MILLA – Multimodal Interactive Language Learning Agent

João Paulo Cabral[1], Nick Campbell[1], Shree Ganesh[2], Emer Gilmartin[1], Fasih Haider[1], Eamonn Kenny[1], Mina Kheirkhah[3], Andrew Murphy[1], Neasa Ní Chiaráin[1], Thomas Pellegrini[4], Odei Rey Orozko[5]

Trinity College Dublin, Ireland[1]; GCDH-University of Goettingen, Germany[2]; Institute for Advanced Studies in Basic Sciences, Zanjan, Iran[3]; Université de Toulouse ; IRIT, France[4]; Universidad del País Vasco, Bilbao, Spain[5]

1 Background

Learning a new language involves the acquisition and integration of a range of skills. A human tutor aids learners by (i) providing tasks suitable to the learner's needs, (ii) monitoring progress and adapting task content and delivery style, and (iii) providing a source of speaking practice and motivation. With the advent of audiovisual technology and the communicative paradigm in language pedagogy, focus has shifted from written grammar and translation to developing communicative competence in listening and spoken production. The Common European Framework of Reference for Language Learning and Teaching (CEFR) recently added a more integrative fifth skill – spoken interaction - to the traditional four skills – reading and listening, and writing and speaking (Little, 2006) . While second languages have always been learned conversationally with negotiation of meaning between speakers of different languages sharing living or working environments, these methods did not figure in formal (funded) settings. However, with increased mobility and globalisation, many learners now need language as a practical tool rather than simply as an academic achievement (Gilmartin, 2008). Developments in Computer Assisted Language Learning (CALL) have resulted in free and commercial language learning material for autonomous study. Much of this material transfers well-established text and audiovisual exercises to the computer screen. These resources greatly help develop discrete skills, but the challenge of providing tuition and practice in the 'fifth skill', spoken interaction, remains. MILLA, developed at the 2014 eNTERFACE workshop in Bilbao is a multimodal spoken dialogue system combining custom modules with existing web resources in a balanced curriculum, and, by integrating spoken dialogue, modelling some of the advantages of a human tutor.

2 MILLA System Components

Tuition Manager: MILLA's spoken dialogue Tuition Manager (Figure 1) consults a two-level curriculum of language learning tasks, a learner record and learner state module to greet and enroll learners, offer language learning submodules, provide feedback, and monitor user state with Kinect sensors. All of the tuition manager's interaction with the user can be performed using speech through a Cereproc Text-to-Speech (TTS) voice and Cereproc's Python SDK (Cereproc, 2014), and understanding via CMU's Sphinx4 ASR (Walker et al., 2004) through custom Python bindings using W3C compliant Java Speech Format Grammars.

Tasks include spoken dialogue practice with two chatbots, first language (L1) focused and general pronunciation, and grammar and vocabulary exercises. Several speech recognition (ASR) engines (HTK, Google Speech) and text-to speech (TTS) voices (Mac and Windows system voices, Google Speech) are used in the modules to meet the demands of particular tasks and to provide a cast of voice characters which provide a variety of speech models to the learner. Microsoft's Kinect SDK ('Kinect for Windows SDK', 2014) is used for gesture recognition and as a platform for affect recognition. The tuition manager and all interfaces are written in Python 2.7, with additional C#, Javascript, Java, and Bash coding in the Kinect, chat, Sphinx4, and pronunciation elements. For rapid prototyping the dialogue modules were first written in VoiceXML, then ported to Python modules.

Pronunciation Tuition: MILLA incorporates two pronunciation modules, based on comparison of learner production with model production using the Goodness of Pronunciation (GOP) algorithm (Witt & Young, 2000). GOP scoring involves two phases: 1) a free phone loop recognition phase which determines the most likely phone sequence given the input speech without

giving the ASR any information about the target sentence, and 2) a forced alignment phase which provides the ASR with the orthographic transcription and force aligns the speech signal with the expected phone sequence. Comparison of the log-likelihoods of the forced alignment and free recognition phases produces a GOP score.

The first module is a focused pronunciation tutor using HTK ASR with the five-state 32 Gaussian mixture monophone acoustic models provided with the Penn Aligner toolkit (Young, n.d.; Yuan & Liberman, 2008) on the system's local machine. In this module, phone specific threshold scores were set by artificially inserting errors in the pronunciation lexicon and running the algorithm on native recordings, as in (Kanters, Cucchiarini, & Strik, 2009). After preliminary tests, we constrained the free phone loop recogniser for more robust behavior, using phone confusions common in specific L1's to define constrained phone grammars. A database of common errors in several L1s with test utterances was built into the curriculum.

The second module, MySpeech, is a phrase level trainer hosted on University College Dublin's cluster and accessed by the system via Internet (Cabral et al., 2012). It tests pronunciation at several difficulty levels as described in (Kane & Carson-Berndsen, 2011). Difficulty levels are introduced by incorporating Broad Phonetic Groups (BPGs) to cluster similar phones. A BFG consists of phones that share similar articulatory feature information, for example plosives and fricatives. There are three difficulty levels in the MySpeech system: easy, medium and hard – the easiest level includes a greater number of BPGs in comparison to the harder levels. The MySpeech web interface consists of several numbered panels for the users to select sentences and practice their pronunciation by listening to the selected sentence spoken by a native speaker and record their own version of the same sentence. Finally, the results panel shows the detected mispronunciation errors of a submitted utterance using darker colours.

Spoken Interaction Tuition (Chat): To provide spoken interaction practice, MILLA sends the user to Michael (Level1) or Susan (Level 2), two chatbots created using the Pandorabots web-based chatbot hosting service (Wallace, 2003). The bots were first implemented in text-to-text form in AIML (Artificial Intelligence Markup Language) and then TTS and ASR were added through the Web Speech API, conforming to W3C standards (W3C, 2014). Based on consultation with language teachers and learners, the system allows users to speak to the bot, or type chat responses. A chat log was also implemented in the interface, allowing the user to read back or replay previous interactions.

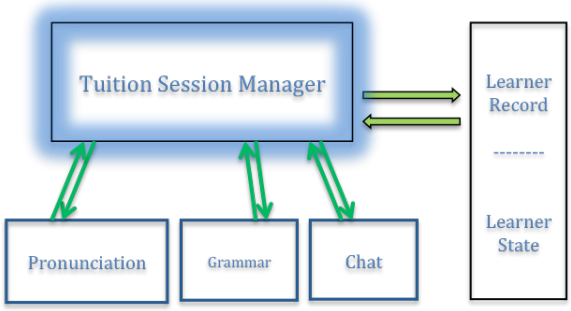

Figure 1 MILLA Overview

Grammar, Vocabulary and External Resources: MILLA's curriculum includes a number of graded activities from the OUP's English File and the British Council's Learn English websites. Wherever possible the system scrapes any scores returned for exercises and incorporates them into the learner's record, while in other cases the progression and scoring system includes a time required to be spent on the exercises before the user progresses to the next exercises. There are also custom morphology and syntax exercises created using Voxeo Prophecy to be ported to MILLA.

User State and Gesture Recognition: MILLA includes a learner state module to eventually infer learner boredom or involvement. As a first pass, gestures indicating various commands were designed and incorporated into the system using Microsoft's Kinect SDK. The current implementation comprises four gestures (Stop, I don't know, Swipe Left/Right), which were designed by tracking the skeletal movements involved and extracting joint coordinates on the x, y, and z planes to train the recognition process. Python's socket programming modules were used to communicate between the Windows machine running the Kinect and the Mac laptop hosting MILLA.

3 Future work

MILLA is an ongoing project. In particular, work is in progress to add a Graphical User Interface and avatar to provide a more immersive version and several new modules are planned. User trials are planned for the academic year 2014-15 in several centres providing language training to immigrants in Ireland.

Acknowledgments

João Paulo Cabral and Eamonn Kenny are supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL (www.cngl.ie) at Trinity College Dublin.Sree Ganesh is supported by GCDH-University of Goettingen. Emer Gilmartin is supported by the Science Foundation Ireland Fastnet Project, grant 09/IN.1/I2631. Neasa Ní Chiaráin and Andrew Murphy are supported by the ABAIR Project, funded by the Irish Government's Department of Arts, Heritage, and the Gaeltacht.

References

Cabral, J. P., Kane, M., Ahmed, Z., Abou-Zleikha, M., Székely, E., Zahra, A., … Schlögl, S. (2012). Rapidly Testing the Interaction Model of a Pronunciation Training System via Wizard-of-Oz. In *LREC* (pp. 4136–4142).

CereVoice Engine Text-to-Speech SDK | CereProc Text-to-Speech. (2014). Retrieved 7 July 2014, from https://www.cereproc.com/en/products/sdk

Gilmartin, E. (2008). Language Training for Adult Refugees: The Integrate Ireland Experience. *Adult Learner: The Irish Journal of Adult and Community Education*, *97*, 110.

Kane, M., & Carson-Berndsen, J. (2011). Multiple source phoneme recognition aided by articulatory features. In *Modern Approaches in Applied Intelligence* (pp. 426–435). Springer.

Kanters, S., Cucchiarini, C., & Strik, H. (2009). The goodness of pronunciation algorithm: a detailed performance study. In *SLaTE* (pp. 49–52).

Kinect for Windows SDK. (2014). Retrieved 7 July 2014, from http://msdn.microsoft.com/en-us/library/hh855347.aspx

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching*, *39*(03), 167–190.

W3C. (2014). *Web Speech API Specification*. Retrieved 7 July 2014, from https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., … Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition.

Wallace, R. S. (2003). *Be Your Own Botmaster: The Step By Step Guide to Creating, Hosting and Selling Your Own AI Chat Bot On Pandorabots*. ALICE AI foundations, Incorporated.

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, *30*(2), 95–108.

Young, S. (n.d.). *HTK Speech Recognition Toolkit*. Retrieved 7 July 2014, from http://htk.eng.cam.ac.uk/

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, *123*(5), 3878.