A Survey about ASR for Children

Felix Claus[1], *Hamurabi Gamboa Rosales*[2], *Rico Petrick*[3], *Horst-Udo Hain*[3], *Rüdiger Hoffmann*[1]

[1]Dresden University of Technology, Chair for System Theory and Speech Technology,
01062 Dresden, Germany
[2]Autonomous University of Zacatecas, 98000, Zacatecas, Mexico
[3]Linguwerk GmbH, Research & Development, 01069 Dresden, Germany
felix_claus@gmx.de, hamurabigr@uaz.edu.mx, Ruediger.Hoffmann@tu-dresden.de
[rico.petrick,udo.hain]@linguwerk.de

Abstract

This paper is intended to survey the state of the art of automatic speech recognition (ASR) for children's speech. Investigating ASR for children is a current trend in research. Therefore databases of children's speech are needed for training and testing of ASR systems. In the first part of this paper the most relevant databases of children's speech are described. There are less speech data of children available than of adults and speech of preschool children is even more rarely available.

In the second part of this paper the common techniques for recognizing children's speech are summarized. Most investigations about children's ASR focus on the acoustic model. The common methods are described and approaches regarding the lexical and speech model are mentioned subsequently.

In an extensive literature research we collected papers investigating ASR for children. Several studies have been carried out investigating children's ASR. Due to the lack of data from preschool children only a few investigations for this age group have been accomplished. This is illustrated by presenting a statistic on the age of the children in past studies.

Index Terms: children's speech, preschool children's speech, ASR for children, child computer interaction, statistics on children's speech, children's speech corpora

1. Introduction

Most applications using speech interfaces are mainly designed for adults and therefore use automatic speech recognition (ASR) systems for adults' speech. Examples are: speech interfaces of navigation systems, mobile phones or dictation systems for the computer. In recent years systems for children, like reading tutors, tools for foreign language learning or computer games came up. Therefore children's speech recognition became more relevant.

Recognition accuracy of children's speech is usually lower than for adults [1, 2, 3], which is caused by the differences between children's and adults' speech [2]. There exist differences due to anatomical differences and differences in linguistic skills. The shorter vocal tracts of children cause higher formant frequencies and the smaller and lighter vocal folds lead to a higher fundamental frequency. Furthermore, linguistic skills of children are poorer than those of adults. Especially young children are not able to articulate all phonemes correctly. Even if they are able to pronounce single phonemes right they pronounce some words wrong. In [4] D'Arcy and Russel investigated the human perception. They compared the recognition accuracy of ASR systems and human listeners for recognizing children's and adults' speech. The results show that for both listeners, ASR systems and humans, recognition accuracy is worse for recognizing children's speech than for recognizing adults'.

Regardless of the difficulties a huge market for potential applications using speech recognition for children can be assumed. In 2002, Narayanan published a study about creating conversational interfaces for children [5]. For the motivation of his study he pointed out how firm children are in using the computer and that they would like to operate with the computer by a speech interface. 60 % of the U. S. children, aged between four and eleven years, use a computer at home, where only 40 % of the adults does. Recent German studies [6] corroborate these results and show the habits of German children in the usage of computers and mobile phones. 80 % of the children, aged six years and older, have access to a computer and 23 % of the four to five years old children have experiences with computers, too. Children are curious to deal with new technologies and consequently they are potential users for applications using speech recognition.

In order to improve ASR for children databases of children's speech and further research are required. Currently there are less databases of children's speech available than of adults'. One reason is that recording children's speech is more difficult than recording adults' and it becomes even more challenging with decreasing age of the children [7]. Several studies were made investigating ASR for children. Due to available databases most of these studies focus on school children and only a few studies were made for recognizing preschool children's speech.

The paper is structured as following. In the first part we survey the most relevant databases consisting of speech from school children as well as from preschool children. Thereafter current methods for recognizing children's speech are described for the acoustic, the lexical and the speech model. Furthermore a statistic about the age of children in past studies about children's ASR is presented and the trends in research are shown.

2. Databases

Databases are needed for training and testing of ASR systems. There exists less data of children's speech than of adults'. One reason is that recording children's speech is more difficult than recording adults' and it gets even more difficult with decreasing age of the children [7]. An extensive overview about existing databases of children's speech can be found in [8].

2.1. Databases of school children

Most databases consist of speech from children aged between 6 and 18 years. The language of the databases is mainly English, German, Italian, Swedish and Dutch. Databases in other languages are more rarely available. The most relevant corpora are:

- Tball corpus (non-native English from native Spanish, 256 children, aged between 5 and 8 years) [7],
- CID children's speech corpus (American English, read speech, 436 children aged between 5 and 17 years) [9],
- CU Kid's Prompted and Read Speech corpus (American English, read speech, 663 children, aged between 4 and 11 years) [10],
- CU Kid's Read and Summarized Story corpus (American English, spontaneous speech, 326 children, aged between 6 and 11 years) [11],
- CMU Kid's speech corpus (American English, read speech, 76 children, aged between 6 and 11 years) [12],
- OGI Kid's speech corpus (English, read speech, 1100 children, aged between 5 and 15 years) [13],
- PF-STAR corpus (multilingual, including English, German, Swedish and Italian, 491 children, aged between 4 and 15 years, including spontaneous and emotional speech corpus FAU-AIBO) [14],
- ChildIt corpus (Italian, 171 children, aged between 7 and 13 years) [15],
- CHOREC corpus (Dutch, read speech, 400 children, aged between 6 and 12 years) [16] and
- JASMIN-CGN corpus (Dutch, read and spontaneous speech of native and non-native speakers, more than 60 hours of speech from children aged between 7 and 16 years) [17].

2.2. Databases of preschool children

As mentioned above recording children's speech is more difficult than recording adults'. Especially recording preschool children is time-consuming since the children are not able to read. Therefore alternative methods have to be applied to obtain the recordings. Furthermore young children can concentrate for only a short period (5…10 min) [3]. Hence less speech than of older children or adults can be recorded within one recording session. Accordingly, there exists significantly less data of young children than of older ones and in most cases the quality of the data is inferior to those of older children or adults.

Most preschool children's speech data exists in the context of the project CHILDES (Child Language Data Exchange System) [18] from the Carnegie Mellon University. CHILDES is part of the TalkBank system for sharing and studying conversational interactions and consists of data from more than 100 corpora of different languages. The multilingual PHON corpus is one of these corpora. It is meant to be used in order to study the phonological development of children. German data attached to PHON is published in [19]. Data from ten children are included and analyzed in detail. Six of these children are recorded from the 5th to the 36th month and four children are recorded from the 36th month to eight years. More details can be found in [19]. Regrettably, for most of the data from CHILDES only the transcript is publicly available, but without media. More data of young children's speech is recorded with the LENA device [20] or in the context of the SpeechHome project [21]. Unfortunately, these databases are not publicly available, too.

3. Acoustic model

Most work dealing with ASR for children focuses on the acoustic model. In this section existing appendages are described. More details can be found in [15, 22, 23].

3.1. Training with children's speech

In order to obtain the best recognition performance the data used for the training of the speech recognizer should be akin to the data which has to be recognized. Therefore recognizers should be trained with children's speech in order to recognize it. But it depends on the language and the considered age group whether enough data is available to train a hidden markov model (HMM) recognizer.

In 1996, there was a key study made by Wilpon and Jacobsen [1]. They wanted to recognize speech of different age groups. Therefore they created different acoustic models, one per each age group. The recognition performance was always the best, when the acoustic model was trained with speech from the same age group. Additionally they noticed that the recognition performance for children's speech is not as good as for adults, even when the acoustic model was trained with children's speech. The word error rates they achieved were 1.9 % for adults from 35 to 59 years and 4.7 % for children from 8 to 12 years, which is remarkable low compared to other studies.

Work of Hagen et al. [24] as well as work of D'Arcy et al. [25] confirmed the results of Wilpon and Jacobsen. They created different acoustic models for different age groups of children and received the best results when training and testing data were from the same age group. The recognition performance also decreased with decreasing age of the children. The disadvantage of this approach is that in most cases there is not enough children's speech data available to train every age group separately. So often it is used to create one acoustic model for children in general [26, 27].

In [28], an automatic reading tutor system is presented, which is trained with children's speech. Later the system is ported on two hand-held devices [29]. The recognizer used for the system is trained with four children's speech databases: CU Kid's Prompted and Read Speech corpus, CU Kid's Read and Summarized Story corpus, OGI Kid's speech corpus and CMU Kid's speech corpus (see section 2). The test data is a subset of Kid's Read and Summarized Story corpus, which was excluded from the training data. With this approach a WER of 11.45 % was achieved.

3.2. Training with adults' speech and VTLN

If there is not enough children's speech data available to train the recognizer, adults' speech is used and the differences in the positions of the formant frequencies are compensated by vocal tract length normalization (VTLN).

This was already done in early studies in 1977 [30]. Many other studies approve that recognition performance of children's speech after training with adults' speech can be increased by applying VTLN methods [26, 27, 31, 32, 33, 34]. For example in [26] the word error rate can be decreased from 15.9 % to 8.7 %. Often the implementation is simple and a general linear, piecewise linear or bilinear distortion function is used. But also more extensive techniques like a phoneme depended distortion function [35] or the distortion with a distortion matrix instead of a simple scale factor [36] are utilized. Further appendages can be found in [37].

3.3. Training with adults' speech and adaptation to children's speech

Due to the fact that there are more differences between children's and adults' speech than only the positions of the formant frequencies [38] adaptation techniques like maximum likelihood linear regression (MLLR), speaker adaptive training (SAT) or maximum a posteriori adaptation (MAP) are used in order to create age depended acoustic models. Several works [31, 33, 39] show that the recognition accuracy could be increased by applying these methods. For example in [39] the word error rate after using VTLN was 10.9 % and could be further decreased to 8.0 % by using adaptation techniques.

4. Lexical and speech model

4.1. Lexical model

A further appendage to improve the recognition performance of children's speech is pronunciation modeling. Young children and children with a poor pronunciation do not use a canonical pronunciation. In these cases it is subsidiary to adapt the lexical model. In [2] user-depended pronunciation dictionaries are employed. Depending on the pronunciation of the child, the recognition performance could be raised in a small range. For a child estimated to have a good pronunciation the word accuracy could be raised from 75.83 % to 76.89 % and for a child estimated to have a poor pronunciation the word accuracy could be raised from 35.47 % to 43.92 %.

In order to improve the recognition performance for recognizing Japanese preschool children, age-depended pronunciation dictionaries are applied in the Takemaru-kun system [40]. They are created by manually adding pronunciation variants of the children from the training data. For example, the word 'Takemaru' is often pronounced as 'Tachimaru', 'Takebaru' or 'Takemau'. These pronunciation variants are added to the pronunciation dictionary and the recognition accuracy could be increased from 51.2 % to 54.7 %. This increase disappears when using acoustic models of children. The authors constitute that in this case, the specifics of children's speech are modeled within the acoustic model already, so there is no room for further improvement through pronunciation modeling.

4.2. Speech model

The use of specific speech models has also been investigated in several studies [5, 27, 40]. Either the speech model is extracted from domain-specific texts or children are recorded during the use of respective systems, wherefrom the speech model is extracted. In [5] the word error rate could be decreased upon 5 to 20 % relatively, according to a word error rate of 22 %.

5. Literature research

5.1. Statistics on the age

In order to obtain the state of the art in children's ASR we did an extensive literature research. We collected over 1000 papers about children's speech recognition and related topics like VTLN and adaption techniques. For this purpose we collected eligible papers from conference proceedings of Interspeech, ICASSP and further conferences with regard to speech processing. Studies, published in the diverse fields of science like medicine or pedagogics, are not considered. Additionally IEEE Xplore and Google Scholar were browsed for papers. In 100 of these 1000 papers aspects of Children's ASR are investigated. These 100 papers are analyzed and used for the statistic presented in this section.

In figure 1 the ages of the children used in the investigations are shown. Every paper is represented by a vertical line. The length of each line specifies the range of the data. Lines next to each other with the same length and the same covered range indicate that the same data is used for the investigations.

Publications are available for children from three years forward with the exception of [41]. In this study analysis and automatic estimation of children's subglottal resonances, which may benefit children's ASR, are investigated for children from birth on. Most studies deal with speech from children in school age. This is a matter of fact of available databases. Some of these studies use data including also speech from preschool children [41, 42, 43, 44]. However most of the data is from older children, and therefore the studies do not focus on preschool children. To the best of our knowledge only a few studies were made for preschool children only (highlighted in figure 1). These studies are [3, 40, 45, 46, 47, 48]. Already in 1993, Strommen published an investigation about simple ASR experiments on preschool children users aged three years [45]. Unfortunately, further investigations of Strommen do not focus on ASR for preschool children. In [3, 46] children's ASR is investigated using a little German database of children aged between three and six years. In [40] Cincarek et al. describe the development of a module for speech recognition and answer generation for preschool children for a speech-oriented guidance system. For their investigations they use data from the Takemaru database consisting of spontaneous speech from Japanese children. Unfortunately, the ages of the children are not documented exactly. Further research on preschool children's speech is published in [47], where the authors investigated reference marking in children's computer-directed speech using a Wizard of Oz scenario for children from three to six years. In this context data was recorded which was used in [48] for automatic detection of disfluency boundaries in spontaneous young children's speech.

5.2. Trends in research

One trend in research is to analyze speech from preschool children. For example Marklund investigated the phonological complexity and vocabulary size in 30-month-old Swedish children [49]. Further research is done on the recognition of children's emotional speech [50, 51, 52]. Most studies dealing with that use the FAU-AIBO corpus. Additionally research is done according to existing applications like reading tutors, tools for pronunciation training or medical assessment [44, 53, 54, 55]. Since the performance of speech recognizers for recognition of children's speech is still lower than for adults, recognition of children's speech in general is also a current field of research. Recent studies are [23, 56, 57].

6. Conclusions

Due to current applications like reading tutors, tools for foreign language learning or computer games children's ASR became more relevant in recent years. But results for recognizing children's speech are worse than for recognizing adults'. Several studies have been carried out in order to improve children's speech recognition. Most work focuses on the acoustic model. The recognition accuracy for children is higher if the recognizer is trained with children's speech. When not enough child data is available VTLN and adaption methods are applied to increase

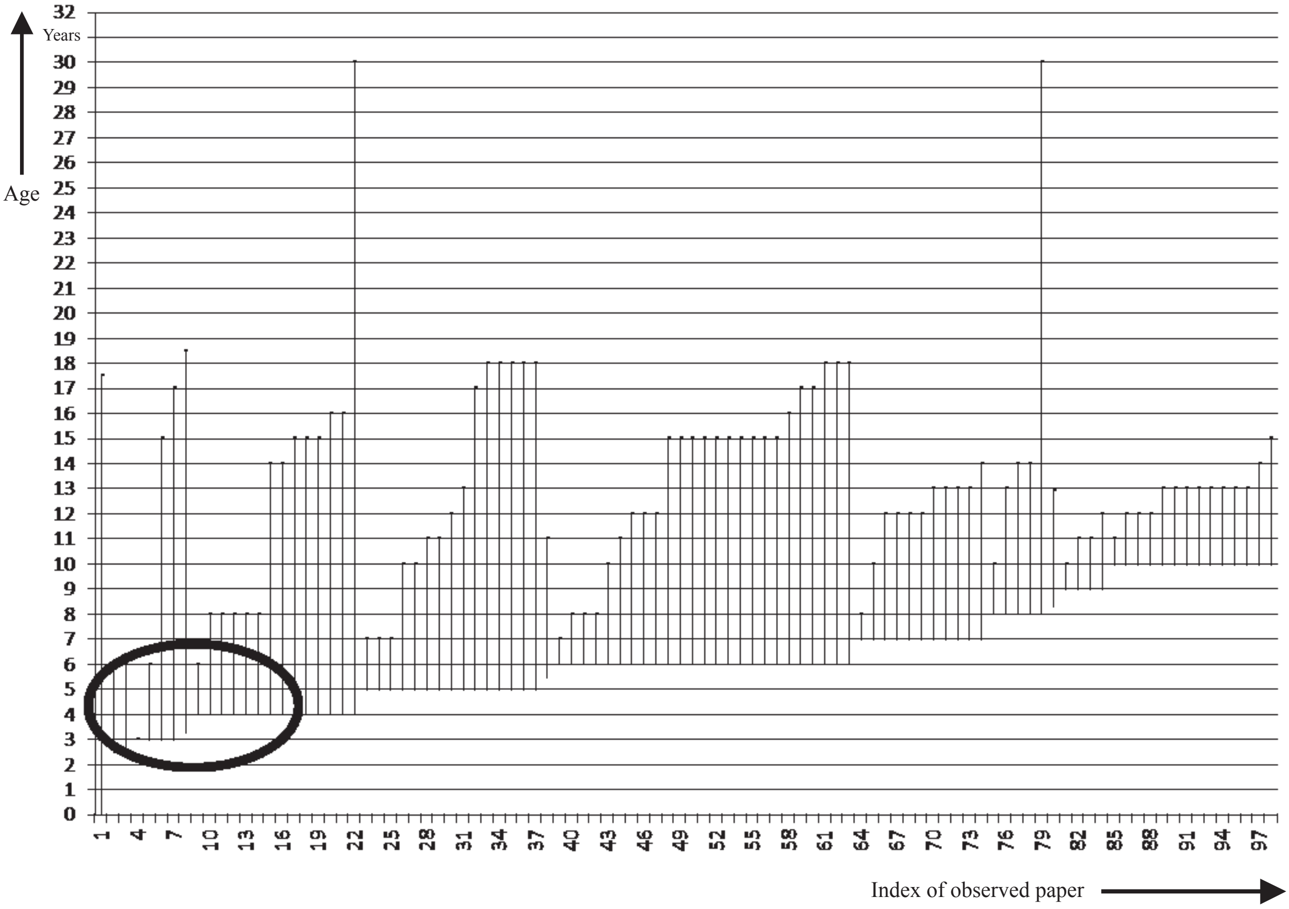

Figure 1: *Considered age group in investigations related to children's ASR.*

the recognition accuracy. Further studies have been carried out on the adaption of the lexical and the speech model. The recognition accuracy could be increased with the described methods but nevertheless results for recognizing children's speech are worse than for recognizing adults'.

Further research is needed in order to improve ASR for children. Therefore databases are required. But databases of children's speech are rare and databases of preschool children are even more rarely available. Our statistic about the age of children in past studies corroborates the lack of young children's speech data (3. . . 6 years) which are eligible for children's ASR. Accordingly, further databases have to be developed in future.

7. References

[1] J.G. Wilpon and C.N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. of ICASSP*, 1996.

[2] Q. Li and M.J. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in *Proc. of ICSLP*, 2002.

[3] K. Matthes, F. Claus, H.-U. Hain, and R. Petrick, "Herausforderungen an Sprachinterfaces für Kinder," in *Proc. of ESSV*, 2010.

[4] S.M. D'Arcy and M.J. Russell, "A comparison of human and computer recognition accuracy for children's speech," in *Proc. of Interspeech*, pp. 2197-2200, 2005.

[5] S.S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65-78, February 2002.

[6] Egmont-mediasolutions, "KidsVerbraucherAnalyse 2012," 2012.

[7] A. Kazemzadeh, H. You, M. Iseli, and B. Jones, "Tball data collection: the making of a young children's speech corpus," in *Proc. of Interspeech*, 2005.

[8] F. Claus, H. Gamboa Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A Survey about Databases of Children's Speech," in *Proc. of Interspeech*, 2013.

[9] S. Lee and A. Potamianos, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455-1468, March 1999.

[10] R. Cole, J.-P. Hosom, and B. Pellom, "University of Colorado Prompted and Read Children's Speech Corpus," *Technical Report TR-CSLR-2006-02*, University of Colorado, 2006.

[11] R. Cole and B. Pellom, "University of Colorado Read and Summarized Stories Corpus," *Technical Report TR-CSLR- 2006-03*, University of Colorado, 2006.

[12] M. Eskenazi, "Kids: a database of children's speech," *Journal of the Acoustical Society of America*, vol. 100, no. 4, 1996.

[13] K. Shobaki, J.-P. Hosom, and R. Cole, "The OGI kids' speech corpus and recognizers," in *Proc. of ICSLP*, 2000.

[14] A. Batliner, M. Blomberg, and S.M. D'Arcy, "The PF-STAR Children's Speech Corpus," in *Proc. of Interspeech*, pp. 2761-2764, 2005.

[15] M. Gerosa, *Acoustic Modeling for Automatic Recognition of Children's Speech*, Ph.D. thesis, University of Trento, 2006.

[16] L. Cleuren, J. Duchateau, P. Ghesquiere, and H. Van Hamme, "Children's Oral Reading Corpus (CHOREC): Description and Assessment of Annotator Agreement," in *Proc. of LREC*, 2008.

[17] C. Cucchiarini, J. Driesen, H. Van hamme, and E. Sanders, ”Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus,” in *Proc. of LREC*, 2008.

[18] B. MacWhinney, ”The CHILDES Project: Tools for Analyzing Talk,” *Lawrence Erlbaum Associates*, 2000.

[19] B. Möbius, ”Ein exemplartheoretisches Modell zum Erwerb der akustischen Korrelate der Betonung,” *DFG-Abschlussbericht*, 2007.

[20] Project LENA, http://www.lenafoundation.org, 2013.

[21] SpeechHome project,
http://www.media.mit.edu/cogmac/projects/hsp.html, 2013.

[22] C. Hacker, *Automatic assessment of children speech to support language learning*, Ph.D. thesis, University of Erlangen-Nuremberg, 2009.

[23] D. Elenius, *Accounting for Individual Speaker Properties in Automatic Speech Recognition*, Ph.D. thesis, KTH Stockholm, 2010.

[24] A. Hagen, B. Pellom, and R. Cole, ”Children’s speech recognition with application to interactive books and tutors,” in *Proc. of Workshop on Automatic Speech Recognition and Understanding*, 2003.

[25] S.M D’Arcy, L.P. Wong, and M.J. Russell, ”Recognition of read and spontaneous children’s speech using two new corpora,” in *Proc. of ICSLP*, 2004.

[26] A. Potamianos, S.S. Narayanan, and S. Lee, ”Automatic speech recognition for children,” in *Proc. of Eurospeech*, 1997.

[27] S. Das, D. Nix, and M. Picheny, ”Improvements in children’s speech recognition performance,” in *Proc. of ICASSP*, 1998.

[28] X. Li, Y.-C. Ju, L. Deng, and A. Acero, ”Efficient and robust language modeling in an automatic children’s reading tutor system,” in *Proc. of ICASSP*, pp. 193-196, 2007.

[29] X. Li, L. Deng, Y.-C. Ju, and A. Acero, ”Automatic children’s reading tutor on hand-held devices,” in *Proc. of Interspeech*, pp. 1733-1736, 2008.

[30] H. Wakita, ”Normalization of vowels by vocal-tract length and its application to vowel identification,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 2, pp. 183–192, April 1977.

[31] D. Elenius and M. Blomberg, ”Adaptation and normalization experiments in speech recognition for 4 to 8 year old children,” in *Proc. of Interspeech*, pp. 2749-2752, 2005.

[32] D.C. Burnett and M. Fanty, ”Rapid unsupervised adaptation to children’s speech on a connected-digit task,” in *Proc. of ICSLP*, pp. 1145-1148, 1996.

[33] M. Gerosa, D. Giuliani, and F. Brugnara, ”Acoustic variability and automatic recognition of children’s speech,” *Speech Communication*, 2007.

[34] O. Jokisch, H.-U. Hain, R. Petrick, and Rüdiger Hoffmann, ”Robustness optimization of a speech interface for child-directed embedded language tutoring,” in *Proc. of Workshop on Child, Computer, and Interaction*, 2009.

[35] A. Potamianos and S.S. Narayanan, ”Robust recognition of children’s speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603-616, November 2003.

[36] D. Saito, R. Matsuura, and S. Asakawa, ”Directional dependency of cepstrum on vocal tract length,” in *Proc. of ICASSP*, pp. 4485–4488, 2008.

[37] S. Molau, *Normalization in the acoustic feature space for improved speech recognition*, Ph.D. thesis, University of Aachen, 2003.

[38] M. Gerosa, D. Giuliani, S.S. Narayanan, and A. Potamianos, ”A review of ASR technologies for children’s speech,” in *Proc. of Workshop on Child, Computer, and Interaction*, 2009.

[39] A. Hagen, B. Pellom, S. Van Vuuren, and R. Cole, ”Advances in children’s speech recognition within an interactive literacy tutor,” in *Proc. of NAACL HLT*, 2004.

[40] T. Cincarek, I. Shindo, T. Toda, H. Saruwatari, and K. Shikano, ”Development of Preschool Children Subsystem for ASR and Q&A in a Real-Environment Speech-Oriented Guidance Task,” in *Proc. of Interspeech*, 2007.

[41] S.M. Lulich, H. Arsikere, J.R. Morton, G.K. Leung, A. Alwan, and M.S. Sommers, ”Analysis and automatic estimation of children’s subglottal resonances,” in *Proc. of Interspeech*, pp. 2817–2820, 2011.

[42] J. Gustafson and K. Sjölander, ”Voice transformations for improving children’s speech recognition in a publicly available dialogue system,” in *Proc. of ICSLP*, 2002.

[43] W.R. Rodríguez and E. Lleida, ”Formant Estimation in Children’s Speech and its application for a Spanish Speech Therapy Tool,” in *Proc. of Workshop on Speech and Language Technology in Education*, 2009.

[44] T. Bocklet, A. Maier, U. Eysholdt, and E. Nöth, ”Improvement of a speech recognizer for standardized medical assessment of children’s speech by integration of prior knowledge”, in *Proc. of Spoken Language Technology Workshop*, pp. 259–264, 2010.

[45] E.F. Strommen and F.S. Frome, ”Talking back to big bird: Preschool users and a simple speech recognition system,” *Educational Technology Research and Development*, vol. 41, no. 1, pp. 5-16, 1993.

[46] F. Claus, *Integrierte Spracherkennung für Kindersprache: Evaluierung phonembasierter Spracherkenner*, diploma thesis, Hochschule für Technik und Wirtschaft Dresden (FH), 2010.

[47] S. Montanari, S. Yildirim, E. Andersen, and S.S. Narayanan, ”Reference Marking in Children’s Computer-Directed Speech: An Integrated Analysis of Discourse and Gestures,” in *Proc. of ICSLP*, 2004.

[48] S. Yildirim and S.S. Narayanan, ”Automatic Detection of Disfluency Boundaries in Spontaneous Speech of Children Using Audio Visual Information,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 2–12, January 2009.

[49] U. Marklund, U. Sundberg, I.-C. Schwarz, and F. Lacerda, ”Phonological complexity and vocabulary size in 30-month-old Swedish children,” in *Proc. of Interspeech*, 2012.

[50] S. Planet and I. Iriondo, ”Spontaneous children’s emotion recognition by categorical classification of acoustic features,” in *Proc. of CISTI*, 2011.

[51] Z. Zhang and B. Schuller, ”Active Learning by Sparse Instance Tracking and Classifier Confidence in Acoustic Emotion Recognition,” in *Proc. of Interspeech*, 2012.

[52] N. Ding, V. Sethu, J. Epps, and E. Ambikairajah, ”Speaker variability in emotion recognition – an adaption based approach,” in *Proc. of ICASSP*, pp. 5101-5104, 2012.

[53] E. Ylmaz, D. Van Compernolle, and H. Van hamme, ”Robust tracking for automatic reading tutors,” in *Proc. of Interspeech*, 2012.

[54] M.P. Black and S.S. Narayanan, ”Improvements in predicting children’s overall reading ability by modeling variability in evaluators subjective judgements,” in *Proc. of ICASSP*, pp. 5069-5072, 2012.

[55] S.-C. Yin, R. Rose, and Y. Tang, ”Verifying session level pronunciation accuracy in a speech therapy application,” in *Proc. of Interspeech*, 2012.

[56] S. Ghai and R. Sinha, ”Analyzing pitch robustness of PMVDR and MFCC features for children’s speech recognition,” in *Proc. of SPCOM*, 2010.

[57] S. Ghai and R. Sinha, ”A Study on the Effect of Pitch on LPCC and PLPC Features for Children’s ASR in Comparison to MFCC,” in *Proc. of Interspeech*, pp. 2589-2592, 2011.