HERAUSFORDERUNGEN AN SPRACHINTERFACES FÜR KINDER

Karina Matthes, Felix Claus, Horst-Udo Hain, Rico Petrick

Projektgruppe Interaktive Sprachlehrmittel Dresden
karina.matthes@gmail.com

Kurzfassung: Der Artikel beschreibt Herausforderungen, die bei der Entwicklung eines Sprachinterfaces für Kinder entstehen. Kindliche Sprachinterfaces stellen eine wissenschaftliche Neuerung dar. Demnach ergeben sich auch für die benötigten traditionellen Forschungsgebiete im Bereich Sprachinterfaces (Spracherkennung, Sprachsynthese und Dialog) neue Herausforderungen. Der Artikel beschreibt dies für die Gebiete Spracherkennung und Dialog. Da Kinder andere Verhaltens- und Interaktionsmuster aufweisen als Erwachsene, sind herkömmliche Dialoggestaltungsprinzipien zu überdenken. Für die Spracherkennung ergeben sich neue Herausforderungen sowohl durch die anatomischen Unterschiede als auch durch die sich mit dem Alter entwickelnden sprachlichen Fähigkeiten. Erkennungsexperimente zeigen das Verhalten von Spracherkennungssystemen mit Kindersprache.

1 Einleitung

Einer von mehreren möglichen Modi für Mensch-Maschine-Schnittstellen ist die Sprache. Nutzer empfinden eine Bedienung von Geräten/Maschinen über Sprache als natürlich und auch sympathisch. Aus dieser Motivation heraus wurde in den vergangenen zwei Jahrzehnten versucht, die Bedienung technischer Geräte durch die Einbeziehung von Sprachinterfaces freundlicher zu gestalten und weniger technik-affinen Personen besser zugänglich zu machen. Sprachinterfaces können auch unterstützend wirken, wenn die Bedienung per Hand eingeschränkt ist (Spracheingabe: Sprachsteuerung bei Automobilelektronik) sowie die Augen vom Gerät abgelenkt sind (Sprachausgabe: sprachliche Nutzerführung beim Navigationssystem).

Abbildung 1 zeigt die drei Komponenten eines Sprachinterfaces Dialog, Spracheingabe sowie Feedbackkanal und Sprachausgabe. Für alle drei Komponenten existieren etablierte Wissenschaftszweige. Der Wissensstand im Bereich Sprachinterfaces beschränkt sich traditionell auf Anwendungen für Erwachsene. Dies ist durch die bisherigen Anwendungen und die zugehörige Nutzergruppe gegeben (z. B. Sprachsteuerung im Kfz). Ein neues Forschungsgebiet ist die Untersuchung von Sprachinterfaces für Kinder. Hier ergeben sich im Wesentlichen drei Anwendungsfelder: elektronisches Spielzeug, elektronische Lehrmittel sowie moderne Medien für die Sprachtherapie. In diesem Artikel werden Unterschiede und Herausforderungen von Sprachinterfaces für Kinder im Gegensatz zu Sprachinterfaces für Erwachsene beschrieben. Dabei werden die beiden Komponenten Dialog und Spracheingabe beleuchtet, auf die Gestaltung des sprachlichen Feedbackkanals wird in diesem Artikel nicht eingegangen. Ziel der geführten Betrachtungen ist ein Spielzeug mit Sprachinterface und pädagogischer Funktionalität [1].

2 Dialog

2.1 Aspekte der Mensch-Maschine-Kommunikation

Technische Geräte werden in der Regel dazu geschaffen, um für den Menschen eine Funktionalität zur Verfügung zu stellen. Damit Menschen diese Funktionalität bedienen können, besitzt jedes Gerät eine Bedienschnittstelle (Userinterface). Der Wissenschaftszweig der Mensch-Maschine-Kommunikation (MMK) oder Mensch-Maschine-Interaktion widmet sich der Gestal-

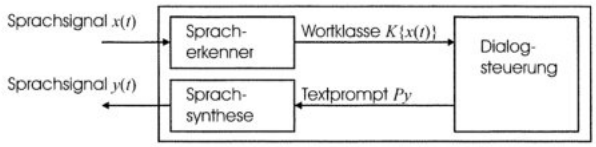

Abbildung 1 - Die Grafik beschreibt die einzelnen Wirkungsblöcke eines Sprachinterfaces.

tung dieser Bedienschnittstelle. Dabei wird das abstrakte System Mensch-Maschine betrachtet und dieses nach verschiedenen Kriterien optimiert. Kriterien können bspw. sein

- Bedienungsergonomie,
- Intuitivität,
- Task Completion Time oder
- Stärke der Aufmerksamkeit, die die Bedienung erfordert.

Die Bedienschnittstelle wird als Mensch-Maschine-Schnittstelle (engl. Human Machine Interface (HMI)) bezeichnet. Es existieren verschiedene Formen des HMIs, was bedeutet, dass Eingaben und Ausgaben des Gerätes auf verschiedenen Wegen oder Modi zwischen Mensch und Maschine ausgetauscht werden können. Beispiele sind

- Eingabemodi – Taster, Pedale, Hebel, Sprache, Berührung oder Touch-Screen,
- Ausgabemodi – funktionale Maschinenreaktionen, Displayanzeigen, Signalleuchten oder LEDs, akustische Signale (Piepen, Hupen) oder Sprache.

Der Wissenschaftszweig der Mensch-Maschine-Kommunikation ist seit Jahrzehnten etabliert, ein Standardwerk ist bspw. [2]. Traditionell sind HMIs für Erwachsene untersucht worden. Jüngere Untersuchungen befassen sich auch mit HMIs für Kinder, z. B. [3].
Bei Sprach-HMIs treten Mensch und Maschine in einen sprachlichen Dialog. Mit der Gestaltung von Sprach-HMIs (Dialogdesign) befasst man sich in den letzten zehn bis zwanzig Jahren, seit praktisch einsetzbare Sprach-HMIs existieren (z. B. maschinelle Telefonauskunft, Sprachsteuerung für Kfz-Elektronik). Ein aktueller Stand wird z. B. in [4] dargestellt. Die Untersuchung von Sprach-HMIs ist ein vergleichsweise junges Teilgebiet der MMK. Untersuchungen zur MMK für kindliche Sprach-HMIs sind den Autoren nicht bekannt und stellen wissenschaftliches Neuland dar. Ein Dialog für ein Kinderspielzeug unterliegt anderen Anforderungen als ein Dialog für eine Steuerung für ein Navigationssystem. Diese Anforderungen zu bestimmen und Lösungsstrategien dafür zu erarbeiten ist Inhalt aktueller und zukünftiger Untersuchungen.

2.2 Implementierung eines Dialogmoduls

Um die Inhalte in einem Sprachinterface umzusetzen, ist es nötig, ein Dialogmodul zu gestalten. Das Dialogmodul verarbeitet Eingaben (Events) und und reagiert situationsspezifisch mit entsprechenden Ausgaben (Actions). Im Falle eines reinen Sprachinterfaces bestehen

- Events aus Erkennungsergebnissen eines Spracherkenners, der als Sensor für gesprochene Äußerungen fungiert, und
- Actions aus Sprachäußerungen des Gerätes.

Bei nicht reinen Sprachinterfaces können Events und Actions wie oben beschrieben außerdem weitere Modi aufweisen. Events und Actions werden in einer Statemachine verarbeitet. Die Statemachine besteht aus Zuständen und Zustandsübergängen. Die einzelnen Zustände stellen Wartepositionen dar, in denen auf bestimmte Events (hier Sprachkommandos/Spracherkennungsergebnisse) gewartet wird. Tritt ein solches Event auf, kommt es zu einem Zustandsübergang, bei dem eine Action ausgelöst werden kann (hier Sprachäußerung). Nach Zustandsübergang besteht wieder die Warteposition.

2.3 Pädagogische Aspekte

Potentielle Einsatzgebiete für Lernspielzeuge in dieser Form liegen u. a. im Bereich der Wortschatzausbildung, Wissensvermittlung und des Aussprachetrainings. Dem entsprechend ergibt sich ein interdisziplinäres Forschungsfeld, das die wissenschaftliche Begleitung und Umsetzung didaktischer Applikationen erfordert. Hier handelt es sich erneut um wissenschaftliches Neuland in der Interaktionsforschung sowie in der Pädagogik unter Benutzung von Lehrmedien.

Eine reine Vermittlung von Sprache bzw. Wissen über einen Sprachdialog ist im frühkindlichen Bereich kaum möglich. Vielmehr ist ein sprachlich geführter Lernkurs mit bildlicher Unterstützung unter Berücksichtigung pädagogischer Gesichtspunkte angebracht. Dieser sollte thematisch und inhaltlich abwechslungsreich gestaltet sein, angepasst an Alter und/oder sprachlicher Entwicklung des jeweiligen Kindes. Die Herausforderungen in der Gestaltung eines solchen Kurses liegen vorwiegend im Erhalt der Aufmerksamkeit und der Förderung der Konzentration von Kindern und in der Vermittelung der konzipierten Lehrinhalte. Dazu gehören u. a. die Auswahl der Themen, der dazugehörigen Lehrinhalte, deren Umfang und die Einbettung des zu vermittelnden Wissens in spielerische Zusammenhänge. Dem entsprechend sind die Auswahl spielzeugspezifischer Vermittlungsformen, z. B. Frage-Antwort-Spiele, und deren kontextgebundener Einsatz in diesem Forschungsgebiet einzubetten. Bezugnehmend auf die visuelle Unterstützung müssen altersgerechte Illustrationen gewählt werden, deren ästhetische Gestaltung die Freude am Spiel und Lernen und den Wunsch nach Wiederholung wecken. Die sprachliche Führung des Dialogs wird altersgerecht angelegt, indem das Lehrmittel Elternsprache, sog. Motherese, benutzt, die in der Kommunikation gegenüber Säuglingen und Kleinkindern angewendet wird. Motherese ist bspw. gekennzeichnet durch langsamere Sprechgeschwindigkeit, Sprechen in höherer Tonlage, viele Fragen usw. [5].

Im Entwicklungsprozess didaktischer Sprachinterfaces für Kinder nimmt der pädagogische Ansatz eine gesonderte Stellung ein, da keine bestehenden Lehrkonzeptionen für diese Formen des Lernens in der Zielgruppe der 3-6jährigen vorhanden sind. Eventuell kann der Rückgriff auf Erkenntnisse aus dem Bereich didaktischer Software oder traditioneller Medien wie Bilderbücher und -geschichten die Forschungsarbeit fördern und die Gestaltung einer Lernkonzeption forcieren. Die Komplexität des Aufgabenbereiches liegt nicht nur in der theoretischen Auseinandersetzung, sondern bezieht sich ebenfalls auf forschungsbegleitende Evaluationsverfahren und anpassende Optimierungsschritte in Abstimmung mit technischen Gegebenheiten.

3 Spracheingabe – Spracherkennung von Kindersprache

Wie in den vorangegangenen Abschnitten bereits beschrieben wurde, werden Geräte mit sprachlichen HMIs zunächst für Erwachsene untersucht, Sprach-HMIs für Kinder spielen außer bei Billigspielzeug bislang keine Rolle. Demzufolge ist auch die Erkennung von Kindersprache wissenschaftliches Neuland. Vereinzelte Experimente existieren in [6] oder in [7]. Kindersprache unterscheidet sich von Erwachsenensprache, demnach müssen auch herkömmliche Erkennungsverfahren an die Besonderheiten von Kindersprache angepasst werden bzw. verschiedene Paradigmen der Spracherkennung für Kindersprache in Frage gestellt werden.

3.1 Besonderheiten von Kindersprache

Unterschiede von Kinder- und Erwachsensprache existieren auf verschiedenen Ebenen.

3.1.1 Anatomische Ursachen

Der Körper eines Kindes ist kleiner als der eines Erwachsenen, das gilt auch für den Vokaltrakt und die Stimmlippen.

Spektrale Hülle Der Vokaltrakt eines Kindes ist kürzer als der eines Erwachsenen (vgl. Abbildung 2). Nach dem Quelle-Filter-Modell nach Fant sorgt der Vokaltrakt für die spektrale

Abbildung 2 - Links: MRT-Bild eines Jungen, Alter: 4 Jahre 4 Monate. Rechts: MRT-Bild eines Mannes, Alter: 54 Jahre. Der Vokaltrakt ist jeweils gelb eingezeichnet und hat bei dem Kind eine Länge von 11,28 cm und bei dem erwachsenen Mann eine Länge von 15,87 cm. Grafik aus [8].

Modulation des Anregungssignals, was u. a. zur Ausprägung der Formanten führt. Der kindliche kürzere Vokaltrakt verschiebt die Formantfrequenzen nach oben, sodass sich die als Merkmalstruktur (für die Spracherkennung) oft benutzte spektrale bzw. cepstrale Hülle im Vergleich zur Erwachsenensprache verändert. Durch das Wachstum eines Kindes besteht eine Altersabhängigkeit der Ausbildung dieser spektralen Hülle und der Formanten, vgl. Abbildung 3.

Grundfrequenz Die kleineren Stimmlippen von Kindern führen zu größeren Schwingfrequenzen und somit zu einer höheren Grundfrequenz F_0 (ca. 300 Hz $< F_0 <$ 1000 Hz). Die Grundfrequenz ist eine sprecherspezifische Größe. Es ist deshalb ein Paradigma in der sprecherunabhängigen Spracherkennung, dass die Grundfrequenz sich in den Merkmalen nicht widerspiegelt. Trotzdem ist eine erhöhte Grundfrequenz für die hier angestellte Betrachtung von Bedeutung, denn die durch den Vokaltrakt erzeugte spektrale Hülle wird von den Harmonischen von F_0 nur abgetastet. Bei sehr großen F_0 stehen demnach nur wenige Stützstellen zur Verfügung. Die für die Merkmalbildung benötigte Rückgewinnung der spektralen Hülle oder etwa das Bestimmen von Maxima, wie der Formanten, gestaltet sich dadurch ungenauer, was zu einer Verringerung der Erkennungsleistung führen kann.

3.1.2 Ursachen in den sprachlichen Fähigkeiten

Im Gegensatz zur Erwachsenensprache ist kindliche Sprache noch im Entwicklungsprozess.

Fehler bei der Artikulation von Lauten Während des Sprachlernprozesses entwickeln sich die motorischen Fähigkeiten der Artikulatoren mit dem Alter weiter. Dabei beanspruchen verschiedene Lautgruppen unterschiedlich komplexe Bewegungen der Artikulatoren, sodass der Umfang der erzeugbaren Laute mit dem Alter zunimmt. Dieser Prozess kann in etwa wie folgt schematisiert werden [9]:

- **2 Monate** – erste Lallphase, gesteuert von taktilen Reizen im Mund,
- **4 Monate** – Schmatz- und Zischlaute, Vokallaute sowie erste Silben,
- **6 Monate** – zweite Lallphase, gesteuert vom Hören,
- **18 Monate** – erste Laute, gezielt zur Wortbildung (/p/, /b/, /m/, /n/),
- **24 Monate** – Laute der ersten und zweiten Artikulationszone[1] (/v/, /f/, /t/, /d/),

[1] zu Artikulationszonen vgl. Tabelle 1.

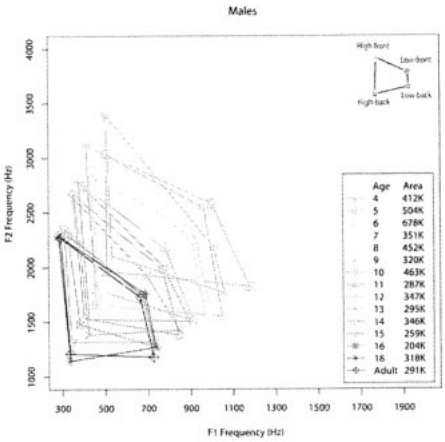

Abbildung 3 - Grafische Darstellung der Altersabhängigkeit der F1-F2 Formantfrequenzen (Durchschnittswerte) für Jungen im Alter von vier Jahren bis zum Erwachsenenalter für die Vokale /i/, /E/, /u/ und /a/ sowie die von F1 und F2 aufgespannte Fläche. Grafik aus [8].

- **30 Monate** – Laute der dritten Artikulationszone (/g/, /k/, /C/, /r/),
- **36 Monate** – erste schwierige Konsonantenverbindungen (/bl/, /kn/, /kr/, /gr/),
- **48 Monate** – die meisten Laute der Muttersprache werden beherrscht, evtl. ausgenommen schwierige Konsonantenverbindungen (/kl/, /gl/, /dr/, /br/) oder Zischlaute (/s/, /ts/, /S/),
- **72 Monate** – Beherrschung aller normgerechten Laute.

Entsprechend wirken sich artikulatorische Fehler in der kindlichen Sprache altersabhängig aus.

Fehler und Variation in der Aussprache von Wörtern Kinder erzeugen oft fehlerhafte Aussprachevarianten sowie eine stärkere Variation der Sprechgeschwindigkeit im Vergleich zur Erwachsenensprache. Die Aussprachevarianten beruhen zum größten Teil auf Sprechfehlern. Die saubere Aussprache von Wörtern nimmt ebenfalls mit dem Alter zu. Parallel zur Entwicklung der Aussprache vollzieht sich die Entwicklung und Erweiterung des Wortschatzes sowie Fähigkeiten des Satzbaus, auf die in diesem Abschnitt nicht weiter eingegangen wird.

3.2 Spracherkennungsexperimente

Aufgrund der im vorigen Abschnitt dargestellten Besonderheiten von Kindersprache ist zu erwarten, dass ein Spracherkennungssystem, das für Erwachsenensprache konzipiert wurde, für Kindersprache eine geringere Erkennungsleistung besitzen wird. Um diese Vermutung zu bestätigen, wurden Erkennungsexperimente durchgeführt. Als Testkorpus wurde eine eigene Sprachdatenbasis generiert, die aus 35 Wortklassen[2] besteht. Die Datenbasis besteht aus den zwei Teilen

- 10 Erwachsene (5 Männer, 5 Frauen) á 3 Realisierungen pro Wortklasse und
- 10 Kinder (5 Jungen, 5 Mädchen) im Alter von $2\frac{1}{2}$ bis 6 Jahren á 3 Realisierungen pro Wortklasse,

die zur Ermittlung der unterschiedlichen Erkennungsleistung von Spracherkennern herangezogen wurden. Zur Spracherkennung wurde mit fünf verschiedenen Systemen gearbeitet:

[2] 35 Wortkassen: Ameise, Apfel, Ball, Bauernhof, Baum, Eichhoernchen, Eimer, Eule, Foermchen, Fuchs, Hahn, hallo, Igel, ja, Katze, Kipper, Kuh, nein, Rutsche, rutschen, Sand, Sandkasten, Schaufel, Schaukel, schaukeln, Schuhe, Sieb, Sonnenblume, Specht, Spinne, Tasche, Trecker, Vogel, Wiese, Wippe.

Tabelle 1 - Artikulationszonen und dazugehörige Laute. Die Tabelle wurde aus [9] entnommen.

Artikulationszone → **Artikulationsart** ↓	**I.** (Lippen, Zähne)	**II.** (Vorderzunge, Gaumen)	**III.** (Hinterzunge, Gaumen)
Reibelaute (Frikative)	*f/v,*	*ts/s, C/j, S/Z*	*r, x*
Verschlusslaute (Plosive)	*p/b*	*t/d*	*k/g*
Nasenlaute (Nasale)		*n*	*N*
sonstige		*l*	

- Erkenner I: DTW[3]-Prinzip, Modelle mit Tracesegmentierung,
- Erkenner II: DTW-Prinzip, Modelle mit Tracesegmentierung und Δ- und ΔΔ-Vektoren sowie einem zweistufigen Klassifikator (grob/fein) [10],
- Erkenner III: phonembasierte HMMs[4], HTK[5]-Training mit HTK-Dekoder,
- Erkenner IV: phonembasierte HMMs, HTK-Training mit eigenentwickeltem Dekoder,
- Erkenner V: phonembasierte HMMs, UASR[6]-System.

Die Experimente haben das Ziel zu ermitteln, wie sich ein Spracherkenner verhält, wenn Erwachsenensprache oder Kindersprache die Trainings- bzw. Testbedingung ist. Der Grund, warum mit einer solchen Anzahl von Erkennern gearbeitet wurde, ist darin zu sehen, dass der Vergleich von zwei Erkennungsraten noch keine wirkliche Aussage enthält, wenn man sie nicht mit anderen Daten in Bezug setzt. Ergibt sich durch die Verwendung von mehreren Erkennern jedoch eine ähnliche Tendenz, kann auf bestimmte Eigenschaften geschlossen werden.

3.2.1 Experimente mit den DTW-Erkennern

Für die DTW-Erkenner gilt, dass pro Klasse eine oder mehrere Realisierungen zum Training benutzt werden. Kommt die Trainingsstichprobe aus dem jeweils anderen Teilkorpus (z. B. Trainingsdaten Erwachsene, Testdaten Kinder), wurde der Mittelwert über die Evaluierungen aller Sprecher gebildet. Zum Training wurde dabei der vollständige Datensatz des jeweils anderen Teilkorpus verwendet, da der zu evaluierende Sprecher in diesem nicht enthalten ist. Kommt die Trainingsstichprobe aus dem Testkorpus (z. B. Trainingsdaten Kinder, Testdaten Kinder), wurde, um eine aussagekräftige Erkennungsrate zu erhalten, das Verfahren der Leave-One-Out-Kreuzvalidierung[7] verwendet. Das bedeutet, dass der Reihe nach jeder Sprecher evaluiert wird. Als Trainingsdaten für den Spracherkenner werden jeweils die restlichen Daten (ohne den zu evaluierenden Sprecher) verwendet. Erkennungsergebnis ist der Mittelwert aller Durchläufe. Tabelle 2 beschreibt die Erkennungsergebnisse für die vier Experimente. Folgende Beobachtungen können dabei gemacht werden (Trainingsdaten/Erkennungsdaten):

- ExpAA: Erwachsene/Erwachsene – die Erkennungsrate ist am größten. Dies entspricht einer herkömmlichen Spracherkennungsanwendung.
- ExpAC: Erwachsene/Kinder – beschreibt die Erkennung von Kindersprache mit einem herkömmlichen System. Die Erkennungsrate bricht drastisch ein, da Trainings- und Testbedingung unterschiedlich sind. Der Erkenner ist nur begrenzt brauchbar.
- ExpCC: Kinder/Kinder – Der Erkenner wird mit ähnlichen Daten trainiert, die er auch im Testfall vorfindet, Trainings- und Testbedingung sind angepasst. Somit steigert sich die Erkennungsrate gegenüber ExpAC. Man sieht jedoch zusätzlich, dass auch bei gleicher Testbedingung die Erkennungsrate deutlich unter der aus ExpAA liegt. Die Autoren

[3] DTW – Dynamic Time Warping [11].

[4] HMM – Hidden Markov Model [11].

[5] Hidden Markov Model Toolkit [12].

[6] UASR – Unified Approach for Speech Synthesis and Recognition [13].

[7] Bei der Leave-One-Out-Kreuzvalidierung (LOOCV, Leave One Out Cross Validation) handelt es sich um ein Testverfahren, bei dem eine Datenmenge, bestehend aus N Instanzen, in N Durchläufen evaluiert wird. Bei jedem Durchlauf wird eine andere Instanz in Bezug auf die restliche Datenmenge evaluiert. Ergebnis ist der Mittelwert aller Durchläufe.

schließen daraus, dass die unterschiedlichen geometrischen Eigenschaften der Vokaltrakte von Erwachsenen und Kindern durch Training mit Kindersprache kompensiert werden können, dass dennoch durch Fehler in der Artikulation die guten Erkennungsraten von ExpAA nicht erreicht werden können.

- ExpCA: Kinder/Erwachsene – Hier wird der umgekehrte Versuch zu ExpAC durchgeführt. Die Ergebnisse sind ähnlich, da Trainings- und Testbedingung unterschiedlich sind.

Tabelle 2 - Erkennungsraten für die Experimente mit den beiden DTW-Erkennern.

Experiment	**RR [%] Erkenner I**	**RR [%] Erkenner II**
ExpAA	78,2	89,8
ExpAC	50,7	63,0
ExpCC	60,5	75,1
ExpCA	47,1	66,9

3.2.2 Experimente mit den HMM-Erkennern

Für das Training von phonembasierten HMM-Erkennern benötigt man große Sprachdatenbasen, um die statistischen Modelle zu formen. Derartig große Datenbasen sind für Kindersprache nicht verfügbar. Deshalb wurden für die HMM Erkenner nur die Experimente ExpAA und ExpAC durchgeführt. Die Ergebnisse der Experimente liegen für alle HMM-Erkenner ähnlich wie für die DTW-Erkenner, abgesehen von geringen prozentualen Unterschieden.

3.3 Datenbasen

Um eine Spracherkennung für Sprach-HMI-Applikation zu erstellen, ist es nötig, zwei Sprachdatenbasen zu generieren:

- Trainingskorpus – bei HMM-Erkennern beinhaltet es große Mengen an Sprache (z. B. > 3 Stunden).
- Testkorpus – beinhaltet exakt die in der Applikation verwendeten Kommandowörter in einer für Testzwecke statistisch relevanten Anzahl (z. B. > 30 pro Klasse).

Wie bereits erwähnt sind, Kinderdatenbasen kaum erhältlich. Die Autoren arbeiten deshalb mit selbst generierten Datenbasen (Testkorpus). Im Gegensatz zum Aufnehmen einer Erwachsenensprachdatenbasis wird das Aufnehmen einer Kindersprachdatenbasis von erheblichen Schwierigkeiten begleitet. Dabei ergeben sich hauptsächlich zwei Herausforderungen. Zum einen sind vor allem kleine Kinder nur für einen recht kurzen Zeitraum von 5 bis 10 Minuten in der Lage, sich auf die Aufgabe zu konzentrieren. Zum anderen ist es oft schwierig, die Kinder dazu zu animieren, genau die Wörter zu sprechen, die für die Aufnahme vorgesehen sind. Das Zeigen und Benennen von Bildern erwies sich als nachteilig. Eine praktikable und letztlich benutzte Methode war das Vorsprechen von Wörtern durch Eltern oder Erzieher, wobei die Kinder das entsprechende Wort nachsprechen sollten. Die Aufnahmen erfolgten meist in gewohnter Umgebung des Kindes (Wohn- bzw. Kinderzimmer). Die Aufnahmen dieser recht geringen Datenmengen haben gezeigt, dass mit vertretbarem Aufwand eine große Trainingsdatenbasis für einen HMM-Erkenner nicht ohne weiteres möglich ist.

4 Schlussfolgerungen

Die Gestaltung von Sprach-HMIs für Kinder ist in allen Bereichen eine Herausforderung, da es sich jeweils um wissenschaftliches Neuland handelt. Insbesondere die Gestaltung der pädagogischen Elemente und des Dialogs können vom aktuellen Wissensstandpunkt noch nicht abgesehen werden. Gleiches gilt für die Erkennung von Kindersprache. Auch hier ist man aktuell

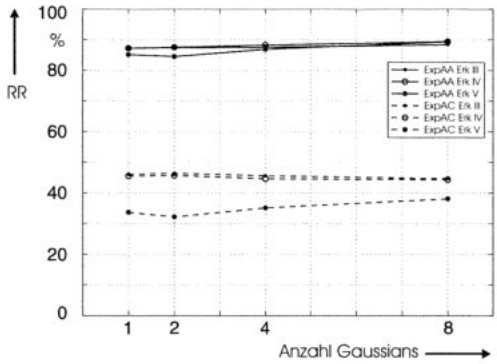

Abbildung 4 - Erkennungsergebnisse der drei HMM-Erkenner für die Experimente ExpAA und ExpAC.

gerade soweit, dass das Problem einer geringen Leistungsfähigkeit von Spracherkennern dargestellt werden kann. Die Erkennungsergebnisse aus Abschnitt 3.2.1 zeigen, dass die Erkennung von Kindersprache mit DTW-Erkennern noch nicht ausreichend gut funktioniert, mit einer intellgenten Dialogstruktur aber im nutzbaren Bereich liegt. Die Benutzung von HMM-Erkennern zeigt aufgrund der nichtvorhandenen Kindersprachdatenbasen für das Training nur geringe Erfolgsaussichten. Da in absehbarer Zeit eine Verfügbarkeit von großen Trainingsdatenbasen für Kinder nicht zu erwarten ist, schlagen die Autoren zunächst die Benutzung von DTW-Erkennern vor. Für diese können mit vertretbarem Aufwand applikatiosspezifische Trainingsdatenbasen generiert werden. Es ist Inhalt aktueller Forschung, die Ansätze für die Erkennung von Kindersprache neu zu überdenken bzw. Lösungsstrategien zu finden.

Literatur

[1] Matthes, L.: Kundenanforderungen und -präferenzen am Produkt LinguBär – Eine qualitative Studie. Bachelorarbeit am SAP-Stiftungslehrstuhl für Entrepreneurship und Innovation, Technische Universität Dresden, 2010, (im Druck).

[2] Geiser, G.: Mensch-Maschine-Kommunikation. Oldenburg, München, 1990, ISBN 3-486-21505-1.

[3] Williams II, R. L.; Chen, M-Y. und Seaton, J. M.: Haptics-Augmented Simple-Machine Educational Tools. Journal of Science Education and Technology, Vol. 12(1), 2003, pp. 1-12.

[4] Schenk, J. und Rigoll, G.: Mensch-Maschine-Kommunikation: Grundlagen von sprach- und bildbasierten Benutzerschnittstellen. Springer, Berlin, 2010, ISBN-10: 3642054560.

[5] Szagun, G.: Sprachentwicklung beim Kind. Weinheim: Beltz, 2000.

[6] Bocklet, T.; Winterholler, C.; Maier, A. ; Schuster, M. und Nöth, E.: An Automatic Screening Test for Preschool Children: Theory and Data Collection. WOCCI2009, Bosten, USA, 2009, CD-ROM.

[7] Jokisch, O.; Hain, H.-U.; Petrick, R. und Hoffmann, R.: Robustness Optimization of a Speech Interface for Child-Directed Embedded Language Tutoring. WOCCI2009, Bosten, USA, 2009, CD-ROM.

[8] Vorperian, H. K. und Kent, R. D.: Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data. J. Speech Lang Hear Res. Author manuscript; available in PCM 2008 December 9.

[9] *http://www.sprachheilberater.de/Sprachentwicklung.htm*, Stand 21. Juli 2010.

[10] Claus, F.: Integrierte Spracherkennung für Kindersprache: Evaluierung phonembasierter Spracherkenner. Diplomarbeit, Fakultät Elektrotechnik, Hochschule für Technik und Wirtschaft Dresden (FH), 2010 (im Druck).

[11] Wolff, M.: Akustische Mustererkennung. Habilitation an der Fakultät Elektrotechnik, Technische Universität Dresden, 2010 (im Druck).

[12] The HTK Book. V3.2., Engineering Department at the Cambridge University, 2002.

[13] Eichner, M.: Sprachsynthese und Spracherkennung mit gemeinsamen Datenbasen, Akustische Analyse und Modellierung. Dissertation, Institut für Akustik und Sprachkommunikation, Technische Universität Dresden, 2007.