Parkinson-Sprachanalyse – Erweiterungen zum Qualitätsmerkmal Formantdreieck

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

[1] *Digitale Signalverarbeitung und Systemtheorie, Christian-Albrechts-Universität zu Kiel, E-mail: {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

Morbus Parkinson ist eine der am weitesten verbreiteten neurodegenerativen Krankheiten weltweit. Häufig geht mit dieser Krankheit eine Sprachstörung einher, die so genannte Dysarthrie. Ein etabliertes Maß, um die Schwere dieser Sprachstörung in der deutschen Sprache zu beurteilen, ist die Fläche des Formantdreiecks, welches aus den ersten beiden Formantfrequenzen der Vokale /a:/, /i:/ und /u:/ gebildet wird. Die Verwendung dieses Maßes führt jedoch zu Nachteilen; so müssen zu analysierende Aufnahmen im Vorwege phonetisch annotiert werden. Diese Arbeit stellt eine Weiterentwicklung auf Basis eines Codebuch-Ansatzes vor, bei dem eine solche Annotation nicht mehr notwendig ist. Dabei werden die ersten beiden Formantfrequenzen aus jedem stimmhaften Signalabschnitt extrahiert, und auf Basis dieser Merkmale wird ein Codebuch trainiert. Aus diesem Codebuch lassen sich verschiedene Maße gewinnen, die auf zum Formantdreieck vergleichbare Weise die Qualität der Artikulation schätzen. Außerdem wird eine Erweiterung des Formantdreiecks durch Einbeziehen aller fünf Basisvokale der deutschen Sprache betrachtet.

Motivation

Die Sprache ist für Menschen eines der wichtigsten und ältesten Mittel, um miteinander zu kommunizieren. Bei der Kommunikation durch Sprache können in kurzer Zeit sehr viele Informationen übermittelt werden. Neben der sachlichen Information, welche konkret verbal geäußert wird, schwingen dabei zusätzlich nonverbale Informationen über den Gemütszustand oder die Gefühle des Sprechers mit, welche zum Beispiel in der Stimmlage, dem Stimmtimbre und dem Sprachrhythmus enthalten sind. Wird diese Sprachfähigkeit infolge einer Erkrankung eingeschränkt oder geht sie sogar verloren, so bedeutet dies eine starke Einschränkung in der Kommunikationsfähigkeit der Betroffenen und somit eine starke Beeinträchtigung des täglichen Lebens. Solche Sprachstörungen können bei allen neurologischen Erkrankungen auftreten; besonders häufig geschieht dies beim so genannten Parkinson-Syndrom. Im Verlauf der Erkrankung entwickeln bis zu 90% der Betroffenen eine Stimm- und Sprachstörung, die auch als Dysarthrie bezeichnet wird [1]. Um den Fortschritt der Dysarthrie zu überwachen, werden die Patienten regelmäßig logopädisch untersucht; dabei sind verschiedenste Sprechaufgaben zu erfüllen, die zur Dokumentation aufgezeichnet werden. Zu diesen Aufgaben zählt unter anderem das möglichst lange Halten der Kernvokale /a:/, /i:/ und /u:/ der deutschen Sprache, auf diese Aufnahme bezieht sich diese Arbeit im weiteren Verlauf.

Um die Qualität der Artikulation und Sprache zu beurteilen, werden etablierte Maße wie das Formantdreieck zur Auswertung der Aufnahmen herangezogen. Das Formantdreieck, auch Vokaldreieck genannt, wird durch das gegeneinander Auftragen der ersten beiden Formanten der eben genannten Kernvokale in einem Koordinatensystem gebildet [2]. Dies ist bisher mit großem zeitlichem Aufwand verbunden, da die einzelnen Vokale dazu von Hand annotiert werden müssen und anschließend die Formanten im betreffenden Abschnitt mit einem Sprachanalyse-Tool, wie beispielsweise Praat [3], extrahiert werden. Aus den extrahierten Formanten eines Vokals wird darauf folgend der Mittelwert gebildet und dieser in die so genannte Formantkarte eingetragen. So entsteht schließlich das für die deutsche Sprache typische Formantdreieck, dessen Fläche eine Aussage über die Artikulationsqualität zulässt [2].

In dieser Arbeit soll eine Methode vorgestellt werden, die eine automatisierte Berechnung dieses Maßes ermöglicht. Dafür werden die Formanten in den Sprachaufnahmen der gehaltenen Kernvokale mit Hilfe einer Sprachaktivitätserkennung und der bekannten Levinson-Durbin-Rekursion aus den Prädiktor-Koeffizienten automatisch bestimmt [4]. Die Menge der extrahierten Formanten wird für ein Codebuch-Training mit dem *k-means*-Algorithmus verwendet [5]. Aus den resultierenden Codebuch-Vektoren kann schließlich das Vokaldreieck gebildet werden. So können die Sprachaufnahmen der Patienten effizient auf die Veränderung der Sprachqualität im Verlauf der Krankheit untersucht werden. Dieses Einzelmaß soll letztendlich in ein Rahmenwerk eingebunden werden, welches automatisch, auf Basis verschiedener, instrumenteller Maße, die Sprachqualität eines Patienten evaluiert. Eine detailliertere Beschreibung dieses Rahmenwerks kann in [6] gefunden werden.

Im Folgenden werden die Berechnung des Formantdreiecks sowie die Umsetzung über den Codebuch-Ansatz genauer beschrieben. Außerdem wird auf mögliche Erweiterungen des klassischen Formantdreiecks mit Hilfe dieses neuen Ansatzes eingegangen.

Berechnung des Formantdreiecks

Die Bestimmung des Formantdreiecks erfolgt, wie bereits erwähnt, bisher meist aufwendig von Hand. Dazu kann ein Sprachanalyse-Tool, wie Praat [3], zur Hilfe genommen werden. Hier wird das gesamte Sprachsignal eingelesen, anschließend können die Formanten durch das Programm automatisch berechnet und graphisch veranschaulicht dargestellt werden. Im nächsten Schritt müssen die Sprachabschnitte, welche die Vokale

/a:/, /i:/ und /u:/ enthalten, von Hand gekennzeichnet und die entsprechenden ersten beiden Formanten gespeichert werden. Über die zu jedem Vokal gespeicherten Formanten wird schleißlich gemittelt, und diese Mittelwerte werden in die so genannte Formantkarte eingetragen; daraus ergibt sich das für die deutsche Sprache charakteristische Formantdreieck, wie in Abb. 1 gezeigt [4].

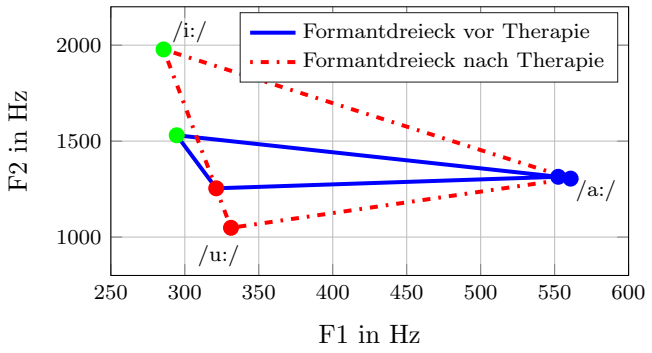

Abbildung 1: Klassisches Formantdreieck eines Patienten vor und nach einer Sprachtherapie.

Die bereits genannten Kernvokale bilden die Eckpunkte eines Dreiecks, aus dessen Fläche sich eine Aussage über die Artikulationsqualität, insbesondere die Deutlichkeit der Aussprache, treffen lässt. Dies stellt ein wichtiges Qualitätsmerkmal in der Bewertung von Parkinson-Sprache dar, da eine verwaschene, nuschelnde Sprache eine häufige Ausprägung der Dysarthrie von Parkinson-Patienten ist. Zum Einen kann man dieses Maß nutzen, um den Fortschritt der Sprachstörung über den Krankheitsverlauf zu beobachten. Zum Anderen werden viele Patienten bereits mit einer von mehreren verschiedenen möglichen Sprachtherapien behandelt; so ist es auch denkbar, dieses Maß einzusetzen, um den jeweiligen Therapieerfolg zu messen, wie in Abb. 1 gezeigt. Es gilt dabei: Je größer die Fläche des Formantdreiecks, desto größer die Ausnutzung des Formantraumes [2] und desto besser die Qualität der Artikulation.

Codebuch-Ansatz

Um die Berechnung der Formantfläche als etabliertes Merkmal bei der Sprachqualitätsbeurteilung zu automatisieren, wird im Folgenden ein codebuch-basierter Ansatz vorgestellt. Das Codebuch ist ein Mustererkenner, der basierend auf einer trainierten Datenbank einen Vergleich zwischen den Datenbankeinträgen und dem aktuellem Merkmalsvektor vollzieht [7]. Für diesen Ansatz wird im ersten Schritt eine einfache Stimmhaft/Stimmlos-Erkennung durchgeführt, daraufhin werden die Formanten berechnet und schließlich über ein Codebuch-Training das Formantdreieck bestimmt.

Stimmhaft/Stimmlos-Erkennung

Basierend auf dem klassischen Quelle-Filter-Modell der Spracherzeugung wird zwischen drei unterschiedlichen Anregungsarten des Sprechtraktes unterschieden. Es handelt sich dabei um die stimmhafte, die stimmlose und die transiente Anregung [4]. Da für die Formantbestimmung nur die stimmhaften Signalabschnitte relevant sind, werden die stimmlose und die transiente Anregung im Folgenden gemeinsam als stimmlose Anregung bezeichnet.

Die angewendete Stimmhaft/Stimmlos-Erkennung basiert auf der Detektion der Sprachgrundfrequenz, im Folgenden Pitch genannt, im betrachteten Signalabschnitt [8]. Dazu wird zunächst die Autokorrelationsfunktion des aktuellen Signalabschnittes berechnet

$$\phi_{xx}(\kappa,k)=\sum_{n=1}^{N}x(n,k)x(n+\kappa,k), \tag{1}$$

wobei N die Länge und k der Index des aktuellen Signalabschnittes ist, $x(n,k)$ der aktuelle Signalabschnitt und κ eine diskrete Zeitverschiebung [8]. Die Autokorrelationsfunktion wird anschließend normiert, so dass sie bei einer zeitlichen Verschiebung von 0 den Wert 1 annimmt

$$\phi_{xx,norm}(\kappa,k)=\frac{\phi_{xx}(\kappa,k)}{\phi_{xx}(0,k)}. \tag{2}$$

Schließlich wird geprüft, ob die normierte Autokorrelationsfunktion ein Nebenmaximum, in einem zur Pitch-Frequenz (zwischen 50 und 500 Hz) passendem Bereich besitzt, dessen Amplitude eine gesetzte Schwelle $\mathrm{VAD}_{\min}$ überschreitet. Somit muss zur Maximums-Suche nur die Autokorrelationsfunktion bei einer zeitlichen Verschiebung zwischen 20 und 200 ms betrachtet werden. Diese Berechnungen erfolgen nach

$$\phi_{max}(k)=\max_{\kappa\in[20\,ms\cdot f_s,200\,ms\cdot f_s]}\{\phi(\kappa,k)\}, \tag{3}$$

und

$$\mathrm{VAD}(k) \quad = \quad \begin{cases} 1 & , \text{ wenn } \phi_{max}(k)>\mathrm{VAD}_{\min} \\ 0 & , \text{ sonst;} \end{cases} \tag{4}$$

dabei ist $\phi_{max}(k)$ das gesuchte Maximum der Autokorrelationsfunktion, f_s die Abtastrate des Signals und $\mathrm{VAD}(k)$ beinhaltet das Ergebnis der Stimmhaft/Stimmlos-Entscheidung, wobei 1 für stimmhaft und 0 für stimmlos steht. Als Schwellenwert wurde $\mathrm{VAD}_{\min}=0,4$ verwendet.

Formantberechnung

Nach der Stimmhaft/Stimmlos-Segmentierung erfolgt die Berechnung der Formanten für alle stimmhaften Signalabschnitte. Zunächst wird hier eine Levinson-Durbin-Rekursion durchgeführt, um die Prediktorkoeffizienten und damit die spektrale Einhüllende des Sprachsignalabschnittes zu bestimmen [4]. Aus der Einhüllenden werden dann die lokalen Maxima bestimmt sowie deren Argumente. Die Position der Maxima bestimmt dabei die Formantfrequenzen [2], wobei das Maximum bei der niedrigsten Frequenz den ersten Formanten darstellt usw. Die ersten beiden Formanten werden dabei für jeden Signalabschnitt gespeichert, da diese für die Bildung des Formantdreiecks relevant sind.

Codebuch-Training

Die über das gesamte Sprachsignal gesammelten Formanten werden als Trainingsdatenset für ein Codebuch-Training verwendet. Als Trainings-Algorithmus kann hier der k-means Algorithmus verwendet werden [5], da im

Vorhinein bekannt ist, dass sich drei Cluster ergeben sollen. Die erhaltenen Codebuch-Vektoren, nach k-means Training, können schließlich als Eckpunkte für das Formantdreieck angenommen werden und daraus die Dreiecksfläche bestimmt werden.

Anschaulich kann das Ergebnis des Codebuch-Trainings wie in Abb. 2 dargestellt werden. Dabei bezeichnen die gestrichelten Linien die Grenzen der durch das Training gefundenen Voronoi-Regionen und die Farben der einzelnen Datenpunkte codieren deren tatsächliche Zugehörigkeit zu den jeweiligen Vokalen. Die anfänglich etwas ungewöhnlich erscheinende Form der Voronoi-Regionen ist auf die ungleiche Skalierung der Achsen zurück zu führen.

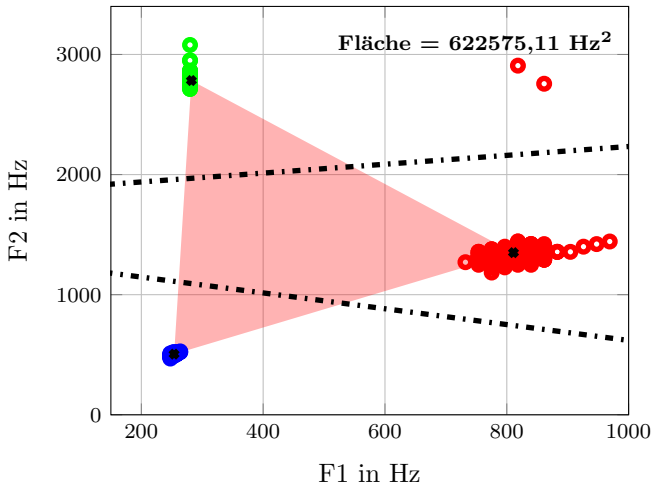

Abbildung 2: Formantdreieck nach Codebuch-Training.

Mögliche Erweiterungen

Aufbauend auf dem beschriebenen Codebuch-Ansatz können verschiedene Erweiterungen eingeführt werden. Das sind zum Einen weitere Maße zur Auswertung des Codebuchs mit drei Vokalen, zum Anderen kann das Codebuch mit den Formanten aller fünf Basisvokale der deutschen Sprache (/a:/, /e:/, /i:/, /o:/ und /u:/) trainiert werden und dementsprechend mit fünf Codebuch-Vektoren trainiert werden, so dass sich das Formantdreieck zu einem unsymmetrischen Fünfeck erweitern würde, dessen Fläche wiederum als Maß für die Artikulationsqualität verwendet werden kann. Auf diese Möglichkeiten zur Erweiterung soll im Folgenden näher eingegangen werden.

Weiterführende Maße

Eine weitere Auswertung des im vorigen Kapitel beschriebenen Codebuchs ist die Betrachtung der durchschnittlichen Varianz innerhalb der Cluster. Mit diesem Maß soll eine Aussage über die Klarheit der Sprache getroffen werden. Die dahinter stehende Idee ist, dass bei gehaltenen Vokalen, bei idealer Aussprache, alle Punkte für den selben Vokal auf dem selben Ort in der Formantkarte liegen würden. Dies ist in der Realität natürlich nicht zu erreichen, selbst bei sprechgesunden Personen, allerdings sollte die Streuung der zu einem Vokal gehörigen Punkte um den zugeordneten Codebuch-Vektor bei einer klaren Aussprache deutlich geringer sein als bei einer nuschelnden, undeutlichen Sprechweise.

Zusätzlich zu der durch die Codebuch-Vektoren aufgespannte Fläche kann der durchschnittliche euklidische Abstand der Codebuch-Vektoren zueinander bestimmt werden. Dieses Maß hat im wesentlichen die selbe Aussagekraft wie die bereits betrachtete Fläche, allerdings können damit erweiterte Laut-Konstellationen erfasst werden. Zur Erweiterung dieses Maßes werden die Kanten, in Abhängigkeit von der Anzahl der Merkmalsvektoren in den verbundenen Clustern, gewichtet, anstatt ein einfaches Mittel über alle Distanzen zu bilden. So wird eine Kante, die ein Cluster verbindet, dem wenige Merkmalsvektoren zugeordnet sind, weniger stark gewichtet als eine Kante, die zwei Cluster mit vielen Merkmalsvektoren verbindet. Die Idee dahinter ist, dass Ausreißer in einem Cluster mit wenigen Merkmalsvektoren stärker zu einer Verschiebung des Codebuch-Vektors beitragen als bei größeren Clustern und somit zu einer Verfälschung der Distanz zu diesem Codebuch-Eintrag führen, weshalb der Beitrag dieser Distanz zum Gesamtdurchschnitt weniger Gewicht bekommen sollte.

Erweiterungen des Codebuchs

Eine weitere Anpassung des Codebuch-Ansatzes kann durch die Erweiterung um die Vokale /e:/ und /o:/ erfolgen. Somit wird aus dem Formantdreieck ein unsymmetrisches Formantfünfeck, da sowohl das /i:/ und das /e:/, als auch das /u:/ und das /o:/ nahe beieinander liegen. Dazu muss die Sprechaufgabe für die Patienten angepasst werden, so dass nun alle fünf Basisvokale der deutschen Sprache vorkommen müssen. Außerdem wird der k-means Algorithmus nun für fünf Codebuch-Vektoren angewendet. Dies führt auf ein Ergebnis des Codebuch-Trainings, wie in Abb. 3 dargestellt.

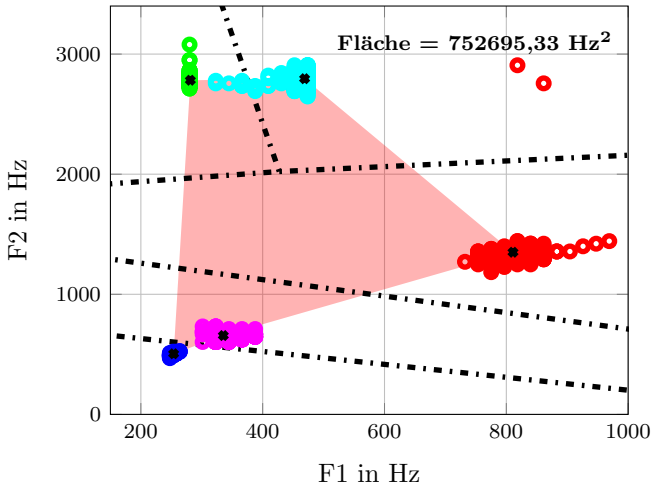

Abbildung 3: Ergebnis des erweiterten Codebuch-Trainings.

Für dieses Codebuch kann ebenfalls eine durch die Codebuch-Vektoren aufgespannte Fläche berechnet werden, die ein Maß für die Artikulationsqualität darstellt. Auch die weiteren Merkmale, wie die gewichtete mittlere Distanz zwischen den Codebuch-Vektoren mit einer ähnlichen Aussagekraft wie die aufgespannte Fläche und auch die Streuung innerhalb der Cluster können hier be-

trachtet werden, als ein Maß für die Klarheit der Aussprache.

Der Vorteil in der Betrachtung aller fünf Basisvokale liegt darin, dass dieses Modell mit einer guten Stimmhaft/Stimmlos-Detektion auch leicht auf andere Sprachaufnahmen anwendbar ist. So könnten Patienten beispielsweise einen Vorlesetext als Sprechaufgabe bekommen, in dem alle fünf Basisvokale in ihre natürliche Sprachumgebung eingebunden sind. Eine andere Möglichkeit ist, die Patienten einem Stresstest zu unterziehen und die Veränderung der Formant-Merkmale unter Stress-Bedingungen zu untersuchen. Eine passende Sprechaufgabe dafür wäre z.B. die abwechselnde und im Tempo immer schneller werdende Wiederholung der Wörter „Ananas, Enten, imitiert, Motor, Unmut“.

Es ist außerdem möglich, dieses Codebuch-Verfahren, wieder in Kombination mit einer guten Stimmhaft/Stimmlos-Detektion, auf einen vorgelesenen Fließtext oder Spontan-Sprache anzuwenden. Dabei sollten folgende Überlegungen in Betracht gezogen werden:

- Verwendung des Linde-Buzo-Gray (LBG)-Algorithmus zum Training der Codebuches [5]. Dies ist sinnvoll, da nun zu den Basisvokalen verschiedene Varianten davon sowie Umlaute im Sprachsignal enthalten sein können und somit mehr Häufungspunkte entstehen.
- Die Betrachtung der durch alle Trainings-Vektoren aufgespannten Fläche, anstelle der durch die Codebuch-Vektoren aufgespannten Fläche, als Maß für die Ausnutzung des Formantraumes und der Artikulationsqualität. Der Grund für diese Überlegung ist, dass durch die hinzukommenden Variationen der Vokale, der von den fünf Basisvokalen aufgespannte Formantraum nahezu komplett aufgefüllt und damit die durch das Codebuch aufgespannte Fläche sehr klein wird. Eine Betrachtung der durch alle Merkmals-Vektoren aufgespannten Fläche stellt hier ein besseres Maß für die Ausnutzung des Formantraumes dar.

Weiterhin ist der Übergang auf mel-gefilterte Cepstral-Koeffizienten (MFCC) als extrahierte Merkmale für das Codebuch-Training denkbar, um mehr Informationen der spektralen Einhüllenden des Sprachsignals beizubehalten [5]. Zur Auswertung können schließlich die bereits beschriebenen Maße, gegebenenfalls mit geringen Modifikationen, verwendet werden.

Zusammenfassung und Fazit

In dieser Arbeit wurde zunächst das Merkmal Formantdreieck als Maß für die Artikulationsqualität eines Parkinson-Patienten vorgestellt, welches aus Sprachaufnahmen der gehaltenen Vokale /a:/, /i:/ und /u:/ gewonnen wird. Anschließend wurde gezeigt, wie dieses Maß automatisch aus einem aufgenommenen Sprachsignal extrahiert werden kann. Dies ist notwendig, um dieses Einzelmaß in ein Rahmenwerk einbinden zu können, welches die Sprachqualität eines Menschen anhand einer Sprachsignal-Analyse evaluiert. Die Analyse geschieht dabei auf der Basis verschiedener, automatisch extrahierter, instrumenteller Merkmale aus dem Sprachsignal.

Dieses Rahmenwerk wird in der Bewertung des Schweregrades der Dysarthrie von Parkinson-Patienten Anwendung finden. Hier soll zum Einen die Entwicklung der Sprachstörung über den Krankheitsverlauf beobachtet werden, zum Anderen soll ein möglicher Therapie-Erfolg nach erhaltener Sprachtherapie messbar werden.

Ausgehend von dem bereits etablierten Maß Formantdreieck und dem hierzu vorgestellten Codebuch-Ansatz sind verschiedene Weiterentwicklungen vorgestellt worden, die ebenfalls in das Rahmenwerk eingebunden werden. Das sind zum Einen weitere Auswertungsmöglichkeiten des Codebuchs, wie die Streuung innerhalb der Cluster oder die gewichtete, mittlere Distanz zwischen den Codebuch-Vektoren. Zum Anderen wurde eine Erweiterung der Mustererkennung auf alle fünf Basisvokale der deutschen Sprache vorgeschlagen sowie die Anwendung dieser Erweiterung auf komplexere Sprechaufgaben. Dabei wurde darauf verwiesen, die Verwendung des LBG Algorithmus, für Fließtext und spontansprachliche Texte, in Betracht zu ziehen ebenso wie eine Modifikation zum Merkmal der Formantfläche.

Abschließend wurde ein Übergang von Formanten zu MFCC's vorgeschlagen, um mehr Informationen der spektralen Einhüllenden des Sprachsignals zu betrachten und in die Auswertung der Sprachqualität einzubeziehen.

Danksagung

Die Autoren danken der Deutschen Forschungsgemeinschaft (DFG) für ihre Unterstützung.

Literatur

[1] A. Nebel und G. Deuschl, *Dysarthrie und Dysphagie bei Morbus Parkinson.* Thieme, 2016.

[2] M. Merk, “Entwicklung und Implementierung PC-gestützter akustischer Analyseverfahren für die klinische Diagnostik neurogener Sprechstörungen,” *Fakultät der Elektrotechnik der Universität der Bundeswehr München*, 2002.

[3] P. Boersma und D. Weenink, “Praat: doing phonetics by computer.” `www.praat.org`, 2015. [Online; accessed 09-December-2016].

[4] P. Vary und U. Heute und W. Hess, *Digitale Sprachsignalverarbeitung.* B.G. Teubner Stuttgart, 1998.

[5] B. Pfister und T. Kaufmann, *Sprachverarbeitung: Grundlagen und Methoden Der Sprachsynthese und Spracherkennung.* Springer, 2008.

[6] C. Baasch und G. Schmidt und U. Heute und A. Nebel und G. Deuschl, “Parkinson Speech Analysis: Methods and Aims,” *ITG Speech, Paderborn, Germany*, 2016.

[7] G. A. Fink, *Markov Models for Pattern Recognition: From Theory to Applications.* Springer London, 2014.

[8] T. Shimamura, “Weighted Autocorrelation for Pitch Extraction of Noisy Speech,” *IEEE Transaction on Speech and Audio Processing*, 2001.