Speech interaction with Saybot player, a CALL software to help Chinese learners of English

Sylvain Chevalier

Saybot Inc.
Shanghai, China
sylvain@saybot.com

Abstract

Saybot is a software application which uses speech processing technology to help users practice spoken English. It is a commercial application available since October 2005 in China with much content designed for learners of many different profiles. In this paper, we present the strategy used and discuss its performance based on data collected from users of the software application in a real practice situation.

Index terms: speech-enabled CALL, dialogue systems, evaluation.

1. Introduction

The need for acquiring proficiency in English became particularly strong in China, especially among students and young professionals where a good level of English is necessary to achieve a good career, starting from proving good results in English tests (TOEFL, GRE, CET, interpretation tests). In China, English training is mainly focused on reading and writing, mainly because of the lack of qualified teachers who can teach spoken English. However, the working world seeks professionals with good oral English and most of the popular English tests tend to include a spoken part in addition to the written ones. This situation makes technologies helping spoken language learning particularly attractive in China.

The use of speech technologies to help language learning and asserting has been investigated for decades [1]. The research community proposed various approaches featuring pronunciation grading [1], recognition of nonnative speech [2], dialogue management [3], prosody [4] and fluency analysis [2]. However, there are still few commercial applications available, often limited to visually displaying the learner's pitch. Most of them are available on Internet[1]. Software and content specifically developed for Chinese learners include 100e[2] and MyET[3].

At Saybot, we have been developing a speech-enabled CALL solution specifically designed for Chinese learners. In addition to a software application featuring speech recognition and pronunciation grading, we designed curricula and an authoring tool to develop these curricula in an effective way. This makes Saybot a complete platform to produce and distribute speech interactive multimedia content.

Saybot platform is divided into:

- Saybot player, a software application running on Windows and including a speech recognizer and pronunciation grading capabilities
- Saybot lessons, multimedia content with speech interaction to be used in the Saybot player
- A web-based authoring tool to be used by curriculum developers to produce Saybot lessons
- A web site where learners can download Saybot player and browse, purchase and download Saybot lessons

The platform is used by Saybot and its partners to produce and distribute curricula for kids, students and young professionals. One of these curricula is now part of the learning process of thousands of English learners in China. Table 1 summarizes the different releases of Saybot player since 2005.

Table 1: Releases of Saybot player

Version number	Release date	Main features
1.0	Oct. 2005	Speech input, audio and text output, 3 feedback types, minimal UI
1.5	Aug. 2006	New UI, audio and graphical output, 4 feedback types, non-linear lesson flow, basic study report
1.6	Feb. 2007	Kids' acoustic model, audio diagnostic tool
1.7	May 2007	Improved scoring capabilities, mouse interaction with the lesson flow
2.0	Aug. 2007 (expected)	Acoustic models adapted to targeted users, more user-friendly UI, web-based study report

A screen-shot of Saybot player 1.7 is shown on figure 1

In section 2, we describe the structure of the speech interaction with Saybot player and in section 3, we present evaluation results of our system for the purpose of spoken language training.

[1] http://www.icsi.berkeley.edu/ gelbart/call
[2] http://www.100e.com/
[3] http://www.myet.com

Figure 1: Saybot player 1.7.

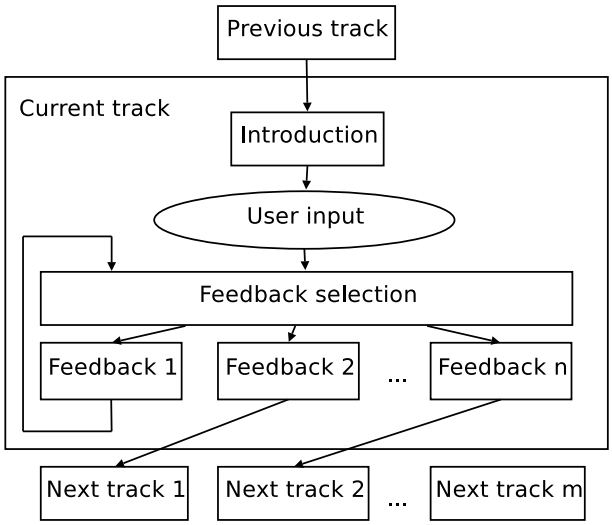

Figure 2: Structure of a track.

2. Speech interaction with Saybot player

Various approaches have been studied to use speech interaction for language learning. Among them were the ideas of giving visual feedback on the pronunciation quality after a read-aloud practice [2], generating audio hints from recognized sentences [5] and producing entirely dynamic feedback using speech synthesis [3].

We assumed that we would maximize the efficiency of a spoken language training software by designing:

1. The simplest possible interaction to limit the time spent learning the software itself
2. Interactions limited to audio input and audio-visual output to keep the learner focused on the dialogue to maximize the time spent in practicing oral English
3. Audio output using only pre-recorded audio data to avoid teaching spoken language using synthetic voices

We defined a rather conservative approach in the sense that, for each defined speaker turn, we focused on a few learning points and designed very specific feedbacks for them. If the learner's input was not among the correct answers and the expected mistakes, he would hear a general feedback. A feedback would embed both qualitative information on the uttered sentences itself and quantitative information on the quality of the pronunciation.

In section 2.1, we give an overview of the structure of Saybot lessons. In section 2.2 we briefly present the recognizer and pronunciation scoring included in Saybot player. Sections 2.3 and 2.4 describe the different types of feedbacks that Saybot player can give to the learner, and the way they are selected and generated.

2.1. Overall structure of a lesson

A Saybot lesson comprises a set of tracks. Each track is one set of introduction recordings followed by a speech input from the user and a feedback which is dynamically picked after analysis of the user's input (see figure 2).

Introduction recordings give instructions to prompt the user to speak. Instructions can be to repeat or read one or several sentences, translate words or sentences, generate a sentence or answer any kind of questions. The learner will then produce a sentence which will be immediately analyzed by the recognizer using the grammar designed for the track. The result of the analysis will be used to pick one of the feedbacks designed for that track. Depending on the feedback, Saybot player will then ask the leaner to practice the same track again or move to another one. This track structure is illustrated on figure 2.

The following sections describe in more details the speech analysis process and definition and selection of the feedbacks.

2.2. Recognition and pronunciation scoring

Saybot player includes a recognition API from to perform recognition and pronunciation scoring. Together with the API, we include two acoustic models, one trained with adults voice and the other one trained with kids' voice. Depending on which lesson is loaded, Saybot player chooses the appropriate acoustic model for that lesson.

The first step is a recognition step. Each track has its own grammar which is designed to recognize expected correct answers and mistakes.

After recognition, the recognizer derives pronunciation scores. These scores are given at the phoneme, word and sentence levels. The value of these scores are computed by force-aligning the phoneme sequence obtained at the recognition step using context-independent models trained on native speakers. The posterior derived from this alignment gives the acoustic match between the learner's pronunciation and the standard one. This acoustic score is associated with a phone duration score and a speech rate score (obtained by matching the detected durations and rates to standard ones of native speakers). The final score is computed using non-linear algorithms trained on nonnative data labeled (graded) by experts.

This pronunciation scores are used to give feedback about the pronunciation quality of the learner as well as

to detect the correctness of the recognized words.

2.3. Feedback types

The recognized sentence and the pronunciation scores are used to define which feedback should be given to the learner. We defined four types of feedbacks associated to possible inputs from learners, they are described in table 2.

Table 2: Feedback types of Saybot player

Feedback Types	
Correct (co)	Users input one of the predefined correct answers. Hearing this feedback, users will know they answered correctly in this practice.
Predefined Error (pe)	Users input one of the predefined mistakes, expected when the lesson was designed. By getting this feedback, users will know exactly what mistakes they made and potentially get very specific instructions to correct the problems.
Mispronunciation (misp)	User input one of the predefined answers but one part of the sentence is detected as being mispronounced. This feedback points to the user where the problem is but cannot tell specifically what the problem is.
General Feedback (gf)	Users did not input any of the predefined answers, or the pronunciation quality was too low to be acceptable. This feedback will give the users general hints about the correct answers.

2.4. Criteria in feedback selection

When a lesson is designed, curriculum developers define a grammar for each track and a set of feedbacks depending on the main teaching points. The goal is to give a maximum of co and pe feedbacks since they are the most specific and potentially the most useful for a the learner. A set of parameters can be set to define in which cases it should be played:

- The recognized sentence
- The pronunciation scores (at word, sentence or cumulative levels)
- scores computed on qualitative results (answers from previous tracks and utterances)
- iteration in one practice (number of times a track has been practiced)
- iteration in one mistake (number of times a specific mistake has been made)

The fact that pronunciation scores are part of the parameters of the feedback allows giving different feedbacks depending on pronunciation quality. Different feedbacks can lead to different tracks so that a user will be guided to tracks aimed at practicing and correcting the problems he has.

In the following section, we give results of the performance evaluation of Saybot player.

3. Evaluation

A speech-enabled CALL application should be evaluated in a number of stages. The recognizer itself can be evaluated in terms of word error rate or sentence error rate, and the scoring functions can be measured in terms of correlation with experts' judgment [2]. At the second level it is necessary to measure how the application can give relevant feedback regarding the learner's input. The third level evaluates the software application together with the content, for example in terms of feedback types (ratio of specific versus general feedbacks) or learning effectiveness. In this section, we give the results of the evaluation of feedback relevance ratio since it is the highest level of evaluation before the content is taken into account. The meaning of the different relevance types is given in table 3.

Table 3: Feedback relevance types

Feedback Relevance	
Relevant (re)	The feedback played is completely relevant regarding the user input.
False Acceptance (fa)	The user input a wrong answer but the player accepted it as a correct answer.
False Reject (fr)	The user input a correct answer but the player rejected it as a mistake.
False Acceptance arguable (fa-arg)	From the transcriber point of view, the sentence should be rejected but the player accepted it as a correct answer (in practice, when one phoneme was mispronounced)
False Reject arguable (fr-arg)	From the transcriber point of view, the sentence should be accepted but the player rejected it as a mistake (in practice, when one phoneme was mispronounced)
Combined (comb)	The feedback contains both fr and fa. Namely, the player pointed one part of a sentence as mispronounced but only another part of the sentence was actually mispronounced.
Confusion (conf)	The feedback gave wrong information although feedback was correctly given as right or wrong answer. In practice, it means a confusion between two co or two pe.

To measure this ratio, we collected logs from real users who practice lessons either at home or at school. The numbers given here are results for a set of lessons developed for kids. We collected 1627 sentences and labeled each feedback given by the recognizer. Each recording was labeled by two transcribers, one native teacher and one nonnative teacher. Non-matching pairs of labels were discussed until a consensus was reached. Results are summarized in table 4.

The raw percentage of relevant feedback is 70% which

Table 4: Feedback relevance

Feedback Relevance			
Total	1627	100%	
Relevant (re)	1135	70%	77%
False Acceptance arguable (fa-arg)	26	2%	
False Reject arguable (fr-arg)	99	6%	
False Acceptance (fa)	56	3%	23%
False Reject (fr)	272	17%	
Combined (comb)	11	1%	
Confusion (conf)	28	2%	

may appear to be pretty low. However, it should be noted that arguable feedback can be classified as relevant by some transcribers. If we include these arguable feedbacks into the relevant case, the percentage of relevant feedbacks increases to 77%. It should also be noted that 17% of the feedbacks were false rejects. When tracks were practiced by a native speaker, false rejects almost never happened. So it is not entirely irrelevant to ask users to practice a track again in that case since the pronunciation is still not standard. The combined score $fa + comb + conf$ is 6% which is low, although it is our goal to get this ratio close to zero.

4. Discussion and future work

This study has been performed on data collected from users practicing lessons that will be adjusted before actual release. When that happens, grammars will be tuned and performance should be significantly increased after that process. Another issue of our current player is that the children's model we use for recognition has been trained with native speakers only. We are currently adapting this model with data collected from nonnative learners which are our real users. We expect significant improvements from this adaptation.

5. About Saybot

Saybot was founded by Dr. Pengkai Pan and Mr. Ho-Ki Au in the fall of 2004. Details can be found at <http://www.saybot.com/>.

6. References

[1] Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price, "Automatic text-independent pronunciation scoring of foreign language student speech," in Proc. ICSLP '96, Philadelphia, PA, 1996, vol. 3, pp. 1457–1460.

[2] Horacio Franco, Victor Abrash, Kristin Precoda, Harry Bratt, Ramana Rao, John Butzberger, Romain Rossier, and Federico Cesari, "The SRI EduSpeak system: Recognition and pronunciation scoring for language learning," in Proc. of InSTIL, Scotland, 2000, pp. 123–128.

[3] Antoine Raux and Maxine Eskenazi, "Using task-oriented spoken dialogue systems for language learning: Potential, practical applications and challenges," in Proc. of InSTIL/ICALL '04, Venice, 2004.

[4] Dorothy M. Chun, "Signal analysis software for teaching discourse intonation," Language Learning & Technology, vol. 2, no. 1, July 1998.

[5] Christopher Waple, Yasushi Tsubota, Masatake Dantsuji, and Tatsuya Kawahara, "Prototyping a call system for students of japanese using dynamic diagram generation and interactive hints," in Proc. of Interspeech 2006 - ICSLP, Pittsburgh, PA, USA, September 2006.