A Survey about Databases of Children's Speech

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

In this paper we survey databases of children's speech. A current trend in research is the investigation of children's automatic speech recognition (ASR). Therefore, databases of children's speech are needed for testing but also for training of ASR systems. However, unlike adult speech corpora, databases for children are rarely available, and in current literature there is no overview of existing databases to be found. Most children's speech databases contain recorded speech in English of children aged between 6 and 18 years. They are described in the first part of this paper. Subsequently databases for German and other languages are mentioned. They are even more rarely available than English databases.

In particular, recordings of preschool children are very rare and therefore regarded separately. Due to the fact that preschool children are not able to read, traditional recording methods cannot be applied, which makes recording of their speech complex. Some ideas covering the difficulties of recordings for speech databases of preschool children are mentioned. Utilizing these methods a small database of German children's speech has been created. Furthermore some statistics about children's speech data are presented.

Index Terms: children's speech, preschool children's speech, children's speech corpora, child computer interaction

1. Introduction

This paper is intended to give an overview about existing databases of children's speech. Databases of children's speech become more and more important, due to the fact, that the young scientific field of children's speech recognition has become more relevant in recent years. Applications, such as speech interfaces of navigation systems, mobile phones or dictation systems for the computer, are mainly designed for adults and therefore use ASR systems for adults' speech. In recent years systems for children, like reading tutors, tools for foreign language learning or computer games have become common. These systems are not as common as systems for adults yet. Nevertheless, they become more relevant and consequently databases of children's speech are required in order to investigate ASR for children.

Several studies prove that recognition accuracy of children's speech is usually lower than for adults [1, 2, 3]. Reasons can be found in the differences between children's and adults' speech [2], which are caused by anatomical differences and differences in linguistic skills. Anatomical differences exist due to the fact that children have shorter vocal tracts (results in higher formant frequencies) and their vocal folds are smaller and lighter (resulting in a higher fundamental frequency). Additionally, children have poorer linguistic skills than adults. Young children are not able to articulate all phonemes correctly. They pronounce some words wrongly, even if they can pronounce the single phonemes correctly. In [4], D'Arcy and Russell accomplish an investigation about the human perception. They compare the recognition accuracy of children's speech by computers (automatic speech recognition (ASR)) and human listeners. Their results show that the recognition accuracy for both listeners, ASR systems and humans, is in general worse for recognizing children's speech than for recognizing adults'.

In order to improve ASR for children, further research but also databases of children's speech are required. Currently, databases of children's speech are significantly less available than of adults' speech. Beyond that, recording of children's speech is much more difficult than recording adults, and it becomes even more difficult with decreasing age of the children [5]. Preschool children are not able to read. Therefore, alternative procedures than reading texts have to be applied. Speech of school children is somewhat easier to acquire since they are able to read. Furthermore, younger children have a very short attention span making recordings rather complicated. Most databases consist of English children's speech, where they are aged between 6 and 18 years. In other languages a few databases exist, too. Data of younger children are even more rarely available due to the difficulties in the recording conditions, and the quality of existing databases is inferior to those of older children or adults. In currently available literature existing databases of children's speech are partially listed [6], but there is no sufficient overview to be found. Therefore, in the present paper, a survey is presented summarizing the current state of the art of available databases of children's speech.

In the first part of the paper we survey the most relevant English and German children's speech databases as well as databases in other languages. Thereafter, we show existing data of preschool children and present a little corpus consisting of speech from German preschool children which was recorded by the authors. Furthermore some summarizing statistics about the range of age of children is presented, whose speech data are applied in published studies on speech processing methods for children's speech.

2. Speech databases of school children

Most children's speech data applicable for speech processing consist of speech from children aged between 6 and 18 years.

Some of the following corpora include a few data of younger children, but basically the data are achieved from children in school age. Most of the corpora we have found consist of English speech and some corpora are available in German, Italian and Swedish. Databases in other languages are more rarely available.

2.1. English children's speech corpora

Following the most applied and most relevant English databases of children's speech are listed:

- CID children's speech corpus (American English, read speech, 436 children aged between 5 and 17 years) [7],
- CMU Kid's speech corpus (American English, read speech, 76 children, aged between 6 and 11 years) [8],
- CU Kid's Prompted and Read Speech corpus (American English, read speech, 663 children, aged between 4 and 11 years) [9],
- CU Kid's Read and Summarized Story corpus (American English, spontaneous speech, 326 children, aged between 6 and 11 years) [10],
- OGI Kid's speech corpus (English, read speech, 1100 children, aged between 5 and 15 years) [11],
- ChIMP corpus (American English, spontaneous speech, 160 children, aged between 8 and 14 years) [12],
- Tball corpus (non-native English from native Spanish, 256 children, aged between 5 and 8 years) [5],
- TIDIGITS corpus (English, 101 children, aged between 6 and 15 years) [13],
- YOUTH corpus (English, 135 children, aged between 8 and 10 years, part of corpus PF-STAR) [14] and
- BIRMINGHAM corpus (British English, 159 children, aged between 4 and 14 years, part of corpus PF-STAR) [14].

Publications about children's speech recognition which are frequently cited were made by Narayanan and Potamianos [15, 16]. For their studies they used data collected over the public switched telephone network:

- DgtII (English, consisting digits, 1234 children, aged between 10 and 17 years),
- DgtIII (English, consisting digits, 501 children, aged between 6 and 17 years),
- SubwII (English, consisting phrases, 1234 children, aged between 10 and 17 years),
- CommI (English, consisting commands, 501 children, aged between 6 and 17 years) and
- CommII (English, consisting commands, 1234 children, aged between 10 and 17 years).

2.2. German children's speech corpora

In this subsection some databases of children's speech in German are listed.

Maier investigated the automatic assessment of speech from children with cleft lip and palate (CLP) [17]. He applied data of children with CLP and as control group data from normal developed children. The data of children with CLP are

- data, recorded in the Oral and Maxillofacial Clinic of the University Hospital of Erlangen (German, 312 children, aged between 4 and 14 years).

For the control group Maier used

- data, recorded at several schools in Erlangen, Nuremberg, Hannover, Karlsruhe and Leipzig (German, 726 children, aged between 5 and 15 years, parts of the data are also included in the German part of the PF-STAR corpus).

In [18], children's speech data was used to interpolate the hidden markov models (HMMs) of an adults' speech recognizer. The corpus consists of data from children reading four different texts: "Nordwind und Sonne" (The North Wind and the Sun) and three texts of the reading test "Züricher Lesetest" [19]:

- Nordwind und Sonne/ Züricher Lesetest corpus (German, 62 children, aged between 10 and 12 years).

Further German databases of children's speech are

- Deutsche Telekom telephone speech corpus (German, prompted and free speech, 106 children, aged between 7 and 14 years) [20],
- VoiceClass Database (Deutsche Telekom Laboratories) (German, free speech, 170 children, aged between 7 and 14 years) [20] and
- FAU-AIBO corpus (British English and German, spontaneous and emotional speech, 81 children, aged between 4 and 14 years, part of corpus PF-STAR) [21].

2.3. Children's speech corpora in further languages

A number of children's speech databases in further languages were published and are listed below:

- 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) [21],
- SpeeCon corpus (multilingual, including Spanish, Russian, Italian, Swedish, German, British English, Danish, Flemish, Hebrew, French, Finnish, Mandarin, Dutch, Japanese, Polish, Portuguese, German from Switzerland, American English, American Spanish and Taiwan, 1000 children (50 per language), aged between 8 and 15 years, part of a database consisting of adults and children's speech) [22],
- ChildIt corpus (Italian, 171 children, aged between 7 and 13 years) [23],
- Tgr-child corpus (Italian, 30 children, aged between 8 and 12 years) [23],
- SponIt corpus (Italian, 21 children, aged between 8 and 12 years) [23],
- NICE corpus (Swedish, 75 children, aged between 8 and 15 years) [24],
- PIXIE corpus (Swedish, 2885 speakers, including many children, ages are not documented, data was recorded during the use of the publicly available spoken dialogue system PIXIE) [25],
- Rafael.0 telephone speech database (Danish, 306 children, aged between 8 and 18 years, was used for the key study about ASR for children made by Wilpon and Jacobsen [1]),
- CHOREC corpus (Dutch, read speech, 400 children, aged between 6 and 12 years) [26],
- SPECO corpus (Hungarian, read speech, 72 children, aged between 5 and 10 years) [27] and
- takemaru-kun corpus (Japanese, 17392 children, ages are not documented, but most children are from lower grades, data were recorded during the use of the public speech-oriented guidance system takemaru-kun) [28].

These are the most relevant databases that may be found in literature. Some more data where acquired while recording in some reading tutors applications, and further, many universities have their own corpora of children's speech, that is not sufficiently described in publications.

3. Speech databases of preschool children

As mentioned above, it is more challenging to record younger children than older ones [5]. Especially data from preschool children are difficult to acquire. Hence, there is less data available. In this section data from children under six years is described.

3.1. Challenges in creating

A lot of studies consider the development of children in the first two years. Therefore, a lot of data has been recorded consisting of speech from children while playing or talking to their mother. With three years of age children pronounce most phonemes correctly and they are able to speak between 800 and 1000 words [29]. The structure of their sentences is simple and they have to learn even more words. Recording specific words from these children is possible yet, but they are not able to read. Therefore, alternative methods to obtain the recordings have to be applied.

A common method which is used by other researchers is: one adult speaks the word first and the child repeats it [30]. This method causes the child to repeat not only the word itself, but also to repeat the emphasis of the adult speaker. In 3.3 a little database of German preschool children's speech is described, which was recorded in our lab. The utilized method is slightly different to the one described above. The children were told to speak the word three times in series. The result is that most children pronounce each word a little bit different, because they emphasize the three words like in a sentence. Another method to get speech from preliterate children is to do a picture naming task, like the German PLAKSS test. This was for example used for the collection of the TBall corpus [5].

Depending on their age, young children are not able to concentrate for a long period, in our experience just for only five to ten minutes [3]. For this reason it is not possible to record the same amount of speech data from a child speaker compared to an adult speaker

Furthermore the quality of recorded data is inferior to those of older children or adults, because of the high variability of different speakers and different recording sessions. Some children slur and often recording conditions are not homogeneous for every child. Most of the recordings take place at the children's home. This guaranties they feel comfortable, which allows them to speak loud and clear.

3.2. Preschool children's speech corpora

Most data of preschool children's speech that could be found exists in the context of the project CHILDES (Child Language Data Exchange System) [31] from the Carnegie Mellon University. This project is part of the TalkBank system, a system for sharing and studying conversational interactions. CHILDES consists of data from more than 100 corpora of different languages. One of these corpora is the multilingual PHON corpus, which is meant to be used in order to study the phonological development of children. In [32], German data is recorded and attached to PHON. The recordings include data from ten children, which are analyzed in detail. Six of the children are recorded from the 5th to the 36th month and four children are recorded from the 36th month to eight years. For more details, refer to [32]. Unfortunately, for most of the data from CHILDES there is only the transcript without media publicly available. More data of young children is recorded in the context of the SpeechHome project [33] or with the LENA device [34]. These data are not publicly available, too.

Furthermore some universities have their own small corpora of young children's speech data, which is not sufficiently described in publications. In summary, there exists significantly less data of young children than of older ones.

3.3. Creation of a small German corpus

Our intention is to investigate speech recognition of German children between three and six years. Due to the lack of data we started to create a small database of German preschool children's speech. Although this data is not enough to train an HMM recognizer it can be used for evaluating the recognition of children's speech running an HMM recognizer.

Our database consists of a child part and an adult part. Ten adults (five men and five women) between 24 and 57 years and ten children (five boys and five girls) between three and six years were recorded. All participants had to speak the same words. Most of the children could concentrate only for five to ten minutes. For that reason only a small vocabulary consisting of 35 commonly used words was recorded. In alphabetical order the German words are: Ameise, Apfel, Ball, Bauernhof, Baum, Eichhörnchen, Eimer, Eule, Förmchen, 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 and Wippe. Each word was recorded three times per speaker.

When young children do not feel comfortable they are too shy to speak loud and clear. For that reason they were recorded in their living rooms. The adults were also recorded in the living rooms and additionally in a studio, an anechoic chamber and in the office. All recordings were performed in 16 bit with a sampling rate of 16 kHz, using a headset microphone. The extension of the database is still in progress.

4. Statistics from literature research

Via an extensive literature review we collected over 1000 papers about children's speech recognition and related topics like vocal tract length normalization (VTLN) and adaption techniques. There are a lot of different studies regarding children's speech in the diverse fields of science like medicine or pedagogics. Papers of these fields are not part of this consideration. To collect eligible papers, conference proceedings of Interspeech, ICASSP and further conferences with regard to speech processing were searched. Additionally IEEE Xplore and Google Scholar was browsed. 200 of the 1000 papers deal only with children's speech, which means they cover information about acoustic and linguistic characteristics of children's speech or children's ASR is investigated.

Figure 1 shows the age of the children in these investigations. It can be seen that most of the studies deal with speech mainly from children in school age. This result reflects the availability of databases. Furthermore the figure points out comprehensively that only a few studies were made for preschool children. Nevertheless, most of these investigations use data covering a big range of age and therefore do not deal with preschool children's speech especially. The studies using data down to the birth of the child are only about features of

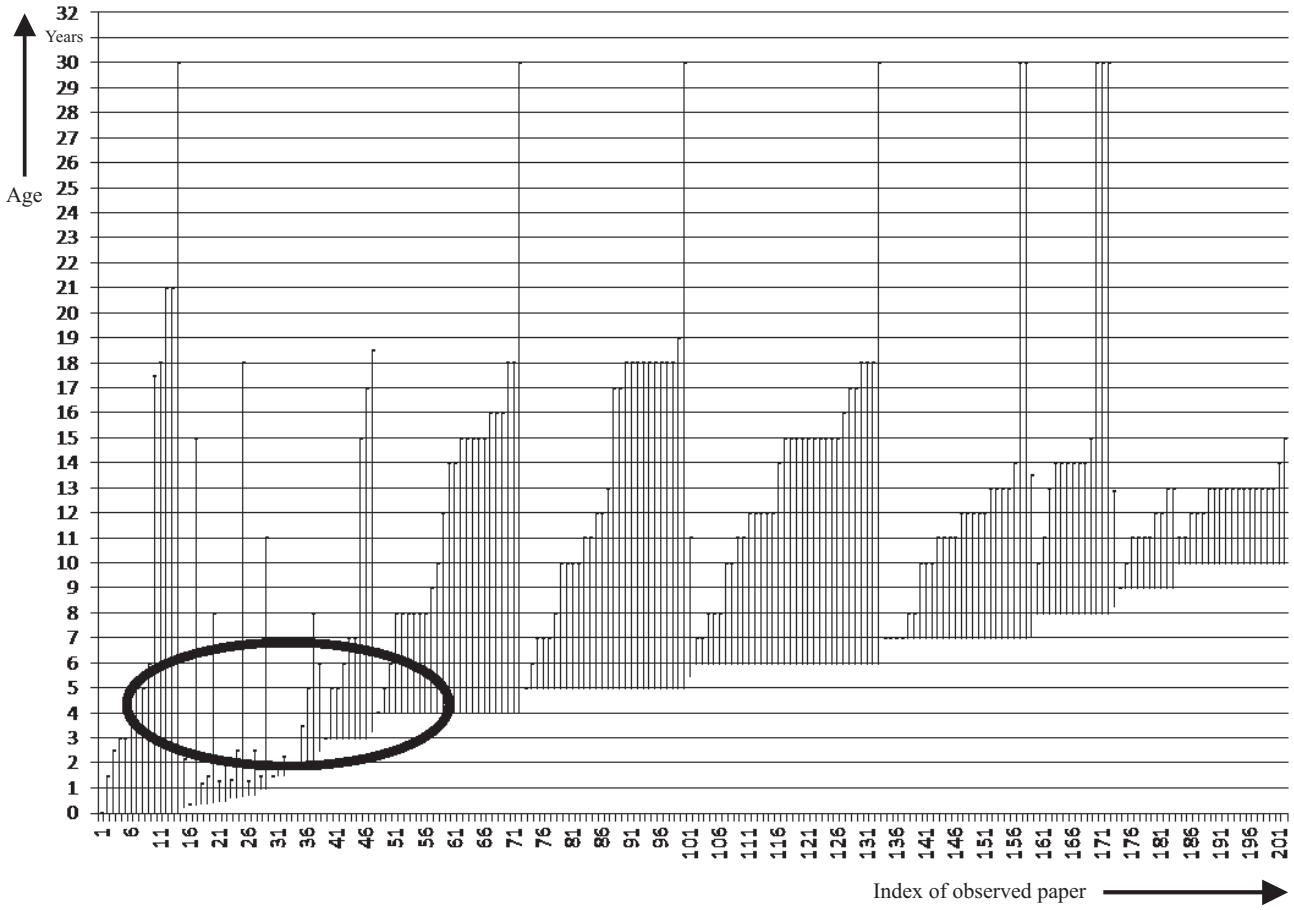

Figure 1: *Age of children in investigations regarding children's speech. On the x-axis the index of the observed paper is illustrated and on the y-axis the age of the children in the studies is shown. Each vertical line represents one paper and the length of the line specifies the range of the data used in the paper. Lines next to each other with the same length and the same covered range indicate that the same data is used for the investigations. The highlighted part of the figure shows investigations using speech of children aged between three and six years. The data in this figure are from the databases listed in the previous sections.*

children's speech, babbling respectively. Investigations about ASR for children do not exist for such young children. Papers using data of children older than four years exist in same amount for both, feature of children's speech as well as ASR for children. As shown in the highlited part of the figure, a limited number of studies (papers) about children aged between three and six years (can speak but not read) were published, which may be related to a limited availability of associated speech data.

5. Conclusions

Current applications like reading tutors, tools for foreign language learning or computer games apply speech recognition methods for children. But recognizing children's speech provides worse results than speech recognition of adults' speech. Further research is required in order to improve existing methods and systems. Databases of children's speech are required in order to train and test ASR systems. Only a limited number of relevant databases of children's speech exist, contrary to adult's speech databases. Furthermore the quality of the data is inferior to those of adults' speech.

Since no overview of existing databases of children's speech can be found in literature, this paper presents a comprehensive survey.

Most of the available databases consist of speech from English speaking children aged between 6 and 18 years. Databases in other languages and data of younger children are much rarer. Especially databases containing preschool children's speech are rare. Recording preschool children is much more complicated than recording older ones, since they cannot read and can concentrate only a short period (5. . . 10 min). Further reasons why recording speech data of young children is difficult are mentioned and a small preschool children's speech database in German is presented, recorded by the authors.

Our statistics about the age of children in existing studies on speech processing for children's speech indicate the lack of speech data from children between three and six years.

In future research on the recognition of children's speech further applicable speech databases are required. Especially data of preschool children is very limited. Recording procedures may be improved or new ones have to be developed.

6. 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] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

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

[12] A. Potamianos and S.S. Narayanan, "Spoken dialog systems for children," in *Proc. of ICASSP*, 1998.

[13] R.G. Leonard, "A database for speaker-independent digit recognition," in *Proc. of ICASSP*, 1984.

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

[15] 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.

[16] 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.

[17] A. Maier, *Speech of children with cleft lip and palate: Automatic assessment*, Ph.D. thesis, University of Erlangen-Nuremberg, 2009.

[18] S. Steidl, G. Stemmer, C. Hacker, E. Nöth, and H. Niemann, "Improving children's speech recognition by HMM interpolation with an adults' speech recognizer," *DAGM Symposium*, 2003.

[19] M. Linder and H. Grissemann, "Züricher Lesetest," Testzentrale Göttingen, 2000.

[20] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," in *Proc. of LREC*, pp. 1562-1565, 2010.

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

[22] D. Iskra, B. Grosskopf, K. Marasek, H. Van Den Heuvel, F. Diehl, and A. Kiessling, "Speecon-speech databases for consumer devices: Database specification and validation," in *Proc. of LREC*, 2002.

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

[24] L. Bell, J. Boye, J. Gustafson, and M. Heldner, "The Swedish NICE Corpus – Spoken dialogues between children and embodied characters in a computer game scenario," in *Proc. of Interspeech*, pp. 1-4, 2005.

[25] L. Bell and J. Gustafson, "Child and adult speaker adaptation during error resolution in a publicly available spoken dialogue system," in *Proc. of Eurospeech*, pp. 613-616, 2003.

[26] 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.

[27] F. Csatari and Z. Bakcsi, "A Hungarian Child Database for Speech Processing Applications," in *Proc. of Eurospeech*, 1999.

[28] 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.

[29] Baby Entwicklungskalender: Kleinkind im 3. Lebensjahr http://www.familie.de/baby-entwicklungskalender/entwicklung-kleinkind-kind-3-jahre/, 2013.

[30] M. Blomberg and D. Elenius, "Collection and recognition of children's speech in the PF-Star project," in *Proc. of Fonetik*, pp. 81-84, 2003.

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

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

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

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