Automatische Spracherkennung für Sprecher mit Dysarthrie

Susanne Rexeis[1], Stefan Petrik[2], Gernot Kubin[3]

[1] *JOANNEUM RESEARCH, 8010 Graz, Österreich, Email: susanne.rexeis@joanneum.at*

[2] *TU Graz, 8010 Graz, Österreich, Email: stefan.petrik@tugraz.at*

[3] *TU Graz, 8010 Graz, Österreich, Email: gernot.kubin@tugraz.at*

Einleitung

Sprachstörungen, die aufgrund neuro-muskulärer Schädigungen und damit zusammenhängender verminderter motorischer Kontrolle über den Sprechapparat auftreten, werden unter dem Begriff Dysarthrie zusammengefasst. Da bei Dysarthrie-Patienten auch Lähmungen anderer Körperteile, zB. der Extremitäten, auftreten, kann die Verwendung von Sprachtechnologie eine wertvolle Erleichterung im Alltagsleben bedeuten. Durch das breite Spektrum an Abweichungen zur Regelsprache ist der Einsatz von Standard-Spracherkennungssystemen stark eingeschränkt. In dieser Arbeit wurden Methoden evaluiert, ein sprecherunabhängiges System mittels akustischer und lexikaler Adaptierung auf Dysarthrie-Patienten anzupassen.

Spracherkennungssystem

Für diese Arbeit wurde ein sprecherunabhängiger (SU) Einzelwort-Spracherkenner trainiert. Die Trainingsdaten sind Teil der Speechdat-AT Datenbank, die Aufzeichnungen von 1000 Sprechern, durchgeführt über das Telefonnetz, enthält.

Als akustisches Modell (AM) wurde ein Hidden-Markov-Modell mit jeweils 3 Zuständen pro Phonem verwendet. Dessen Beobachtungswahrscheinlichkeiten wurden durch Gauß'schen Mischverteilungen mit 16 Komponenten modelliert. Sowohl das Training und die Evaluierung des Erkenners, als auch die akustische Adaptierung erfolgten mit Werkzeugen aus dem HTK-Toolkit [5].

Zur Evaluierung des Erkenners wurde ein Datensatz von 69 Kommandoworten verwendet, die für einfache Interaktionen mit einem Computer geeignet sind. Das Lexikon des Erkenners basiert auf dem der Speechdat-Datenbank.

Sprachdaten der Dysarthrie-Patienten

Die Daten für die Evaluierung stammen von fünf männlichen Sprechern, im Alter von 17 bis 38 Jahren, die an mittlerer bis schwerer Dysarthrie leiden. Die Aufnahmen wurden vom Verein „Simon Listens"mit einem Laptop und einem Headset durchgeführt und haben eine Abtastrate von 16kHz. Um die Aufnahmen mit dem verwendeten Erkenner evaluieren zu können wurden sie auf 8kHz heruntergesampelt und Bandpass begrenzt.

Von jedem Sprecher wurden zwischen 5 und 16 Sessions der Kommandoworte aufgenommen. Zur Bestimmung von Aussprachefehlern wurden zusätzlich 100 Worte aus dem Sotschek-Reimtest [2] ausgewählt und aufgenommen. Dieser Reimtest besteht aus 100 Ensembles einsilbiger Wörter, die sich entweder im Silbenkern, An- oder Auslaut unterscheiden.

Eine akustische Analyse der Aufnahmen [6] zeigte, dass Dysarthrie-Patienten langsamer und gedehnter sprechen, als Vergleichspersonen, die nicht an Dysarthrie leiden. Auch konnten die Sprecher bestimmte Phoneme nicht korrekt aussprechen. Diese Phoneme variierten nach Sprecher, jedoch hatten alle Probleme mit Plosivlauten.

Akustische Adaptierung

Das SU-AM wurde mittels Maximum Likelihood Linear Regression (MLLR) auf jeden der dysarthrischen Sprecher angepasst. Bei diesem Verfahren werden die Mittelwert-Vektoren der Gaußverteilungen, zur Modellierung der Beobachtungswahrscheinlichkeiten, lineartransformiert, sodass die Wahrscheinlichkeit der Sprachdaten des jeweiligen Sprechers maximiert wird.

Für die Adaptierung wurde jeweils eine Session der Kommandoworte verwendet. Da die Zahl der zu adaptierenden Parameter im Vergleich zur Anzahl der verwendeten Daten sehr groß ist, wurden die Vektoren vor der Adaptierung nach ihrer euklidischen Distanz geclustert und jeweils eine Transformation für jeden Cluster ermittelt.

Lexikale Adaptierung

Zur Anpassung des Spracherkenners an die individuelle Ausdrucksweise der Dysarthrie-Patienten wurde das Aussprachelexikon (PL) um neue Aussprachevarianten erweitert. Wie in [1] beschrieben, wurden dazu Weighted Finite State Transcducer (WFST) Repräsentationen für das AM und PL des Erkenners verwendet.

Das PL-FST L übersetzt jedes Wort w_i aus dem Lexikon in eine Phonem-Kette. Die WFST-Repräsentation des AMs C wurde aus der Evaluierung des Spracherkenners auf den Reimtestdaten abgeleitet. Das WFST hat nur einen Zustand, die Transitionen stellen alle möglichen Kombinationen zwischen den Phonemen dar. Die Kosten für jede Transition entspricht dem Logarithmus des entsprechenden Eintrags der Konfusionsmatrix aus der Reimtest-Evaluierung.

Durch Komposition dieser WFSTs werden für jedes Wort w_i im Lexikon neue Varianten p_{cand} generiert und nach ihren Kosten gereiht.

$$p_{cand} = w_i \circ L \circ C \tag{1}$$

Die auf diese Weise generierten Aussprachevarianten werden durch Komposition mit der Inversen des PL-FST L^{-1} auf Wörter aus dem Lexikon zurückgeführt.

$$w_{conf} = w_i \circ L \circ C \circ L^{-1} \quad (2)$$

Durch Reihung der Ergebnisworte nach ihren Kosten wird die Verwechselbarkeit der Aussprachevariante für w_i mit anderen Wörtern aus dem Lexikon bewertet.

Zur Evaluierung des beschriebenen Ansatzes wurde die Implementierung aus [7] verwendet.

Evaluierung

Für einen Sprecher ohne Sprachstörung konnte das SU Basissystem bei der Kommandowort-Erkennung eine Wort-Erkennungsrate (WEKR) von über 97% erzielen. Im Gegensatz dazu lag die WEKR für die Dysarthrie-Patienten bei unter 30%. Die Ergebnisse der Evaluierung sind in Tabelle 1 dargestellt.

Durch die MLLR-Adaptierung konnte die WEKR um ein Vielfaches verbessert werden. Jedoch lag die maximale WEKR nur bei rund 70%. Eine Analyse der Ergebnisse ergab, dass Teile des verwendeten Vokabulars, zB. das Wort „Prozent“, eine komplexe Aussprache haben und den dysarthrischen Sprechern deshalb Probleme bereiteten. Auffällig ist die besonders niedrige WEKR für Sprecher 2. Dieser Sprecher hatte in der Evaluierung die undeutlichste Aussprache, die zum Teil auch für das menschliche Ohr kaum verständlich ist.

Die Adaptierung des PLs führte bei drei Sprechern zu einer leichten Verbesserung gegenüber dem Basissystem. Eine Kombination der lexikalen und akustischen Adaptierung erzielte nur bei einem der Sprecher ein besseres Ergebnis als die MLLR-Adaptierung.

Ein Grund für das schlechte Abschneiden der lexikalen Adaptierung ist, dass nur die 100 Reimtest-Erkennungsergebnisse verwendet wurden, um die Aussprachefehler abzuleiten. Eine Möglichkeit ohne die Aufnahme weiterer Daten mehr Ergebnisse zu generieren, wäre eine Erweiterung des Reimtests um zusätzliche Reime für jede Aufnahme. Zudem enthält der verwendete Reimtest nur einsilbige Wörter, die Kommandoworte sind aber meist mehrsilbig. Damit werden zahlreiche Phonem-Kombinationen nicht abgebildet. Die Verwendung eines mehrsilbigen Reimtests, wie in [3] beschrieben, wäre eine interessante Alternative zum Sotschek-Reimtest, weil er auch speziell für Dysarthrie-Patienten konzipiert wurde.

Zusammenfassung

Die unterschiedlichen Merkmale der Dysarthrie wirken sich massiv auf die Erkennungsrate von Spracherkennungssystemen aus.

Durch die akustische Adaptierung lässt sich die Erkennungsrate bereits um ein Vielfaches steigern. Jedoch ist der Erkenner aufgrund der absoluten Erkennungsrate von unter 70% für den praktischen Einsatz nicht geeignet. Ein Problem ist, dass einige der aufgenommenen Worte phonetisch zu schwierig auszusprechen sind.

Tabelle 1: Wort-Erkennungsrate in [%] des sprecherunabhängigen Spracherkenners vor und nach Anwendung von akustischer und/oder lexikaler Adaptierungsmethoden

Sprecher	Spracherkenner			
	Basis-System	MLLR	WFST	MLLR+ WFST
1	30,27	68,12	28,99	66,67
2	9,21	19,20	9,21	20,47
3	9,42	49,64	10,87	49,28
4	28,62	71,38	29,71	70,29
5	28,26	69,57	31,16	69,57

Eine zusätzliche Adaptierung des Aussprachemodells des Erkenners auf die dysarthrischen Sprecher konnte in dieser Evaluierung keine wesentliche Verbesserung der Erkennungsrate erzielen. Der in dieser Arbeit beschriebene Ansatz, basierend auf dem Sotschek-Reimtest, konnte nicht genug Daten für die korrekte Modellierung der Aussprachefehler der individuellen Patienten liefern. Eine weitere Einschränkung ist, dass der Reimtest auf Phoneme ausgelegt ist, die in der Regelspache leicht verwechselt werden und spezielle Aussprachepobleme von Dysarthrie-Patienten nicht vollständig abdeckt.

Literatur

[1] Fosler-Lussier, E, Amdal, I. und Kuo, H.-K. J.: A framework for predicting speech recognition errors. Speech Communication (2005), 153-170

[2] Sotschek, J: Ein Reimtest für Verständlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachübertragungsgüte. Der Fernmeldeingenieur (1982), 345-353

[3] Ziegler, W., Hartmann, E. und von Cramon D.: Word identification testing in the diagnostic evaluation of dysarthric speech. Clinical Linguistics & Phonetics (1988), 291-308

[4] simon listens - Gemeinnütziger Verein für Forschung und Lehre, URL: `http://www.simon-listens.org`

[5] Hidden Markov Model Toolkit, URL: `http://htk.eng.cam.ac.uk`

[6] Susanne Rexeis: Automatic Speech Recognition for Dysarthric Speakers. Masterarbeit, TU Graz (2011)

[7] Stefan Petrik: Phonetic Similarity Matching of Non-Literal Transcripts in Automatic Speech Recognition. Doktorarbeit, TU Graz (2010)