WebGrader™: A Multilingual Pronunciation Practice Tool

Leonardo Neumeyer, Horacio Franco, Victor Abrash, Luc Julia, Orith Ronen, Harry Bratt, Jehan Bing, Vassilis Digalakis, and Marikka Rypa

Speech Technology and Research Laboratory, SRI International, Menlo Park, USA
{leo,hef,victor,julia,orith,harry,jehan,vas,marikka}@speech.sri.com

Abstract

WebGrader™ is a pronunciation grading tool designed for practicing pronunciation in a second language. The system uses SRI's speech recognition [1] and pronunciation scoring [2][3][4] technologies. The application client was implemented by using the Java platform to facilitate deployment and updates of software and content over the World Wide Web. We present the overall system architecture, user-interface design, scoring algorithms, and a preliminary user study.

1. Introduction

Most foreign language instruction courses focus on teaching reading, and writing, and on listening comprehension. Much less effort is dedicated to teaching speech production because it is sometimes considered less critical for communicating in a foreign language, or simply because of a lack of resources such as private tutors who are native or near-native speakers of the target language.

We believe that an interactive system capable of grading pronunciation can facilitate the pronunciation learning process. We developed an interactive pronunciation tool, the WebGrader™, for practicing pronunciation in multiple languages. WebGrader™ is a research prototype capable of scoring sentences and words by using text-independent algorithms developed at SRI [2][3][4]. WebGrader™ requires a speech recognition engine [1] installed locally in the user's computer and, optionally, an Internet connection for downloading the graphical user interface, the scoring algorithms, and the lesson materials. A more distributed architecture for speech recognition over the World Wide Web is also possible. For example, the client could simply capture the audio and extract the acoustic features that are sent to a recognizer located on the server side. A more detailed discussion of alternative architectures can be found in [5][6].

WebGrader™ is organized in lessons. A lesson is a collection of related sentences organized by themes such as transportation or eating in a restaurant. Students can listen to natives saying the phrases, part of the phrases, or individual words. They can also record themselves and obtain pronunciation scores for the phrase and for individual words. Words that are hard to produce can be practiced by selecting the target word and obtaining scores for that particular word. The content can easily be updated, and additional lessons can be downloaded from a content server.

2. System Description

The WebGrader™ system has three main components: the speech recognition engine, the client-side application, and the server-side application. A block diagram is shown in Figure 2.

The hidden-Markov model (HMM) based speech recognition engine is used to align the student's speech signal to the sentence text. The speech engine is installed locally at the student's computer and provides, to the client application, services such as audio capturing, feature extraction, Viterbi search that generates time alignments, phone and word durations, and associated scores such as phone posterior probabilities. Communication between the client application and the speech engine is established through a Java application user interface (API). This allows the client application to be written entirely in the Java programming language, facilitating portability as well as network interoperability and software updates.

The client-side application includes the following modules: the graphical user interface (GUI), the pronunciation scoring unit which implements the scoring algorithms, the content database which includes sentence text, recordings by native speakers, and alignments for the native speech (which is used to display prototypical duration information to the student). The client application is also capable of communicating with the speech recognition engine by means of its Java API and to the server-side applications by means of the Internet protocols.

The server-side application provides services for software download and updates as well as access to content material, which includes text, audio, and alignments in the supported target (L2) languages, as well as the associated help information in the supported native (L1) languages.

3. Graphical User Interface

The user-interface design is constantly evolving as more user feedback is obtained and algorithmic improvements are achieved. In its current form, the WebGrader™ consists of various user-selectable panels to control its operation. These panels include audio controls, sentence selection table, and results screens.

The audio control screen allows the user to start recognition; the end of speech is detected automatically by means of an endpointer. The user can practice using the whole sentence or a subset of contiguous words. The subset of words is selected by dragging the mouse. The user can check audio levels by observing a vu-meter, and can modify input and output volume levels. Warning panels appear when an audio error condition is detected, such as low audio level, clipping, and truncation. Special care has been taken to control audio quality to guarantee accurate scores.

The sentence selection screen allows the user to navigate the topic tree. Sentences are organized in a tree structure where the parent is the main topic, followed by a subtopic, and a list of practice sentences. The sentences can be read in any of the supported L1 and L2 languages.

The results can be observed in the sentence and word score panel. A bar graph indicates the resulting score for each word. In addition, a numeric result for the whole sentence is displayed. A sentence-level average score is updated and displayed after each attempt. Word durations can be observed and compared to a prototypical native realization of the same sentences by using two approaches: (1) an absolute duration display shows scaled word segments for the native and nonnative utterances, and (2) a relative duration display aligns the start and end of speech of the native and nonnative utterances indicating the relative durations of words and pauses regardless of the total duration.

4. Pronunciation Scoring Algorithms

4.1 Phonetic segmentation

The pronunciation scoring algorithms are based on phonetic time alignments generated by SRI's speech recognition system. In this application, the transcription of the utterance is known because the student is prompted to read a sentence from the screen. By using the alignments and the native-trained HMMs, the system computes various scores that rely on the phone-level statistics. Because no word- or sentence-level statistics are used, the result is a flexible (text-independent) system that can easily be customized for new lesson materials by developers with no speech recognition background.

4.2 Pronunciation scores

To produce accurate pronunciation grades, the WebGrader™ system computes scores based on phone posterior probabilities and phone duration models. The calculation and validation of these scores is described in detail elsewhere [3][4].

4.3 Calibration using nonnative data

To perform accurately, machine scores generated by the WebGrader™ must correlate well with human judgement of pronunciation. Machine scores should also be calibrated to a human-readable scale. For the Spanish version of the WebGrader™, we used human judgements provided by a panel of five raters on speech read by 206 nonnative speakers. Five native Spanish graders were selected among eleven as the most consistent. They graded each nonnative sentence on a scale of 1 to 5, ranging from strongly foreign to almost native quality. The raters also had the option to *reject* sentences that had poor audio quality, serious disfluencies, truncation, and other serious deficiencies. To prevent the raters from clustering the data in native and nonnative classes, no native examples were mixed with the presented material. There was some overlap in the speech material rated by the teachers for consistency checking. The consistency across raters was assessed in a subset of the database consisting of 2,800 sentences. Average inter-rater correlation was r=0.68 at the sentence level and r=0.9 at the speaker level.

The mapping of machine scores to human grades can be defined as a classification problem. Given the set of machine scores obtained from a sentence, we try to classify the sentence as belonging to one of N classes, the classes being defined by the discrete pronunciation grades assigned by the human raters. We implemented a Bayes classifier to map the machine scores to human grades. We estimated the class conditional probability distributions of the machine scores for each human grade by using smoothed histograms of data from 7,000 nonnative sentences graded by human raters, plus an additional 15% of native data, which was assigned grade 6. We assumed equal priors for the grade classes in the Bayes classifier. As an example, in Figure 1 we

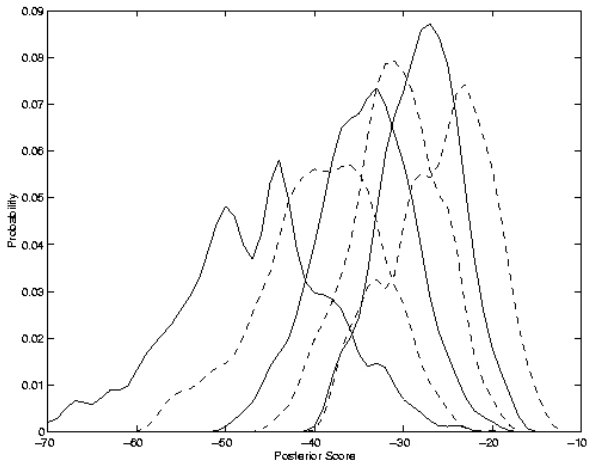

Figure 1. Distribution of posterior scores for grade classes 1 to 6 ordered from left to right. Grade 6 corresponds to the native speakers.

see the distribution of the posterior machine scores for each grade. The points where the adjacent curves cross each other define the boundaries between the grade classes. With a single sentence and a single machine score there is considerable superposition among the grade classes. The discrimination improves by averaging scores among more sentences and using more machine

features to assign a grade to a speaker [3][4]. There is also some level of "noise" in the assignment of grades to sentences by human raters. This noise contributes to the overlap among grade classes.

4.4 Word-level pronunciation scores

Presenting word scores to a student is important because it provides more detailed feedback than the sentence grades. To compute scores at the word level, we compute the average phone score for the target word as we do for the whole sentence. Because the number of phones at the word level is much smaller than at the sentence level, the resulting estimate presents much greater variability at the word level. The lack of target human grades at the word level makes it even more difficult to validate the estimation algorithm. (This problem will be overcome in the future by using detailed phonetic transcriptions of the nonnative data.) An experimental word-level scoring algorithm was implemented. We computed the mapping for word-level machine scores as we did for the sentence-level scores. Since human grades at the word-level were not available, we approximated the human word grades by using the corresponding sentence-level grades. To avoid extreme variations at the word level, we smoothed the raw word-level estimates with the sentence-level estimate. The resulting word-level scores were only informally evaluated with various test users.

4.5 Calibration without using nonnative data

So far, we have collected data for two languages: Spanish and French. In both databases, we rated the pronunciation of nonnative speakers. These ratings were used to validate the algorithms and calibrate the scale of the machine scores. We observed that the relative positions of the grade class boundaries for the mapping of the log posterior scores to human grades were very similar. This occurs in spite of the fact that different human graders have been used in the two systems, and different acoustic models were used to compute the posterior scores. This result suggests that it is possible to extrapolate the mapping from machine scores to human grades to other languages without using an annotated nonnative speech database. We only need to use the relative position of the grade class boundaries with respect to the native distribution. By doing this, we only need a small native test set to calibrate the mapping. Informal tests were carried out for English resulting in acceptable performance.

5. Preliminary User Study

An informal user testing procedure was carried out on a Spanish version of the WebGrader™ to perform initial evaluation of the tool's overall perceived usefulness, the interface design, and the performance of the scoring algorithms.

5.1 Subjects

Five subjects were recruited internally at SRI as preliminary testers of the software. For gender balance we selected three male and two female subjects. All were native English speakers, so that we could map to the expected user base. All sessions were recorded, and utterances were identified by speaker in a database of session logs.

5.2 Analysis

The purpose of the initial WebGrader™ evaluation was to identify the most salient areas in interface design and tool performance for further research and development. In the course of administering both written questionnaires and oral interviews in subject tests, a qualitative picture of the perceived value and strong points of the overall tool also emerged.

In general, the WebGrader™ tool was received enthusiastically, and all subjects (except one) considered the tool fun and engaging as well as useful in helping them improve their pronunciation. One subject indicated that she would use it if additional refinements were made. The strengths of the tools were judged to be the following:

- The pronunciation scoring feedback and the ability to repeat until some improvement was perceived was beneficial.
- The ability to hear a native speaker in combination with playback of the subject's sentence/word recording was helpful.
- In particular, the ability to break down a sentence into words or phrases, record, and receive scoring feedback was judged to be a key factor in promoting improvements in pronunciation.
- Four out of the five subjects indicated that they perceived the tool in its current initial state to be very helpful, although they cited several areas for further improvement.

The areas for further development fell into three major categories: consistency of scoring feedback, user interface, and further refinement of diagnosis and repair strategies. For brevity we will not report, in this paper, comments related to the user interface.

5.2.1 Scoring Feedback

- Further work needs to be carried out in returning more consistent scoring. Although users felt that there was a general correspondence to their performance, they felt that some of the scoring was not always consistent. It is anticipated that ongoing research into scoring and calibration with human raters will be integrated into the tool, with more consistent feedback to the user. Particularly, having more data on words in isolation will help.
- A scoring scale from 1 to 10 or from 1 to 100 was judged to map more closely than the original scale of 1 to 6 to a scale meaningful to most users.

5.2.2 Diagnosis and Repair

- Although users like seeing a score for a whole sentence, they all felt that scoring individual words was most helpful to them. Furthermore, users wanted to see scores for individual problem sounds so that they could better target areas for improvement.
- All users wanted more help in how to improve their pronunciation. Although receiving a score was motivational in the desire to improve, users wanted specific help in targeting problem areas and how to improve them. Future plans include incorporation of targeted feedback and assessment of the efficacy of different types of pronunciation training. We are working on the detection of specific problems, such as letter to sound rules, dipthongization, etc.

6. Summary

We described a software tool for practicing pronunciation in a second language. Preliminary user tests showed that students find the tool useful for improving pronunciation. More detailed and accurate feedback, as well as remedial excercises, need to be included. Future version of WebGrader™ will include improved algorithms for detection of specific pronunciation errors as well as the associated remedial excercises.

Acknowledgements

We gratefully acknowledge support from the U.S. Government under the Technology Reinvestment Program (TRP). The views expressed here do not necessarily reflect those of the Government.

References

[1] V. Digalakis and H. Murveit, "Genones: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer," *IEEE Trans. Speech Audio Processing*, pp. 281-289, July 1996.

[2] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic Evaluation and Training in English Pronunciation," *Proceedings ICSLP*, 1990.

[3] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech," *Proceedings ICSLP*, 1996.

[4] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic Pronunciation Scoring for Language Instruction," *Proceedings ICASSP*, 1997.

[5] V. Digalakis, L. Neumeyer, and M. Perakakis, "Quantization of Cepstral Parameters for Speech Recognition over the World Wide Web," *Proceedings ICASSP*, Seattle, WA, 1998.

[6] L. Julia, A. Cheyer, L. Neumeyer, J. Dowding and M. Charafeddine, "http://www.speech.sri.com/demos/atis.html," *Proceedings AAAI*, Stanford, CA, 1997.

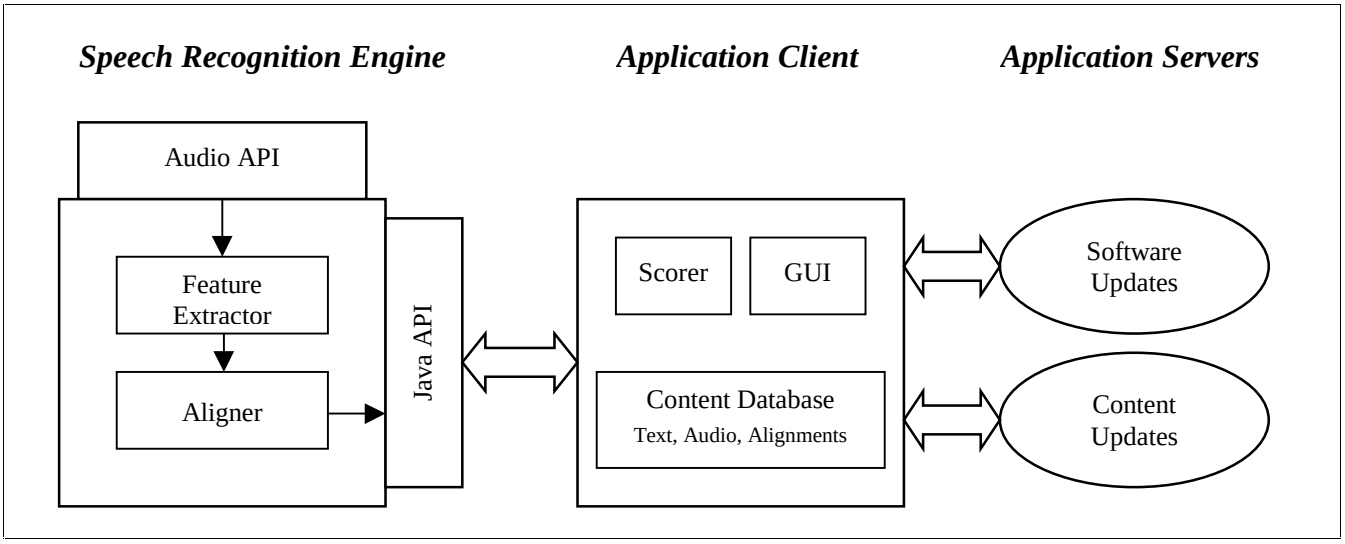

Figure 2. WebGrader™ system block diagram.