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

Christian Hacker, Andreas Maier, Andre Hessler, Ute Guthunz, Elmar Nöth

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 students' pronunciation are discussed: acoustic modelling, prosodic features, and pronunciation features.

1 Introduction

Commercial systems for computer-aided language learning (CALL) are nowadays available in every bookshop for different L1/L2 pairs. They are useful for learners 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 systems for non-native speech, 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; furthermore, 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. It focuses on German pupils learning English and allows to integrate complex scoring algorithms on the server. It was developed in cooperation with a grammar school (grade 5-13) and tested there in class. The modular concept of the software makes it easily extendable and allows to exchange all contents easily. 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 based on a text book which addresses students learning English as a foreign language in the first year (age 10 and 11).

Major systems on CAPT (computer assisted pronunciation training) have been developed in the USA at the SRI international (VILTS™, Autograder™, EduSpeak®) [1, 2, 3], at the Carnegie Mellon University (Fluency pronunciation trainer, NativeAccent™) [4], and at the Center for spoken Language Research, Colorado (WriteToLearn™) [5]. Several systems were brought on the market through spin-off companies like *Carnegie Speech*. An example of a European system is ARTUR, developed at the KTH, Stockholm [6]. Research on non-native speech from German learners of English, however, basically took place 1998-2000 in the context of the ISLE[1] project (Interactive spoken Language Education). Some systems based on

1 http://nats-www.informatik.uni-hamburg.de/~isle/

simple algorithms have been launched for German learners, in particular by digital publishing[1] (based on Intellispeech), Auralog[2] (based on the S.E.T.S.® technology), and Pons[3], which includes external technology developed by Acapela.

In the European ISLE project an approach to pinpoint pronunciation errors was investigated [7] that is now integrated in many commercial systems: Acoustic models with wrong pronunciation are built and added to the speech decoder, using a database with typical mispronunciations of Germans speaking English. 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 forced Viterbi alignment, it can be determined which model better fits the speech signal. This induces a decision, whether "w" is pronounced correctly or not. Now, hints can be given, how to improve the pronunciation. A critical review on CAPT systems including ISLE is given in [8].

2 Description of the CALL-System

At the University of Erlangen the automatic scoring of the pronunciation of non-native speech is being investigated [9, 10]. Similar algorithms are also applied to objectively evaluate people with speech disorders [11]. Our CALL application is the client/server system *Caller*. 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.

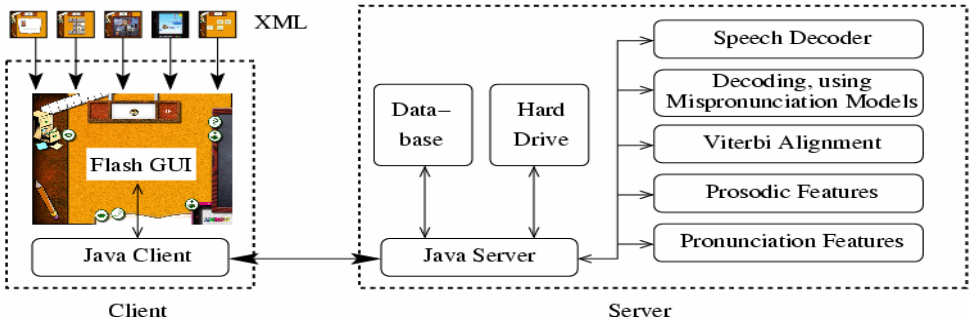

***Fig.1**: The client/server architecture of Caller*

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. First, each exercise has 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.

1 http://www.digitalpublishing.de
2 http://www.auralog.com
3 http://www.pons.de, http://www.acapela-group.com

In Fig. 1 the server, including speech technology, is shown on the right. User data like name, grade, and password and a superuser-flag for the teachers are stored in the database. 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 e.g. in the reading exercise to recognise what the student has said. The Viterbi algorithm aligns the reference sentence, which is known in the case of a reading exercise, and the recorded speech signal. To evaluate the pronunciation, several approaches are provided, e.g. automatic classification with prosodic and pronunciation features. Those algorithmic approaches are described in Sect. 3.

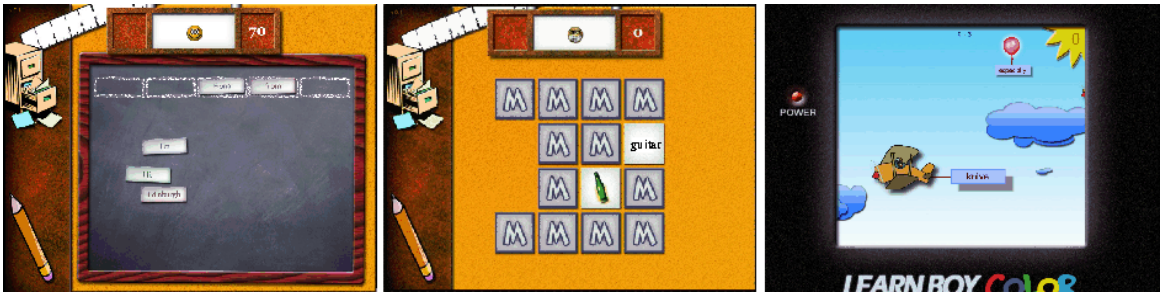

Fig.2: *Caller, selected exercises: Build the sentence, the Memory Game, and the bonus game Moorwords*

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” (Fig.2, left) 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" (Fig.2, middle). The "learnboy" that has been designed like a Gameboy™ is an appropriate environment for the bonus games, like “Moorwords” (Fig.2, right), where the learner has to shoot off wrongly spelled words flying past. Another bonus game allows the user to navigate a bouncing ball with voice commands through a maze passing several obstacles. The reading test is provided by the notebook environment (cf. Fig.3, right). This notebook module simply allows to display text which is also used for the written vocabulary test.

Speech technology is integrated into several exercises: reading test, vocabulary test, and bonus games. Fig.3 illustrates the speech analysis applied for pronunciation scoring of read sentences. As feedback, up to now all mispronounced words and vowels are marked. By clicking on the wrongly pronounced words, the user can listen to his own recording and a reference speaker.

3 Automatic Pronunciation Scoring

Similar as in ISLE, the first approach integrated in *Caller* (overview of modules: Fig. 1, right) is to enrich the speech decoder with possible mispronunciation models. The underlying knowledge base is illustrated in Fig.3, and is easily extendable, e.g. by the teacher monitoring the student. For each word in the reference text, different mispronunciation models are provided. The speech recogniser (decoder) decides which word sequence was the one uttered most likely. In this process, each desired word of the reference text can be substituted by one or many other words. If it decides for a mispronunciation model instead of a real word model, e.g. for “wery” instead of “very”, for “sis” instead of “this”, or for “bake-pipes” instead of “bagpipes”, one can directly localise the wrong pronunciation and give hints how to improve it. However, this way usually too many words are rejected or marked as mispronounced; a combination with other approaches is therefore required.

One further approach is based on pronunciation features. If additionally the word sequence that has to be read is known (and this is true in the reading task) the speech recogniser 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 recogniser has decided for. Comparing both phones and the corresponding acoustic scores, about 60 word-based pronunciation features are calculated [9]. However, each mismatch between reference and automatic recognition can be caused either by a mispronunciation or by an error of the speech recogniser. Therefore for each observed phone confusion the priori probabilities of both events are compared (phone confusion features). Other features result from the phoneme accuracy and confidence scores of the recogniser, or measure the rate-of-speech.

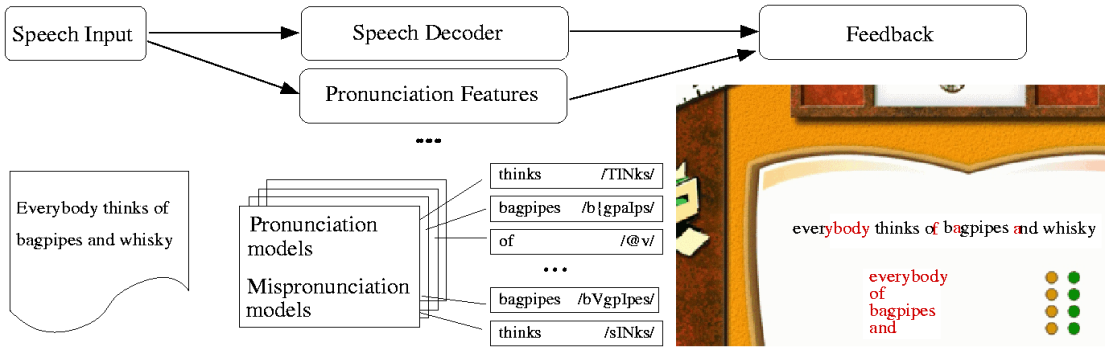

Fig.3: *Reading exercise with pronunciation scoring. Transcriptions given in Sampa [13]*

The third and last approach is to calculate about 100 prosodic features [12] per word in the text that has to be read. Prosodic features describe energy, fundamental frequency, jitter, and shimmer of the signal as well as word duration and length of pauses obtained from the Viterbi alignment. Prosodic and pronunciation features are the input of a statistical classifier that maps words onto the categories "correctly" or "incorrectly" pronounced.

4 Data and Results

The corpus recorded in Erlangen contains 3.4 h of realistic speech data from 57 children of a local grammar school (Ohm-Gymnasium) and a general-education secondary school (Montessori-Schule). The recordings include reading errors, repetitions of words, word fragments, and non-verbals. The size of the vocabulary is 942 words. The age of most children is 10-13; they had been learning English in their first or second year. All the data has been rated by a German university student of English (graduate level, rater S) on the word level (wrongly vs. correctly pronounced) and on the sentence level (marks 1=best to 5) . The Ohm-data (28 pupils) was additionally rated by 13 teachers of English on the word and text-level, with a text consisting of around 11 short sentences. One of the teachers was a native speaker of English and 5 were student teachers who have less than two years teaching experience. 5 teachers re-evaluated the data half a year later again. On the word level, all teachers marked those words, were they would have stopped and corrected the student in class, whereas rater S marked all phone deviations.

To measure the agreement of the 8 experienced teachers a leave-one-rater-out approach was chosen. In each of the iterations one teachers is tested against the others; then the results of all iterations are averaged. However, what does *"testing against the others"* mean? For this purpose ratings from all other teachers have to be combined to one reference rating. On the word level, this *combination* means, that a word is only mispronounced, if at least 3 teachers marked this word as mispronounced; otherwise it is correct. On the text level, which has been evaluated with marks 1-5, simply the average mark is calculated. *Testing* against this new reference means in the word-level case that the percentage agreement of wrongly pronounced

words (sensitivity) and the percentage of agreement of correctly pronounced words (specifity) are calculated. The mean of both values is the class-wise averaged recognition rate (CL). On the text level, testing means calculating the correlation (Pearson correlation). The Spearman correlation calculates ranks for each mark and correlates the rank values. This way, the marks 1-5 are not any more assumed to be equidistant as it is the case in Pearson's correlation coefficient.

	8 experienced teachers	***5 student teachers***	***rater S***	***Caller***
word-level	80 (±2) % CL	74 (±4) % CL	69 % CL	69 % CL cf. [9]
text-level (Pearson)	0.79 (±7)	0.73 (±8)	0.77	0.58
text-level (Spearman)	0.72 (±4)	0.68 (±5)	0.67	0.57

Tab. 1: Agreement between teachers, between student teachers and teachers, between rater S and teachers, and between the automatic system and teachers: CL on the word level, correlation on the text level, standard deviation in brackets.

Tab.1 shows the agreement of the teachers with more than 2 years teaching experience (2nd col.) and the agreement between each of the 5 student teachers and the combination of the expert teachers' rating (3rd col.). This agreement is also shown for rater S (4th col.). Finally, the agreement between the automatic classifier and the teachers can be found in the 5th col. First, it can be seen, that even the human raters do not agree 100%. However, they show very high agreement in the correctly pronounced words (specifity) but differ in the way which of many mispronounced words they reject. They do not reject all wrongly pronounced words, since they do not want to frustrate the pupils. Second, it can be seen that our automatic system reaches in two cases the worst human rater.

The intra-rater agreement for the teachers who re-evaluated the data is between 74 % CL and 81 % CL on the word level and between 0.62 and 0.83 (Pearson correlation) on the text level. Note, that those values are partially worse than the values in Tab.1, but here no such reference that is robustly estimated from many teachers can be used. However those values are markedly better than the CL or correlation between any *pairs* of *different* teachers.

The results in Tab.1, right, are calculated from prosodic and pronunciation features. In the system *Caller,* this approach is further improved by adding mispronunciation models to the speech recogniser (Fig. 3). A small set of evaluation data recorded from pupils using *Caller* in school shows that the combination of both approaches results in a noticeable increase of classification rate CL. An evaluation of our test data that is rated by multiple experts will be done next.

5 Conclusion

In this paper different approaches for pronunciation scoring are combined and integrated into the system *Caller*. The automatic scoring is evaluated on realistic but difficult data, that also contains reading errors. The data has been annotated by 14 experts, among them 12 German teachers of English and a native English teacher. *Caller* is a client/server system. For the student's PC only minimal installation is required on the client; complex scoring algorithm run on a server. Teachers are allowed to monitor the student's activities. The exercises of Caller can be easily modified, since all content is stored in xml-files that are located on a web server and separated from the structure. This way it was possible, that in Ohm-Gymnasium Erlangen, students of the 11th grade could design new exercises for beginners of English in the 5th grade.

Acknowledgements:

The authors want to thank the headmasters and all the teachers of the Ohm-Gymnasium Erlangen and the Montessori-Schule Erlangen who made our recordings possible. Special thanks also to all the English teachers who evaluated the data.

References:

[1] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub. “Automatic Scoring of Pronunciation Quality”. Speech Comm., Vol. 30, pp. 83–93, 2000.
[2] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen. “Combination of Machine Scores for Automatic Grading of Pronunciation Quality”. Speech Comm., Vol. 30, pp. 121-130, 2000.
[3] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sonmez. “Prosodic features for automatic textindependent evaluation of degree of nativeness for language learners”. In: Proceedings of the Int. Conf. on Spoken Language Processing (ICSLP), Bejing, 2000.
[4] M. Eskenazi, Y. Ke, J. Albornoz, and K. Probst. “The Fluency Pronunciation Trainer: Update and User Issues”. In: Proc. of InSTIL, Dundee, 2000.
[5] A. Hagen, B. Pellom, S. van Vuuren, and R. Cole. “Advances in Children's Speech Recognition within an Interactive Literacy Tutor”. In: Proc. Human Language Technology conference - North American chapter of the Association for Computational Linguistics annual meeting (HLT NAACL), pp. 25-28, Boston, 2004.
[6] B. Granström. “Speech Technology for Language Training and e-Inclusion”. In: Proc. 9th European Conference on Speech Communication and Technology (Interspeech), pp. 449-452, 2005.
[7] W. Menzel, D. Herron, P. Bonaventura, and R. Morton. “Automatic detection and correction of nonnative English pronunciation”. In: Proc. of InSTIL, pp. 49 – 56, Dundee, 2000.
[8] A. Neri, C. Cuchiarini, and C. Strik. “Feedback in Computer Assisted Pronunciation Training: technology push or demand pull?” In: Proceedings of the Int. Conf. on Spoken Language Processing (ICSLP), pp. 1209–1212, Denver, 2002.
[9] C. Hacker, T. Cincarek, A. Maier, A. Heßler, and E. Nöth. “Boosting of Prosodic and Pronunciation Features to Detect Mispronunciations of Non-Native Children”. In: Proc. Int. Conf on Acoustics, Speech, and Signal Processing (ICASSP), pp. 197 – 200, Honolulu, 2007.
[10] C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann. “Pronunciation Feature Extraction”. In: 27th DAGM Symposium, pp. 141-148, Springer, Berlin, 2005.
[11] A. Maier, C. Hacker, E. Nöth, E. Nkenke, T. Haderlein, F. Rosanowski, and M. Schuster. “Intelligibility of Children with Cleft Lip and Palate: Evaluation by Speech Recognition Techniques”. In: 18th Int. Conf. on Pattern Recognition, pp. 274–277, 2006.
[12] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. “How to Find Trouble in Communication?” Speech Comm., Vol. 40, pp. 117-143, 2003.
[13] “Sampa, a computer readable phonetic alphabet”. http://www.phon.ucl.ac.uk/home/sampa/home.htm

Authors:

Dipl.-Inf. Christian Hacker
Institute for Pattern Recognition
University of Erlangen-Nuremberg
Martensstr. 3
91058 Erlangen, Germany
hacker@cs.fau.de

StRin Ute Guthunz
Ohm-Gymnasium
Am Röthelheim 6
91052 Erlangen
Erlangen, Germany
ute.guthunz@web.de

Dipl.-Inf. Andreas Maier
Institute for Pattern Recognition
University of Erlangen-Nuremberg
maier@cs.fau.de

PD Dr.-Ing. habil. Elmar Nöth
Institute for Pattern Recognition
University of Erlangen-Nuremberg
Martensstr. 3
91058 Erlangen, Germany
noeth@cs.fau.de

Andre Hessler
University of Erlangen-Nuremberg
ahessler@gmx.de