MILLA – Multimodal Interactive Language Learning Agent

João P. Cabral, Nick Campbell, Shree Ganesh, Emer Gilmartin, Fasih Haider, Eamonn Kenny, Mina Kheirkhah, Andrew Murphy, Neasa Ní Chiaráin, Thomas Pellegrini and Odei Rey Orozko

Abstract—**The goal of this project was to create a multimodal dialogue system which provides some of the advantages of a human tutor, not normally encountered in self-study material and systems. 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**

MILLA is a prototype language tuition system comprising tuition management, learner state monitoring, and an adaptable curriculum, all mediated through speech. The system enrols and monitors learners via a spoken dialogue interface, provides pronunciation practice and automatic error correction in two modalities, grammar exercises, and two custom speech-to-speech chatbots for spoken interaction practice. The focus on speech in the tutor's output and in the learning modules addresses the current deficit in spoken interaction practice in Computer Aided Language Learning (CALL) applications, with different text-to-speech (TTS) voices used to provide a variety of speech models across the different modules. The system monitors learner engagement using Kinect sensors and checks pronunciation and responds to dialogue using automatic speech recognition (ASR). A learner record is used in conjunction with the curriculum to provide activities relevant to the learner's current abilities and first language, and to monitor and record progress.

Index Terms—**language learning, CALL, spoken dialogue system.**

I. Introduction

Language learning is an increasingly important area of human and commercial endeavour as increasing globalisation and migration coupled with the explosion in personal technology ownership expand the need for well designed, pedagogically oriented language learning applications.

While second languages have long been learned conversationally with negotiation of meaning between speakers of different languages sharing living or working environments, these methods did not figure in formal settings. In contrast, traditional formal language instruction followed a grammar-translation paradigm, based largely on the written word. The advent of more communicative methodologies in tandem with increased access to audio-visual media in the target language had led to much greater emphasis on use of the language in both the spoken and written forms. 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 [1]. The nature of language curricula is also undergoing change as, with increased mobility and globalisation, many learners now need language as a practical tool rather than simply as an academic achievement [2].

The language learning sector has been an early adopter of various technologies, with video and audio courses available since the early days of audiovisual technology, and developments in Computer Assisted Language Learning (CALL) have resulted in freely available and commercial language learning material for autonomous study. Much of this material provides learners with reading practice and listening comprehension to improve accuracy in syntax and vocabulary, rather like exercises in a textbook with speech added. These resources greatly help develop discrete skills, but the challenge of providing tuition and practice in the ever more vital "fifth skill", spoken interaction, remains.

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) [3] or statistically driven (e.g. Cleverbot) [4] architectures.

An excellent overview of uses of speech technology in language education is given by [5], covering the use of ASR and TTS to address specific tasks and implementations of complete tutoring systems. Ellis and Bogart [9] outline theories of language education / second language acquisition (SLA) from the perspective of speech technology, while Chapelle provides an overview of speech technology in language learning from the perspective of language educators [10]. Simple commercial pronunciation tutoring applications range from "listen and repeat" exercises without feedback or with auto-feedback. In more sophisticated systems the learner's utterance is recorded and compared with a target or model, and then feedback is given on errors and strategies to correct those errors. Interesting examples of spoken production training systems based on speech technology, where phoneme recognition is used to provide corrective feedback on learner input, include CMU's Fluency [6], KTH's Arthur [7] and MySpeech [8].

Dialog systems using text and later speech have been

João P. Cabral, Nick campbell, Emer Gilmartin, Fasih Haider, Eamonn Kenny, Andrew Murphy and Neasa Ní Chiaráin are with Trinity College Dublin, Ireland

Mina Kheirkhah is with Institute for Advanced Studies in Basic Sciences, Zanjan, Iran

Shree Ganesh is with University of Goettingen

Thomas Pellegrini is with Université Toulouse, France

Odei Rey Orozko is with University of the Basque Country, Bilbao, Spain

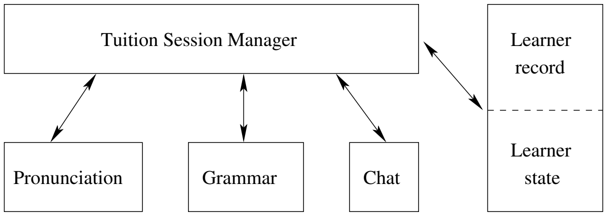


Fig. 1: General architecture of the MILLA system.

successfully used to tutor learners through a natural language interface in science and mathematics subjects. For example, relevant paradigms are the AutoTutor [11] and ITSPOKE [12] systems. In language learning, early systems such as VILTS [13] presented tasks and activities based on different themes which were chosen by the user, while other systems concentrated on pronunciation training via a conversational interface [7].

The MILLA system developed in this project is a multimodal spoken dialogue system combining custom language learning modules with other existing web resources in a balanced curriculum, and offering some of the advantages of a human tutor by integrating spoken dialogue both in the user interface and within specific modules.

II. MILLA System Overview

Figure 1 shows the general architecture of the MILLA system. MILLA's spoken dialogue Tuition Manager (Figure 1) consults a curriculum of language learning tasks, a learner record and learner state module to greet and enrol learners. It also offers language learning submodules, provides feedback, and monitors user state. The tasks comprise spoken dialogue practice with two chatbots, general and focussed pronunciation practice, grammar and vocabulary exercises. All of the tuition manager's interaction with the user can be performed using speech and gestures.

The tuition manager and all interfaces are written in Python 2.6, with additional C#, Javascript, Java, and Bash in the Kinect, chat, Sphinx4, and pronunciation elements respectively.

III. Tuition Management

MILLA's spoken dialogue Tuition Session Manager is a Spoken Dialog System (SDS) that guides the user through the system. The SDS is rule-based, i.e. depending on the answer of the user, the system provides one answer or another. As shown in Figure 2 the Tuition Session Manager first welcomes the user and checks if they already have an account. If the user does not have an account, the system offers to create one. Then, the system consults the curriculum of language learning tasks, the learner record and learner state associated to the user. The way the Tuition Manager updates the learner record is explained in SectionV. The user is asked to choose a language learning submodule and she is redirected to the selected learning module. Meanwhile, the Tuition Manager monitors the user state so that it can offer another alternative tasks ifsignals of frustration or lack of interest are detected by the system (as planned for a future version). The way the Tuition Manager monitors the user state is explained in Section IV.

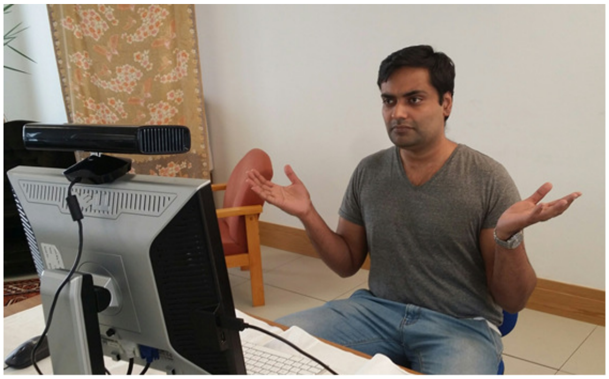


Fig. 3: Example of a gesture ("I don't know") which is detected by kinect in the learner state monitor module.

Spoken interaction with the user is performed through TTS and ASR. The first is implemented using the Cereproc's Python SDK [15], while the second is based on the CMU's Sphinx4 ASR [16] through custom Python bindings using W3C compliant Java Speech Format Grammars. During the design phase the dialogue modules were first written in VoiceXML for rapid prototyping purposes, and then ported to Python.

IV. Learner State Monitor

Microsoft's Kinect SDK [17] is used for 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" and "Swipe Right". Figure 3 shows a snapshot of the "I don't know" gesture. They were modelled by tracking the skeletal movements associated with these gestures and extracting joint coordinates on the x, y, and z planes to train the gesture classifier. Python's socket programming modules were used to communicate between the Windows machine running the Kinect and the Mac laptop hosting MILLA.

V. Learner progression - Curriculum and Learner Record

MILLA creates a learner record for each user which is used in conjunction with the curriculum and user state model to guide users to relevant activities, monitor in-session performance and learner state through the predefined gestures or any other infered information such as boredom and frustration. It also record the learner's progress along the curriculum. The curriculum consists of activities for each of the modules tagged with level, first language suitability, duration, and any other information needed to run the activity. As an example, there is a curriculum entry for a focussed pronunciation activity based on the difference between the [ɪ] and [iː] sounds in "fit" and "feet" respectively. It contains information on the

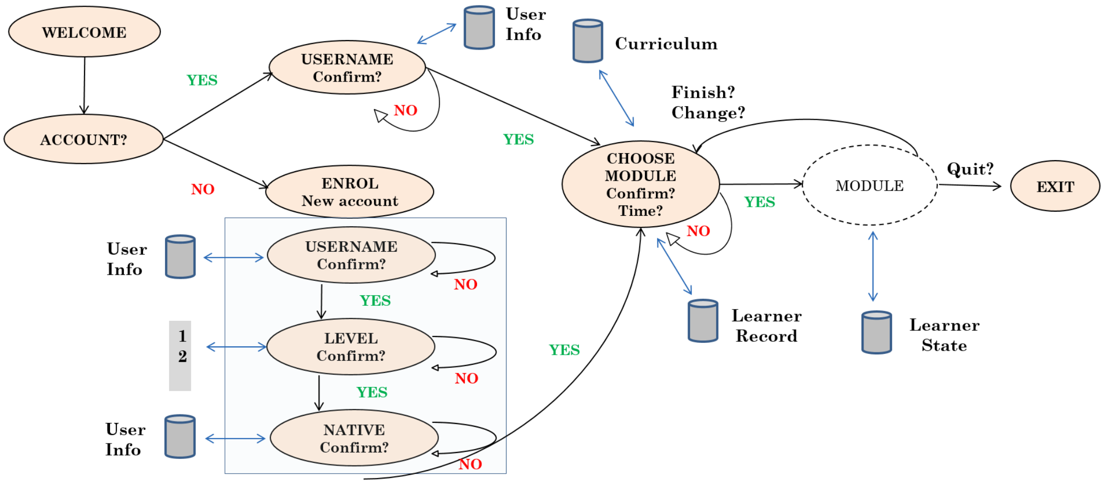


Fig. 2: Diagram of the tuition manager component.

sentence used in this exercise ("These shoes fit my feet"), including suitable explanatory graphics and tips to display on the user interface, level, and the first languages where these sounds are frequently confused. It also contains automatically extracted phonetic renderings of the target phrase and an "erroneous" version for use as parameters in the focussed pronunciation module. For the general pronunciation, chat and grammar modules, which are web-based, the system stores the relevant *urls* for different levels and activities plus the score or accumulated time needed to progress to the next activity or level. This abstraction will facilitate curriculum authoring and editing in the future.

When a learner logs on to the system and chooses their learning module, the learner record is queried to check the user's level, first language, and which activities have been completed. The curriculum is then searched for the next relevant activities in the module, and the learner is directed to suitable activities.

When the module progression is based on time accumulated, the system allows the user to choose how long they will stay on the activity. On completion of the activity the system updates the learner record and prompts the user to choose a new activity or quit. The curriculum and learner records are currently stored as JSON lists. The plan is to port them to an SQL database as the system develops.

VI. Pronunciation Tuition

MILLA incorporates two pronunciation modules. They are both based on comparison of learner production with model production using the Goodness of Pronunciation (GOP) algorithm [18]. However one is first language (L1) focused by taking into account common pronunciation errors from L1 learners, whereas the other provides general pronunciation error feedback independently of L1.

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 phonetic transcription and force aligns the speech signal with the expected phone sequence. Then, the GOP score is computed by comparison of the log-likelihoods obtained from the forced alignment and free recognition phases.

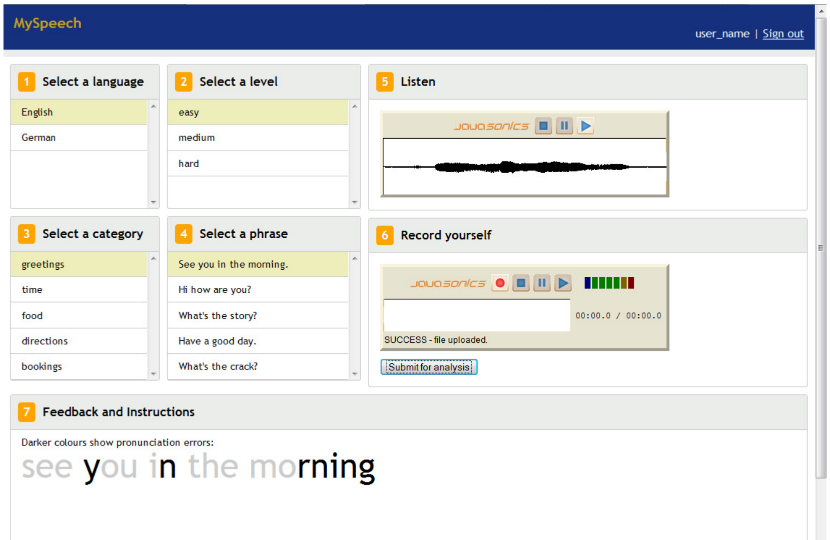
A. Focussed tuition based on common L1 specific errors

The first module was implemented using the HTK toolkit [19] and is defined by five-state 32 Gaussian mixture monophone acoustic models provided with the Penn Aligner toolkit [20], [21]. In this module, phone specific threshold scores were derived by artificially inserting errors in the pronunciation lexicon and running the algorithm on native recordings, as in [22]. After preliminary tests, we constrained the free phone loop recogniser for more robust behaviour, using phone confusions common in specific L1's to define constrained phone grammars. A database of utterances with common errors in several L1's was built into the curriculum (for the learner to practice), so that the system offers relevant pronunciation training based on the learner's first language, which is obtained from the learner record.

B. General Phrase Level Tuition

The second pronunciation training module is a phrase level trainer which is accessed by MILLA via the MySpeech web service [8]. It tests pronunciation at several difficulty levels as described in [23]. 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.

Figure 4 shows the web interface of MySpeech. It 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.

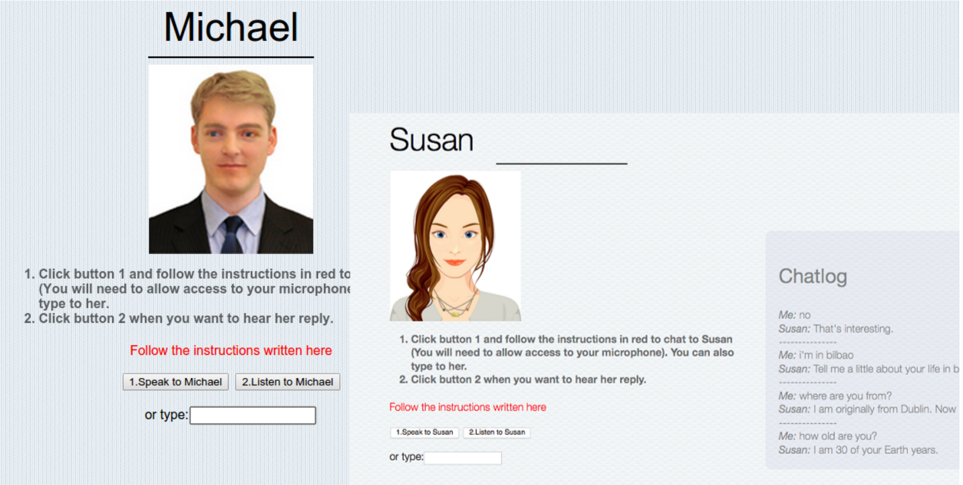


SUCCESS - file uploaded.

Darker colours show pronunciation errors:

see you in the morning

Fig. 4: Web interface of the MySpeech system for pronunciation training.



1. Click button 1 and follow the instructions in red to
(You will need to allow access to your microphone
type to her.
2. Click button 2 when you want to hear her reply.

Follow the instructions written here

1. Click button 1 and follow the instructions in red to chat to Susan (You will need to allow access to your microphone). You can also type to her.
2. Click button 2 when you want to hear her reply.

Follow the instructions written here

Me: no
Susan: That's interesting.

Me: i'm in bilbao
Susan: Tell me a little about your life in b

Me: where are you from?
Susan: I am originally from Dublin. Now

Me: how old are you?
Susan: I am 30 of your Earth years.

Fig. 5: Web interfaces for the "Michael" and "Susan" chatbots.

VII. Spoken Interaction Tuition (Chat)

In order to provide spoken interaction practice, MILLA sends the user either to Michael (Level 1) or to Susan (Level 2), two chatbots created using the Pandorabots web-based chatbot hosting service [24]. Figure 5 shows the web interface for the chatbot service with Michael and Susan.

In this work, these bots were first implemented in text-to-text form in AIML (Artificial Intelligence Markup Language). Then, TTS and ASR were added through the Web Speech API, conforming to W3C standards [25]. The system design is based on previous research in the field of CALL as well as consultation with language teachers and learners [26]. The system allows users either to speak to the bot, or to type chat questions/responses. The user receives text feedback from the bot and can also listen to these utterances pronounced in the accent of the bot (Michael: British-English and Susan: American-English). A chat log was also implemented in the interface, allowing the user to read back or replay previous interactions.

VIII. 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 by these web services 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 (as explained in Section V). During the project custom morphology and syntax exercises created using VoiceXML, which will be ported to MILLA.

IX. Future Work

MILLA is an ongoing project. In particular work in progress includes the development of a Graphical User Interface and avatar to provide a more immersive version. We also have a plan to incorporate several new modules into MILLA. Finally, user trials are planned in several centres providing language training to immigrants in Ireland.

ACKNOWLEDGMENT

João P. 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. Thomas Pellegrinis participation was partially supported by an IDEX "Nouveaux Entrants" fellowship. Fasih Haider is supported by the METALOGUE (Multiperspective Multimodal Dialogue: dialogue system with metacognitive abilities) project under FP7-ICT, grant 611073. For details on METALOGUE visit http://cordis.europa.eu/project/rcn/110655_en.html. 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. We would like to thank the eNTERFACE organisers for providing a richly collaborative setting for this work.

REFERENCES

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

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

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

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

[5] Eskenazi, M. (2009). An overview of spoken language technology for education. Speech Communication, vol. 51, no. 10, 832–844.

[6] Eskenazi, M. and Hansma, S. (1998). The fluency pronunciation trainer, in Proc. of the STiLL Workshop.

[7] B. Granström. (2004). Towards a virtual language tutor. in Proc. of InSTIL/ICALL Symposium 2004.

[8] Cabral, J. P., Kane, M., Ahmed, Z., Abou-Zleikha, M., Szkely, E., Zahra, A., Ogbureke, K., Cahill, P., Carson-Berndsen, J. and Schlögl, S. (2012). Rapidly Testing the Interaction Model of a Pronunciation Training System via Wizard-of-Oz. In Proc. of International Conference on Language Resources and Evaluation (LREC), Istanbul.

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

[10] Chapelle, C. A. (2009). The Relationship Between Second Language Acquisition Theory and Computer-Assisted Language Learning, in Mod. Lang. J., vol. 93, no. s1, 741–753.

[11] Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A. and Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. Behav. Res. Methods Instrum. Comput., vol. 36, no. 2, 180–192.

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

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

[14] Jokinen, K. and McTear, M. (2009). Spoken Dialogue Systems. in Synth. Lect. Hum. Lang. Technol., vol. 2, no. 1, 1–151.

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

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

[17] Kinect for Windows SDK. (2014). Retrieved 7 July 2014, from http://msdn.microsoft.com/enus/library/hh855347.aspx

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

[19] Young, S., Evermann, G., Gales, M., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). The HTK book version 3.4. http://htk.eng.cam.ac.uk/.

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

[21] Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. Journal of the Acoustical Society of America, 123(5), 3878.

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

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

[24] 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.

[25] W3C. (2014). Web Speech API Specification. Retrieved 7 July 2014, from https://dvcs.w3.org/hg/speech-api/rawfile/tip/speechapi.html

[26] Ní Chiaráin, N. (2014). Text-to-Speech Synthesis in Computer-Assisted Language Learning for Irish: Development and Evaluation. Unpublished doctoral dissertation, Trinity College, Dublin.