MILLA

Multimodal Interactive Language Learning Agent

Prof. Nick Campbell
Dr Joao Cabral
Dr Eamonn Kenny
Emer Gilmartin

1. Abstract

Learning a language involves the acquisition and integration of a range of skills. A human tutor aids learners by

- providing a framework of tasks suitable to the learner's needs
- continuously monitoring learner progress and adapting task content and delivery style
- providing a source of speaking practice and motivation

The goal of this project is to create a multimodal dialogue system which will provide some of the advantages of a human tutor which are not normally encountered in self-study material and systems.

We envisage a system which combines language learning tasks with a chat module (chatbot) providing fun practice in the target language. Speech recognition will be used to monitor learner pronunciation and provide error detection. We would incorporate multimodal sensors in this system to aid in monitoring learner interest and affect. The information gathered by these sensors along with ASR (automatic speech recognition) results can then be used to inform the virtual tutor's output and behavior, tailoring the activities to the individual learner's needs. We will provide a dialogue platform incorporating TTS (Text-To-Speech) synthesis, ASR (automatic speech recognition), cameras, and biosensors. We can also provide sample language activities but welcome participants' own ideas. The participants will work together on design, implementation, and testing of the system.

2. Project objectives

Languages cannot be taught – rather they are learned or acquired. Thus, the function of a learning environment is to provide a setting and materials in which a learner can most efficiently acquire communicative competence in the target language. Tutors select and mediate activities and provide scaffolding and monitoring for the learner.

This way, learner autonomy is fostered while learners are provided with suitable tasks to acquire all of the skills needed to successfully communicate in a new language [1].

This project involves modelling aspects of a language tutor and learning environment as a computer aided language learning (CALL) system. The system will be implemented as a spoken dialogue system with multimodal inputs. In this project the participants will be able to:

- Get experience in working with a variety of advanced input sensors, such as those provided by Kinect and arousal measurement sensor
- Learn how to use a combination of multiple input modalities to infer about behavior and affective state of the user
- Learn about spoken dialogue modelling, spoken interaction components (ASR and TTS), and synchronization of different components in a complex human-computer interaction system
- Design and implement individual CALL activities
- Design and implement a tuition manager to monitor and guide the user through multiple learning activities
- Learn to implement spoken dialogue agents
- Design and conduct and experiment for evaluation of a CALL system

Overall, the project will result in the implementation of several new language learning modules and a learner management system which can direct the learner to appropriate activities, monitor progress, and adapt ongoing activity depending on the learner's current state.

3. Background information

Computer assisted language learning (CALL) is used to create an artificial environment containing tasks and activities to help learners attain their goals of improving language skills. An excellent overview of uses of speech technology in language education is given by Eskenazi [2], covering the use of ASR and TTS to address specific tasks and implementations of complete tutoring systems. Ellis and Bogart outline theories of language education / second language acquisition (SLA) from the perspective of speech technology [3] while Chapelle provides an overview of speech technology in language learning from the perspective of language educators [4]. A broad introduction to spoken dialogue systems is given in Jokinen and McTear [5].

Language learning is an increasingly important area of human and commercial endeavour, and has been an early adopter of various technologies, with video and audio courses available since the early days of audiovisual technology. Increasing globalisation and migration coupled with the explosion in personal technology ownership have increased the need for well designed, pedagogically oriented CALL applications.

Many existing CALL activities provide learners with reading practice and listening comprehension to improve accuracy in syntax and vocabulary, rather like exercises in a textbook with speech added. Simple commercial pronunciation tutoring applications range from 'listen and repeat' exercises without feedback or with auto-feedback. On the other hand, in more sophisticated systems the learner's utterance is compared with the target and feedback is given on errors and strategies to correct those errors. Interesting examples of spoken production training based on speech technology where phoneme recognition is used to provide corrective feedback on learner input include CMU's Fluency, KTH's Arthur and Cabral et al's MySpeech [5]. Much effort has been put into creating speech activities which allow learners to engage in spoken interaction with a conversational partner, the most difficult competence for a learner to acquire independently, with attempts to provide practice in spoken conversation (or texted chat) using chatbot systems based on pattern matching (e.g. Pandorabots) [6] or statistically driven (e.g. Cleverbot) [7] architectures.

Dialog systems using text and later speech have been successfully used to tutor learners through a natural language interface in science and mathematics subjects, relevant paradigms are AutoTutor [8]–[10] and ITSPOKE [11]. In language learning, early systems such as VILTS presented tasks and activities based on different themes which were chosen by the user [12], while other systems concentrated on pronunciation training via a conversational interface [13]. The use of gamification in educational software is receiving a lot of attention as a method of increasing learner motivation. In this project, MILLA's existing core activities are expected to take advantage of gamification by building a scoring and user record system.

4. Detailed technical description

4.1. Overview of proposed system

Participants will design and develop the CALL system upon an existing interaction platform based on the Semaine platform (http://www.semaine-project.eu) and taking advantage of tools already provided by the team. Figure 1 shows the general block diagram of the system.

A user interface will be developed by participants for the learner to interact with the system, which includes an avatar and other GUI's associated with the various learning activities. The idea is for the avatar to act as the tutor. For example, she would login or register the user, suggest and mediate activities (e.g. pronunciation training tasks and chat), and monitor user progress. In order to perform these tasks through speech, the system will incorporate open source and ready to use ASR and TTS components. Biometric sensors and cameras will play a role on monitoring the learner's behavior and affect. Other types of input modalities such as the recorded speech itself could also be used to infer high-level information about the learner's state. The system will also

incorporate a dialogue manager component which participants will use to design and build the dialogue activities. The input data obtained during the interaction of the learner with the system will be stored in a database as well as other information resulting from the interaction such as the results obtained in the activities. This information can then be used in monitoring learner's progress within and across sessions. Finally, the interaction platform links all these components and enables the synchronisation of the flow of data and the decision about what processes to use in each stage.

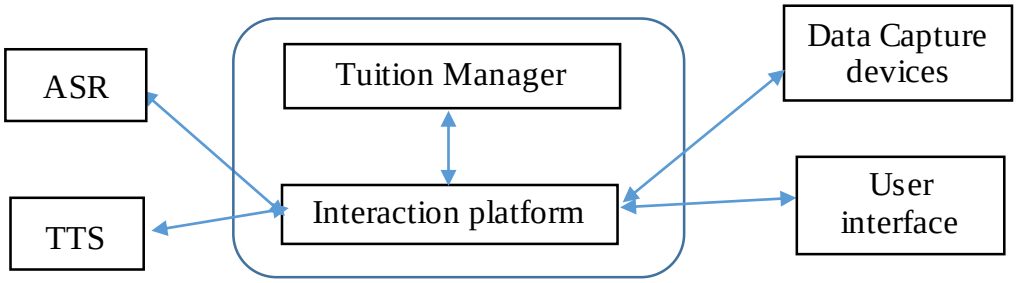

Figure 1 – General block diagram of the Milla system

Components Description:

- **Tuition manager**

  The Tuition Manager (TM) guides the learner using a spoken dialogue system and suggests learning activities based on their progress (chat is a possible activity) at any point. These activities would be also monitored by the avatar using dialogues. The leaner records include information about the learner affective state and user progress across sessions, such as frustration and failed answers respectively.

  Below is an example list of tasks of the TM:

  1. Greet learners and get their identity.
  2. Direct learner to specific interactive language learning activities.
  3. Continue to receive updates from learner record and perform any necessary actions when warranted (for example, if session time is up then inform user of that).
  4. Use learner record to trigger changing prompts, instructions, and explanations on return visits to activities. For example, the first time a user

attempts a particular type of learning task, instructions will necessarily be verbose, later on shorter prompts will be used.
5. Monitoring learner's state during activities and acting to avoid frustration, or to give hints.
6. Give progress report and sign learner out.

- **Interaction platform**

We encourage participants to use the Semaine platform as we can provide a working version with the basic functionalities and support on how to use the platform.

- Speaking avatar
- OpenSmile and OpenCV for multiple face detection with expression
- Active MQ Messaging service for synchronization of messages transmitted between components of the platform.
- Expressive speech synthesis using Voice XML as input to an expressive TTS (includes Mary TTS).
- Basic spoken dialogue using XML.

- **Data capture devices**

- Q sensor measures skin conductance, temperature, and motion to detect user engagement, stress or excitement.
- Kinect which includes webcam, microphone array, depth sensor, and software for gesture recognition, facial recognition and voice recognition (http://www.xbox.com/en-IE/Kinect).
- Leap Motion, a sensor for hand gesture (https://www.leapmotion.com)

- **ASR and TTS**

Ready-to-use open-source ASR and TTS systems will be provided. We use the Kaldi speech recognition toolkit and the HTS speech synthesis system. However, participants can use alternative systems such as Mary TTS.

4.2. Work plan and implementation schedule:

A tentative timetable detailing the work to be done during the workshop is given next.

Week 1: Familiarisation of participants with the hardware and software, set goals, discuss ideas, draft plans and get to know each other.

Week 2: Design and implementation of tutoring framework and the language learning activities - multilevel interaction with a tutor GUI/avatar coordinating and monitoring the various learning activities available. The team will provide existing activities to kickstart the system but participants will be actively encouraged to add modules of their own design.

Week 3: Design of user state monitor and adaptation of the system to the user's state using multimodal input devices and tools/software to infer information about user affect and engagement.

Week 4: Testing and evaluation - the multilingual environment at eNTERFACE should prove ideal for user evaluation.

4.3. Benefits of the research:

This project will enable participants to:

- Learn and gain practical experience on designing and building dialogue system modules, multimodal user interfaces and multimodal signal processing.
- Learn about CALL systems, gamification and gain experience in the design, implementation, and evaluation of CALL activities
- Have valuable experience in working on a multidisciplinary multinational team

4.4. Profile of team

Nick Campbell (nick@tcd,ie) is SFI Stokes Professor of Speech & Communication Technology at Trinity College Dublin (The University of Dublin) in Ireland. He received his Ph.D. degree in Experimental Psychology from the University of Sussex in the U.K., and was previously engaged at the Japanese National Institute of Information and Communications Technology, (as nick@nict.go.jp) and as Chief
Researcher in the Department of Acoustics and Speech Research, Advanced Telecommunications Research Institute International (as nick@atr.jp), Kyoto, Japan, where he also served as Research Director for the JST/CREST Expressive Speech Processing and the SCOPE "Robot's Ears" projects. He was first invited as a Research Fellow at the IBM U.K. Scientific Centre, where he developed algorithms for speech synthesis, and later at the AT&T Bell Laboratories, where he worked on the synthesis of Japanese. He served as Senior Linguist at the

Edinburgh University Centre for Speech Technology Research before joining ATR in 1990. His research interests are based on large speech databases, and include nonverbal speech processing, concatenative speech synthesis, and prosodic information modeling. He spends his spare time working with postgraduate students as Visiting Professor at the School of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan, and was also Visiting Professor at Kobe University, Kobe, Japan for 10 years.

Dr João P. Cabral is a Postdoctoral Researcher at Trinity College Dublin, as part of the Centre for Next Generation Localisation (CNGL). He also worked as a Postdoctoral Researcher at University College Dublin (UCD), as part of CNGL, from January 2010 to February 2013. He received the Ph.D. degree in Computer Science and Informatics at The University of Edinburgh, in 2010. He has a BSc and MSc from Instituto Superior Técnico (I.S.T.)/Technical University of Lisbon in Electrical and Computer Engineering. His main areas of expertise are text-to-speech synthesis and speech signal processing. His research interests also include machine learning, automatic speech recognition, glottal source modelling and Computer-Assisted Language Learning (CALL).

Dr Eamonn Kenny joined Trinity College Dublin as a research assistant in the FP4 EU project STORMS producing the propagation models for UMTS systems. In 2003 he completed a PhD in Telecommunications funded by the Enterprise Ireland Informatics Programme. From 2003-2010 he worked on the portability of the Grid middleware for the Large Hadron Collider in CERN. In 2007, he was invited to become the portability coordinator for the gLite middleware FP7 consortium in EGEE-II/III. From 2010-2013 he worked as the metrics task leader for the quality assurance team of the European Middleware Initiative. His main areas of research to date have been drug delivery problem solving and radio wave propagation solutions using numerical computation, algorithms and statistics. He is currently working on interaction systems for dialogue.

Emer Gilmartin is a Ph.D. candidate at the Speech Communication Lab at Trinity College Dublin. Her work is on modelling real human spoken interaction beyond the simplified task-based dialogues which have formed the basis for current dialogue technology. She holds degrees in Engineering (B.E.(Mech)), Linguistics (M.Phil.), and Speech and Language Processing (Post Grad. Dip.). She has twenty years of experience in provision of second language learning at all levels - teaching, teacher training, testing, and curriculum and materials design and distribution. She was the Executive Manager of IILT, Ireland's national programme for provision of language support to refugees, with direct involvement at national level in the development of language provision to migrants with all levels of language proficiency and needs ranging from basic literacy to language competence for professional or academic purposes.

References

[1] N. Garrett, 'Computer-Assisted Language Learning Trends and Issues Revisited: Integrating Innovation', *Mod. Lang. J.*, vol. 93, no. s1, pp. 719–740, 2009.

[2] M. Eskenazi, 'An overview of spoken language technology for education', *Speech Commun.*, vol. 51, no. 10, pp. 832–844, 2009.

[3] N. C. Ellis and P. S. Bogart, 'Speech and Language Technology in Education: the perspective from SLA research and practice', *Proc. ISCA ITRW SLaTE Farmington PA*, 2007.

[4] C. A. Chapelle, 'The Relationship Between Second Language Acquisition Theory and Computer-Assisted Language Learning', *Mod. Lang. J.*, vol. 93, no. s1, pp. 741–753, 2009.

[5] K. Jokinen and M. McTear, 'Spoken Dialogue Systems', *Synth. Lect. Hum. Lang. Technol.*, vol. 2, no. 1, pp. 1–151, 2009.

[6] M. Eskenazi and S. Hansma, 'The fluency pronunciation trainer', in *Proceedings of the STiLL Workshop*, 1998.

[7] B. Granström, 'Towards a virtual language tutor', in *InSTIL/ICALL Symposium 2004*, 2004.

[8] J. P. Cabral, M. Kane, Z. Ahmed, M. Abou-Zleikha, E. Székely, A. Zahra, K. U. Ogbureke, P. Cahill, J. Carson-Berndsen, and S. Schlögl, 'Rapidly Testing the Interaction Model of a Pronunciation Training System via Wizard-of-Oz.', in *LREC*, 2012, pp. 4136–4142.

[9] 'Pandorabots - A Multilingual Chatbot Hosting Service'. [Online]. Available: http://www.pandorabots.com/botmaster/en/home. [Accessed: 14-Jun-2011].

[10] 'Cleverbot.com - a clever bot - speak to an AI with some Actual Intelligence?' [Online]. Available: http://www.cleverbot.com/. [Accessed: 18-Apr-2013].

[11] A. C. Graesser, S. Lu, G. T. Jackson, H. H. Mitchell, M. Ventura, A. Olney, and M. M. Louwerse, 'AutoTutor: A tutor with dialogue in natural language', *Behav. Res. Methods Instrum. Comput.*, vol. 36, no. 2, pp. 180–192, 2004.

[12] D. J. Litman and S. Silliman, 'ITSPOKE: An intelligent tutoring spoken dialogue system', in *Demonstration Papers at HLT-NAACL 2004*, 2004, pp. 5–8.

[13] M. E. Rypa and P. Price, 'VILTS: A tale of two technologies', *Calico J.*, vol. 16, no. 3, pp. 385–404, 1999.