ZUM STAND DER TECHNIK IM AUTOMATISCHEN ERKENNEN VON KINDERSPRACHE

Felix Claus, Rico Petrick und Horst-Udo Hain

Projektgruppe Interaktive Sprachlehrmedien Dresden
[felix.claus,rico.petrick,udo.hain]@linguwerk.de

Kurzfassung: Der Artikel gibt einen Überblick über den Stand der Technik im automatischen Erkennen von Kindersprache. Begonnen wird mit einer kurzen Einführung und einer Darstellung verfügbarer Kindersprachdatenbasen. Anschließend werden verschiedene Verfahren zum Erkennen von Kindersprache betrachtet und mit Beispielen unterlegt. Der Fokus richtet sich dabei auf das Erstellen des akustischen Modells, indem die drei Fälle betrachtet werden: Training der Spracherkenner mit Kindersprache, Training mit Erwachsenensprache und anschließende Vokaltraktlängennormierung sowie Training mit Erwachsenensprache und anschließende Adaption an Kindersprache. Danach werden einige Beispiele für Systeme gegeben, in denen Spracherkennung von Kindersprache angewendet wird und als Abschluss folgt eine kurze Zusammenfassung.

1 Einführung

Obwohl Spracherkennung seit mehreren Jahrzehnten umfangreich wissenschaftlich untersucht wird, sind Ansätze zur Erkennung von Kindersprache ein vergleichsweise wenig bearbeitetes Forschungsgebiet. Herkömmliche Ansätze befassen sich vorwiegend mit der Erkennung von Erwachsenensprache, da hierfür zahlreiche Anwendungen existieren, bspw. Sprachwahl am Telefon, maschineller Sprachdialog beim Telefonbanking, Sprachsteuerung eines Navigationsgerätes oder Mobiltelefons (Bsp.: Spracherkennung von Apple für iPhone: Siri).

Kindersprache hat im Vergleich zu Erwachsenensprache unterschiedliche Eigenschaften. Zum einen gibt es anatomische Unterschiede zwischen Erwachsenen und Kindern und zum anderen gibt es Unterschiede aufgrund der sprachlichen Fähigkeiten [1].

1. **Anatomische Ursachen:** Kinder haben einen kürzeren Vokaltrakt sowie kleinere und leichtere Stimmlippen. Das führt zu anderen Eigenschaften des Sprachsignals in Form von höheren Formantfrequenzen sowie einer höheren Grundfrequenz.
2. **Ursachen aufgrund der sprachlichen Fähigkeiten:** Neben semantischen und grammatikalischen Unterschieden zwischen Erwachsenen- und Kindersprache sind phonetische und phonologische Unterschiede zu beobachten. Mit steigendem Alter wächst die Fähigkeit des Kindes, Laute phonetisch richtig und in phonologisch richtiger Abfolge zu produzieren.

Aufgrund der unterschiedlichen Eigenschaften, lässt sich vermuten, dass die Erkennung von Kindersprache mit herkömmlichen Methoden weniger gut funktioniert. Dies wurde durch erste Experimente [2, 3] bestätigt. Es hat sich zusätzlich gezeigt, dass die Erkennungsleistung vom Alter der Kinder abhängig ist, je geringer das Alter umso schwieriger wird die Erkennung [2, 3, 4, 5]. Um Lösungsansätze zu dieser Problematik zu entwickeln, nehmen sich bereits einige Wissenschaftler dieser Thematik an. Dieser Artikel gibt einen Überblick zum Stand der Wissenschaft auf diesem Gebiet. Die existierenden Ansätze werden genannt und in Klassen eingeteilt.

2 Sprachdaten

Eine Schwierigkeit beim Untersuchen, Entwickeln und Testen von Spracherkennungssystemen für Kindersprache besteht in der geringen Verfügbarkeit von Datenbasen mit Kindersprache. Diese Datenbasen werden sowohl zum Training als auch zum Test der Spracherkenner benötigt. Das Erzeugen von Kindersprachdatenbasen ist aufwendiger als das Erzeugen von Erwachsenensprachdatenbasen. Dies gilt insbesondere je jünger die Kinder sind, die aufgenommen werden sollen. In [1] wird eine kleine Sprachdatenbasis von Kindern im Alter von 2 ½ bis 6 Jahren aufgenommen. Dabei wird deutlich, wie schwierig es ist, eine Sprachdatenbasis von Kindern in diesem Alter zu erstellen. Neben der Schwierigkeit, die Kinder dazu zu animieren, die gewünschte Äußerung zu sprechen, ist die relativ kurze Aufmerksamkeitsspanne von nur 5 bis 10 Minuten eine Randbedingung, die die Aufnahmen im Vergleich zu Aufnahmen mit Erwachsenen erheblich erschwert. Daher ist die Anzahl an verfügbaren Kindersprachdatenbasen sehr gering und sie beinhalten hauptsächlich Daten von Kindern im Alter von 6 und 18 Jahren. In [6] wird ein Überblick über verfügbare Kindersprachdatenbasen gegeben. Demnach sind die am häufigsten verwendeten:

- das CID Kinder Korpus (amerikanisches Englisch, 436 Kinder im Alter von 5 bis 17 Jahren) [7],
- das KIDS Korpus [8],
- das CU Kids‘ Audio Speech Corpus (663 Kinder, Kindergarten bis 5. Klasse) [4] und
- das PF-STAR Korpus (britisches Englisch, Italienisch, Deutsch und Schwedisch, 63 Stunden, Kinder im Alter von 4 bis 15 Jahren) [9].

Weiterhin gibt es einige wenige Korpora, bestehend aus Spontansprache, aufgenommen bei der Interaktion zwischen Kind und Maschine [6]. Dazu zählen:

- das NICE Korpus (Kinder im Alter von 8 bis 15 Jahren) [10],
- das FAU-AIBO Korpus (51 Kinder im Alter von 10 bis 13 Jahren, Teil des PF-STAR Korpus) [11],
- sowie weitere, verwendet in [12].

Das Projekt CHILDES (Child Language Data Exchange System) ist ein Teil des TalkBank Systems zum Austausch und für Untersuchungen zur Gesprächsinteraktion. Es enthält Daten von Kindern unter 6 Jahren und besteht aus über 100 verschiedenen Korpora unterschiedlicher Sprachen [13]. Darin sind ebenfalls deutsche Sprachdaten enthalten, wobei oftmals nur die Transkriptionen ohne Audio-Dateien öffentlich verfügbar sind. Weitere Daten von Kindern wurden innerhalb des SpeechHome Projekts [14] oder mittels LENA [15] aufgenommen. Diese Daten sind ebenfalls nicht öffentlich verfügbar.

3 Automatisches Erkennen von Kindersprache

Stand der Technik in der automatischen Spracherkennung (ASR) ist die Modellierung von Spracheinheiten durch Hidden Markov Modelle (HMMs) unter Verwendung von Mel-Frequenz-Kepstral-Koeffizienten (MFCCs) und eines Sprachmodells basierend auf n-Gramm Statistiken. Diese Verfahren sind für das Erkennen von Erwachsenensprache bewährt und meist auch Ausgangspunkt zum Erkennen von Kindersprache [6]. Versuche zum Erkennen von Kindersprache werden bereits in den 90er Jahren des vergangenen Jahrhunderts durchgeführt, Beispiele sind [2, 3, 16, 17]. Seitdem ist die Anzahl der Artikel, die sich mit diesem Thema beschäftigen, stetig gewachsen.

3.1 Akustisches Modell

Zum Erstellen eines akustischen Modells, welches auf das Erkennen von Kindersprache ausgerichtet ist, gibt es eine Vielzahl an Techniken. An dieser Stelle wird ein Überblick über die wichtigsten Ansätze gegeben. Details sind aus [18, 19, 20] zu entnehmen.

3.1.1 Kindersprachdaten als Referenzdaten

ASR Systeme erzielen eine hohe Erkennungsgenauigkeit, wenn die Trainingsdaten und die zu erkennenden Daten möglichst ähnliche Merkmale besitzen. Wie in Abschnitt 1 bereits beschrieben, hat Kindersprache andere Eigenschaften als Erwachsenensprache. Deshalb besteht ein Ansatz zum Erkennen von Kindersprache darin, die Spracherkenner mit Kindersprache zu trainieren.

1996 veröffentlichen Wilpon und Jacobsen eine Studie zur Spracherkennung [2], in der sie verschiedene akustische Modelle für unterschiedliche Altersgruppen erstellen. Dabei kommen Sie zu dem Ergebnis, dass die Erkennungsgenauigkeit einer Altersgruppe am höchsten ist, wenn das akustische Modell derselben Altersgruppe zum Training des Spracherkenners verwendet wird. Weiterhin stellen sie fest, dass die Erkennungsleistung von Kindersprache geringer ist als bei Erwachsenensprache, auch wenn das Modell mit Sprachdaten desselben Alters trainiert wird. Die erzielten Wortfehlerraten (WER) liegen dabei für Erwachsene zwischen 35 und 59 Jahre bei 1,9 % und für Kinder zwischen 8 und 12 Jahren bei 4,7 %.

Diese Ergebnisse werden in späteren Arbeiten zum Erkennen von Kindersprache von Hagen et al. [4] sowie D'Arcy et al. [21] bestätigt, in denen verschiedene akustische Modelle für Kinder unterschiedlicher Altersgruppen erstellt werden. In den Experimenten wird beobachtet, dass die Erkennungsleistung für Kinder einer Altersgruppe am höchsten ist, wenn das Modell mit Daten von Kindern derselben Altersgruppe trainiert wird und dass die Erkennungsleistung mit steigendem Alter der Kinder anwächst. Ein solcher Ansatz hat den Nachteil, dass eine begrenzte Menge an Trainingsmaterial für jede Altersgruppe zur Verfügung steht. Da in den meisten Fällen ohnehin schon wenig Kindersprachdaten verfügbar sind, wird oftmals trotzdem nur ein akustisches Modell für Kindersprache im Allgemeinen generiert [3, 17].

Die These, dass Kindersprache umso schwerer zu erkennen ist, je jünger die Kinder sind, wird in Experimenten zur menschlichen Wahrnehmung von D'Arcy und Russel [22] bestätigt. In diesen erkennen Menschen die Sprache von Kindern ebenfalls schlechter je jünger die Kinder sind.

3.1.2 Erwachsenensprachdaten als Referenzdaten und Vokaltraktlängennormierung (VTLN)

Ein wesentlicher Grund für die akustischen Unterschiede zwischen der Sprache von Erwachsenen und Kindern ist die unterschiedliche Vokaltraktlänge, wodurch sich verschiedene Lagen der Formantfrequenzen ergeben.

Die begrenzte Menge an verfügbaren Kindersprachdaten hat dazu geführt, dass mithilfe von Vokaltraktlängennormierung (VTLN) und Adaptionsmethoden versucht wird, die für das Training im größerem Maße verfügbare Erwachsenensprache an Kindersprache anzunähern, um damit Kindersprache besser zu erkennen.

Bereits in frühen Experimenten [23] wird versucht, die akustischen Unterschiede zwischen Erwachsenen- und Kindersprache mithilfe einer VTLN zu kompensieren. In [3] kann die WER durch VTLN von 15,9 % auf 8,7 % gesenkt werden. Zahlreiche weitere Experimente belegen den Zuwachs der Erkennungsleistung zum Erkennen von Kindersprache bei Training der HMMs mit Erwachsenensprache und VTLN [5, 16, 17, 24, 25].

Neben einfachen Verzerrungsfunktionen mit linearer, stückweise linearer oder bilinearer Verzerrung der Frequenzachse werden auch aufwendigere Ansätze wie bspw. eine phonembezogene VTLN [26] sowie eine VTLN mit einer Matrix anstelle eines einfachen Verzerrungsfaktor untersucht [27]. Details zu Verfahren zur VTLN sowie weiteren Ansätzen finden sich in [28].

3.1.3 Erwachsenensprachdaten als Referenzdaten und Adaption an Kindersprachdaten

Es gibt weitere Unterschiede zwischen der Sprache von Erwachsenen und der Sprache von Kindern als nur die unterschiedliche Lage der Formantfrequenzen aufgrund der kürzeren Vokaltraktlänge der Kinder [6].

Um die Sprache von Erwachsenen noch besser an Kindersprache anzupassen, werden deshalb Verfahren zur Sprecheradaption angewendet. In [5, 24, 29] wird nachgewiesen, dass Sprecheradaptionsverfahren, die für Erwachsene bewährt sind, auch gut geeignet sind, um altersabhängige akustische Modelle zum Erkennen von Kindersprache zu erzeugen. In den Versuchen werden Verfahren wie Maximum Likelihood Linear Regression (MLLR), sprecheradaptives Training (SAT) sowie Adaption mit der Maximum A Posteriori Methode (MAP) durchgeführt, um die Erkennungsleistung von Kindersprache zu steigern. So kann z. B. in [29] die WER nach VTLN durch Adaptionsmethoden weiter von 10,9 % auf 8,0 % gesenkt werden.

3.2 Lexikalisches Modell

Kanonische Aussprachen des phonetischen Lexikons sind nicht geeignet zur Anwendung bei sehr jungen Kindern sowie Kindern mit schlechter Aussprache. Deshalb wird in [30] die Verwendung nutzerspezifischer Aussprachewörterbücher untersucht und festgestellt, dass sich die Erkennungsrate in den Versuchen innerhalb eines gewissen Rahmens steigern lässt. So konnte die Erkennungsrate für ein Kind mit normaler Aussprache von 75,83 % auf 76,89 % und für ein Kind mit schlechter Aussprache von 35,47 % auf 43,92 % gesteigert werden. In [31] wird die Verwendung altersgruppenspezifischer Aussprachewörterbücher untersucht. Dafür wird ein spezielles Aussprachewörterbuch für Kinder im Vorschulalter erstellt, indem manuell Aussprachevarianten der Kinder aus den Trainingsdaten dem Aussprachewörterbuch hinzugefügt werden. Im Vergleich zum Standard-Aussprachewörterbuch können dadurch Steigerungen der Erkennungsrate erzielt werden. Diese Steigerungen verschwinden jedoch, wenn die Modelle mit Sprache von Kindern im Vorschulalter trainiert werden. Die Autoren begründen dies damit, dass in diesem Fall die Besonderheiten der Aussprache bereits aufgrund des Trainings in den akustischen Modellen enthalten sind.

3.3 Sprachmodell

In [12, 17, 31] wird gezeigt, welchen Nutzen es hat, speziell auf Kinder angepasste Sprachmodelle zu verwenden. Dafür werden die Sprachmodelle entweder aus domainspezifischen Texten extrahiert oder es werden Daten von Kindern bei der Benutzung des jeweiligen Systems aufgenommen und daraus das Sprachmodells erstellt. In [12] kann die WER somit relativ um 5 bis 20 % reduziert werden, ausgehend von einer WER von 22 %.

3.4 Kombinationen verschiedener Verfahren

In [29] wird ein System zum Lesen lernen für Schulkinder vorgestellt. Mit diesem System wird gelesene Sprache bei einem Wortschatz mittlerer Größe von 1000 bis 2000 Wörtern erkannt. Durch Anpassung des Sprachmodells und Bearbeitung des akustischen Modells mit VTLN, SAT und MLLR wird die WER des Basissystems von 18 % auf 8 % gesenkt.

Gerosa, Giuliani und Brugnara erzielen beachtliche Ergebnisse mit einem System, das sie in [32, 33] vorstellen. Zum Training des Spracherkenners stehen hier 57 Stunden Erwachsensprache und 9 Stunden Kindersprache (7 bis 13 Jahre) zur Verfügung. Bei einem Vokabular von 64.000 Wörtern und einem Trigramm-Sprachmodell, gilt es gelesene Sprache von Kindern und Erwachsenen zu erkennen. Die WER liegt für das Erkennen von Erwachsenensprache bei Training mit Erwachsenensprache bei 10,1 % und für Kindersprache bei Training mit Kinderdaten bei 13,8 %. In einem weiteren Schritt wird ein allgemeines akustisches Modell mit den Daten der Erwachsenen und der Kinder erstellt. Dieses wird anschließend durch VTLN, SAT, MLLR sowie weitere Verfahren auf die jeweilige Nutzergruppe (Erwachsene/Kinder) angepasst. Die erreichte WER liegt für Erwachsene bei 8,2 % und für Kinder (8 bis 12 Jahre) bei 10,2 %.

3.5 Dialogsysteme

Die Funktionsfähigkeit eines Dialogsystems hängt nicht nur von der reinen Spracherkennung ab. Eine anwendungsspezifische Dialoggestaltung sorgt für eine intuitive Bedienung des Gerätes und kann diese somit erleichtern. Weiterhin kann das Spracherkennungssystem unterstützt werden, indem es zusätzliche Informationen von der Dialogsteuerung erhält wie z. B. eine Auswahl zu erwartender Wörter, wodurch die Spracherkennungsaufgabe vereinfacht wird. Beispiele, konzipiert für Kinder zwischen 6 und 15 Jahren, sind [4, 10, 12]. Ein System für Kinder im Vorschulalter wird in [34] beschrieben.

4 Anwendungen

Im Folgenden werden ausgewählte Systeme zum Erkennen von Kindersprache vorgestellt, die mögliche Anwendungsfelder aufzeigen. Weitere Beispiele sind in [6] zu finden.

4.1 Moderne Medien für die Sprachtherapie

In [35] wird das SPECO System beschrieben. Dieses wird im Rahmen des INCO-Copernicus Programms entwickelt und bietet ein audio-visuelles Aussprachetraining für taubstumme Kinder. Ein weiteres System, konzipiert für hörgeschädigte Kinder, wird in [36] vorgestellt. In diesem kommunizieren hörgeschädigte Kinder mit dem animierten Charakter Baldhi, der ihnen Feedback über die Aussprache der Wörter gibt. In [37] wird das System PEAKS dargestellt, welches sich an Kinder mit Lippen-, Kiefer-, Gaumenspalten richtet und ihnen beim Aussprachetraining helfen soll.

4.2 Elektronische Lehrmittel

Aussprache- und Leseübungen sind auch für Kinder ohne besondere Auffälligkeiten sinnvoll. In diesem Rahmen bewegt sich das Projekt LISTEN [38], welches seit vielen Jahren an der Universität von Colorado entwickelt wird und Aufschluss über die Lesefähigkeiten der Kinder geben soll. Ein weiteres Projekt, das zur Beurteilung der Sprachfähigkeiten eingesetzt wird, ist das TBALL-Projekt [39], welches sich speziell an englischsprechende Kinder aus mexikanischen Familien richtet.

4.3 Elektronisches Spielzeug und Computerspiele

Es gibt einige elektronische Spielzeuge, die in der Lage sind, einen begrenzten Wortschatz zu erkennen [6]: Sonys AIBO, Mattels Diva Petz, MGAs Commando-Bot. In Computerspielen ist Sprachsteuerung bisher meist nur als weitere Modalität zur Menübedienung eingezogen. Ein Beispiel, in welchem Sprachsteuerung als Teil des Spiels angewendet wird, ist NICE [10]. Hier kommuniziert der Nutzer mit Charakteren aus einer Märchenwelt.

5 Zusammenfassung

In diesem Artikel wird ein Überblick zum Stand der Technik im automatischen Erkennen von Kindersprache gegeben. Nach einer kurzen Einführung sowie einem Abriss über die Besonderheiten von Kindersprache und verfügbare Kindersprachdatenbasen werden gängige Methoden zum besseren Erkennen von Kindersprache vorgestellt und anschließend mit ausgewählten Anwendungsbeispielen unterlegt. Es wird festgestellt, dass einige Systeme existieren, in welchen Spracherkennung von Kindersprache angewendet wird. Die meisten Systeme adressieren Kinder zwischen 6 und 18 Jahren. Für jüngere Kinder gibt es wenige Anwendungen. Verfügbare Systeme zum Erkennen von Kindersprache kommen nicht an die Leistungsfähigkeit von Systemen zum Erkennen von Erwachsenensprache heran. Des Weiteren nimmt die Erkennungsleistung der Systeme mit absteigendem Alter der Kinder ebenfalls ab.

Literatur

[1] Matthes, K.; Claus, F.; Hain, H.-U. und Petrick, R.: Herausforderungen an Sprachinterfaces für Kinder. In Proc. of Konferenz für elektronische Sprachsignalverarbeitung, ESSV 2010, Berlin, 2010.

[2] Wilpon, J. und Jacobsen, C.: A Study of Speech Recognition for Children and the Elderly. In Proc. of ICASSP 1996, pages I-349 – 352, Atlanta, GA, 1996.

[3] Potamianos, A.; Narayanan, S. und Lee, S.: Automatic Speech Recognition for Children. In Proc. of Eurospeech 1997, pages 2371 – 2374, Rhodos, Griechenland, 1997.

[4] Hagen, A.; Pellom, B. und Cole, R.: Children's Speech Recognition with Application to Interactive Books and Tutors. In Proc. of IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, pages 186 – 191, St. Thomas, US Virgin Islands, 2003.

[5] Elenius, D. und Blomberg, M.: Adaption and Normalization Experiments in Speech Recognition for 4 to 8 Year old Children. In Proc. of Interspeech 2005, pages 2749 – 2752, Lissabon, Portugal, 2005.

[6] Gerosa, M.; Giuliani, D.; Narayanan, S. und Potamianos, A.: A Review of ASR Technologies for Children's Speech. In Proc. of ICMI-MLMI 2009 Workshop on Child, Computer and Interaction (WOCCI), Cambridge, MA, USA, 2009.

[7] Lee, S.; Potamianos, A. und Narayanan, S.: Acoustics of children's speech: Developmental changes of temporal and spectral parameters. Journal of the Acoustical Society of America, pages 1455 – 1468, 1999.

[8] Eskernazi, M.: A database of children's speech. Journal of the Acoustical Society of America, Vol. 100, No. 4, pages 2759 – 2759, 1996.

[9] Batliner, A.; Blomberg, M.; D'Arcy, S.; Elenius, D.; Giuliani, D.; Gerosa, M.; Hacker, C.; Russel, M.; Steidl, S. und Wong, M.: The PF_STAR Children's Speech Corpus. In Proc. of Interspeech 2005, pages 2761 – 2764, Lissabon, Portugal, 2005.

[10] Bell, L.; Boye, J.; Gustafson, J.; Heldner, M.; Lindström, A. und Wirén, M.: The Swedish NICE Corpus – Spoken dialogues between children and embodied characters in a computer game scenario. In Proc. of Interspeech 2005, pages 2765 – 2768, Lissabon, Portugal, 2005.

[11] Batliner, A.; Hacker, C.; Steidl, S.; Nöth, E.; D'Arcy, S.; Russel, M. und Wong. M.: "You stupid tin box" – children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In Proc. of the 4th Intern. Conf. of Language Resources and Evaluation, Lissabon, Portugal, 2004.

[12] Narayanan, S. und Potamianos, A.: Creating Conversational Interfaces for Children. IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 2, pages 65 – 78, 2002.

[13] MacWhinney, B.: The CHILDES Project: Tools for Analyzing Talk (3rd Ed.). Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

[14] http://www.media.mit.edu/cogmac/projects/hsp.html, Stand 29. Juli 2012.

[15] http://www.lenafoundation.org/, Stand 29. Juli 2012.

[16] Burnett, D. und Fanty, M.: Rapid Unsupervised Adaption to Children's Speech on a Connected-Digit Task. In Proc. of ICSLP 1996, volume 2, pages 1145 – 1148, Philadelphia, PA, 1996.

[17] Das, A.; Nix, D. und Picheny, M.: Improvements in Children's Speech Recognition Performance. In Proc. of ICASSP 1998, pages 433 – 436, Seattle, Washington, 1998.

[18] Gerosa, M.: Acoustic Modeling for Automatic Recognition of Children's Speech. Dissertation, International Doctorate School in Information and Communication Technologies, DIT – University of Trento, Trient, 2006.

[19] Hacker, C.: Automatic Assessment of Children Speech to Support Language Learning. Dissertation, Technische Fakultät der Universität Erlangen – Nürnberg, 2009.

[20] Elenius, D.: Accounting for Individual Speaker Properties in Automatic Speech Recognition. Dissertation, KTH School of Computer Science and Communication, Stockholm, Schweden, 2010.

[21] D'Arcy, S.; Wong, L. und Russel, M.: Recognition of Read and Spontaneous Children's Speech Using Two New Corpora. In Proc. of ICSLP 2004, pages 1473 – 1476, Jeju Island, Korea, 2004.

[22] D'Arcy, S. und Russel, M.: A Comparison of Human and Computer Recognition Accuracy for Children's Speech. In Proc. of Interspeech 2005, pages 2197 – 2200, Lissabon, Portugal, 2005.

[23] Wakita, H.: Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification. IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. ASSP-25, No. 2, 1977.

[24] Gerosa, M.; Giuliani, D. und Brugnara, F.: Acoustic Variability and Automatic Recognition of Children's Speech. Speech Communication 49, pages 847 – 869, 2007.

[25] Jokisch, O.; Hain, H.-U.; Petrick, R. und Hoffmann, R.: Robust Optimization of a Speech Interface for Child-Directed Embedded Language Tutoring. In Proc. of ICMI-MLMI 2009 Workshop on Child, Computer and Interaction (WOCCI), Cambridge, MA, USA, 2009.

[26] Potamianos, A.; Narayanan, S.: Robust Recognition of Children's Speech. IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 6, pages 603 – 616, 2003.

[27] Saito, D.; Matsuura, R.; Asakawa, S; Minematsu, N; Hirose, K.: Directional Dependency of Cepstrum on Vocal Tract Length. In Proc. of ICASSP 2008, 4485 – 4488, Las Vegas, Nevada, USA, 2008.

[28] Molau, S.: Normalization in the Acoustic Feature Space for Improved Speech Recognition. Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, Fakultät für Mathematik, Informatik und Naturwissenschaften, 2003.

[29] Hagen, A.; Pellom, B.; Van Vuuren, S. und Cole, R.: Advances in Children's Speech Recognition within an Interactive Literacy Tutor. In Proc. of HLT/NAACL, pages 25 – 28, Boston, MA, 2004.

[30] Li, Q. und Russel, M.: An Analysis of the Causes of Increased Error Rates in Children's Speech Recognition. In Proc. of ICSLP 2002, pages 2337 – 2340, Denver, CO, 2002.

[31] Cincarek, T.; Shindo, I.; Toda, T.; Saruwatari, H. und Shikano, K.: Development of Preschool Children Subsystem for ASR and Q&A in a Real-Environment Speech-Oriented Guidance Task. In Proc. of Interspeech 2007, Antwerpen, Belgien, pages 1469 – 1472, 2007.

[32] Gerosa, M.; Giuliani, D. und Brugnara, F.: Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech. In Proc. of Interspeech 2005, pages 2193 – 2196, Lissabon, Portugal, 2005.

[33] Gerosa, M.; Giuliani, D. und Brugnara, F.: Towards Age-Independent Acoustic Modeling. Speech Communication, Vol. 51, pages 499 – 509, 2009.

[34] Kannetis, T. und Potamianos, A.: Towards adapting fantasy, curiosity and challenge in multimodal dialogue systems for preschoolers. In Internat. Conf. on Multimodal Interfaces, Cambridge, MA, 2009.

[35] Vicsi, K.; Roach, P.; Öster, A.; Kacic, Z.; Barczikay, P. und Sinka, I.: SPECO A Multimedia Multilingual Teaching and Training System for Speech Handicapped Children. In Proc. of Eurospeech 1999, Budapest, Ungarn, 1999.

[36] Cole, R.; Massaro, D.; Rundle, B.; Shobaki, K.; Wouters, J.; Cohen, M.; Beskow, J.; Stone, P.; Connors, P; Tarachow, A. und Solcher, D.: New Tools for Interactive Speech and Language Training: Using Animated Conversational Agents in the Classrooms of Profoundly Deaf Children. In Proc. of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education, 1999.

[37] Maier, A.; Haderlein, T.; Eysoldt, U.; Rosanowski, F.; Batliner, A.; Schuster, M. und Nöth, E.: PEAKS – A system for the automatic evaluation of voice and speech disorders. Speech Communication Vol. 51, No. 5, pages 425 – 437, 2009.

[38] Beck, J.; Jia, P. und Mostow, J.: Automatically assessing oral reading fluency in a computer tutor that listens. In Technology, Instruction, Cognition and Learning, Vol 1, pages 61 – 81, 2004.

[39] Alwan, A.; Bai, Y.; Black, M.; Casey, L.; Gerosa, M.; Heritage, M.; Iseli, M.; Jones, B.; Kazemzadeh, A.; Lee, S.; Narayanan, S.; Price, P.; Teppermann, J. und Wang, S.: A System for Technology Based Assessment of Language and Literacy in Young Children: the Role of Multiple Information Sources. In Proc. of International Workshop on Multimedia Signal Processing, Chania, Creete, Griechenland, 2007.