Application and evaluation of speech technologies in language learning: experiments with the Saybot Player

Sylvain Chevalier and Zhenhai Cao

Saybot, Inc.
Wise Logic International Center, 66 North Shanxi Road, 200041 Shanghai, China
{sylvain,zhenhai.cao}@saybot.com

Abstract

In this paper we present Saybot Player, an application of speech technologies to help Chinese learners of English. It is a dialogue system based on a speech recognizer and a pronunciation scoring engine. We compare our approach with existing similar systems, propose metrics to evaluate such systems and give evaluation results obtained with our users in real learning situations.

Index Terms: language learning, dialogue systems, pronunciation scoring

1. Introduction

Foreign language learning is one of the long-awaited applications of speech technologies. Researchers hope that speech science can be used in the context of practicing, acquiring or assessing a foreign language, mainly by automatically assessing pronunciation quality [1, 2, 3] and guiding a learner through a spoken dialogue simulating a conversation with a human tutor [4, 5]. Obvious advantages of such systems include personalized teaching style and curriculum, low cost and unlimited practice time.

Since 2004, we have been developing a software called Saybot Player aimed at helping learners of English in China acquiring better spoken skills. Five versions of Saybot Player have been publicly released since 2005 and a fair amount of curriculum is being developed for different audiences. Since then, a wide and diverse population of learners has been using it and sampled data collected from them has enabled analysis, evaluation and re-factoring of the different component of our system.

In this paper, we explain our approach, compare it with existing systems, propose paradigms to evaluate such systems and give results for our current level of development.

A number of systems using speech technologies have been described in the literature. They share many basic components and differ in others, especially in the type of interactivity they give to the learners. Some fundamental differences revolve around:

- Focus of the system: Some systems focus on assessing the learner's level of spoken English (A) while others are designed for training and practicing spoken skills (T).
- Input Constraint: Different degrees of constraints are set on the speech input, from very constrained read-aloud practices (RA) to less restricted speech recognized with a grammar (G) to near-free speech, usually associated with a statistical language model in the recognizer (F).
- Pronunciation Scoring: Systems differ in whether or not they attempt to score pronunciation and give feedback to the learners.
- Types of feedback and Interaction: Existing systems mainly, give a scoring report (SR), which can be cumulative or instant, or simulate a conversation with predefined (CPD) or dynamically generated (CG) feedback.

In Table 1, we attempt to summarize approaches implemented by some of the systems described in the literature[1], including Saybot Player. Figure 1 shows a screenshot of Saybot Player.

Table 1: *Existing ASR-enabled language learning systems*

System	Focus	Input Const.	Pron. Scoring	Interac.
SpeechRater [6]	A	F	√	SR
Saybot Player [7]	T	G	√	CPD
MIT - Spoken Language Systems Group [5]	T	G	×	CG
Tactical Language [8]	T	G	×	CPD
Ordinate [9]	A	RA/G	√	SR
NativeAccent [10]	A	RA	√	SR
LET'S GO (modified version) [4]	T	F	×	CG

2. Structure of Saybot Player

Saybot Player is a spoken dialogue system whose main components are a speech recognizer, a pronunciation scoring engine and a dialogue engine. Its structure is summarized in Figure 2.

As with most of the other systems, it combines the output of the recognizer (what the learner has said) with the output of pronunciation scoring (how well it has been said) to give feedback to the learners and guide them through other practices or exercises.

2.1. Recognizer

The core component of our system is the speech recognizer. The goal is to let the dialogue engine know what the learner

[1] Content is subject to interpretation and there might be differences depending on different versions of one system.

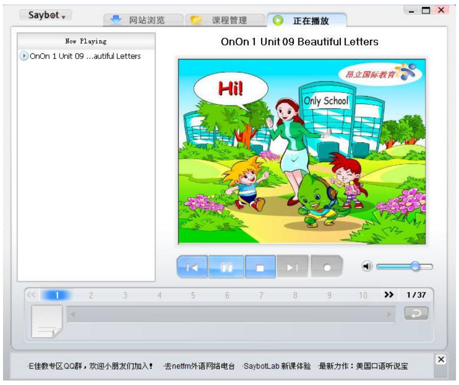

Figure 1: *Saybot Player 2.5*

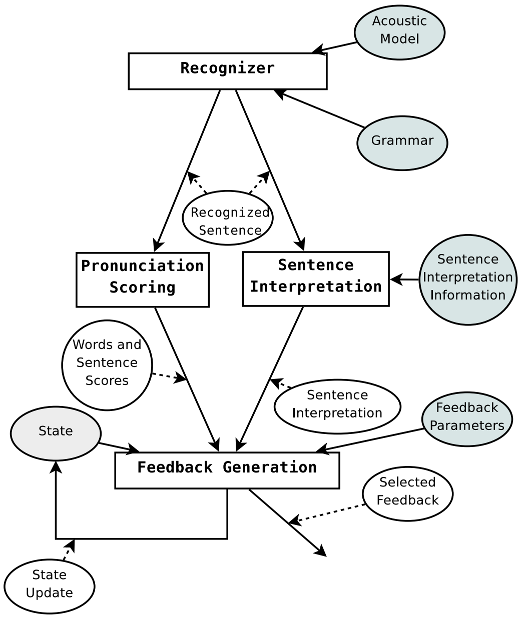

Figure 2: *Speech Analysis in Saybot Player.*

said, so it uses an acoustic model trained on real users. Earlier versions of Saybot Player were based on a licensed proprietary recognition engine, our latest version (expected at the end of May 2008) is built upon the Sphinx 3 open-source recognizer[2], which gave the best results of the possible engines we evaluated.

We use grammars as language models. They are defined by curriculum developers to recognize possible answers for one practice, as well as selected expected errors from Chinese learners of the level targeted by each curriculum. Expected errors can be defined in the syntax or in the pronunciation by modeling incorrect pronunciations in the pronunciation dictionary.

2.2. Pronunciation scoring

Unlike transcription, evaluating pronunciation is a multi-dimensional task. Intonation, fluency and realization of individual phonemes are different parameters affecting pronunciation quality. Common approaches to pronunciation scoring combine scoring of these different parameters and attempt to map the combination of these scores onto scores given by human experts. However, there is currently no agreement on a definition of a composite pronunciation score (referred to as nativeness [11], goodness of pronunciation [2]). The concept of pronunciation score is closely related to confidence scoring which is widely used by dialogue systems to measure the reliability of the recognized sentence. Confidence scoring usually includes scoring from the language model, but pronunciation scoring is only based on acoustic clues.

In Saybot Player, we combine likelihood from the recognizer with phoneme duration scoring trained on native speakers to be used by the dialogue engine as a confidence measure (to assure that the recognized sentence is reliable) and a pronunciation score (to allow different feedback to the learner depending on the pronunciation quality).

2.3. Dialogue engine

The dialogue engine processes lower-level outputs the recognizer and pronunciation scoring engine to define the flow of the dialogue. Feedback depend on these outputs and on other cumulative information defined during the dialogue.

We classify the different types of feedback into four categories: Correct (CO), Pre-defined Error (PE), Mispronunciation (MISP) and General Error (GE). This classification system should be applicable to most ASR-enabled language learning software. The definitions are given with examples of feedback for a translation from Chinese of: *ta feichang xihuan yingyu (He likes English very much).*

- Correct (CO): All recognized words are considered as correctly pronounced by the pronunciation scoring engine and the recognized sentence is among the expected correct inputs. For example: *Correct, he likes English very much.*
- Pre-defined Error (PE): All recognized words are considered as correctly pronounced by the pronunciation scoring engine and the recognized sentence is among the expected incorrect inputs. In this case, the system knows exactly which error was made and can give a very specific feedback. For Example: *You said "he very much likes English", you should have said "he likes English very much".*
- Mispronunciation (MISP): Some of the recognized words are considered as badly pronounced by the pronunciation scoring engine. In this case, the system detected an error and can point where it is but it can not give a specific feedback on the type of error though. For example: *Right, he likes English, but I did not understand the end of your sentence. "feichang" can be translated into "very much".*
- General Error (GE): Any other case, the input was too far from any of the expected answers: *I did not understand your sentence, try again.*

The dialogue is managed by a Finite State Machine: It is created by curriculum developers in an authoring tool developed by Saybot. Apart from the uttered sentence and associated scores, the flow of the dialogue can depend on a number of other parameters such as cumulative pronunciation or syntax-related scores, previous states in the dialogue and choices made by the learner during the dialogue.

[2] http://cmusphinx.sourceforge.net

3. Evaluation of Saybot Player

Evaluation is often seen as an important factor that drove the success of speech technologies. Applications such as transcription, language recognition or speaker recognition are clear tasks with metrics and publicly available evaluation corpus. Evaluation of a dialogue system is more complex and contingent on the evaluation of individual components with their associated models and content (for example curriculum design).

3.1. Recognizer

Evaluation of a speech recognizer is a well-known task with standard metrics, mainly the word error rate (WER) and sentence error rate (SER). We give the error rates for different recognizers, acoustic models and testing data in Table 2 using the following abbreviations:

- Recognizer: Proprietary (P), Sphinx 3 (S).
- Training data: Native adults (NA), native children (NC), native children adapted with non-native children voice (NCA), non-native adults (NNA), non-native children (NNC), mixed non-native adults and children (NNM). For the Sphinx recognizer, we also give different results depending on the type of Gaussian mixtures used in the acoustic model (continuous or semi-continuous) and the number of senones kept in the model.
- Testing data: Non-native adults (NNA), non-native children (NNC).

Native data was collected with native American English speakers reading aloud; non-native data come from sampled recordings of Saybot users. Both the adults' and children's testing corpora have around 3,000 sentences. Non-native adults training corpus has 60,000 sentences, non-native children training corpus has 25,000 sentences.

Table 2: *Error rates for Saybot users*

Recognizer	Training Data	Testing Data			
		NNC		NNA	
		WER	SER	WER	SER
S/semi/2000	NNM	13.2	25.4	9.2	19.9
S/semi/1000	NNM	13.4	25.7	9.1	20.1
S/cont	NNM	13.6	25.5	9.5	21.0
S/semi	NNC	13.4	25.4	×	×
S/cont	NNC	13.9	26.0	×	×
P	NC	17.9	31.8	×	×
P	NCA	14.8	26.6	×	×
S/semi	NNA	×	×	9.3	20.2
S/cont	NNA	×	×	9.7	21.2
P	NA	×	×	11.9	26.4

These results have been derived using grammars corresponding to each specific practice. It shows the performance of the player in a real situation. For development purposes, this protocol is not very satisfactory because for some of the recordings, the uttered sentence could not be realized by the grammar. It means that the ideal error rate can not be 0, and the room for improvement is difficult to extrapolate. So, we also evaluate the recognition using a statistical language model, trained on the whole transcribed corpus, including the testing corpus. The purpose of such an experiment is to measure the relative performance of different acoustic models. Table 3 gives an example of results.

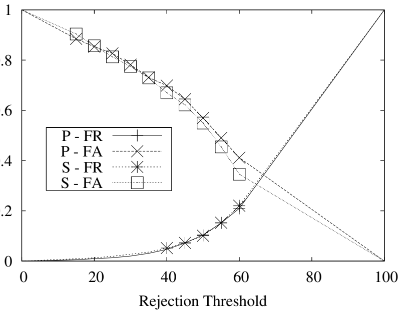

Figure 3: *Confidence Scoring Performance: ratio of false acceptance (FA) and false rejects (FR) for proprietary (P) and Sphinx (S) recognizers.*

Table 3: *Error rates, using statistical language modeling with Sphinx recognizer, on children and adults data.*

Test Set	WER	SER
NNC	28.6%	48.4%
NNA	12.1%	27.8%

3.2. Pronunciation Scoring

Although automatic pronunciation scoring has been studied for decades, the task is not clearly defined and there are no standard metrics or public evaluation database.

Most of the research conducted in this domain evaluates results with the correlation of machine scores with human scores, or by computing the ratio of incorrectly classified words or sentences.

At this point, we evaluate our pronunciation scoring by aligning the recognized sentences and their associated transcriptions and looking at the false acceptance and rejects. False Acceptance (FA) and False Reject (FR) are defined by:

$$FR = \frac{\#\, rejected\ words\ which\ are\ actually\ correct}{\#\, correct\ words} \qquad (1)$$

$$FA = \frac{\#\, accepted\ words\ which\ are\ actually\ wrong}{\#\, wrong\ words} \qquad (2)$$

Figure 3 gives these rates for different values of the rejection thresholds and the two recognizers (Proprietary, P, and Sphinx, S).

3.3. Dialogue Engine

Dialogues are based on instructions and feedback, and the quality (effectiveness) of a dialogue depends on how specific the received feedback are. Table 4 gives the ratio of each feedback type received by Saybot users. They are computed from 25,600 feedback received by around 3300 users (child users, using a player with an acoustic model trained on native kids and adapted with non-native data from Saybot users.)

Table 4: *Statistics of feedback: ratio of each feedback type actually received by Saybot users.*

Feedback Types	
Not Available (NA)	0.3%
Correct (CO)	52.1%
General error (GE)	38.4%
Pre-defined error (PE)	5.7%
Mispronunciation (MISP)	3.5%

The most valuable type is PE, where very specific feedback can be given to the learner after a mistake. The value of 5.7% is probably higher than what children can expect in a classroom in China. The least valuable type is GE, where no specific feedback on the utterance can be given. The value in this experiment can appear high, but it includes random answers made by learners, mostly unsupervised kids (practicing at home) and noisy environments. It also includes frequent poor hardware configurations for such software (low quality or mis-configured microphones and soundcards).

Apart from their type it is also important to measure how relevant feedback are. In previous experiments, we asked experts to listen to the recordings from users and mark whether the received feedback was relevant or not [7]. In this research, we proposed a more objective labeling: experts listening to the recordings select which of the pre-defined feedback would be relevant for that recording. Then, we check whether the given feedback was in the list of relevant ones. Results are computed on a subset of 2500 sentences from the previous set, labelled by two experts (or more, until at least two of them specify the same list of relevant feedback). They are given on Table 5.

Table 5: *Relevance of feedback received by users*

Feedback Relevance			
Relevant	Relevant	74.3%	75.2%
	MPE	0.9%	
Not Relevant	FA	7.6%	24.8%
	FR	15.6%	
	Conf.	0.7%	
	Comb.	0.1%	
	Irr. MISP	0.8%	

The meaning of the different cases are: Relevant when the received feedback is in the list of relevant feedback for that recording; Missed PE (MPE) when the user received a GE or MISP feedback, but one PE was defined for that input; False Acceptance (FA) when the sentence was accepted (CO) but it should have been rejected (PE, MISP or GE); False Reject (FR) if the sentence was rejected (PE or GE) but it should have been accepted (CO); Confusion (Conf.) if the feedback type was specific (PE, CO or MISP) but not pointing towards the correct mistake or correct answer; Combination (Comb.) for the case of several problems combined, mainly confusion between MISP feedback; Irrelevant MISP (Irr. MISP) if the specified mispronounced part was actually correctly pronounced.

Although almost 25% of the received feedback were not relevant, we can notice that most of the irrelevant feedback are false rejects, which we consider the less harmful for the learners: a very brief study of the relevance with native speakers using the player showed a relevance of almost 100%.

4. Summary and discussion

Although spoken dialogue systems for language learning and assessment show very promising results, they lack a method of objective evaluation and mainly suffer from the facts that:

1. Performance in terms of recognition is significantly lower for non-native speakers because of greater acoustic variability. It is even lower in the case of children speech, which is often the target of such systems.
2. Recognition performance is also affected by the fact that the language models include grammatically incorrect sentences. The cost of a confusion between two phonetically close sentences can be very high (can change a correct utterance into an incorrect one).
3. Defining and implementing a reliable pronunciation scoring paradigm is still under way.
4. Sentence transcription is not sufficient to evaluate the quality of pronunciation. Fluency and intonation can only be scored if the correct breaks and intonations are defined, and this is a very complex task.

We are currently working on these four issues.

5. Acknowledgment

Hoki Au, the main architect of Saybot Player, and Chris Yan, implemented most of the speech technology in Saybot Player. Alison McLaughlin and Tracy Li took care of an important part of the experiments presented in this paper.

6. References

[1] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, “Automatic text-independent pronunciation scoring of foreign language student speech,” in *Proc. ICSLP ’96*, vol. 3, Philadelphia, PA, 1996, pp. 1457–1460.

[2] S. Witt and S. Young, “Performance measures for phone-level pronunciation teaching in CALL,” in *Proc. of STiLL*, 1998, pp. 99–102.

[3] M. Eskenazi, “Using automatic speech processing for foreign language pronunciation tutoring some issues and a prototype,” *Language Learning & Technology*, vol. 2, no. 2, pp. 62–76, 1999.

[4] A. Raux and M. Eskenazi, “Using task-oriented spoken dialogue systems for language learning: Potential, practical applications and challenges,” in *Proc. of InSTIL/ICALL ’04*, Venice, 2004.

[5] S. Seneff, C. Wang, and J. Zhang, “Spoken conversational interaction for language learning,” in *Proc. of InSTIL/ICALL*, 2004.

[6] K. Zechner, D. Higgins, and X. Xi, “SpeechRaterTM: A construct-driven approach to scoring spontaneous non-native speech,” in *Proceedings of SLaTE ITWR Workshop*, 2007.

[7] S. Chevalier, “Speech interaction with Saybot player, a CALL software to help Chinese learners of English,” in *Proceedings of SLaTE ITWR Workshop*, 2007.

[8] J. Meron, A. Valente, and L. Johnson, “Improving the authoring of foreign language interactive lessons in the tactical language training system,” in *Proceedings of SLaTE ITWR Workshop*, 2007.

[9] J. Balogh, J. Bernstein, J. Cheng, and B. Townshend, “Automatic evaluation of reading accuracy: Assessing machine scores,” in *Proceedings of SLaTE ITWR Workshop*, 2007.

[10] M. Eskenazi, A. Kennedy, C. Ketchum, R. Olszewski, and G. Pelton, “The NativeaccentTM pronunciation tutor: measuring success in the real world,” in *Proceedings of SLaTE ITWR Workshop*, 2007.

[11] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari, “The SRI EduSpeak system: Recognition and pronunciation scoring for language learning,” in *Proc. of InSTIL, Scotland*, 2000, pp. 123–128.