--extended abstract--
Full paper: Int.Conf ICL 2007, Ed. M.E. Auer, 6 pages, no pagination, ISBN 978-3-90058-279-6

Caller: Computer Assisted Language Learning from Erlangen - Pronunciation Training and More

Christian Hacker[1], Andreas Maier[1], Andre Hessler[1], Ute Guthunz[2], Elmar Nöth[1]

[1]Institute for Pattern Recognition, University of Erlangen-Nuremberg, Germany
[2]Ohm-Gymnasium, Erlangen, Germany

Key words: *Computer aided language learning, Pronunciation Scoring, Evaluation*

Abstract:

In school, oral examination and the ability to speak a foreign language properly have become more important. Yet, individual time per pupil to train the correct pronunciation is extremely short. Caller is a program to support learning English including pronunciation training in class and at home. Its client/server architecture allows to run complex analysis programs like speech recognition on the server without having to consider computational restrictions of PCs. At the same time, the teacher has privileged access to monitor the students' progress. In this paper, technologies to evaluate the student's pronunciation are discussed: acoustic modeling, prosodic features, and pronunciation features.

1 Extended Abstract

Commercial systems for computer-aided language learning (CALL) are nowadays available in every bookshop for different L1/L2 pairs. They are useful for people who do not have the time to attend regular evening classes and for students as additional tuition. Most products focus on reading, listening comprehension, and writing. Speaking is an emerging aspect that requires robust speech recognition for non-natives, robust scoring algorithms, and an appropriate feedback on how to improve the pronunciation. Unfortunately, even in school individual time per pupil to train the spoken language and its correct pronunciation and intonation is extremely short; further, some students do not have the courage to speak aloud unless they feel confident with the foreign sounds. In this paper the client/server system *Caller* (**C**omputer **a**ssisted **l**anguage **l**earning from **Er**langen) is described that focuses on German pupils learning English and allows to integrate complex scoring algorithms on the server. It was developed in a cooperation with a grammar school (grade 5-13) and tested there in class. The modular concept of the software makes it easily extendable and makes all contents easily exchangeable. There was even a project for pupils of the 11th grade to design new exercises for the 5th grade students. The current system implements several exercises that are motivated by a text book which addresses students learning English in the first year (age 10 and 11).
As shown in Fig. 1, the client is programmed in Flash and Java. Only a minimal installation is required locally to run the program; exercises that are independent from speech input run in every browser. Besides the low effort to install the clients and the easy maintenance and update possibilities of the complex speech technology on the server, one of the greatest

advantages of a client/server architecture is that students can access the system from home. Additionally, a control tool allows the teacher to log into the database in order to monitor the students' activities. All students' utterances are recorded, so that a protocol of their mistakes is provided and teachers even can listen to their spoken utterances.
All content is separated from structure and defined in xml-files that are located on a web-server and can easily be modified. Text, images, and sound are loaded dynamically. Speech technology runs on a Linux-server, together with a database that contains e.g. user-information. Each chapter consists of several exercises; the structure is also defined in xml. Each exercise has first to be performed by the student, then he/she is allowed to play a bonus exercise. The student can collect points and an avatar (smiley) reacts positively or negatively depending on the user's input. The student can improve his score by repeating the respective chapter. Four Flash-modules are provided for the different kinds of exercises. The magnet board allows to drag and drop magnets, which is used e.g. in the exercise “Build the sentences” and the listening test. The desktop provides cards that can be used as file cards or game cards e.g. in the spoken vocabulary test or the “Memory game". In the reading test, text is displayed which is provided by the notebook environment. At last, the "learnboy" that has been designed like a "Gameboy" is an appropriate environment for the bonus games.

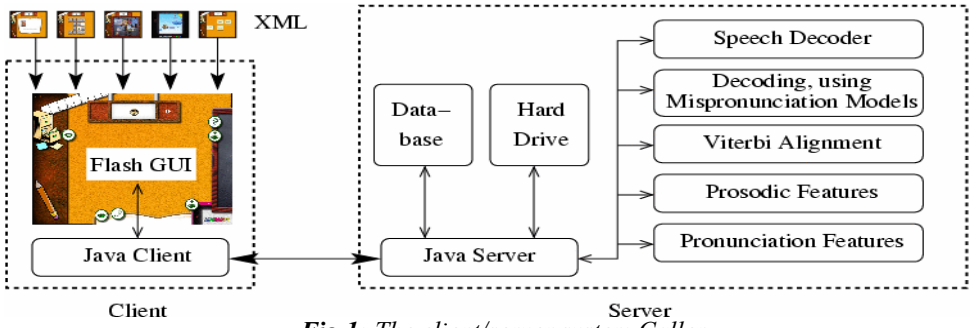

***Fig.1**: The client/server system Caller*

Fig. 1, right shows the server, including speech technology. In the database user data like name, grade, and password and a superuser-flag for the teachers are stored. Progress and mistakes of the learners are logged in this database, too. Speech input is stored on the hard drive. The server provides a speech decoder trained on children data that is invoked in the spoken vocabulary test. To evaluate the pronunciation in the reading exercises, several approaches are provided, that are discussed in the following.
In the European ISLE project an approach to pinpoint pronunciation errors was investigated that is now integrated in many commercial systems: From a database with typical mispronunciations of Germans speaking English, acoustic models with wrong pronunciation are built and added to the speech decoder. Since e.g. the semi-vowel in the word "where" is often wrongly pronounced like "very", both pronunciation variants exist as acoustic models, the correct one and the wrong one. Using Viterbi alignment, it can be determined which model better fits the speech signal. This induces a decision, whether "w" is pronounced correctly or not.
Similar as in ISLE, the first approach integrated in Caller is to enrich the speech decoder with possible mispronunciation models. Then it decides which word sequence was the one uttered most likely. If additionally the word sequence that has to be read is known (and this is true in the reading task) the speech recognizer can be re-run in alignment mode. Now, at each time frame the phone that was to be read by the user can be compared with the phone the recognizer has decided for. Comparing both phones and the corresponding acoustic scores, about 60 word-based pronunciation features are calculated. The third and last approach is to calculate about 100 prosodic features per word in the text that has to be read. The features describe energy, fundamental frequency, jitter, and shimmer of the signal as well as word duration and length of pauses obtained from Viterbi alignment. Prosodic and pronunciation features are the input of a statistical classifier that maps words onto the categories "correctly"

or "incorrectly" pronounced. The combination of different scoring approaches resulted in improved classification rates. The references to evaluate the system are markings of 12 German teachers of English, a native British teacher, and a university student of English (graduate level). 5 teachers re-evaluated the data half a year later again. The automatic classification is compared to the teachers in the same range as the university student compared to the teachers.