iTalk2Learn
2013-04-30

Deliverable 3.1

Requirements and state-of-the-art of speech production and voice recognition for ITS

30 April 2013

Project acronym: iTalk2Learn

Project full title: Talk, Tutor, Explore, Learn: Intelligent Tutoring and Exploration for Robust Learning

Work Package: **3**

Document title: **Requirements and state-of-the-art of speech production and voice recognition for ITS**

Version: 1.0

Official delivery date: 30 April 2013

Actual publication date: ______________

Type of document: Report

Nature: Restricted to a group specified by the consortium

Authors: Norbert Pfannerer (SAIL) and Gerhard Backfried (SAIL) with contributions from other partners

Approved by: __________________________

Version	**Date**	**Sections Affected**
1.0	30/04/2013	Initial version

Executive Summary

Automatic speech recognition for children is used in a growing number of commercial and academic systems, ranging from entertainment, like toys and games, to educational systems. This deliverable presents an analysis of the state-of-the-art of speech recognition systems applied to tutoring systems for children. There is a reasonable number of such systems in the area of helping children to improve their language skills, like the various reading tutor systems. A consensus among researchers seems to be that speech recognition for children is still a hard task.

The deliverable therefore, reports on the various difficulties and problems in this area as well as different approaches and techniques towards a solution to these problems, especially with respect to their relevancy and applicability to iTalk2learn. Various adaptation and normalization methods have been tried, but it seems that the best way to achieve good results in this area is to use adequate training material for children's voices. Important requirements for this training material are that it should be as close as possible to the target environment of the final system, especially with regards to the children's age and accent.

The exact requirements of voice recognition in terms of expected error rate as well as in terms of required speed and latency depend strongly on the details of the user interface and the planned child-computer interaction scenarios, including specific fall-back solutions to deal with recognition errors and on the overall architecture of the system. Therefore it was decided to extend deliverable 4.1, "Technical report on the first prototype platform" with a section detailing those requirements instead of stating them in the current document.

For speech production the project decided to use off-the-shelf products. As there will consequently not be any research activities associated with this task, but rather only engineering and architectural issues will be addressed, we will likewise report the technical details as part of deliverable 4.1.

Table of Contents

List of Abbreviations

ASR	Automatic speech recognition
AM	Acoustic model
LM	Language model
HMM	Hidden markov model
SAT	Speaker adaptive training
MAP	Maximum a posteriori
MLLR	Maximum likelihood linear regression
VTLN	Vocal tract length normalization
CU	University of Colorado in Boulder
CMU	Carnegie Mellon University
LDC	Linguistic Data Consortium

1. Introduction

In this deliverable we will present an analysis of the state-of-the-art of speech recognition systems with the application to tutoring systems for children, reporting on the various difficulties and problems in this area as well as different approaches and techniques towards a solution to these problems, especially with respect to their relevancy and applicability to iTalk2learn.

Automatic speech recognition (ASR) can be applied in a variety of ways in the context of serving children. Teaching and tutoring systems, gaming, or medical analysis and therapy form just a few of these fields. In the area of tutoring systems, speech recognition can be used to measure and improve children's reading and pronunciation skills. Problem areas such as oral reading miscues can be detected and tracked in the process. By involving children in natural dialogs or by allowing them to interact with animated agents, assessment or training of comprehension in diverse scenarios can be accomplished.

Despite a long history in trying to use automatic speech recognition for children, see for example [1, 2, 3, 8, 10, 16], many problems and open questions still remain to be solved. ASR, in particular models and algorithms in the area of statistical modeling of language which form the core of today's academic as well as commercial systems, has made substantial progress over the last decades. Dictation systems, which had to be trained to a particular speaker's voice and required the speaker to utter words in a word-by-word fashion, pausing in between different words, have long given way to systems allowing users to speak continuously and naturally. Large research-driven efforts were undertaken in the area of ASR to further the state-of-the-art and technologies in the context of security-related application areas. Results of successful developments carried over into such diverse areas as smart-phone navigation or in-car automotive-applications. Meanwhile remarkably little progress was made in the area of children's ASR where many open problems still remain to be resolved.

ASR systems typically employ statistical models for both, the acoustic modeling of speech (AM) as well as for the lexical (syntactic) modeling, so called language models (LMs). Whereas AMs deal with the audio-related aspects of speech, the LM deals with vocabularies and the use of words, typically not via syntactical models but rather through models of co-occurrence and sequencing of lexical items (typically word-based or sub-word-based n-gram models are employed which model the probabilities of sequences of words occurring one after another).

Large progress, both in the area of AM as well as LM has been made, however, the basic stack of technologies and models used has remained largely constant over the recent years (certainly so in commercial systems).

Children's speech and use of language, grammar, vocabulary and sentence-structure, pose a number of problems for existing technologies.

Acoustic and linguistic characteristics of children's speech change rapidly as a function of age and context. Anatomical and morphological differences, e.g. regarding vocal-tract geometry and control of articulators lead to children's speech varying largely from adults speech. Even children of different ages and background themselves vary largely with respect to the above factors. One major problem on the acoustic side is the high spectral and temporal variability in children's speech and higher fundamental and format frequencies. Greater spontaneity and linguistic flexibility form further problematic areas.

ASR systems are typically trained using corpora in order to estimate the parameters of the various statistical models employed. As a result of this process, ASR systems work best when dealing with similar (ideally identical) types of data during runtime. In case the training corpus differs from the data encountered at runtime, a model-mismatch takes place, hurting system performance or rendering system performance completely unusable. The level of performance depends on both, providing adequate technology (algorithms) as well as employing adequate training material (corpus) to avoid such mismatches.

2. ASR Challenges for Children's Speech

Today's ASR systems fail to work reliably on children's data due to one or both of the above factors. Systems trained on adult's voices, in different setups, fail to recognize children's voices properly. A number of mechanisms (such as vocal-tract-normalization) exist to counter this problem, but the basic mismatch remains. LMs trained on *proper language* (complete sentences, terminology,..) likewise fail to capture children's use of language and model it poorly, resulting in inferior performance.

A lot of the relevant literature is devoted to measure the differences and to explore the reasons for the difficulty [1, 4, 5, 6, 7]. Possible reasons for inferior performance of ASR systems on children's speech have been identified and ways to (at least partially) ameliorate these difficulties have been proposed. We summarise them below. The given reasons include on one hand things related to physiological differences, like the shorter length of the vocal tract, other reasons are related to the rate of development of speech acquisition by children. There is even a high variability within children. The recognition accuracy of ASR systems seems to be correlated with the children's pronunciation ability as judged by teachers [5].

2.1 Specific reasons for the difficulty of ASR for children

2.1.1 **Fundamental Frequency and Spectral Variability**

Due to the different geometry of the vocal tract, the fundamental frequency and formant frequencies, as well as the spectral variability within the same phone are different for children than for adults [8, 10]. Even within children the variability is larger than within adults.

The fundamental frequency for male speakers drops significantly from age 11 to 13 and between age 12 to 15, due to pubertal pitch change. For female speakers a significant pitch drop between age 7 and 12 can be observed. For intra-subject pitch variability the effect of age is again significant, with younger children displaying higher variability.

As children grow older, vowel formant frequencies decrease, and in general a compaction of the vowel space can be observed. Consequently acoustic models for adult speech don't fit the spectral composition of children's voices. This effect is much greater at reduced bandwidth, such as telephone speech, because higher formants drop out of the frequency range [4, 5, 7].

The average spectral differences between two repetitions of the same vowel show higher variability for younger children. This variability decreases to adult level at about age 14. The change after age 11 is not significant.

2.1.2 **Vowel durations**

Vowel durations have been found to be significantly longer for children between the age of 5 and 6 years. Durations decrease between the ages of 10 and 15 and then remain largely constant. Gender has been found to exert no significant influence on vowel duration. Variability of vowel durations, intra- and inter-subjects has been found to be significantly higher at younger ages. Duration and intra-subject variability can be expected to reach adult levels at about the age of 12.

2.1.3 **Fricative durations**

Fricatives show similar properties to vowels. The duration of fricatives decreases significantly between the ages of 10 to 12. Intra and inter subject variability in duration also decrease up to age 13 and does not change significantly after that.

2.1.4 **Sentence structure**

Due to the increasing reading ability as well as pause durations, and of course the speaking rate, the average sentence duration decreases almost linearly from age 7 to age 14. After age 14 the sentence duration still varies but does not significantly decrease. The sentence structure can be assumed to be different and less complex than in adult speakers.

The children’s reading ability is obviously more relevant to literacy tutors than in a math tutor. In this setting the spontaneous nature of children’s speech is of more concern. Phrases uttered by children are often ungrammatical, and contain a lot of words that appear out of context. Other disfluencies, like false starts, filled pauses, broken off words, or frequent repetitions of words additionally increase this effect.

2.1.5 **Vocabulary**

The vocabulary (terms) used and grammatical structures vary dramatically and can be expected to be largely different from those of adults. Likewise, pronunciation, mis-pronunciations and fillers used can be expected to yield different behavior and scope.

3. ASR Methods to improve ASR for Children's Speech

3.1 Techniques to improve ASR performance

Several methods to improve ASR performance for children's speech have been investigated and put into action. The improvements are geared towards better acoustic modeling as well as better modeling of the language use of children.

3.1.1 **Acoustic Model**

Possible approaches to deal with these difficulties include various adaptation and normalization techniques on the preprocessing and acoustic modeling side, such as speaker adaptive training (SAT), the maximum a posteriori (MAP) algorithm, maximum likelihood linear regression (MLLR) or vocal tract length normalization (VTLN) and variations [12, 13, 14, 15]. To deal with the problem of the longer duration of certain phones, normalization of the speaking rate has also been applied [15].

These acoustic adaptation techniques all reduce the errors made by the ASR system, but may require additional passes of adaptation. Their use is therefore restricted to cases where online adaptation data for a test subject is available and time is not critical.

Any normalization technique cannot recreate information that has been lost, like higher formants that drop out of the speech signal in situations with a reduced bandwidth, such as telephony [5]. It is therefore of vital importance to make use of the full bandwidth whenever possible.

3.1.2 **Language Model**

Adaptation of n-gram models can be carried out by specifically training n-grams on data corresponding to sentences and vocabulary in the context at hand. Methods such as cache-LM or interpolating different types orders of LMs can be applied. Word- or sub-word-based models can be employed to allow for flexibility in the lexical domain [19], pronunciations can be adapted and hand-crafted to best match a particular student (or group of students).

3.1.3 Other approaches

Various other approaches to deal with the difficulty of children's speech, and with the differences to adults have been explored. Especially in the area of reading tutors, where the relevant disfluencies contain partial words or syllables, it makes sense to try to recognize subword units instead of whole words [1, 19]. In [18] methods to interpolate HMMs from children's speech with those from a recognizer for adults' speech are explored. Other technologies include a flexible architecture of the speech engine, where in a layered structure the first step uses an unconstrained phoneme recognizer and only the next step makes use of the constrained vocabulary of the current task [20].

3.2 Training corpora

The sum total of the above mentioned variabilities indicates why ASR is relatively inaccurate for young children. Even with adaptation and normalization methods a significant performance gap remains. Despite these technologies, it turns out that a system, which was trained on adult speech and adapted to children's voices, still performs poorer than what would be expected from adults.

This shows the necessity for children's corpora to train adequate models and avoid the train/test-model-mismatch. There are already a few efforts that aim at creating such corpora. Some of the mentioned systems created their own database of children's speech, see e.g. [5, 26, 28]. In [11] a few Italian corpora of children speech are mentioned. A database of children's speech while playing video games in Japanese, the CIAIR-VCV Video Game Command Voice can be found in [27].

Of more relevance to the iTalk2Learn project is the PF-STAR project, a EU Framework 5 project, grant number IST-2001-37599, which collected a multi-lingual corpus of children speech, including recordings of Italian, German, Swedish and British English children [15, 22, 23, 24]. The English part consists of 14 hours of read speech from 158 British English children as well as emotional speech of about 1.5 hours from 30 children. The German data consists of recordings of 62 children reading different texts, in a total of about 3.5 hours.

For American English, the CMU Kids corpus is of interest, which is available through LDC (see http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S63). It consists of read speech of 76 children and contains 5180 utterances in total.

The CU Prompted and Read Speech Corpus as well as the CU Read and Summarized Story Corpus were recorded by the University of Colorado at Boulder [21, 29]. The first one consists of speech of 633 children from kindergarten through 5th grade of isolated words, sentences and spontaneous speech. The second database contains data from 106 children reading and summarizing stories.

We think that for the iTalk2Learn project the CMU Kids corpus can be of help. There is obviously a mismatch of accents, but even getting the pitch and formant frequencies of children can be useful. Even more relevant to the project is the PF-STAR corpus. We already bought the CMU Kids corpus and plan to acquire the PF-STAR data. In addition due to our constant contact with CMU we will try to get access to any data that might be available for research purposes there.

4. Existing systems and commercial products

An increasing number of commercial and academic systems make use of automatic speech recognition for children. Some of them are products for entertainment, such as toys and games [26, 27]. More relevant to the iTalk2Learn project however, is the use of speech technology in the area of education.

A lot of effort has been spent in the area of literacy tutors [1, 19, 20, 21, 25]. Generally speaking, these systems aim at helping children to improve their language skills, e.g. reading, or help with learning correct pronunciations. Various techniques are employed, like tracking the reading, or taking attention to mis-pronunciations. According to these observations, the systems then can give various kinds of feedback, hints, or trigger other pedagogically appropriate interventions. There are a lot of differences between the various systems of this kind in relation to user interfaces, child-computer interactions, pedagogical tools, etc. The challenges and techniques that are relevant for ASR have been given in detail in chapters 2 and 3 of this document.

Academic projects in this area include MIT's Literacy Tutor, the STAR project, which is a pronunciation tutor, and, more significantly, the LISTEN project at CMU, which is a reading tutor [25]. Additional systems include the University of Colorado's Foundation to Literacy Project [1, 21, 29], and the University of Pittsburgh's ITSPOKE. The LISTEN and ITSPOKE projects are already referenced in the Description of Work.

The LISTEN system uses CMU's Sphinx recognizer. It was trained with a speech recorded from children in elementary schools in Pittsburgh. The language model is based on the words of the current sentence. Confidence measures for the correctness of the recognized words are employed to enhance the reliability of the reading tracking.

CU's Foundation to Literacy Project uses speech recognition for reading tracking as well as for summaries of stories. In addition to the speech recognition technology it makes use of semantic analysis to gather information about the spoken summaries [1]. In [1] the author reports on experiments using n-gram models that are sensitive to the reading position in order to better predict the next word of a read story. Using these models together with better modeling of the context at beginning and end of sentences as well as across pauses, the author reports up to 34.8% relative reduction in error rate. He also employs LMs based on sub-word units, and reports on various experiments using different algorithms for the creation of the sub-word units.

Additionally [1] mentions two commercial products, IBM's Watch-me!-Read and Soliloquy's Reading Assistant. Both are intended to help children to improve their reading ability. These products automatically track the reading position and give feedback and help. [1] already mentions that not much technical information is available about these products. Meanwhile it seems as if those products are no longer available, at lease the listed webpages can no longer be reached.

Software in the educational domain for topics like mathematics and science has existed at least since the late 1990s. A suite of such software was developed by Henry Gray at the Sothern Methodist University in Dallas [30, 31, 32, 33]. The programs are based on dictation software by Dragon. With the exception of a small part, they are not targeted at children, but at adults, e.g. university students, and at persons with disabilities, helping them to input mathematical expressions. No detailed technical information seems to be available.

Another software in the area of mathematics education is a system called Mathifier [34]. The target users are not children, but students and professionals. It uses CMU's speech recognizer sphinx to recognize spoken mathematical formulas and translates them to LaTeX.

Most of the systems deal with speech recognition for the English language, American English as well as British English, and there are a few other languages, like Italian [2, 11, 26] or Dutch [20]. Research in German include the studies cited in [15] and [28].

Recent developments focus on integrated systems, where automatic speech recognition is only a small part, and use standard training and adaptation techniques for the speech recognition part, often building the speech recognition models using a database of children voices. One such project is ALIZ-E, an EU-FP7 project [26] It is a conversational system of child-robot interaction in Italian.

In other areas of research, that are of less relevance to the project at hand, speech technology is used in the context of assessing the speech of children with certain disabilities or malformations, e.g. cleft lip and palate [28].

5. Conclusions

Automatic speech recognition for children's voices is still a hard task. A number of promising adaptation and normalization techniques, as well as structural, architectural and algorithmical enhancements of speech recognizers exist, that can gain substantial improvements in error rate. Despite that, the performance of speech recognizers on children speech that are trained on adult speech does not reach adult levels.

It is therefore of importance to use training material that match as closely as possible the target environment of the final system. It is necessary to record children's speech that has the same or similar characteristics with respect to acoustic environment and children's accent and age as will be encountered in the final setting. To be able to make use of various adaptation techniques it is necessary to record the voices using a high bandwidth and without any lossy compression.

For the training of the language model, data is needed with the same sentence structure as is expected to be used in the final target system. This is not as critical as the acoustic training material, because it is text only and can therefore be created when the final usage scenarios are known, without the need to record children in classrooms. Additionally Sail's speech recognizer contains a toolset that allows to re-train the language model within hours, and it is therefore possible to make use of additional data very late, even after the final system is ready.

Although the various adaptation techniques don't provide the same results for children as for adults, these methods still yield substantial improvements and should therefore be taken into consideration. As time permits we will strive for implementing some of these enhancements into Sail's speech recognizer.

References

[1] Andreas Hagen. Advances in children's speech recognition with application to interactive literacy tutors. A thesis submitted to the Faculty of the Graduate School of the University of Colorado in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Department of Computer Science, 2006.

[2] Subrata Das, Don Nix and Michael Picheny. Improvements in children's speech recognition performance. In Proc. ICASSP, pages 433-436, May 1998.

[3] Martin Russell, "Speech technologies for children", IEEE Speech and Language Processing. Technical Committee Newsletter, July 2010 http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2010-07/speech-tech-for-children/

[4] Qun Li and Martin Russell. Why is automatic recognition of children's speech difficult? In Proc. EUROSPEECH, pages 2671-2674, 2001.

[5] Qun Li and Martin Russell. An analysis of the causes of increased error rates in children's speech recognition. In Proc. ICSLP, Denver, Colorado 2002.

[6] Martin Russell and Shona D'Arcy. Challenges for computer recognition of children's speech. In Proceedings of SLaTE-2007, pages 108-111.

[7] M J Russell, S. D'Arcy and Li Qun, The effects of bandwidth reduction on human and computer recognition of children's speech, IEEE Signal Processing Letters, Vol. 14, Number 12, pp 1044-1046, December 2007

[8] S. Narayanan and A. Potamianos. Creating conversational interfaces for children. IEEE Trans. Speech and Audio Proc., Vol. 10, No. 2., 65-78, 2002.

[9] Daniel Elenius and Mats Blomberg. Comparing speech recognition for adults and children. In Proc. FONETIC 2004, pages 156-159.

[10] A Potamianos, S Narayanan, S Lee. Automatic speech recognition for children. In Proc. EUROSPEECH 1997.

[11] Matteo Gerosa, Diego Giuliani, Fabio Brugnara. Acoustic variability and automatic recognition of children's speech. In Speech Communication, volume 49, issues 10-11, pages 847-860, October-November 2007.

[12] Gauvain J. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing, vol. 2, no. 2, pp. 291–298, April 1994.

[13] Leggetter C. and Woodland P. Maximum likelihood linear regression for speaker adaptation of

continuous density hidden Markov models. Computer Speech and Language, vol. 9, no. 2, pp. 171–185, April 1995.

[14] L. Lee, R C Rose. Speaker normalization using efficient frequency warping procedures. In Proc. ICASSP 1996, pp. 353–356, May, 1996.

[15] Stemmer, G., Hacker, C., Steidl, S., and Nöth, E. Acoustic normalization of children's speech. In Proc. European Conf. on Speech Communication and Technology (Vol. 2, pp. 1313-1316). September 2003.

[16] Potamianos, A. Robust recognition of children's speech. IEEE Transactions on Speech and Audio Processing, 11(5), pages 603-616, November 2003.

[17] Daniel Elenius and Mats Blomberg. Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year old Children. In Proc. INTERSPEECH 2005.

[18] Stefan Steidl, Georg Stemmer, Christian Hacker, Elmar Nöth, and Heinrich Niemann. Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer. Lecture Notes in Computer Science, volume 2781, pages 600-607, 2003.

[19] Andreas Hagen, Bryan Pellom and Ronald Cole. Highly accurate children's speech recognition for interactive reading tutors using subword units. Speech Communications 49, pages 861-873, 2007.

[20] Jacques Duchateau, Mari Wigham, Kris Demuynck and Hugo Van hamme. A flexible recogniser architecture in a reading tutor for children.In Proc. ITRW on Speech Recognition and Intrinsic Variation, pages 59-64, 2006.

[21] Andreas Hagen, Bryan Pellom, and Ronald Cole. Children's speech recognition with application to interactive books and tutors. In IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, St. Thomas, USA, December 2003.

[22] Martin Russell. The PF-STAR British English Children's Speech Corpus. http://www.thespeechark.com/pf-star-page.html 2006.

[23] Shona M D'Arcy, Lit Ping Wong, Martin J Russell. Recognition of Read and Spontaneous Children's Speech Using Two New Corpora. In Proc. ICSLP, 2004.

[24] Mats Blomberg and Daniel Elenius. Collection and recognition of children's speech in the PF-Star project. PHONUM 9, pages 81-84, 2003.

[25] Project LISTEN. http://www.cs.cmu.edu/~listen/

[26] Ivana Kruijff-Korbayová, Heriberto Cuayáhuitl, Bernd Kiefer, Marc Schröder, Piero Cosi, Giulio Paci, Giacomo Sommavilla, Fabio Tesser, Hichem Sahli, Georgios Athanasopoulos, Weiyi Wang, Valentin Enescu and Werner Verhelst. Spoken Language Processing in a Conversational System for

Child-Robot Interaction. In Workshop on Child, Computer and Interaction, 2012.

[27] Center for Integrated Acoustic Information Research, http://db.ciair.coe.nagoya-u.ac.jp/eng/dbciair/dbciair2/kodomo.htm

[28] Andreas Maier, Elmar Nöth and Anton Batliner. Fully Automatic Assessment of Speech of Children with Cleft Lip and Palate. Informatica 30, pages 477-482, 2006.

[29] Andreas Hagen, Bryan Pellom, Sarel Van Vuuren, and Ronald Cole. Advances in children's speech recognition within an interactive literacy tutor. In Proc. of HLT-NAACL 2004, short papers pages 25-28, 2004

[30] http://www.mathtalk.com/index.htm

[31] Catherine Greenman. New help for disabled math students. The New York Times, 30 March 2000, http://www.nytimes.com/2000/03/30/technology/new-help-for-disabled-math-students.html

[32] Dr. Henry L. Gray, Charles F. Frensley and Nancilu McClellan. Speech recognition + voiced mathematics = accessability. In Proc. CSUN - Technology and persons with disabilities conference 1999, http://www.csun.edu/cod/conf/1999/proceedings/session0224.htm

[33] Nancilu McClellan, Dr. Henry L. Gray and Charles F. Frensley. Speech recognition + voiced mathematics = opportunity. In Proc. CSUN - Technology and persons with disabilities conference 2000, http://www.csun.edu/cod/conf/2000/proceedings/0150McClellan.htm

[34] Batlouni, S.N. Karaki, H.S.; Zaraket, F.A.; Karameh, F.N. Mathifier – speech recognition of math equations. In Proc. International Conference on Electronics, Circuits and Systems (ICECS), pages 301-304, 2011.