Spoken Conversational Interaction for Language Learning

Stephanie Seneff, Chao Wang, and Julia Zhang
Spoken Language Systems Group
MIT Computer Science and Artificial Intelligence Laboratory
Stata Center, 32 Vassar street, Cambridge, MA 02139
{seneff,wangc,juliaz}@sls.csail.mit.edu

Abstract

This paper describes our efforts towards utilizing multilingual spoken dialogue systems as an aid to second language acquisition. We argue that it is important for language students to have the opportunity to practice communication in a non-threatening environment, something that a computer can naturally provide. We envision a three-stage interaction focused around a specific topic of a lesson plan. The first stage would familiarize the student with the vocabulary and syntax by presenting simulated dialogues at a Web page. The second stage would involve spoken dialogue interaction with the computer, either at a workstation or on the telephone. The third stage would provide feedback to the user on the quality of the utterances they recorded during the dialogue exchange. We have thus far concentrated on Mandarin and English as the two languages, where the system can be configured reversibly to support either learning English or Mandarin. We have begun to develop dialogue interaction capabilities for a number of domains centered around the scenario of a traveler to a foreign city.

1 Introduction

Over the past 15 years, members of the Spoken Language Systems group at MIT have been developing multilingual spoken conversational systems typically supporting access to on-line information sources (Zue and Glass, 2000). We have long believed that such systems could benefit foreign language learning, because conversational interaction is a critical part of acquiring fluency, and the computer provides an entertaining and non-threatening environment. The capabilities of our core technology components, including speech recognition and understanding, speech generation and synthesis, and discourse and dialogue modeling, are sufficiently mature now such that we feel the time is right to move towards turning our belief into a reality.

To this end, we are developing a Web-based interface for language learning, which we call SCILL (Spoken Conversational Interaction for Language Learning). It supports an envisioned three-stage activity associated with any particular lesson plan (Lau, 2002). The first stage involves examining a simulated dialogue on the topic of the lesson, as well as practicing speaking sentences within the dialogue. Every time a student visits the webpage, a new simulation dialogue is automatically generated, in both the native language and the target language. The system includes support for playing spoken examples of words and sentences from the simulated dialogue, produced by a speech synthesis system.

In the second stage, the student would engage in an interactive spoken conversation with the system, where, at any time during the dialogue, the student could speak a sentence in their native language to obtain a spoken translation, which they could then immediately repeat to push the conversation forward. Thus both a conversational partner and a translation assistant are accessible to the student during the conversation.

The third stage would involve Web-based analysis of their conversation, where the student could review their own speech and contrast it with native renderings of the same utterances. We envision that automatic scoring algorithms would provide feedback on their phonetic, prosodic, and syntactic productions.

We have thus far concentrated on the scenarios of a native English speaker learning Mandarin or a native Mandarin speaker learning English. We have been developing dialogue interaction in a number of domains, mostly related to travel, including scenarios involving booking a hotel or a flight (Wang et al., 1997), inquiring about the weather (Zue et al., 2000a; Zue et al., 2000b; Wang et al., 2000), making an acquaintance (Lau, 2002), navigation assistance in a city (Wang et al., 1997), or asking for a reminder as a call back at a designated time (Chuu, 2002).

The remainder of this paper is organized around the three stages of interaction depicted above. We first describe briefly the core technology that makes it possible to design dialogue systems for language learning. The next section introduces the idea of simulated dialogues in the domain of booking a hotel room. Section 4 illustrates the stage of spoken dialogue interaction within the weather domain, and elaborates on some of the underlying technology. In Section 5 we discuss the assessment stage, which is only in the early phase of development at this time. After a short section on related work, we conclude with a discussion of our future plans.

2 Core Technology

Dialogue Systems: Our dialogue systems are all

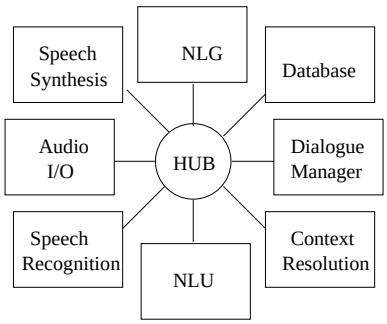

Figure 1: Block diagram of a spoken dialogue system.

developed based on the Galaxy Communicator architecture (Seneff et al., 1998). As illustrated in Figure 1, a central hub mediates interactions among a suite of technology and interface servers, and control flow is specified in a simple scripting language. We have found that it is relatively straightforward to manipulate the dialogue interaction to *respond to* queries spoken in the language being acquired and *translate* queries spoken in the native language. Meta questions such as "please translate that" are also redirected appropriately in the hub script.

Language Translation: Central to our research is the critical requirement for near-perfect translation between the native and second language, which, we believe, can be achieved given the restricted scope of the domain. We require not only translation from a *query* spoken in the native language (henceforth L1) to the acquired language (L2), but also of the *system responses* from L2 to L1. Our approach is based on an interlingual representation of both the input queries and output responses. Details of our translation framework are provided in (Wang and Seneff, 2004).

3 Simulated Dialogues

We believe it is important that the language learner experience variety in their exposure to language usage. Thus, we imagine that, every time they enter a Web page associated with a particular lesson, a new dialogue appears, which they would then be able to study in order to gain fluency in speaking about the topic of the lesson. Each utterance in the dialogue, both user and system side, would be associated with a synthesized waveform that could be played via a mouse click.

We have implemented this concept in the context of a hotel booking domain, with the goal of learning English. An example English dialogue is shown in Figure 2, and a block diagram of the framework that generates such a dialogue is shown in Figure 3. The simulated user is implemented as a separate server within the Galaxy framework, and it decides stochastically what to say on the basis of the system's previous reply. Its query is represented as a simple structured object with keys and associated values. GENESIS converts the query into well-formed strings, in both English and Mandarin. Variability is achieved in the string-generation stage through a random selection among alternative rules and lexical entries. A speech synthesizer then transforms the English strings into speech waveforms. The resulting simulated conversations exhibit variability both in the dialogue flow and in the linguistic con-

SU:	I'd like a room next Thursday and Friday
SYS:	I have found 48 rooms. I have rooms with double, queen, king or twin beds.
SU:	I'd like a queen bed.
SYS:	I have found 13 rooms with a queen bed. There are both smoking and non-smoking rooms.
SU:	How about a non-smoking room?
SYS:	I have found 7 non-smoking rooms with a queen bed. Many of them have a kitchen, a hot tub, internet access or a view.
SU:	I want a room with a view.
SYS:	I have found 4 nonsmoking rooms with a queen bed and a view. You have a choice of penthouse or room.
SU:	Standard room.
SYS:	How about room 210? It is a non-smoking room with a queen bed. It comes with a view. The room costs 115 dollars.
SU:	That would be fine.
SYS:	You have reserved a room for Thursday and Friday. It is a non-smoking room with a queen bed. It comes with a view. The room costs 115 dollars.
SU:	I'd like a wake up call tomorrow at 9 a.m.
SYS:	Okay, you'll get a wake up call at 9 a.m.
SU:	What time is the restaurant open for dinner?
SYS:	It is open for dinner from 5 p.m. to 10 p.m.
SU:	Is there a business center?
SYS:	I'm sorry, but the hotel does not have a business center.
SU:	Good bye now.
SYS:	Have a nice day!

Figure 2: Simulated dialogue in the hotel room domain. NOTE: SU = Simulated User; SYS = Dialogue system.

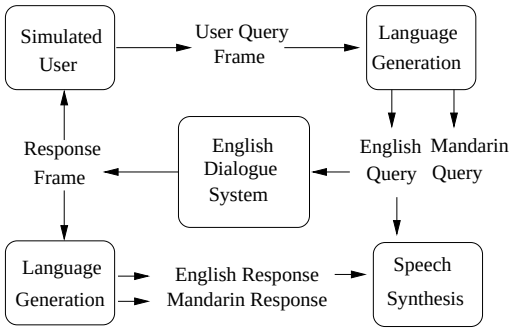

Figure 3: Procedure to generate simulated conversation.

User asks in Chinese:	bo1 shi4 dun4 ming2 tian1 hui4 xia4 yu3 ma5? (Is it going to rain in Boston tomorrow?)
System replies in Chinese:	tian1 qi4 yu4 bao4 ming2 tian1 bo1 shi4 dun4 mei2 you3 yu3. (The forecast calls for no rain in Boston tomorrow.)
User asks in English:	What is the temperature?
System translates into Chinese:	qi4 wen1 shi4 duo1 shao5?
User imitates in Chinese:	qi4 wen1 shi4 duo1 shao5?
System replies in Chinese:	bo1 shi4 dun4 ming2 tian1, zui4 gao1 qi4 wen1 er4 she4 shi4 du4, ming2 tian1 ye4 jian1, zui4 di1 qi4 wen1 ling2 xia4 wu3 she4 shi4 du4.
User requests translation:	Could you translate that?
System translates:	In Boston tomorrow, high 2 Celsius, tomorrow night, low -5 Celsius.

Figure 4: A spoken dialogue interaction between a user and the system, in the weather domain. The system is configured for a native English speaker to practice Chinese. Chinese inputs are represented using tone-marked pinyin. English translations are provided in parentheses for reference.

structs of the individual utterances. They are useful not only to the language student but also to the system, which can automatically harvest thousands of sentences for training the language models of the speech recognizer and natural language understanding (NLU) components from the simulated user data.

One issue is the quality of the synthetic speech. To address this, we have available to us a concatenative speech synthesizer called ENVOICE (Yi and Glass, 2002), which is capable of producing extremely high quality synthesis in restricted domains. It depends critically upon recorded speech for a set of utterances carefully selected to cover every word of the domain in every appropriate prosodic context. ENVOICE searches a weighted finite state network to find the optimal solution, where preference is given to matching context conditions.

For the hotel domain, we have created two voices for the user side of the conversation within ENVOICE. Responses are currently synthesized using the commercially available Dectalk speech synthesizer.

4 Conversational Interaction

Once the student feels prepared to converse with the system, they would attempt to carry on a conversation with the system, as illustrated in Figure 4. They would be able to either type the queries into a type-in window or speak them into a microphone or at a telephone. For typed input, they would use tone-marked pinyin (as illustrated in the figure). However, the system is very robust against tone errors. It achieves this robustness by building a word graph, expanding each syllable into every possible tone substitution, based on the system's vocabulary (Peabody et al., 2004). It parses the graph, following the rules of a stochastic context-free grammar, seeking the highest probability solution. The tone-corrected pinyin string is then displayed to the student.

To recognize spoken inputs, the system makes use of the SUMMIT landmark-based speech recognition system (Glass et al., 1996). A word graph produced by the recognizer is parsed by the NLU component into candidate semantic frames, and final selection is based mainly on a combination of word-based confidence scores (Hazen et al., 2002) and language model scores. One issue we faced was the problem of acquiring adequate training data for the language model statistics of both the recognizer and NLU systems. Our strategy was to utilize the existing language translation capability to acquire a Mandarin training corpus by translating a pre-existing English corpus. Details of this process can be found in (Wang and Seneff, 2004). We do not yet have in place a simulated user capability for the weather domain, although we could conceivably by-pass it by exploiting the thousands of real-user interactions we have from our Jupiter English-based system (Zue et al., 2000a), using our translation capabilities to map them to Mandarin.

In the dialogue interaction, whether spoken or typed, the user can switch freely between English and Mandarin. English queries are translated by the system, and Mandarin queries are answered. If the user does not understand the response, they can ask, in either English or Mandarin, for a translation, which will cause the system to speak the weather report in English.

5 Assessment

Once the user has completed their spoken dialogue, the system would be able to present it to them at a Web page. They would see the transcript of the interaction, with the hypothesized spoken words displayed individually as clickable icons. The recognizer retains timing information for each word, so that the user could play back their own words in isolation or as a complete utterance. We are hoping that word-based confidence scores will correlate with the quality of the user's pronunciations.

While we have yet to integrate assessment mechanisms into the SCILL framework, we have available to us some tools that we believe will be useful for this, particularly with respect to tone assessment. We have previously implemented a pitch detection

algorithm designed specifically for telephone quality speech (Wang and Seneff, 2000b), and have developed a tone recognition system based on parameterizing the pitch contour over a syllable final using an orthonormal decomposition of the F_0 contour (Wang and Seneff, 2000a). We achieved 18% error rate on tone classification for a digits task, applied to native speakers. We are hoping that this technology can be applied to the task of identifying incorrect tones, where the system would highlight the one or two syllables giving the worst scores for tone. Furthermore, a framework for speech transformations (Tang et al., 2001) can be adopted to correct the student's tones, while preserving the voice quality. Our plan is to use ENVOICE to synthesize the student's utterance, then warp the fundamental frequency contour so that it corresponds with the F_0 contour of the synthetic speech. Thus, students would be able to listen to speech that retains their own voice quality but has improved tone productions.

6 Related Work

There are many research projects that involve providing pronunciation training using a speech recognizer in a forced recognition mode (Dalby and Kewley-Port, 1999; Neri et al., 2001), etc., but few systems that allow the user to engage in some form of meaningful dialogue. Perhaps the most closely related research is the Fluency project at CMU, where carefully constructed questions are intended to solicit with extremely high likelihood a small number of possible answers. (Eskenazi, 1999) has emphasized the benefits of giving the student an *active* rather than a passive role in the exercise, in order to improve language retention.

Another example where students experience a limited but effective dialogue interaction is the multimodal microworld for teaching Arabic described in (Holland et al., 1998). An on-screen agent navigates a virtual space, and students can speak one of three sentence choices presented at each branch in a dialogue tree. An example sentence would be "open the drawer." The entire space encompassed a total of 72 Arabic sentences. This strategy severely limits the scope of the recognition task, but allows the student to experience a game-like interaction.

7 Summary and Future Work

This paper describes some of the research activities we are currently pursuing, aimed at the ambitious goal of providing spoken conversational interaction with a computer as an aid to second language acquisition. Our research is still in an early stage, in that we have not yet run user studies to see if the system is actually useful to language learners. In a companion paper (Peabody et al., 2004), we describe our first data collection efforts, involving students learning Mandarin, which includes a typed-input drill exercise, followed by solicited recordings of spoken queries in the weather domain.

8 Acknowledgement

This work was supported in part by the Cambridge MIT Institute and by ITRI Research Labs.

References

C. Chuu. 2002. LIESHOU: A mandarin conversational task agent for the galaxy-II architecture. *MIT MEng Thesis*.

J. Dalby and D. Kewley-Port. 1999. Explicit pronunciation training using automatic speech recognition technology. *CALICO Journal*, 16(3):425–445.

M. Eskenazi. 1999. Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning and Technolgy*, 2(2):62–76.

J. Glass, J. Chang, and M. McCandless. 1996. A probabilistic framework for feature-based speech recognition. In *Proc. ICSLP*, volume IV, pages 1–4, Philadelphia, PA.

T. J. Hazen, S. Seneff, and J. Polifroni. 2002. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech and Language*, 16:49–67.

V. M. Holland, J. D. Kaplan, and M. A. Sabol. 1998. Preliminary tests of language learning in a speech-interactive graphics microworld. *Calico Journal*, 16(3):339–359.

J. Lau. 2002. SLLS: An online conversational spoken language learning system. *MIT MEng Thesis*.

A. Neri, C. Cucchiarini, and H. Strik. 2001. Effective feedback on L2 pronunciation in ASR-based CALL. In *Proc. of the workshop on Computer Assisted Language Learning*, pages 40–48, San Antonio, Texas.

M. Peabody, S. Seneff, and C. Wang. 2004. Mandarin tone acquisition through typed dialogues. In *These Proceedings*.

S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue. 1998. Galaxy-II: A reference architecture for conversational system development. In *ICSLP '98*, pages 931–934, Sydney, Australia.

M. Tang, C. Wang, and S. Seneff. 2001. Voice transformations: from speech synthesis to mammalian vocalizations. In *Proc. Eurospeech*, pages 357–360, Aalborg, Denmark.

C. Wang and S. Seneff. 2000a. Improved tone recognition by normalizing for coarticulation and intonation effects. In *Proc. ICSLP*, Beijing, China.

C. Wang and S. Seneff. 2000b. Robust pitch tracking for prosodic modeling in telephone speech. In *Proc. ICASSP*, Istanbul, Turkey.

C. Wang and S. Seneff. 2004. High-quality speech translation for language learning. In *These Proceedings*.

C. Wang, J. Glass, H. Meng, J. Polifroni, S. Seneff, and V. Zue. 1997. YINHE: A mandarin chinese version of the GALAXY system. In *Proc. Eurospeech*, pages 351–354, Rhodes, Greece.

C. Wang, D. S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, and V. Zue. 2000. Muxing: A telephone-access mandarin conversational system. In *Proc. ICSLP*, volume II, pages 715–718, Beijing, China.

J. Yi and J. Glass. 2002. Information-theoretic criteria for unit selection synthesis. In *Proc. ICSLP*, pages 2617–2620, Denver, Colorado.

V. Zue and J. Glass. 2000. Conversational interfaces: Advances and challenges. In *Proc. IEEE, Special Issue on Spoken Language Processing*, volume 88.

V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington. 2000a. JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1).

V. Zue, S. Seneff, J. Polifroni, M. Nakano, Y. Minami, T.J. Hazen, and J. Glass. 2000b. From JUPITER to MOKUSEI: Multilingual conversational systems in the weather domain. In *Proc. MSC*, pages 1–6.