Träger des Hörgeräts gegeben werden.
Für die Sprachqualitätsschätzung werden die Daten zunächst vorverarbeitet, um Störungen zu minimieren und Rahmen der Länge 20 ms mit 50% Überlappung zu extrahieren. Es werden Daten aus dem Spektrum

des Signals, sowie Mel-Frequenz-Cepstrum-Koeffizienten (MFCCs) verwendet, um verschiedene Merkmale zu berechnen. Um die Analyse ausführen zu können, wird das System durch eine Eigenspracherkennung (OVD, engl.: own voice detection) gesteuert und es wird das Umgebungsrauschen geschätzt. Die OVD wird mithilfe eines Pegelvergleichs zwischen dem aktuellen Pegel und dem der Rauschschätzung und der Feststellung der Richtung der aufgenommenen Sprache realisiert. Somit wird eine für die Laborumgebung ausreichende OVD realisiert. In Hörgeräten ist oftmals bereits eine Erkennung der Sprache des Trägers implementiert, zum jetzigen Zeitpunkt konnte jedoch kein solches System verwendet werden. Es wird angenommen, dass ein späteres System Zugriff auf diese Informationen hat.
Nachdem in einem bestimmten Zeitintervall 5 s Sprachaktivität detektiert und neue Daten gesammelt wurden, wird ein Codebuch berechnet, bzw. ein bereits bestehendes verbessert. Dieses wird ausgewertet und, zusammen mit den gesammelten Merkmalen, wird mithilfe eines Klassifikators der analysierte Sprachabschnitt einer NTID-Stufe zugeordnet. Der Klassifikator besteht entweder aus mehreren Gaußschen Mischmodellen (GMMs) oder einem neuronalen Netz (NN).

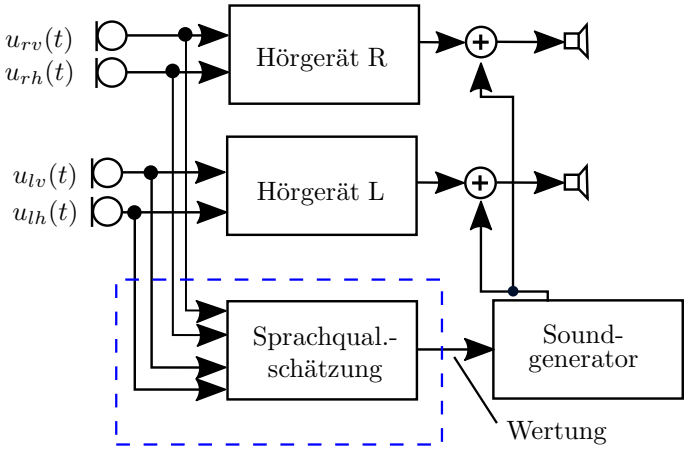

Abbildung 2: Systemüberblick

Merkmalsextraktion

Zur Sprachqualitätsanalyse werden 4 verschiedene Qualitätsattribute untersucht, um abschließend eine NTID-Wertung schätzen zu können. Alle mit x bezeichneten Variablen werden als Eingänge für den Klassifikator verwendet.

Lautstärke

Zu jedem analysierten Spektrum eines Sprachblocks $U(\mu, k)$ wird mithilfe eines Filters zur A-Gewichtung $X_A(\mu, k)$ der Eingangspegel

$$p_{U,A}(k) = 20 \log_{10} \sum_{\mu=1}^{N_\mu} |U(\mu, k)|\, X_A(\mu, k) \qquad (1)$$

in dBA berechnet. Die Spektren werden für N_μ Teilbänder berechnet, wobei μ der Teilbandindex sei. Der Lautstärkepegel über dem Rauschen $p_{\Delta,A}(k)$ zu jedem Zeitrahmen k ergibt sich als die Differenz zwischen dem A-gewichteten Pegel des Rauschens $p_{R,A}(k)$ und der aufgenommen Sprache:

$$p_{\Delta,A}(k) = p_{U,A}(k) - p_{R,A}(k). \qquad (2)$$

Am Ende eines Auswertungsintervalls wird berechnet, wie oft der Lautstärkepegel als ausreichend laut eingestuft wurde und dieses Verhältnis $\bar{x}_{Pegel}(\lambda)$ wird im Folgenden vom Klassifikator verwendet. λ beschreibt einen Zeitindex für jedes Auswertungsintervall.

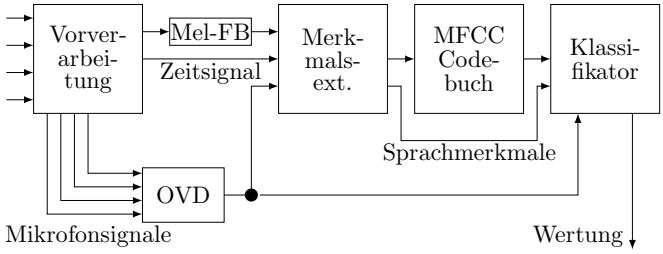

Abbildung 3: Sprachqualitätsschätzung

Sprachpausen

Zur Erkennung von Sprachpausen wird die Eigenspracherkennung genutzt. Es werden Sprachpausen erkannt, wenn ein Sprachabschnitt endet und der nächste Sprachabschnitt innerhalb eines bestimmten Zeitfensters beginnt. Startet nach dem Verstreichen des Zeitintervalls kein neuer Sprachabschnitt, so wird der Satz des Sprechers als beendet angesehen und es wird keine Pause detektiert.
Es kann sowohl eine Ober- als auch eine Untergrenze der Länge der Sprachpausen festgelegt werden, die nicht verletzt werden darf. Ansonsten kann eine Rückmeldung an den Sprecher gegeben werden, um ihm mitzuteilen, dass er zu langsam oder zu schnell spricht. Außerdem wird die Anzahl der Sprachpausen $N_{Pausen}(\lambda)$ in dem analysierten Abschnitt vom Startzeitpunkt T_A bis zum Endzeitpunkt T_E festgehalten. Innerhalb eines Auswertungsintervalls wird die Dauer aller Sprachpausen $T_{Pausen}(k)$ zeitlich gemittelt:

$$\bar{x}_{Pausen}(\lambda) = \frac{1}{N_{Pausen}(\lambda)} \sum_{k=T_A}^{T_E} T_{Pausen}(k). \qquad (3)$$

„Nuscheln“

Bei der Erzeugung unterscheidbarer Phoneme nimmt der Vokaltrakt des Sprechers unterschiedliche Formen an und moduliert so die spektrale Einhüllende des Sprachsignals. Dies wird bei der Erkennung von „nuscheliger“ Sprache genutzt, indem die spektrale Einhüllende, bzw. die MFCCs des Sprechers analysiert werden. Sind die MFCCs ausreichend unterschiedlich zueinander, kann davon ausgegangen werden, dass sich der Vokaltrakt beim Sprechen ausreichend stark verändert, um klar artikulierte Sprache zu erzeugen.
Die MFCCs aus jedem Sprachabschnitt werden während

des Analyseintervalls gesammelt und es wird zum Zeitpunkt der Auswertung ein Codebuch trainiert. Die Trainingsdaten des Codebuchs setzen sich aus allen MFCC-Vektoren zusammen. Diese werden auf den Maximalwert der MFCCs für jedes Analyseintervall normiert. Die Anzahl der MFCCs eines Vektors wird als $N_{MFCC} = 13$ gewählt.
Im ersten Auswertungsschritt wird ein Codebuch generiert und zu jedem neuen Auswertungsschritt wird dieses Codebuch auf die neuen Daten angepasst. Dabei wird eine Modifikation des LBG-Algorithmus verwendet, dessen Funktionsweise den Quellen [5] und [6] entnommen werden kann.
Insgesamt werden 6 Merkmale aus dem Codebuch extrahiert und später in dem Klassifikator verwendet. Die Anzahl der Codebuchvektoren kann direkt als $x_{N_C}(\lambda)$ ausgelesen werden. Es erfolgt außerdem die Berechnung der durch die Codebuchvektoren aufgespannte Fläche $x_F(\lambda)$, der Standardabweichung, zum aktuellen Zeitpunkt $\bar{x}_\sigma(\lambda)$ und über der Zeit geglättet $\bar{x}_{\bar{\sigma}}(\lambda)$, und der Varianz, zum aktuellen Zeitpunkt $\bar{x}_{\sigma^2}(\lambda)$ und über der Zeit geglättet $\bar{x}_{\bar{\sigma}^2}(\lambda)$.

„Stottern“

Das 4. Qualitätsattribut dient zur Überprüfung der Sprachstörung „Stottern“. An dieser Stelle wird die grundsätzliche Idee skizziert. Es soll überprüft werden, ob sich eine einzelne Silbe in kurzer Folge mehrfach wiederholt, da dies auf ein Stottern des Sprechers hinweisen würde. Um die ungefähre Dauer einer Silbe abzubilden, werden Signalfenster einer Länge von 150 ms mit einer Überlappung von 50% verglichen [4].
Es werden die bereits berechneten MFCCs verwendet, um die einzelnen Phoneme einer Silbe abzubilden. Eine Silbe besteht somit aus ca. 8 MFCC-Vektoren, die in einer Matrix gespeichert werden. Wurde eine neue Silbe aufgenommen, wird die dazugehörige Matrix mit den Matrizen der vorherigen Silben aus den letzten 3 s verglichen. Dies geschieht, durch einen Vorgang der ähnlich der Korrelation beider zu vergleichender Matrizen ausgeführt wird. Dazu wird schrittweise die Distanz der MFCC-Vektoren ausgewertet, indem eine Matrix um Vielfache der Länge der Rahmenlänge k verzögert wird.
Durch diese Auswertung wird zu jeder vorherigen Silbe im Auswertungsintervall ein Vergleichswert berechnet und in einem Vektor $\boldsymbol{x}_S(k_S)$ gespeichert. Es sei k_S ein Zeitindex, der die Anzahl der bisherigen Silben angibt. Für den Klassifikator wird der Maximalwert $x_{S,max}(\lambda)$ aus dem Vektor der mittleren Vergleichswerte der Silben $\boldsymbol{x}_S(k_S)$ in dem aktuellen Auswertungsintervall verwendet. Dieser Wert spiegelt die mittlere Ähnlichkeit der Silben in dem Auswertungsintervall wieder, die die größte Ähnlichkeit zu den vorherigen Silben im Auswertungsintervall besitzt.

Mustererkenner

Im Folgenden wird erläutert, wie mehrere GMMs oder ein neuronales Netz verwendet werden, um eine NTID-Wertung abzuschätzen. Als Eingang dienen die 10 zuvor berechneten Merkmale zu jedem Auswertungsschritt λ, zu dem eine NTID-Wertung berechnet wird.

NTID-Wertung durch GMMs

Es werden 6 GMMs verwendet, um jeweils eine NTID-Stufe abzubilden [1]. Dazu wird für jedes der GMMs die aktuelle Beobachtungswahrscheinlichkeit $p\big(n, \boldsymbol{x}(\lambda)\big)$ berechnet. Dabei sei $n \in \{1, ..., 6\}$ die Nummer des aktuell ausgewerteten GMMs und $\boldsymbol{x}(\lambda)$ der aktuelle Merkmalsvektor.
Die abschließende Bewertung $\hat{w}_{\mathrm{GMM}}(\boldsymbol{x}(\lambda))$ berechnet sich als gewichtete Summe über die normierten Beobachtungswahrscheinlichkeiten $\tilde{p}\big(n, \boldsymbol{x}(\lambda)\big)$ der GMMs:

$$\hat{w}_{\mathrm{GMM}}\big(\boldsymbol{x}(\lambda)\big) = \frac{1}{\sum_{n=1}^{6} \tilde{p}\big(n, \boldsymbol{x}(\lambda)\big)} \sum_{n=1}^{6} n \cdot \tilde{p}\big(n, \boldsymbol{x}(\lambda)\big). \tag{4}$$

Es ergibt sich die Schätzung einer NTID-Wertung $\hat{w}_{\mathrm{GMM}}\big(\boldsymbol{x}(\lambda)\big)$, die nicht auf ganze Zahlen beschränkt ist, sondern den reellen Raum von 1 bis 6 komplett nutzt.

NTID-Wertung durch ein NN

Es wird ein neuronales Netz als Alternative zu den GMMs genutzt, welches die selben Daten zum Training und zum Testen verwendet. Dazu wird ein Netzwerk mit 4 versteckten Schichten bestehend aus 15, 12, 9, und 6 Neuronen mithilfe des BFGS Quasi-Newton Backpropagation-Algorithmus trainiert [3]. Das vollständig trainierte neuronale Netz nimmt als Eingang den Merkmalsvektor $\boldsymbol{x}(\lambda)$ und gibt direkt die Schätzung der NTID-Wertung $\hat{w}_{\mathrm{NN}}\big(\boldsymbol{x}(\lambda)\big)$ aus, die sich ebenfalls in dem Raum der reellen Zahlen von 1 bis 6 befinden kann.

Evaluation

Verwendete Daten

Um ausreichend Daten zum Training und zur Evaluation der Modelle zur Verfügung zu haben, wurde eine Datenbank verwendet, welche verschiedene Sprachaufnahmen von Parkinson-Patienten bei Sitzungen mit Logopäden enthält. Die Datenbank besteht aus Aufnahmen von 225 verschiedenen Patienten, welche alle von einer oder mehreren Personen bewertet wurden. Die Aufnahmen sind nach verschiedenen Sprachübungen der Patienten geordnet und umfassen Aufgaben, wie einen Vokal möglichst lange zu halten, einen Text vorzulesen, bestimmte Silben zu wiederholen und spontane Sprache. Letztere werden aus dieser Datenbank ausgewählt, um eine Alltagssituation nachstellen zu können, in der auch das Hörgerät in einem normalen Gespräch eingesetzt werden würde.
Nach dieser Voraussetzung werden 426 Aufnahmen, die alle eine NTID-Wertung besitzen, von 222 Patienten ausgewählt, deren Länge in der Regel zwischen 20 s und 60 s liegt. Es sei an dieser Stelle erwähnt, dass die Aufnahmen von unterschiedlicher Qualität sind, da sie teilweise starkes Rauschen enthalten und der Abstand der Patienten zum Mikrofon nicht genau festgelegt wurde. Es stehen außerdem verschiedene Anzahlen an Aufnahmen für jede NTID-Stufe zur Verfügung und gerade für

die unteren Stufen sind relativ wenige Daten vorhanden. Die Anzahl an ausgewerteten Sprachsignalabschnitten pro NTID-Stufe sind in folgender Tabelle aufgeführt.

Tabelle 2: Aufgenomme Sprachintervalle von Parkinson-Patienten

NTID-Stufe	1	2	3	4	5	6
Anzahl	14	62	102	178	369	182

Evaluationsmaße

Zur Evaluation der Mustererkenner werden die folgenden Qualitätsmaße verwendet, es sei N_T die Anzahl aller ausgewerteten Sprachintervalle der Testdaten:

- Mittlere Betragsdistanz $\bar{d}$ der Klassifikatorenwertungen $\hat{w}(n)$ zu den Referenzwertungen $w(n)$ der Testdaten (der erwünschte Optimalwert ist 0):

$$\bar{d} = \frac{1}{N_T} \sum_{n=1}^{N_T} |\hat{w}(n) - w(n)|. \tag{5}$$

- Lineare Korrelation r der Klassifikatoren- und Referenzwertungen (der erwünschte Optimalwert ist 1):

$$r = \frac{\sum_{n=1}^{N_T} \left(\hat{w}(n) - m_{\hat{w}}\right)\left(w(n) - m_w\right)}{\sqrt{\sum_{n=1}^{N_T} \left(\hat{w}(n) - m_{\hat{w}}\right)^2 \sum_{n=1}^{N_T} \left(w(n) - m_w\right)^2}}. \tag{6}$$

Es seien $m_{\hat{w}}$ und m_w die Mittelwerte der Klassifaktoren- und der Referenzwertungen der Testdaten. Beide Qualitätsmaße werden für die Ergebnisse der GMMs und der neuronalen Netze ausgewertet.

Ergebnisse

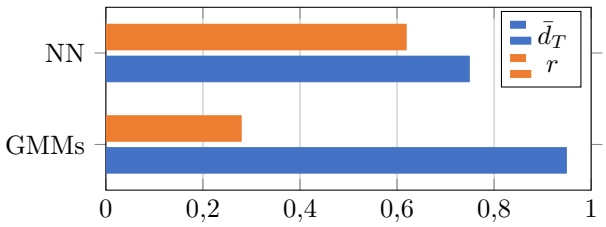

Abbildung 4: Ergebnisse

In Abb.4 sind die Ergebnisse für die beste Konfiguration der GMMs und des neuronalen Netzes dargestellt. Während beide Klassifikatoren ähnliche Korrelationen der Trainingsdaten gezeigt haben, fällt die Korrelation der Testdaten der GMMs stark ab (auf ca. 0,28). Die Korrelation der Ergebnisse des neuronalen Netzes liegt bei ca. 0,62. Vergleicht man nun den mittleren Fehler der beiden Klassifikatoren lässt sich, wie zu erwarten, ein deutlicher Unterschied erkennen. Während das neuronale Netz einen mittleren Fehler von 0,75 erreicht, weisen die GMMs einen mittleren Fehler von 0,95 auf. Insgesamt erreicht das neuronale Netz eine deutlich bessere Klassifikation.

Es ist zu beachten, dass deutlich weniger Daten für die NTID-Stufen im Bereich 1-2 ausgewertet wurden, als für den Rest der NTID-Skala. Dies kann dazu geführt haben, dass diese Modelle in diesem Bereich nicht ausreichend trainiert wurden.

Fazit

Es wurde ein System zur Echtzeitanalyse der Sprache eines Hörgerät-Trägers vorgestellt, welches an Sprachstörungen, verursacht durch die Parkinson-Krankheit, motiviert wurde. Es werden verschiedene Sprachmerkmale im Zeit- und Frequenzbereich berechnet, um die 4 Sprachqualitätsattribute Lautstärke, Sprachpausen, Nuscheln und Stottern zu untersuchen. Die Auswertungsergebnisse der Sprachqualitätsattribute werden genutzt, um die Sprache in festgelegten Zeitintervallen von 5 sek aktiver Sprache mithilfe eines Klassifikators einer NTID-Stufe zuzuordnen. Dabei hat sich gezeigt, dass mit einem neuronalen Netz ein mittlerer Abstand von 0,75 NTID-Stufen zu den subjektiven Bewertungen durch Testpersonen erreicht werden kann.
Zusammenfassend zeigt das entwickelte System in der hier vorgestellten Realisierung bereits verwertbare Ergebnisse. Es besteht noch deutliches Potential um das System weiterzuentwickeln, um zukünftig die Lebensqualität von Parkinson-Patienten verbessern zu können. Dies wäre durch die Verwendung weiterführender Sprachqualitätsanalysen möglich, welche z.B. die Formanten der Sprache oder spezifischere Codebücher analysieren könnten. Ebenfalls wäre eine spezifischere Evaluation der vorgestellten Qualitätsattribute denkbar.

Literatur

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer, 2011.

[2] Raymond D. Kent, editor. *Intelligibility in Speech Disorders.* John Benjamins Publishing Company, 1992.

[3] David Kriesel. *Ein kleiner Überblick über Neuronale Netze.* 2007.

[4] Hartmut R. Pfitzinger. *Phonetische Analyse der Sprechgeschwindigkeit.* Inst. für Phonetik und Sprachliche Kommunikation München, 2001.

[5] Lawrence Rabiner und Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[6] Yu-Chen Hu und Chin-Chen Chang. A progressive codebook training algorithm for image vector quantization. In *Fifth Asia-Pacific Conference on ... and Fourth Optoelectronics and Communications Conference on Communications,*. IEEE, 1999.

[7] Adelheid Nebel und Günther Deuschl. *Dysarthrie und Dysphagie bei Morbus Parkinson.* Thieme, 2016.