Towards an Embedded Language Tutoring System for Children

Oliver Jokisch
Dresden University of Technology
Laboratory of Acoustics and Speech Communication
D-01062 Dresden, Germany
oliver.jokisch@tu-dresden.de

Ruediger Hoffmann
Dresden University of Technology
Laboratory of Acoustics and Speech Communication
D-01062 Dresden, Germany
ruediger.hoffmann@tu-dresden.de

ABSTRACT

Computer-aided language learning (CALL) applications for adult learners are well-established. Tutoring applications for children require special didactic and ergonomic concepts. Except for low functionality (toy-like) products for children and PDA-based solutions, embedded systems seem to be not present in the market yet. The authors currently adapt baseline tutoring technology for adult learners to the special requirements of young children aged 3+ years and they therefore shortly discusses other preliminary studies and the own tutoring system AzAR. The paper introduces didactic and technological concepts of the embedded target system for young children. A German language prototype is still under development. The speech technology and audio components were already tested in comparable application contexts regarding dialogue control, robust recognition and hardware design.

1. INTRODUCTION

The sufficient infantile language tutoring plays an increasing role within the member states of European Community assuming different aspects, such as:

- The native language deficits e. g. in near-illiterate social classes which affect later school education, see Programme for Intern. Student Assessment (PISA),
- The integration constraints of migrant families and
- The earlier children begin learning foreign languages, the better their progress tends to be.

1.1 Computer Aided Language Learning

Computer-aided language learning (CALL) methods for adult learners have been established during the last two decades. Prevalent language tutoring applications are PC-based audio-visual language courses which focus on vocabulary acquisition and pronunciation training. Advanced systems like Pronunciation Power, American Sounds or Eyespeak (see also [1]) include:

- Speech analyzing windows,
- Internet-based features like email answering, online help and chat sessions with human tutors,
- Animated views of the articulatory mechanics, video clips showing jaw, lip and tongue movement and
- Waveform patterns of sound samples.

The users are able to record sound files and to compare acoustic and graphic representations of their utterances with the reference sounds of the instructor. A few systems, such as Fonix iSpeak and ProNunciation, include synthesized speech or TTS solutions [2]. Furthermore, speech recognition technology was implemented into innovative interactive systems like ISTRA and PRONTO [3] and within the project Interactive Spoken Language Education (ISLE [4]). Considering the advances of computer-assisted language tutoring systems for adults, applications and underlying technology for (young) children are underdeveloped.

1.2 Tutoring Systems for Children

According to [5], the intensive practice in reading aloud helps students to establish the conventional association between sounds and their written forms, a skill that requires years of practice in young children and students of languages with non-phonetic writing, such as Japanese or Chinese. Ehsani and Knodt [5] argued that teaching children how to read their own native or even a foreign language is thus an area where speech recognition technology can make a significant difference. The concept behind is a reading tutor that not only listens to children and students reading aloud a story presented on the screen, but intervenes when needed and to correct mistakes.

In [5], they further examined voice-interactive tutoring approaches in the early 1990s and concluded that designing a basic recognition network for a voice-interactive reading tutor is relatively straightforward. There is only one correct spoken response to any given written prompt, and the system “knows” in advance what the student will be trying to say. However, the technical challenge is to recognize and respond adequately to the dysfluencies of inexperienced readers. Such dysfluencies include hesitations, mispronunciations, false starts, and self-corrections.

One of the first prototype systems for teaching reading to young children was developed by Kantrov at the Center for Teaching and Learning (CTL) already in 1991 [6]. This simple multimedia application used an isolated, speaker-dependent recognizer and limited reading vocabulary (18+ words). The system was designed to expand children's reading vocabulary by embedding new words within the context of a goal-oriented game: children are called upon to help a bear overcome obstacles on his way home; reading the word correctly removes the obstacle. Results of field trials in public schools indicated that the problems with the application were related to the human interface and input mode (microphones), rather than the speech recognition component per se. The application was described as robust and ironically,

recognition errors, especially misrecognition of correctly read words, contributed positively to the pedagogical effect of the application: the children got additional reading practice, because they had to repeat the words several times until the machine responded appropriately.

In the following time, an ambitious automated reading trainer was developed within the LISTEN project at Carnegie Mellon University (CMU). Designed to combat illiteracy, the fully automated prototype by Mostow et al. [7] used continuous speech recognition to listen to children read continuous text and automatically trigger pedagogically appropriate interventions. As summarized in [5], the system featured the personalized agent "Emily," who provided feedback and assistance when necessary. The system incorporated expert knowledge on individual reading assistance that was both pedagogically relevant and technically feasible. Emily intervened when the child misread one or more words in the current sentence, got stuck, or clicked on a word to get help. To reduce frustration in children with reading difficulties, the system deliberately refrained from treating false starts, self-corrections, or hesitations as "mistakes." Instead, errors of this type were modeled and included into the recognition grammar as acceptable. An experimental trial of the system was conducted among 12 second graders at an urban school. Results showed that the children could read at a reading level 0.6 years more advanced when using the automated reading coach, and the average number of reading mistakes dropped down from 12.3 % (without assistance) to 2.6% (with assistance) in texts with similar difficulty.

An improved version of CMU's reading coach running real-time on an affordable PC platform was fielded in 1996 among poor third grade readers of American English to measure improvements in reading performance over an eight month period of using the system [8]. While earlier studies measured reading performance only in terms of student word error rates, the improved system implemented algorithms for measuring reading fluency in young children. Relevant performance variables include reading rate, inter-word latency (silence), dysfluency (false starts, self-corrections, omissions) and time spent with the assistant. Comparing subjects' reading fluency levels at the beginning of using the system with those at the end, the experiments showed an overall improvement in reading accuracy of 16% and a 35% decrease in inter-word latency. After using the system for only eight months, the student' reading levels improved by an average of two years.

These early results in CALL were encouraging in that they show how careful system design and evaluation based on user data can lead to useful and practical applications.

Recently, interactive CALL systems and the underlying methodology for children have made considerable progress. Silverman developed a multi-dimensional vocabulary program incorporating storybook reading and opportunities to say vocabulary words aloud in kindergarten classrooms [9].

Analyzing rates of target word acquisition and overall vocabulary development, the study of Silverman found that students learning English as a second language picked up general vocabulary more quickly and target vocabulary words at the same rate as native English-speaking kindergarteners. Other studies deal with animated computer tutors for remedial readers and autistic or hearing-impaired children. Massaro et al.'s facial animation software features a realistic tongue and palate that students can access in dynamic side view cutaways of tongue, jaw and teeth [10]. The majority of interactive tutoring methods focus on PC-based or online-based tools involving multimodal channels and on children of 7+ years. On the other hand, plenty of electronic toys with speech technology extensions occur in the market, which provide low functionality, e. g. playing back simple statements while a touch sensor is activated. Embedded speech recognizer (ASR) algorithms usually operate near to a random process.

The authors develop a concept for adapting language tutoring technology which has been formerly directed to adult learners to the special requirements of young children, starting at 3 years. This paper focuses on the technological parameters of a standalone embedded tutoring system but also discussing some didactic and ergonomic aspects. The children-directed prototype is not completed or evaluated in a real user scenario yet. The speech technology and audio components were already tested within other tutoring contexts, and the presented concept was discussed and confirmed by kindergarten pedagogues.

2. OBJECTIVE

Toys involving sound or speech feedback are widely spread but a serious language tutoring application for children requires a sophisticated approach, mainly because:

a) The common PC-oriented graphical or textual user interfaces are usually not operated by infants,

b) A playful achievement motivation and a positive reinforcement strategy are needed,

c) A special robustness against faulty operation and non-cooperative or misuse is required,

d) A toys-like, embedded system only provides limited resources for underlying speech technology and

e) From the technical point of view, children voices and respective speech databases are less examined.

There are numerous empirical studies within the area of educational psychology. The age-depending, expected educational level of children is well-defined (see e. g. overview German [11]) and national standards like the ‘Active Vocabulary Test’ (AWST 3-6 for German [12]) are given. The paper therefore mainly deals with technical feasibility, concept and a functional model of embedded language tutoring for children.

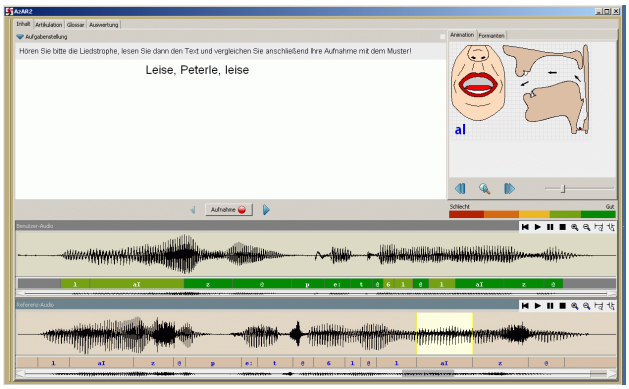

Figure 1: AzAR II tutoring template (screen shot).

3. BASELINE TUTORING SYSTEM

Teaching local Russian migrants in improving their German skills, in particular their pronunciation, the authors were dissatisfied with existing computer-assisted tools. Common systems do not include a distinct user feedback and partly marking wrong phenomena or pronunciation positions.

Supported by companies dealing with further education and speech technology, the authors established a network of teachers, phoneticians and technicians aiming at an "Automat for Accent Reduction" (German acronym: AzAR [13]). The regarding two AzAR projects 2004-2007 enabled the interactive training of second language pronunciation for adult and adolescent learners (native speakers from the Slavonic language group, mainly Russians). The AzAR system includes a curriculum for the production and perception training of difficult segmental contrasts. Animated articulatory organs and recorded time function can be displayed. Gradually mispronounced phonemes are marked using a color scale. Furthermore, formant trajectories are measured and can be compared with a female reference voice. All learning advances are logged in detail. Figure 1 shows a template from AzAR.

Since November 2007, a new version of the AzAR software is under development. Beside German and Russian, it will include new databases for source and target languages (L1/L2) involving Polish, Czech and Slovak - supported by several partners within the project "Euronounce" granted by European Community [14].

4. DIDACTIC CONCEPT OF „LISA"

Initiated by inquiries from companies which are active in the education market and also by encouraging statements of educationalists and kindergarten teachers, the authors drafted a technical study and a business plan [15]. The concept aims at a didactic design model which encourages children to learn their age-related basic vocabulary and which is conform to standard curricula. The concept contains three different tasks and actions:

- To learn short or long vocabulary lists,
- To listen to a story from a cloze text which is only continued if a missing words is pronounced adequately,
- To mimic a rhyme or a song (advanced concept).

The instructions are presented age-related, i.e. for children aged 3+ years by a medium sized doll (standard size of 40+ cm) personifying an elder sibling to achieve an appropriate level of authority (educator versus funny doll). The doll is able to sit e.g. on a table enabling a face-to-face operation. The current German prototype version LISA includes laminated boards from a conventional children's book extended by a marker symbol (a cloud) to activate a barcode scanner. The scanner (hidden in the doll arm) synchronizes doll's instructions with the non-verbal user action. The according example instruction of LISA reads as follows (in English): "Hello, I am Lisa, do you want to look at your book with me? Put my hand on the cloud at the top of the page." Figure 2 shows a board example including the vocabulary list: lamp, glasses and book. The design was adapted from a children's book [16] and is used for scientific evaluation, only.

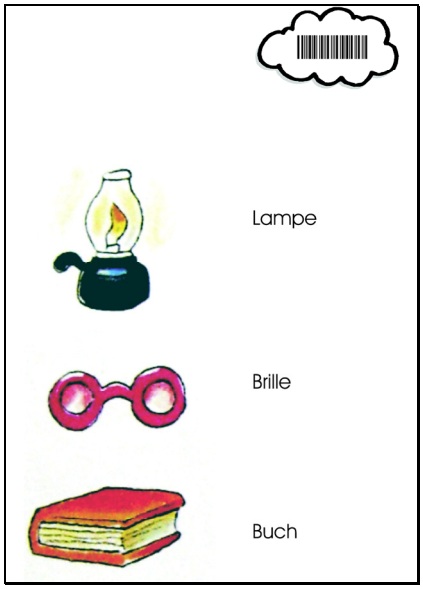

Figure 2: Exemplary picture board adapted from [16].

The program sequence to synchronize scanner and instructions is presented in Figure 3. A tutoring example and the pronunciation validation are shown in Figure 4. The dialogue model focuses on the correct pronunciation and meaning of terms.

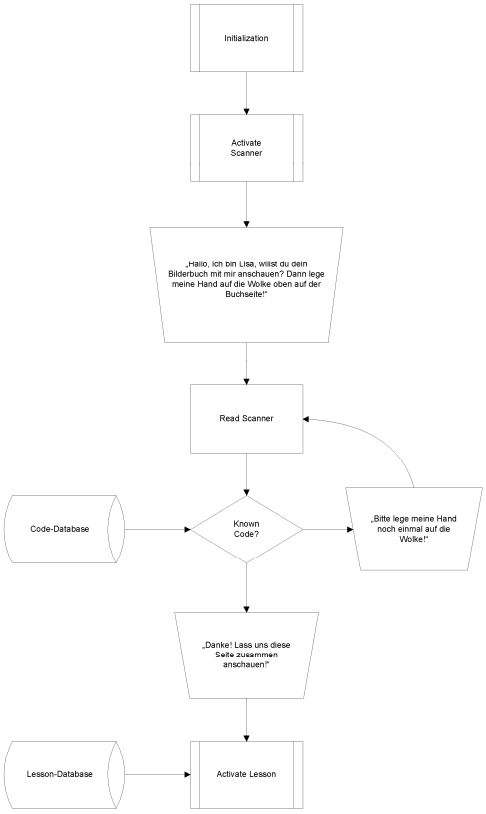

Figure 3: Scanner-based activation of instructions.

Figure 4: Tutoring sequence example of LISA.

5. EMBEDDED FUNCTIONAL MODEL

Figure 5 shows the main program components: dialogue control, tutoring module, recognizer and synthesizer. Additionally, an acoustic frontend processor and audio hardware is required.

5.1 Robust Speech Recognition

Considering the technical feasibility, room acoustics (potential reverberation and noise) is important since the application design does not guarantee close-talk microphone distance or specified environmental characteristics. In a real user scenario, the word recognition rate (RR) can fall below 30% (far below usability) versus a RR of close-to 100% for clean speech. A performance experiment on speaker-independent, small vocabulary recognition using the UASR recognizer of TU Dresden showed that, for certain conditions like noisy training data simulating the target scenario and advanced frontend analysis methods, RRs of 90+% can be basically achieved. Evaluating 1,020 command phrases (17 classes) of the APOLLO corpus, assuming a reverberation time of T_{60}=1s and a speaker-microphone distance of SMD=1m, Petrick et al. obtained a RR of 96.3% on the laboratory scale [17]. Nevertheless, advanced frontend analysis leads to additional calculation complexity and complicates the integration onto embedded platforms. Due to the product availability, the authors currently use the embedded vicCONTROL recognizer of voiceINTERconnect GmbH. API and hardware are optimized for automotive and home automation solutions [18]. The underlying hardware is based on the BlackFIN Digital Signal Processor (DSP) series of Analog Devices. The circuit boards dimensions are 60x60 mm. The board includes several audio and control connectors. The flash memory is configurable according to the target vocabulary. The software package also contains frontend algorithms. The speech recognizer produces an n-best list of correctly or mispronounced word hypothesizes. The final selection and pronunciation scoring is based on phonemic confidence measures as implemented in the AzAR system [13]. Preliminary testing is based on adult voices. A new database for the adaption and training of infantile voices is under development.

5.2 Speech Dialogue and Synthesis

The dialogue model (based on the described didactic concept and synchronized by a barcode scanner) will be tested and optimized in further steps empirically. Hypothesizes are described by a BNF grammar. The tutor's response is generated by pre-recorded and synthesized speech depending on the respective voice quality requirements and vocabulary limitations. From the technical viewpoint, small footprint synthesis like microDRESS of TU Dresden is available [19] but educational experts argue that the synthesis quality might not be sufficient in a tutoring context.

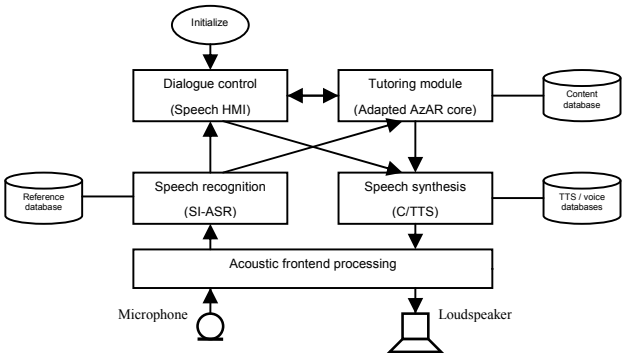

Figure 5: Block diagram of speech technology components.

5.3 Prototype Design

The prototype is compiled from a commercial doll with a height of 54cm (sitting 37cm) and a net weight of 580g. The head (Ø 13cm) is made from medium-soft polyethylene providing low resonance characteristics which is a challenge for the inbuilt loudspeaker. The plastic head also carries the microphone (mouth position). The prototype includes the handheld-scanner CipherLab 1090 (weight 148g) which is partly integrated in the doll's arm. The textile torso contains DSP board and battery pack. Power consumption and operation temperature have been optimized during previous application projects and are less critical. Two general purpose switchers (on/off, test modus) and an USB interface for functionality upgrades (or as scanner interface) are provided.

There are several proposals regarding the doll design which is not yet agreed among the educationalists. The authors plan a survey in some regional kindergartens and preschools.

6. CONCLUSIONS

The study discussed the adaptation of pronunciation tutoring technology for adult learners to the special requirements of young children. The authors develop respective speech technology algorithms and specialized voice databases for human-machine dialogue, speech recognition and synthesis. The paper introduced a didactic and a technological concept. A functional model for German is being developed. Furthermore, multilingual prototypes are planned.

There are several high-tech toy products, partly comprehending speech or image processing but also different sensors and robotics. These products are usually placed in a high-price segment and their operation is directed to gaming and adventure. Additionally, low-cost toys with simple interactive scenarios are coming up. By contrast, the introduced embedded system combines tutoring software, speech technology and electronic hardware with a special focus on teaching aids. The hardware is based on a digital signal processor (DSP). The design enables a medium price range which is usually represented by less elaborate products for speech and music storage. Preferable design and look for young children are currently surveyed.

7. ACKNOWLEDGEMENT

The authors would like to thank Tobias Katz for his valuable assistance regarding the application design and the functional model implementation.

8. REFERENCES

[1] J. Finley et al., Pronunciation power, Educational Software Review, Learning Village. http://www.learningvillage.com/html/guide.html

[2] J. Burston (Ed.), The CALICO software review, Computer Assisted Language Instruction Consortium. http://calico.org/CALICO_Review/

[3] J. Dalby and D. Kewly-Port, Explicit pronunciation training using automatic speech technology, CALICO Journal, vol. 16, no. 3, 425-445, 1999.

[4] Interactive Spoken Language Education (ISLE), project homepage of Hamburg University. http://nats-www.informatik.uni-hamburg.de/~isle/

[5] F. Ehsani and E. Knodt, Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm, Journal of Language Learning & Technology Vol. 2, No. 1, pp. 45-60, July 1998.

[6] I. Kantrov, Talking to computer: A prototype speech recognition system for early reading instruction, Technical Report No. 91-3, Education Development Center, Newton, MA, 1991.

[7] J. Mostow, S. Roth, A. G. Hauptmann and M. Kane, A prototype reading coach that listens. Proc. 12th Nation. Conf. on Artificial Intelligence, Aug., 785-792, 1994.

[8] J. Mostow and G. Aist, G., The sounds of silence: Towards automated evaluation of student learning in a reading tutor that listens, Proc. 14th Nation. Conf. on Artificial Intelligence, July, 355-361, 1997.

[9] R. D. Silverman, Vocabulary development of english-language and english-only learners in kindergarten," Elementary School Journal, 107(4), 365-383.

[10] D. W. Massaro, M. M. Cohen, M., Tabain, J. Beskow and R. Clark, Animated speech: Research progress and applications. In E. Vatiokis-Bateson, G. Bailly, & P. Perrier (Eds.), Audiovisual Speech Processing, MIT Press, Cambridge, MA (in print).

[11] L. Fried, Expertise zu Sprachstandserhebungen für Kindergartenkinder und Schulanfänger - Eine kritische Betrachtung. Deutsches Jugendinstitut, 2004 (in German). http://cgi.dji.de/bibs/271_2232_ExpertiseFried.pdf)

[12] C. Kiese, P.M. Kozielski, Aktiver Wortschatztest für 3- bis 5-jährige Kinder, Hogrefe, Goettingen 1996 (in German).

[13] O. Jokisch, U. Koloska, D. Hirschfeld and R. Hoffmann, Pronunciation learning and foreign accent reduction by an audiovisual feedback system, Proc. ACII, Beijing, Oct. 2005, LNCS-3784, 419-425, Springer, 2005.

[14] O. Jokisch, R. Jäckel, M. Rusko, G. Demenko, N. Cylwik, A. Ronzhin, D. Hirschfeld, U. Koloska, L. Hanisch, R. Hoffmann, The EURONOUNCE project – An intelligent language tutoring system with multimodal feedback functions: Roadmap and specifications. Proc.19th Conf. on Electronic Speech Signal Processing (ESSV), Frankfurt, September 2008.

[15] O. Jokisch, R. Hoffmann, W. Borkenstein, S. Greiner-Sachs, Business plan: Learning aid for infantile language promotion, Dresden / Herzogenaurach, Germany, November 2007 (in German).

[16] W. Färber and M. Schober, Markus Maul hat viel zu tun. German children's book, Loewe publisher, Bindlach, Germany, 1998 (ISBN 3785532091).

[17] R. Petrick, X. Lu, M. Unoki, M. Akagi, R. Hoffmann: Robust Front End Processing for Speech Recognition in reverberant environments: Utilization of speech characteristics, Proc. INTERSPEECH, Brisbane, 2008 (in print).

[18] J. Ploennigs, O. Jokisch, U. Ryssel, D. Hirschfeld, and K. Kabitzsch, Generation of adapted, speech-based user interfaces for home and building automation systems. Proc. 7th IEEE International Workshop on Factory Communication Systems (WFCS 2008), Dresden, Germany, May 2008.

[19] R. Hoffmann, O. Jokisch, D. Hirschfeld, G. Strecha, H. Kruschke, U. Kordon, A multilingual TTS system with less than 1 megabyte footprint for embedded applications. Proc. ICASSP, vol. I, 532-535, Hong Kong, 2003.