# Automatisierte Diagnose prosodischer Störungen bei Aphasie mittels künstlicher neuronaler Netze

J. Haring[1], C. J. Werner[2], C. Kohlschein[1], U. D. Peitz[2], B. Schumann-Werner[2], J. Niehues[3]

[1] HotSprings GmbH, Aachen, Deutschland

[2] Klinik für Neurologie, Medizinische Fakultät RWTH Aachen, Deutschland

[3] Department of Data Science and Knowledge Engineering, Maastricht University, Maastricht, Niederlande

**Zusammenfassung**

Der Aachener Aphasie Test (AAT) erfasst auch die spontansprachlichen Leistungen eines Menschen mit Aphasie. Dieser Teil ist jedoch nur manuell durch geschultes Personal auswertbar. Die vorliegende Arbeit fokussiert auf die automatisierte Bewertung einer der sechs AAT-Spontansprachskalen. Im der vorgestellten Studie wird die Möglichkeit untersucht, künstliche neuronale Netze zur automatisierten Identifikation von Auffälligkeiten der Dimension »Prosodie und Artikulation« zu implementieren und verschiedene Ansätze hierzu werden verglichen. Ziel des Studienprogramms ist es, die Durchführung des AAT durch computergestützte Methoden unter Aufrechterhaltung bestehender Qualitätsanforderungen zu automatisieren.

**Schlüsselwörter:** Aphasie, Aachener Aphasie Test, künstliche neuronale Netze, Prosodie, Dysarthrie

## Einleitung

Im Rahmen des Aachener Aphasie Tests (AAT) [3], der den Goldstandard der Diagnostikinstrumente aphasischer Störungen im deutschsprachigen Raum darstellt, werden mittels verschiedener Untertests alle sprachlichen Modalitäten getestet. Zusätzlich wird auch eine Quantifizierung der spontansprachlichen Leistungen der Patient*innen vorgenommen, welche ein wertvoller Bestandteil der Diagnose und Graduierung sowie der darauffolgenden Therapiekontrolle ist. Da die Durchführung und Auswertung des AAT jedoch einen hohen Zeit- und Ressourcenaufwand und gut geschultes Personal erfordert, eröffnet eine eventuelle automatisierte Diagnose die Möglichkeit eines erhöhten Standardisierungsgrades unter gleichzeitiger Einsparung wertvoller Personalressourcen. Dies könnte auch personelle Ressourcen für die Therapie freisetzen und zu vereinheitlichter Diagnostik, beispielsweise im Rahmen multizentrischer Studien, führen. Diese Arbeit befasst sich mit dem initialen Problem der automatisierten Bewertung einer der sechs AAT-Spontansprachskalen. Ziel ist, die Dimension »Artikulation und Prosodie« auf Auffälligkeiten zu untersuchen. Das untersuchte Merkmal beschreibt hierbei unter anderem Abweichungen in Ton und Betonung (Dysprosodie) sowie ungenaue Artikulation (Dysarthrie), welche korrelieren und hier zusammengefasst werden, da keine getrennten Indikatoren bestehen. Es wird die Klassifikation roher Sprachaufnahmen der Patient*innen durch künstliche neuronale Netze, eine Methode des maschinellen Lernens, beschrieben.

## Material

Der Arbeit von Kohlschein [5] entstammt die informationstechnische Betrachtung einer Automatisierung des AAT mittels Machine Learning (ML) und eine Datenbank von sprecher*innengetrennten Sprachaufnahmen samt Transkriptionen und Aphasiediagnose, auf die wir in der vorliegenden Arbeit zugreifen konnten. Diese Datenbank besteht aus den semi-standardisierten Spontansprachinterviews des AATs zwischen Patient*innen und Therapeut*innen. Sie enthält spontansprachliche Äußerungen von 240 Aphasiepatient*innen für das Training der künstlichen neuronalen Netze, deren Anzahl nach Aphasiesyndrom in **Tabelle 1** dargestellt ist.

**Tab. 1:** Anzahl der Aphasiepatient*innen nach den vier Standardsyndromen

| | n |
|---|---|
| Amnestische Aphasie | 38 |
| Broca-Aphasie | 72 |
| Globale Aphasie | 83 |
| Wernicke-Aphasie | 47 |

Insgesamt fallen 44 Stunden Rohaufnahmen an. Jedes zuvor sprecher*innengetrennte Segment enthält eine anonyme einzigartige Kennung, eine Diagnose des Aphasiesyndroms und des Schweregrades sowie Bewertungen der sechs Spontansprachdimensionen. Die AAT-Auswertungen stammen von geschulten Logopäd*innen der Aphasiestation des Universitätsklinikums Aachen. Zehn Prozent der so entstandenen 66.935 Äußerungssegmente wurden für die Evaluation reserviert, nachdem nach Klassengröße sowie Schweregrad der prosodi-

Neurol Rehabil 2022; 28(2): 69–72 | https://doi.org/10.14624/NR2202003 | © Hippocampus Verlag 2022

## Automated diagnosis of prosodic disorders in aphasia using artificial neural networks

J. Haring, C. J. Werner, C. Kohlschein, U. D. Peitz, B. Schumann-Werner, J. Niehues

**Abstract**

The Aachen Aphasia Test (AAT) also records the spontaneous speech performance of a person with aphasia. However, this part can only be evaluated manually by trained personnel. The present work focuses on the automated scoring of one of the six AAT spontaneous speech scales. The possibility of implementing artificial neural networks for the automated identification of abnormalities of the dimension »prosody and articulation« is investigated and different approaches to this are compared. The aim of the study program is to automate the performance of the AAT using computerized methods while maintaining existing quality requirements.

**Keywords:** aphasia, Aachen Aphasia Test, artificial neural networks, prosody, dysarthria

schen Einschränkung, Patient*in und Aphasiediagnose stratifiziert wurde. Die Anzahl der Testsegmente war gering, da ein Ungleichgewicht der Äußerungslängen pro Patient*in und Aphasiesyndrom besteht, und damit ein Kompromiss zwischen ausreichenden Trainings- und Testbeständen.

## Methoden

Zunächst wurden die AAT-Werte dichotomisiert und als neue abhängige Variablen der Klassen »beeinträchtigt« (Wertung < 5) und »nicht beeinträchtigt« (Wertung = 5) beigefügt, um angesichts der begrenzten Datengrundlage verbesserte Precision und Recall Werte zu erreichen.

### Maschinelles Lernen

Maschinelles Lernen entspringt der Schnittmenge von Informatik und Mathematik. Im Gegensatz zur traditionellen Programmierung wird hier nicht anhand von Eingaben und programmierter Regeln eine Menge von Ausgaben erzeugt, sondern durch Betrachtung von Ein- und Ausgabebeispielen Wissen erzeugt, welches auf neue Eingangsdaten zu übertragen ist. Im Fall dieser Arbeit wurden demnach Audioaufnahmen mit der dazugehörigen Prosodie-Wertung verknüpft und durch maschinelles Lernen der Prozess repräsentiert, welcher diese erzeugte, um anschließend ungesehene Aufnahmen klassifizieren zu können. Fehler bei der ursprünglichen Bewertung des AATs wurden demnach auch von den Algorithmen des maschinellen Lernens abgebildet und somit propagiert. Die verwendeten Algorithmen sind als Blackbox Modelle zu verstehen, es konnte demnach nicht ohne Weiteres der genaue Entscheidungsprozess der Modelle, jedoch deren Lernerfolg auf Validierungsdaten untersucht werden.

### Einbettungsformat

Um maschinelles Lernen zu ermöglichen, war es zunächst notwendig, die Audiodaten in ein durch Computer lesbares Format einzubetten. Es wurde ein Spektrogramm in die Mel-Skala [8] überführt, welche die nicht-lineare Funktionsweise des menschlichen Gehörs in Bezug zur Hertz-Skala abbildet. Außerdem wurde durch Logarithmieren des Spektrums die Multiplikation des Anregungssignals und dessen Impulsantwort in eine Addition umgewandelt, was Log-Mel Filterbanken der Dimension [n, t, 80] als Einbettungsmethode erzeugte. Hier beschreibt n die Anzahl der Segmente, t die Zeit und 80 die Anzahl der Filterbanken.

### Netzarchitektur

Die Klassifizierung der prosodischen Einschränkung geschah mittels eines einschichtigen »Convolutional Neural Networks« (CNN), welches sich in einem Auswahlprozess von 13 verschiedenen Referenzarchitekturen bewährte. Es bestand aus 40 Filterkernen der Größe 5, deren Ergebnisse durch Global Maximum Pooling aggregiert und in eine Softmax-Schicht mit zwei Neuronen zur finalen Klassifizierung geleitet wurden. Das neuronale Netz wurde mittels Kreuzentropie unter Verwendung von Nadam [2] bei einer Batch Size von 20 und unter Ausschluss von 3.129 Validierungssegmenten, einer Teilmenge der Trainingsdaten, trainiert. Es wurden maximal 100 Epochen trainiert, jedoch wurde bereits nach zehn Epochen ohne Verbesserung der Validierungs-Lossfunktion abgebrochen.

### Verwendete Metriken

Zur Auswertung wurde die F1-Metrik – also das harmonische Mittel aus Precision und Recall – verwendet. Hier erklärt »Precision« den prozentualen Anteil der korrekten positiven Instanzen unter der Gesamtheit der positiven Klassifizierungen, und »Recall« den Anteil der tatsächlich positiven Instanzen, welche auch als solche erkannt werden. Die F1-Metrik kann somit Werte zwischen 0 % und 100 % annehmen, wobei höhere Werte bessere Leistung beschreiben. Im binären Fall, wie hier vorliegend, ist der Recall der positiven Klasse auch als Sensitivität, der Recall der negativen Klasse als Spezifität bekannt. Die F1-Metrik ist jedoch vollkommen unabhängig von richtig-Negativen, was zu einer übermäßig positiven Bewertung des Algorithmus führen kann, sofern diese nicht durch die Betrachtung von anderen Metriken ergänzt wird, wie in diesem Fall der Sensitivität oder anderen Metriken, wie dem Matthews Correlation Coefficient (MCC) [1]. Da eine falsch negative Diagnose potenziell schwerwiegender als eine falsch positive ist, wurden falsch positive den falsch negativen in der Klassifikation vorgezogen, was eine hohe Sensitivität voraussetzt und die Precision weniger einbezieht.

### Feinjustierung des Modell-Trainings

Um eine Klassifikation auf Patient*innenebene zu erreichen, wurden abschließend die Klassifikationsergebnisse der zugehörigen Segmente aus der Netzarchitektur durch Mehrheitsvotum aggregiert. Um einen stabileren Gradienten zu erreichen, wurde die Learning Rate auf $\eta_0 = 0{,}0005$ reduziert und die Batch Size auf 256 erhöht. Außerdem wurde die Learning Rate um einen Verfallmechanismus in der Epoche t unterstützt: $\eta_t = \eta_0 * e^{-0{,}01*t}$. Die Gewichte der Filterkerne wurden zudem regularisiert, indem ihre Quadratsumme multipliziert mit 0,05 an die Lossfunktion angefügt wurde, um eine Überanpassung zu vermeiden. Die Gewichte der Filterkerne wurden zudem durch LeCun Initialisierung [6] und der Bias mit 0,01 initialisiert, wie von Li et al. [7] empfohlen.

### Anpassung der Zielparameter

Aus der klinischen Erfahrung ist bekannt, dass bei der Bewertung der Spontansprachdimensionen insbesondere an der Grenze zwischen den Werten »4« und »5« die Interrater-Reliabilität sinkt. Daher kann informationstheoretisch nicht zwingend davon ausgegangen werden, dass eine klare latente Grenze in den Daten besteht, welche das Modell erlernen könne. Es wurden daher in einem zweiten Experiment alle Patient*innen mit einer Wertung von 4 für »Artikulation und Prosodie« aus dem Trainingsdatensatz, jedoch nicht aus dem Validierungsdatensatz, entfernt.

## Ergebnisse

Nach Feinjustierung des Modell-Trainings wurden 79 % F1-Metrik auf Patient*innenebene erreicht, es wurden jedoch besonders häufig Falsch-Negative ausgegeben. Wie in **Tabelle 2** dargestellt, betrug demnach der Recall auf der beeinträchtigten Klasse nur 64 %.

**Tab. 2:** Metriken initialer Experimente

| Klasse | Precision | Recall | F1 |
|---|---|---|---|
| beeinträchtigt | 100 % | 64 % | 78 % |
| nicht beeinträchtigt | 67 % | 100 % | 80 % |
| Durchschnitt | 83 % | 82 % | 79 % |

**Abbildung 1** verdeutlicht, dass besonders Patient*innen mit einer Bewertung nahe der Dichotomisierungsgrenze zwischen einer Wertung von 4 und 5 falsch klassifiziert wurden.

### Anpassung der Zielparameter

Die Anpassung der Zielparameter verbesserte die Klassifikationsergebnisse besonders für die beeinträchtigte Prosodieklasse, wie in **Tabelle 3** und **Abbildung 1** dargestellt. Die Sensitivität betrug nun 100 %. Auch die F1-Metrik verbesserte sich um 7 % auf insgesamt 86 %, und es traten keine falsch negativen Klassifikationen mehr auf.

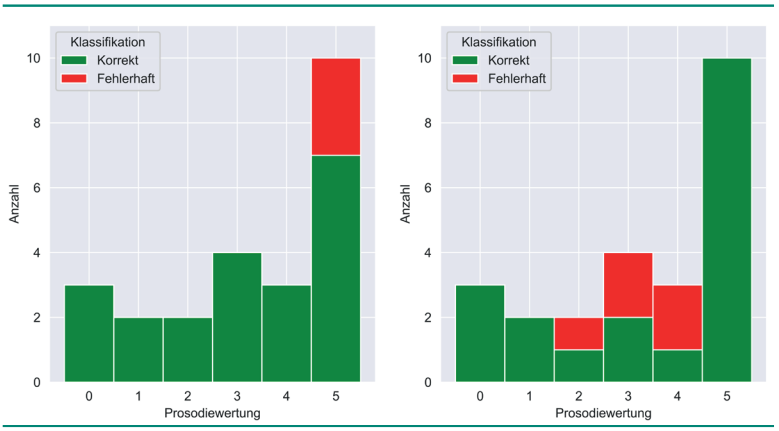


**Abb. 1:** Korrekte und fehlerhafte Klassifikationen vor und nach Abzug von Grenzfällen nach Prosodiewertung

**Tab. 3:** Metriken nach Abzug von Grenzfällen

| Klasse | Precision | Recall | F1 |
|---|---|---|---|
| beeinträchtigt | 82 % | 100 % | 90 % |
| nicht beeinträchtigt | 100 % | 70 % | 82 % |
| Durchschnitt | 91 % | 85 % | 86 % |

## Diskussion

Die vorliegende Arbeit demonstriert, dass rohe Sprachaufnahmen sich für die Klassifikation einer Spontansprachdimension des AAT eignen. Vergleichbar ist die beschriebene Herangehensweise mit der Arbeit von Kim et al. [4], welche eine binäre Klassifikation der Verständlichkeit bei pathologischer Sprache vornehmen. Es werden drei verschiedene Submodule zur Einbettung der Sprachinformationen vorgeschlagen, welche sich audio- und sprachspezifischer Merkmale (beispielsweise Formanten) sowie spektraler Einbettungen (MFCCs) bedienen und mittels Supportvektormaschinen klassifizieren. Der erreichte ungewichtete Recall liegt bei 73,5 % und somit 11,5 % unter dem in dieser Arbeit erreichten Recall von 85 %. Jedoch ist ein Vergleich nur eingeschränkt möglich, da die Patient*innen aus Kim et al. [4] keine Personen mit Aphasie sind. Sie leiden unter fortgeschrittenen Kopf-Hals-Karzinomen mit Zerebralparese und Amyotropher Lateralsklerose und somit Einschränkungen der Stimme und Sprechmotorik, im Gegensatz zur Einschränkung der gesamten Sprachverarbeitung bei Aphasie. Nichtsdestotrotz wird durch die vorliegende Arbeit demonstriert, dass auch die Verwendung einer einzigen Methode, in diesem Fall Log-Mel Filterbanken, mittels künstlicher neuronaler Netze existierende Ansätze zur Klassifikation von Sprachdaten übertreffen kann. Es liegen außerdem keine vergleichbaren Algorithmen für die Bewertung der Prosodie bei Aphasie vor, wes-

halb 86 % F1-Metrik als initiales Ergebnis für besagte Problemstellung zu sehen ist.

Durch die Notwendigkeit menschlicher Intervention in der Diagnostik sowie Therapie ist keine direkte klinische Anwendbarkeit zu erwarten, da ein Informationsverlust durch die Dichotomisierung der Wertungsskala stattfindet, welche jedoch bei der vorliegenden Datengrundlage notwendig ist, um das Training neuronaler Netze zu ermöglichen. Außerdem kann die Leistungsfähigkeit des Modells ohne Vergleichsgrößen seitens des AATs nicht eingeordnet werden, um die klinische Anwendbarkeit zu beweisen.

Erfahrungswerte aus dem klinischen Alltag bescheinigen dem AAT eine Zeitaufwendigkeit, welcher der präsentierte Algorithmus nicht unterliegt. Die Klassifikation auf Patient*innenebene dauerte im Schnitt 6,3 Sekunden bei einer durchschnittlichen Aufnahmelänge von 647,2 Sekunden pro Interview, im Gegensatz zur mehrstündigen Auswertung durch Fachpersonal.

## Fazit und Ausblick

Die Einbettung roher Sprachdaten mittels Log-Mel Filterbanken und deren Klassifikation durch Convolutional Neural Networks erwiesen sich als fähig, die Spontansprachdimension der »Artikulation und Prosodie« des Aachener Aphasie Tests zu klassifizieren, deren Aufnahmen einem neuen Sprachkorpus für deutschsprachige Aphasiepatient*innen entnommen wurden. Bei einem initialen Experiment mit binärer Auswertung wurden 86 % F1-Metrik erreicht, bei vollständiger Elimination von falsch negativen Diagnosen, also ohne verfehlte Diagnosen von real bestehenden Einschränkungen. Die Klassifikation anderer Spontansprachdimensionen mittels Sprachaufnahmen sowie die Kombination mit anderen Modalitäten, beispielsweise Transkripten, sind Gegenstand zukünftiger Forschung, mit dem Ziel, die Durchführung des Aachener Aphasie Tests durch computergestützte Methoden unter Aufrechterhaltung bestehender Qualitätsanforderungen zu automatisieren. Ein größerer Datenbestand kann die Notwendigkeit der Dichotomisierung aufheben und eine Klassifikation auf der vollen Werteskala, und somit auch den Vergleich mit bestehenden Instrumenten des AAT, ermöglichen. Dies hätte eine Verschiebung der Rollen des klinischen Personals zur Folge, welche näher an therapeutischen Maßnahmen als an der Diagnostik zum Einsatz kämen, da computergestützte diagnostische Verfahren Zeit einsparen können. Außerdem kann die Nutzung von Convolutional Neural Networks die Analyse der Klassifizierungsfilter auf lokale Signaleigenschaften, welche zum Klassifizierungsergebnis geführt haben, ermöglichen, indem die errechneten Filterkerne untersucht werden. So könnten eventuelle Muster, beispielsweise Änderungen im Artikulationstempo, mit spezifischen Klassifikationsergebnissen in Verbindung gebracht werden.

**Glossar**

| | |
|---|---|
| Batch Size | Anzahl der Proben, welche zur Berechnung des Gradienten auf einmal herangezogen werden |
| Epoche | Vollständiger Trainingszyklus auf dem Gesamtkorpus |
| Filterkern | Faltungsmatrix, welche über das Eingangssignal iteriert |
| Learning Rate | Faktor, welcher die Änderung des Gradienten skaliert |
| Loss Funktion | Funktion, welche die Distanz zwischen optimalem Wert und Punktschätzung berechnet |
| Softmax | Normalisierte Exponentialfunktion |
| Überanpassung | Verhalten eines Klassifikators, bei dem zum Training verwendete Instanzen besonders gut klassifiziert werden, die Übertragbarkeit auf ungesehene Instanzen jedoch minimal ist |

## Literatur


1. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics 2020; 21(1): 6.
2. Dozat T. Incorporating Nesterov Momentum into Adam. 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings 2016. https://openreview.net/pdf/OM0jvwB8jIp57ZJjtNEZ.pdf. Zugriff: 04.04.2022.
3. Huber W, Poeck K, Wilmes K. Aachener Aphasie-Test. Göttingen: Hogrefe 1983.
4. Kim J, Kumar N, Tsiartas A, Li M, Narayanan SS. Automatic intelligibility classification of sentence-level pathological speech. Computer speech & language. January 2015; 29: 132—44.
5. Kohlschein CP. Automatische Verarbeitung von Spontansprachinterviews des Aachener Aphasie Tests mittels Verfahren des maschinellen Lernens: Shaker; 2019.
6. LeCun Y, Bottou L, Orr G, Müller KR. Efficient BackProp. 2000 August.
7. Li FF, Krishna R, Xu D. Lecture notes on CS231n: Convolutional Neural Networks for Visual Recognition. 2020 April 28. Zugriff: 2020.12.27.
8. Stevens S, Volkmann J, Newman E. A Scale for the Measurement of the Psychological Magnitude Pitch. Journal of the Acoustical Society of America. 1937; 8: 185–90.


**Interessenvermerk**
Der korrespondierende Autor erklärt, dass die präsentierten Ergebnisse unterstützt durch HotSprings GmbH erarbeitet und publiziert wurden.

**Korrespondenzadresse:**
Julius Haring
HotSprings GmbH
Am Kraftversorgungsturm 3
52070 Aachen
julius.haring@umlaut.com