Automatische Stottererkennung und Klassifikation mit Hilfe von Hidden-Markov-Modellen

Tino Haderlein, Thomas Wittenberg, Michael Decher, Elmar Nöth

1 Einleitung

In der phoniatrischen und logopädischen Praxis werden im täglichen Routinebetrieb u.a. Patienten mit Stotter-Syndrom untersucht und mit Hilfe logopädischer Übungen behandelt. Die Diagnose solcher 'Redeflussstörungen' basiert in der Regel auf einer überwiegend subjektiven Überprüfung der Sprechfunktion mit einer möglichst genauen qualitativen und quantitativen Erfassung der Redeflussstörungen nach Art und Häufigkeit sowie der Sprechgeschwindigkeit [Wen96]. Insbesondere wird bei der Art der Redeflussstörung sog. klonisches und tonisches Stottern unterschieden. Klonisches Stottern äußert sich durch rasch aufeinanderfolgende Laut- und Silbenwiederholungen, insbesondere am Wortanfang. Tonisches Stottern manifestiert sich dagegen durch Wortdehnungen und Verzögerungen mit einer Tonanhebung [1].

Die derzeit häufigste klinische Praxis zur Diagnose und Therapieverlaufskontrolle von Patienten mit Stotter-Syndrom besteht in einer Video- oder Audioaufnahme der betroffenen Personen während des Lesens eines phonetisch ausgewogenen Standardtextes (i.d.R. „Nordwind und Sonne“). Neben der reinen Dokumentation der Diagnose bzw. des Therapiefortschrittes dienen diese Aufnahmen zusätzlich als Basis von objektiven und quantitativen Auswertungen der Sprachunflüssigkeit. Eine solche Quantifizierung und Auswertung der Aufnahmen ist jedoch sehr zeitraubend, da der Auswertende nicht alle Stotterphänomene und –ausprägungen beim ersten Durchlauf erfassen kann. Zudem ist diese Vorgehensweise rein subjektiv, da die Resultate sehr stark von der individuellen Ausbildung und Erfahrung des Untersuchenden abhängen.

In dieser Arbeit wird daher eine neuer, erfolgsversprechender Ansatz zur automatisierten Auswertung und Quantifizierung des Schweregrades des Stotterns vorgestellt. Dieses Verfahren basiert auf einer automatischen Analyse der digitalisierten Stotter-Aufnahmen mittels Methoden aus der automatischen Spracherkennung, den sog. Hidden-Markov-Modellen (HMM´s). Durch diesen Ansatz lässt sich der Aufwand der quantitativen Stotter-Analyse erheblich vereinfachen, es lässt sich zusätzlich eine statistische Analyse des analysierten Textes durchführen (z.B. Häufigkeiten von tonischem und klonischem Stottern) und das Verfahren arbeitet rein objektiv.

2 Stotter- bzw. Sprachanalysesystem

Aufbauend auf das am Lehrstuhl für Mustererkennung entwickelte HMM-basierte Spracherkennungssystem ISODORA [2] wurde im Rahmen einer interdisziplinären Arbeit [3] ein spezielles System implementiert und trainiert, mit dessen Hilfe es möglich ist, Stotterphänomene aus einem dem System bekannten Text (Nordwind und Sonne) zu bestimmen, zu zählen und zu klassifizieren.

Dieses System basiert auf einer Grammatik des fehlerfrei gesprochenen Nordwind-Textes. Diese Grammatik besteht zunächst aus einer Abfolge der korrekten Phonemketten des Textes [3]. Es wird davon ausgegangen, dass mit Hilfe des Spracherkennungssystems eine Abfolge von Phonemen erkannt und einer logischen Kette von bekannten Wörtern bzw. Phrasen zugeordnet wird. Diese bekannte Abfolge von Phonemen wird von dem Spracherkennungssystem intern als ein sog. Aussprache-Graph dargestellt. Dieser Aussprache-Graph ist ein gerichteter Graph, bei dem jeder Knoten ein Phonem und jede Kante zwischen aufeinanderfolgenden Knoten eine Übergangswahrscheinlichkeit zu einem möglichen Nachfolgephonem darstellt. Die Kombination aller Knoten und Kanten, gekoppelt mit den zugehörigen Übergangswahrscheinlichkeiten von Knoten zu Knoten wird auch als Hidden-Markov-Modell (HMM) bezeichnet. Sowohl einzelne Wörter, als auch Phrasen und Sätze lassen sich mit so einem HMM beschreiben und erkennen. Implizit beschreibt jedes einfache oder komplexe HMM zugleich die Grammatik der zu erkennenden Sprache. Aufbauend auf der bekannten Wortfolge wird automatisch ein Ausgabegraph erzeugt, der mögliche gesprochene Realisierungen des Nordwind-Textes enthält, wenn Stotterphänomene berücksichtigt werden. Der Basis Nordwind-Graph wurde im Rahmen dieser Arbeit derart erweitert, dass zusätzlich speziell stotterspezifische Phänomene wie Ton- und Lautrepetitionen, Wortdehnungen, Unterbrechungen sowie sog. „gefüllte“ Pausen von dem System erkannt werden [4].

Da Stotterphänomene überwiegend an Wort- bzw. Phrasengrenzen auftreten, wurde im derzeit realisierten System [2] innerhalb des Aussprache-Graphs die wortweise Erkennung mittels eines speziellen Wort-Aussprachelexikons realisiert. Im Gegensatz zum generellen Standard-Spracherkennungssystem können nun zusätzlich an jeder Stelle stotterspezifische Äußerungen detektiert werden. Zur Erkennung von tonischem Stottern wurden zudem in dem Aussprache-Graph an jedes Phonem sog. aktive Pause- bzw. Schweigeknoten angefügt. Für die Erkennung von

klonischen Stottersyndromen wie Wort- bzw. Silbenwiederholungen oder gar Neuanfängen von Sätzen oder Phrasen wurden zusätzliche Schleifen auf den vorhergehenden Phonemknoten eingeführt.
Zusammen mit den genannten Ergänzungen ist das erweiterte System in der Lage, auftauchende Stotterphänomene nicht nur zu erkennen und zu klassifizieren, sondern zudem zu zählen und eine Statistik aufzubauen.

3 Patienten und Probanden:

In der Abteilung für Phoniatrie und Pädaudiologie bzw. der Berufsfachschule für Logopädie der Universität Erlangen wurden in einem Zeitraum von 36 Monaten von 37 untersuchten und behandelten Patienten (27 männlich, 10 weiblich) insgesamt 69 Video-Aufnahmen (U-MATIC) gemacht. Alle Patienten lasen den phonetisch ausgewogenen Standardtext `Nordwind und Sonne` teils vollständig (52x), teils in gekürzter Form (17x). Die Tonspuren aller Videoaufnahmen wurden am Lehrstuhl für Mustererkennung mit 16 Bit digitalisiert. Zum Training des Stotteranalysesystems wurden zusätzlich 16 Aufnahmen des vollständig gesprochenen Nordwind-Textes von flüssigsprechenden Probanden aus dem BAS Strange-Corpus 1 aus dem Bayrischen Archiv für Sprachsignale verwendet. Nachfolgende Tabelle zeigt einige Charakteristiken der Stotter- Aufnahmen und Patienten. Über das Alter der Normal-Sprecher machen die BAS-Unterlagen keine Angaben.

	Männlich	Weiblich	Gesamt
Anzahl der Sprecher	27	10	37
Davon subjektiv schwere Stotterer	5	1	6
Anzahl der Aufnahmen	49	20	69
Kurzer Text	13	4	17
Vollständiger Text	36	16	52
Alter des ältesten Sprechers	45.1 J.	31.0 J.	
Alter des jüngsten Sprechers	12.2 J.	10.3 J.	
Durchschnittsalter	23.4 J.	17.11 J.	21.8 J.
Standardabweichung	7.6 J.	5.3 J.	7.4 J.
Dauer des kürzesten (vollst.) Textes	0:35 min	0:36 min	
Dauer des längsten (vollst.) Textes	3:09 min	5:41 min	
Durchschnittliche Dauer des kompl. Textes	1:35 min	1:38 min	1:36 min
Standardabweichung	0:44 min	1:20 min	0:58 min
Durchschnittliche Dauer des kurzen Textes	0:24 min	0:22 min	0:23 min
Standardabweichung	0:10 min	0:03 min	0:08 min

Tabelle 1: Charakteristiken der verwendeten Aufnahme-Datenbank

4 Experimente und Ergebnisse

Die Experimente, die auf dem obigen Datensatz durchgeführt wurden, behandelten zwei unterschiedliche Fragestellungen:

1. Wie gut kann das vorgestellte automatische System Unflüssigkeiten in gestotterter Sprache detektieren, und
2. Welche berechenbare Merkmale lassen klinische Aussagen über den Schwergrad des Stotterns zu?

Korrektheit des Systems

Um die erste Fragestellung zu bearbeiten, wurden zunächst die 16 Referenzaufnahmen von flüssigsprechenden Probenden sowie 16 repräsentative Stotteraufnahmen aus der digitalisierten Stichprobe von geschultem Personal segmentiert, d.h. auftretende Unflüssigkeiten wurden manuell markiert. Die Anzahl der manuell detektierten Unflüssigkeiten wurde anschliessend mit der korrespondierenden Anzahl der zugehörigen automatischen extrahierten Unflüssigkeiten verglichen. Obwohl das automatische System eine Tendenz dazu zeigte, wesentlich mehr Unflüssigkeiten zu detektieren als zuvor manuell segmentiert worden waren, betrug der Korrelationskoeffizient zwischen beiden Ergebnissen 0,99 über alle Aufnahmen, 0,988 für die Stotternden und 0,978 für die Normalsprecher. Abbildung 1 zeigt dieses Verhältnis zwischen beiden Segmentierungsarten. Abbildung 2 zeigt dagegen die Anzahl der manuell segmentierten Unflüssigkeiten in Relation zu der prozentualen Anzahl der daraus resultierenden Phonemfehler.

Untersuchte Merkmale

Zur Beantwortung der zweiten Frage wurden zunächst das tonische Stottern untersucht, in dem die Länge der Pausen zwischen Wörtern bzw. Phrasen von Stotterern bzw. Nicht-Stotterern in Abhängigkeit der Pausenlänge einander gegenübergestellt wurden. Das Ergebnis, dargestellt in Abb. 3, zeigt, dass nicht nur tonische Stotterer vermehrte Pausen während des Lesens einlegen sondern die untersuchten Stotterer allgemein eine längere Zeitspanne zum Lesen des Textes

benötigen. Speziell kurze Unterbrechungen mit einer Dauer von 200 ms tauchen zehnmal häufiger in der Klasse der Stotterer als in der Klasse der Nicht-Stotterer auf. Im Bereich der Pausendauern von 500 ms beträgt der Unterschied nur noch den Faktor zwei.

Um klonisches Stottern - also Wortwiederholungen und Repetitionen - zu quantifizieren, wurde die Anzahl der bekannten - manuell segmentierten - Unflüssigkeiten pro Wort in einer Aufnahme mit der Gesamtsumme der darin auftretenden Pausen einander gegenübergestellt. Dabei ergab sich bei einem Vergleich von 16 Aufnahmen von Nicht-Stotterern mit 16 repräsentativen Aufnahmen von Stotterern ein relativ ausgewogenes Verhältnis zwischen Unflüssigkeiten und Pausenlänge und somit ein Häufung im Bereich von 0.1 Fehler pro Wort und 1 Sekunde Pause, s. Abb. 4. Die Ausreißerpunkte außerhalb dieses Clusters zeigen speziell die klonischen Stotterer, bei denen durch die Repetitionen überproportional viel Pausen auftraten.

5 Ausblick

Die beschriebenen Unterscheidungs- , Klassifikations- und Bewertungskriterien für klonisches bzw. tonisches Stottern gegenüber flüssigem Sprechen sind in diesem Ansatz noch relativ grob, sie zeigen jedoch eine positive Tendenz. Insbesondere bedarf es für weitere Untersuchungen des vorgestellten Systems einer größeren Stichprobe, die von geschultem Personal manuell segmentiert werden muss. Bisher wurde das HMM der zugrundeliegenden Grammatik ausschließlich mit Spontansprache von Normalsprechern trainiert. Es ist jedoch zu erwarten dass ein Nachtrainieren des Systems mit einer klassifizierten Stotterstichprobe die Erkennungsleistung um ein wesentliches erhöht. Langfristig ist auch geplant, im Rahmen einer erweiterten Studie, die Klassifikation von Stotterphänomenen auf spontane Sprache zu erweitern, da aus vielen Studien bekannt ist, dass die Symptomatik des Stotterns während des Lesens nicht bei allen Patienten gleichmäßig ausgeprägt ist.

Literatur

[1] J. Wendler, W. Seidner, G. Kittel, U. Eysholdt: *Lehrbuch der Phoniatrie und Pädaudiologie*, Thieme Verlag, Stuttgart, 1996.

[2] E. G. Schukat-Talamanzini: *Automatische Spracherkennung*, Viehweg Verlag 1995

[3] T. Haderlein: *Quantitative Stotteranalyse: Erstellung eines Diagnoseprogramms zur Feststellung von Unflüssigkeiten in gesprochener Rede*. Studienarbeit, Lehrstuhl für Mustererkennung (Informatik 5) Univ. Erlangen-Nürnberg, 2000.

[4] E. Nöth, H. Niemann, T. Haderlein, M. Decher, U. Eysholdt, F. Rosanowski, Th. Wittenberg: *Stuttering Recognition using Hidden Markov Models*. Erscheint in *Procedings Int. Conference on Spoken Language Processing, Beijing, China, 2000.*

Autoren

Dr.-Ing. Thomas Wittenberg, Fraunhofer Institut für Integrierte Schaltungen -- Angewandte Elektronik, Projektgruppe Medizinische Bildverarbeitung, Am Weichselgarten 3, 91058 Erlangen, Tel.: +49-(0)9131-776-512, Fax: -598, Email:wbg@iis.fhg.de

Dr.-Ing. Elmar Nöth, Cand.-Inf. Tino Haderlein, Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen-Nürnberg, Martensstr.3, 91058 Erlangen, Tel.: +49-(0)9131-852-7888, Email: noeth@informatik.uni-erlangen.de

Michael Decher, Lehr-Logopäde, BFS für Logopädie, Universität Erlangen-Nürnberg, Universitätsstr. 44, 91054 Erlangen, Email: decher@phoni.imid.uni-erlangen.de

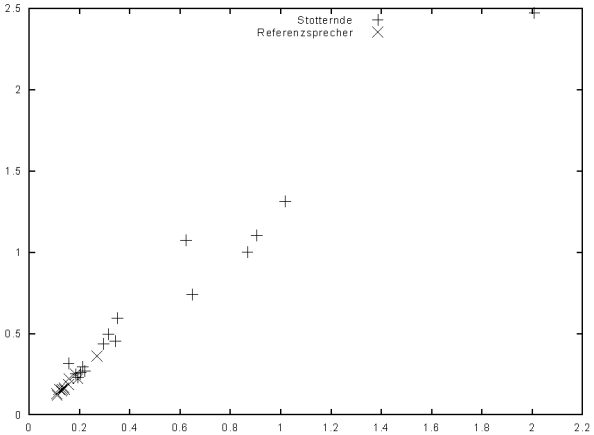

Abbildung 1:Erkannte Unflüssigkeiten: Tatsächliche Unflüssigkeiten (X-Achse), Berechnete Unflüssigkeiten (Y-Achse)

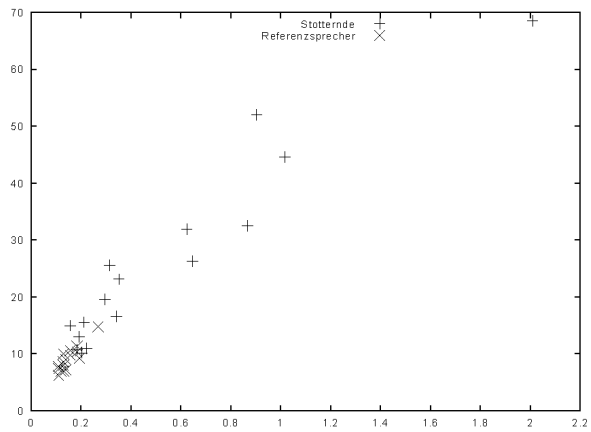

Abbildung 2: Lautfehlerrate

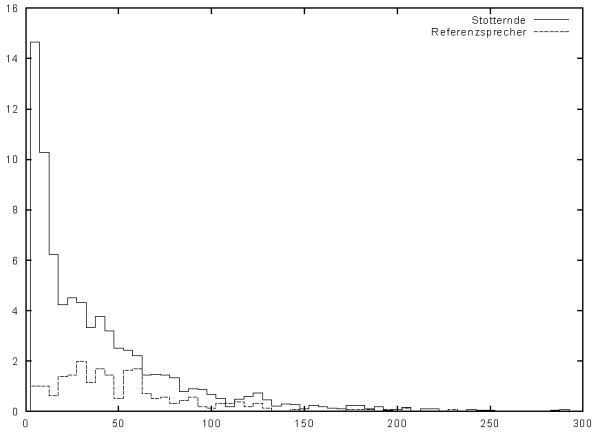

Abbildung 3: Absolute Häufigkeit von Pausen bezogen auf die Dauer in Frames

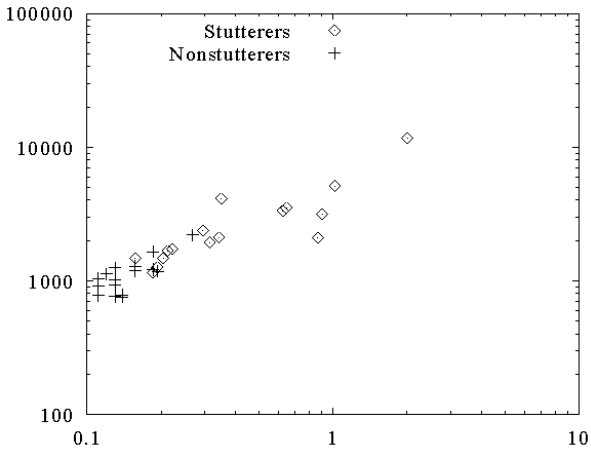

Abbildung 4: Anzahl von Unflüssigkeiten pro Wort (X-Achse) und Summe aller Pausen (logarithmische Y-Achse)