Duenna—An experimental language teaching application

Balázs Zsigmond Horváth\textsuperscript{a,}*, Bence Blaske\textsuperscript{b}, Anita Szabó\textsuperscript{c}

\textsuperscript{a} Pogány Frigyes Vocational School, 11 Thőkőly út, 1183 Budapest, Hungary
\textsuperscript{b} SAP Hungary, 7 Záhony utca, 1031 Budapest, Hungary
\textsuperscript{c} ELTE University Department of Arts and Humanities, 4/A Múzeum korát, 1088 Budapest, Hungary

Received 10 April 2016; received in revised form 31 July 2016; accepted 2 August 2016

Abstract

The presented TTS (text-to-speech) application is an auxiliary tool for language teaching. It utilizes computer-generated voices to simulate dialogs representing different grammatical problems or speech contexts. The software is capable of producing as many examples of dialogs as required to enhance the language learning experience and thus serve curriculum representation, grammar contextualization and pronunciation at the same time. It is designed to be used on a regular basis in the language classroom and students gladly write materials for listening comprehension tasks with it. A pilot study involving 26 students (divided into control and trial groups) practicing for their school-leaving exam, indicates that computer-generated voices are adequate to recreate audio course book materials as well. The voices used were able to involve the students as effectively as if they were listening to recorded human speech.

© 2016 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Keywords: Computer aided language learning; Text to speech; Language teaching

Metadata

Current Code version: 2.1
Permanent link to code/repository used of this code version: https://github.com/ElsevierSoftwareX/SOFTX-D-16-00039
Legal Code License: MIT
Code Versioning system used: none
Compilation requirements, Operating environments & dependencies: ActionScript, MXML
If available Link to developer documentation/manual: https://github.com/ElsevierSoftwareX/SOFTX-D-16-00039/blob/master/README.md
Support email for questions: bence.blaske@yahoo.com

Software metadata

Current software version: 2.1
Permanent link to executables of this version: https://github.com/ElsevierSoftwareX/SOFTX-D-16-00039/tree/master/build
Legal Software License: MIT
Computing platform/Operating System: web based, Flash Player
Installation requirements & dependencies: none (built-in)
If available Link to user manual - if formally published: include a reference to the publication in the reference list
Support email for questions: bence.blaske@yahoo.com

* Corresponding author.
E-mail addresses: zsmondi@gmail.com (B.Z. Horváth), bence2000@yahoo.com (B. Blaske), plussmacskanita@gmail.com (A. Szabó).

http://dx.doi.org/10.1016/j.softx.2016.08.002
2352-7110/© 2016 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Introduction

Text-to-speech (TTS) applications read out any written text produced by various types of software. As long as the characters are recognizable as typable texts and they are not converted into an image, they can be read out by these applications. TTS technology is identified and used mostly in the form of screen readers for blind or visually impaired people, but there are many more possibilities and TTS should be applied in the broader field of language education. Instead of restricted use for reading out chunks of longer texts, it is possible to produce stand-alone learning materials which demonstrate parts of the curriculum by using dialogs to put grammatical elements into communicative contexts. The dialogs can also be illustrated and animated in order to make them more user-friendly.

2. Problems and background

Voice emulation or TTS research is a branch of applied linguistics and the technology can indeed be used for dialog simulation and beyond simple text-reading. It is not difficult for a teacher who already has to use media in the classroom to compose stand-alone listening comprehension tasks with it. The success of the application depends more on the creativity of the teaching person and than on technological issues. As early as 1998, Warschauer [1] noted that any computer application in a classroom needs an “amusement factor” to get the students involved on both on individual and group levels. The main characteristic of the software described here is that all the senses needed for language learning are employed while the students are using it, just as they are in an illustrated role playing game, which it resembles. Nowadays teaching materials sometimes need to compete with the gaming industry to get their message through and communicative language learning methods, combined with the possibility to technically simulate any speech content, can reach this goal.

Communicative language teaching in the 21st century demands the enhancement of performance in order to help learners develop efficient command of any foreign language. Savignon [2] points out that modern technological development and the new focus on learner autonomy mean that contextualization needs to be represented to the learner by as many situations as possible.

It might be desirable to build data sets for different contexts but at present such sets would require too much storage space. By utilizing TTS technologies, the theoretical background is provided for the purpose of this study as it is a valuable aid to make interaction contexts more tangible for learners and it makes use of computerized tests which in general take fewer computing resources than any other teaching material. The need for solutions for representing practices or strategies for teaching designed curricula to reflect local needs and experiences can indeed be met.

The connection between Savignon’s [2] reasoning and different uses of teaching materials was clarified by Kilickaya [3]. The most important consideration here is to examine whether TTS voice technology can create databases of sufficient quality similar to those that Savignon mentions. TTS technologies improved in quality with the introduction of the Microsoft Speech Application Programming Interface (SAPI) 5.0 speech engine, which made better accent pronunciation available to personal computer users. The author lists TTS functions as the following: reading any text, at any speed and any speaking quality, with different accents (male, female, British English and American English, etc.)

Kilickaya [3] admitted the limitations as well: although speech programs can create realistic, human-sounding voices, there is always a difference in terms of intonation and stress. These assumptions were well founded, but the technology of voice emulation has gone through many changes since 2006. Sobkowiak [4] noted that with time, all technologies find their way into education, so it is only a matter of time—dependent on the speed of technological research, until TTS gains ground.

In 2008, Azuma [5] already spoke highly of technological developments in TTS research and went so far as to claim that possibilities may open a whole new horizon in every form of language pedagogy. Azuma drew attention to the low costs of material production, which is also the aim of this software publication. The researcher described his experience using the Pentax “VoiceText” system both in classroom and in social network-based e-Learning environments.

TTS scenarios are also supported by McKeeman and Oviedo [6] who mention different computer applications for enhancing communicative competence and their scope recognizes TTS as one possibility. The authors mention that students were able to compose dialogs on their own with the help of an on-line environment allowing them, for example, to edit animations from their own materials. Yet, the authors do not go as far as to emphasize the true benefit of contextualization and the structuring of materials. The composed dialogs should simulate various outcomes of speaking situations.

As a different approach, a longitudinal study conducted for the Department of Education of Iowa in classroom environments by Hodapp and colleagues [7], showed increasing favor for the use of screen readers among learners who have mild disabilities. The study focused on reading comprehension and showed that between 80% and 95% of the students reported that TTS helped their independent learning and made them understand their materials better. The study supported these claims by presenting the academic test results of the students.

LaRoussa and colleagues [8] drew attention to the fact that TTS applications in general also indicate, by highlighting, the words and sentences the computer is pronouncing and so they are also beneficial for dyslexic students. Highlighting sections of text as they are pronounced proved to be successful even for eighth-grade students as Socol [9] pointed out in his article. Altogether 76% of the children said that even longer passages read with TTS were “more interesting”. This preference for the TTS-read passage was reported no matter what the reading material was. This suggests the possibility that the dual-sensory system provided by highlighting TTS systems holds the attention of certain students more than the traditional mode of reading. Socol’s findings are true for reading out situations as well, as highlighting can be implemented anywhere.
Linking back to the topic of language learning, the most widespread use of computer emulated voices is the pronunciation of words in on-line dictionaries. Ruedas [10] conducted a study in which different lexical items from the field of chemistry were read out by TTS voices and then recorded for university students. Results showed – and most of the students agreed – that drilling via the audio materials helped them to learn the expressions, especially in cases of missed lectures.

TTS technology is without doubt useful for language learning and the presented software is intended to extend its use even further by illustrating the read texts with graphic novel-style photographs of the speakers, concentrating on their facial expressions and body language. The challenge was to combine the experiences listed in the literature into a new and creative form which is useful for the language learner.

3. Software framework

3.1. Software architecture

The overall software architecture is depicted above (Fig. 3.1): the actor is the ESL learner. He/she accesses the software functions via the Duenna User Interface (UI). The said UI allows the ESL learner to select a conversation from various topics that illustrate a grammatical area, e.g. past tense, etc. Then the software plays back the conversation using comic-like imagery, and utilizing TTS voices. The application thus makes use of an external web service (Acapela VAAS) and also software resource files such as facial expression images, background audio and background images.

3.2. Software functionalities

The principal software functions are shown in the image below (Figs. 3.2–3.8). Again, the Duenna graphical User Interface enables the learner to observe an imaginary situation that is played back using photos of actors and utilizing TTS voices. Additional functionalities include a window where grammatical explanations are shown and the software can also show the conversation transcript. It also allows the user to pause and resume the conversation being played. The learner can select from a variety of topics and conversations belonging to a certain topic.

3.3. Software structure

The following class diagram depicts the overall structure of the software (Fig. 3.9). The Main class provides services and functions to the overall application, such as tracking the actual conversation, controlling the state of the application and triggering additional sub-functions. Such sub-functionality includes for example the configuration window (Settings), the conversation Transcript window and the grammatical explanation window. The Main class also accesses the external third party TTS webservice (Acapela VAAS). The software only needs time to load the voices which, in turn, depends only on the speed of the Internet connection.

4. Implementation and empirical results

4.1. Implementation details

The presented TTS application is an auxiliary tool for language teaching. It utilizes computer-generated voices to simulate dialogs representing different grammatical problems or speech contexts. The software is capable of demonstrating as many examples of language as required to enhance the language learning experience.

Duenna is implemented using a technology called Adobe Flash. This provides a unique advantage compared to other competitive products. Competitive products in many cases rely on local computer installations, and thus need to have an additional step to be able to use the product.

Additionally Duenna is running directly in the Web Browser. This has the following benefits compared to many competitive products:
1. Loads almost instantly, no installation is needed,
2. Product content (conversations, voices) loads very quickly from fast web servers,
3. Product content can be updated on the web server side, thus providing a seamless upgrade experience,
Fig. 3.2. Duenna User Interface overview detailing the places of controls on the screen.

Fig. 3.3. Duenna User Interface overview in action.

Fig. 3.4. Details of how the grammar rules are represented which are connected to the dialogs.
4. Adobe Flash is a cross-platform solution, therefore Duenna can be used on wide range of operating systems and devices,

5. Thanks to the run time environment, it runs perfectly on low-end machines. It proved to be loading at a rate of 1.55 s faster compared to the closest avatar-oriented TTS teaching
solution found on the Internet at the time of submitting this article (based on Gtmetrix page performance diagnostics results).

4.2. Empirical results

The software in its present form is used on a regular basis in the language classroom and students gladly write dialogs for listening comprehension tasks with it. Earlier research and piloting in 2013 with students practicing for their school-leaving exam proved that computer-generated voices are adequate to recreate audio course book materials for 26 students, divided for control and trial groups. The voices employed were able to involve the students as much as if they were listening to recorded human speech. Results showed that listening comprehension tasks could be solved with success. The results of the control and trial groups were always comparable with each other. No student in any group received less than 30% for his or her work, which indicates that we can accept that the voices were understandable.

The task sheets were composed of short-answer, pairing and true–false items. Short-answer items proved to be easier to solve than pairing and true–false tasks. Pairing activities peaked at an 80% success rate and were never lower than 56%. True–false questions were never solved perfectly, but this phenomenon was due to the sounds echoing in the classroom where the tests took place. The students themselves did not consider voice quality as a deciding factor.

Since then, there have been two further analyzed testing occasions: in September 2015 and in February 2016. On the first occasion 11 new 11th grade students watched three demonstrative dialogs for the sake of introducing the software. Each dialog was played twice. While watching the demonstrations, they had to answer short-answer questions. Eight members out of the whole group of students got all the answers right. After the session the students were asked to write dialogs as homework representing either use of English or contextual issues. On the second occasion again two of their original written texts were chosen after correction for another round of short answer questions. The samples were again played twice. Three students managed to understand the first text after the first playback and
the rest after the second. The second dialog was understood by everyone after the first playback. 80% of the students managed to answer the questions acceptably and acknowledged the usefulness of the exercise. The students expressed their need for further practice with the software as it demonstrated that any speaking situation can be simulated. Further testing will take place as the year progresses.

5. Illustrative examples

The video accompanying this submission shows a recording of some of the sample conversations found in Duenna. It exemplifies some of the additional features that the software offers. It is basically a walkthrough of the application, giving a glimpse on the user interface and functions.

6. Conclusions

In conclusion we can observe that there is little indication to date that TTS is used for any purpose besides reading longer segments of texts or purposefully by bringing everyday speaking situations closer to the learners. Vocabulary use is promising but realization seems very occasional. Google Translator makes mistakes and is not yet sufficiently reliable for academic purposes. The researchers explored various possibilities for TTS utilization, but there is indeed space for more empirical research. The idea of building an entire language course around this technology is a dream for the future that deserves serious consideration; the use of simplified or full-version books is another interesting prospect, because of course they usually contain interactions between characters,—but producing stimulating and representative dialogs for language learners is an important first step.

The software framework is adoptable to any TTS system using SAPI voices. The technology has proved to be more than adequate for listening comprehension tasks. The results confirm, on the practical level, the concepts of Kilickaya [3] and Azuma [5] and encourage continuing research and development with TTS technologies. Further research is underway with the intellectual help of the IT Foundation for the Visually Impaired (INFOALAP Budapest, Hungary).

Acknowledgments

The authors would like to express their thanks to Christopher Ryan of ELTE—Department of English Language Pedagogy, Budapest and all the students who participated in the project.

Formatting of funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.softx.2016.08.002.

References