DEMO - MILLA – A Multimodal Interactive Language Agent

Joao Cabral[1], Nick Campbell[1], Shree Ganesh[2], Emer Gilmartin[1], Fasih Haider[1], Eamon Kenny[1], Mina Kheirkhah[3], Andy Murphy[1], Neasa Ní Chiaráin[1], Thomas Pellegrini[4], Odei Rey Orosko[5]

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

Abstract

We describe the motivation behind, design, and implementation of MILLA, a prototype speech-to-speech English language tutoring system.

1 Background

Learning a new language involves the acquisition and integration of a range of skills. A human tutor aids learners by (i) providing a framework of tasks suitable to the learner's needs, (ii) monitoring learner progress and adapting task content and delivery style to suit, 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 activities focusing more on communicative competence in listening and spoken production. In recent years the Common European Framework of Reference for Language Learning and Teaching (CEFR) officially 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 interactively through 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 formal learners now need language as a practical tool for everyday life and business rather than simply as an academic achievement. Developments in Computer Assisted Language Learning (CALL) have resulted in a range of free and commercial language learning material for autonomous study. Much of this material transfers long-established paper-based and audiovisual exercises to the computer screen. Pronunciation training exercises have been developed which provide feedback either through the learner listening back to their own efforts and comparing to a model, or the system providing a score or other feedback. These resources are very useful in development of discrete skills, but the challenge of providing spoken interaction tuition and practice remains. The MILLA system, developed at the 2014 eNTERFACE workshop ('The 10th International Summer Workshop on Multimodal Interfaces - eNTERFACE'14 ISCA Training School', 2014) 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

MILLA's spoken dialogue Tuition Manager consults a two-level curriculum of language learning tasks, a learner record, and a learner state module to greet and enroll learners, direct them to language learning submodules, provide feedback and scoring, 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 TTS voice using Cereproc's Python SDK ('CereVoice Engine Text-to-Speech SDK | CereProc Text-to-Speech', 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 different chatbots, first language (L1) focused and general pronunciation training, 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 incorporated 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 2.6, with additional C#, Javascript, Java, and Bash coding in the Kinect, chat, Sphinx4, and pronunciation elements. For rapid prototyping many of the dialogue modules were first written in VoiceXML, and then ported to Python modules.

2.1 Pronunciation Tuition

MILLA incorporates two pronunciation modules, based on comparison of learner production with model production: (i) 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 and (ii) MySpeech a phrase level trainer hosted on University College Dublin's cluster and accessed by the system via Internet (Cabral et al., 2012).

For the focused pronunciation system, we used the baseline implementation of the Goodness of Pronunciation 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 system with the orthographic transcription of the input sentence and force aligns the speech signal with the expected phone sequence. For each phone realization aligned to the speech signal, comparison of the log-likelihoods of the forced alignment and thevfree recognition phases, produces a GOP score where zero reflects a perfect match and increasing positive scores correspond to inaccuracies. Phone specific threshold scores were set to decide whether a phone was mispronounced (``rejected") or not (``accepted"), by artificially inserting errors in the pronunciation lexicon and running the algorithm on native recordings, as in (Kanters, Cucchiarini, & Strik, 2009). After preliminary testing, 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 module.

2.2 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 . 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 directly to the chat bot, or enter chat responses using text input. A chat log was also implemented into the interface, allowing the user to read back or replay several of their previous interactions with the chat bot.

2.3 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 (REF). 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 a number of custom morphology and syntax exercises designed for MILLA using Voxeo's Prophecy platform and VoiceXML which will be ported to MILLA in the near future.

2.4 User State and Gesture Recognition

MILLA includes a learner state module which will eventually infer boredom or involvement in the learner. 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. As MILLA is multiplatform (Unix and Windows), Python's socket programming modules was used to communicate between the Windows machine running the Kinect and the Mac laptop hosting MILLA.

3 Future work

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

Acknowledgments

Do not number the acknowledgment section. Do not include this section when submitting your paper for review.

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

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.

The 10th International Summer Workshop on Multimodal Interfaces - eNTERFACE'14 ISCA Training School. (2014). Retrieved 7 July 2014, from http://aholab.ehu.es/eNTERFACE14/

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. Retrieved from http://dl.acm.org/citation.cfm?id=1698193

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, *30*(2), 95–108. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167639399000448

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.