TEXT-TO-SPEECH APPLICATIONS TO DEVELOP EDUCATIONAL MATERIALS

Raquel Sanchis, Beatriz Andrés, Raúl Poler

Research Centre on Production Management and Engineering (CIGIP),
Universitat Politècnica de València (SPAIN)

Abstract

There are several ways to develop educational materials and several different types of educational materials depending on the audience, objectives, topics or themes, type of education, among others. One of the most common educational materials developed is the use of presentations slides where to shape the information that the trainer wishes to share. The most used presentation graphics packages are Microsoft PowerPoint, OpenOffice Impress and Apple KeyNote. These systems enable word processing, outlining, drawing, graphing, and displaying different presentation management tools to design and configure a presentation. This educational material is usually used to be shown during an explanation in a master class or online through an e-learning platform. In the case that the education material is available through an online resource, it is important not only to present the information in a readable manner but: (i) to add the explanation as a spoken sound version in order to give to the receiver more information than the one that is displayed in the slides and (ii) to avoid fatigue due to reading all the information of the slides.

Currently, there are different text-to-speech applications that allow to play sound files based on text without the interaction of humans. This paper focuses on these applications, which their main characteristics are and which their benefits and weaknesses are in order to select the most appropriate one to develop the different types of educational materials.

Keywords: Educational Material, Presentations, Text-to-Speech applications.

1 INTRODUCTION

According to Gardiner [1], speech is quadrilateral, requiring a speaker, listener, words, and things to be spoken about. The speech arises from a desire of a speaker to acquaint a listener with the "thing-meant." Speech is the oldest means of communication between people and it is also the most widely used [2]. For this reason, most of the educational materials such as videos, tutorials… are based on providing the information about the training contents built around the speech. In face-to-face training, the speech is the most used method to provide to the audience the explanation about the concepts that the speaker wishes to transmit. However in the e-learning context, this interaction between the speaker and the listener is more difficult to achieve. In the e-learning world, the delivery formats of the education materials are mainly static. Slide presentations are one of the most popular e-learning formats due to the fact that allows to easily add text, images, videos, animations, and graphs [3]. However this type of educational material presents the disadvantage that does not provide a spoken sound version of the content what could make the audience to lose the interest in the matter and get bored.

One of the solutions to overcome this drawback is the use of the available technologies such as text-to-speech applications. The area in charge of converting text into understandable voice is the speech synthesis, that is defined as the artificial production of human speech. This area was born in order to help people with visual problems or reading disabilities to allow them to 'listen' documents or any written work through their personal computers [2].

The applications that transform the text into understandable sounds are known as text-to-speech (TTS) systems. As aforementioned, these systems were created to support people with visual and/or reading problems, but currently its use have been widespread and they are not only addressed to these target users but these systems are widely used with other purposes, such as for example as support to develop educational materials.

Currently there are several tools/applications that "read" the text and convert it into "sound". Their main characteristics will be shown in the following sections. There are open source tools and also commercial ones. The open source tools are freely available (also the code) to the public, so its main

advantage is that these open text-to-speech tools can be used without the need of high investments or other type of costs. In this paper, an exhaustive review of the main open source tools is described in section 2. However, some of the text-to-speech applications analysed are not open source but they are freeware and users can use them without such a high investment but user are not able to customize or change the application as the programming code is not freely available. Section 3 offers a characterization of the main features of each of the text-to-speech applications and a comparison among them in order to support trainers in the selection of the most adequate one. Finally, the conclusions section that offers an overview of the main outcomes of the paper.

2 TEXT-TO-SPEECH APPLICATIONS

Some research about the most popular and free text-to-speech applications has been developed. It is worth to mention than the speech synthesis science, more specifically the text-to-speech research has grown in the last 20 years, with 114 publications related to the keywords “text-to-speech” and “education” as it is shown in Fig. 1 [4]. With regard to the percentage of research of text-to-speech aspects, the area that has invested more in investigating is computer science with 32,2% followed by educational research with 31,3% (Fig. 2 [4]).

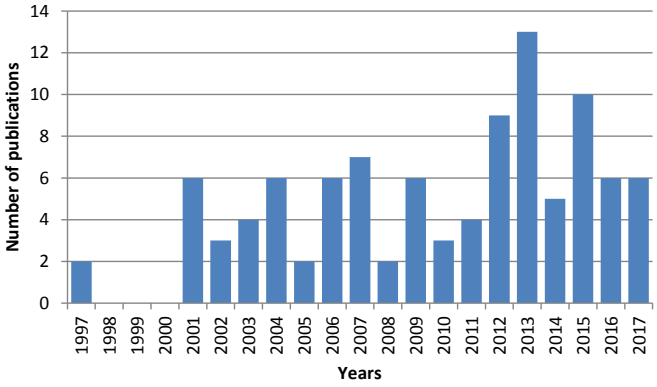

Figure 1. Number of publications related to the text-to-speech and education in the Web of Science.

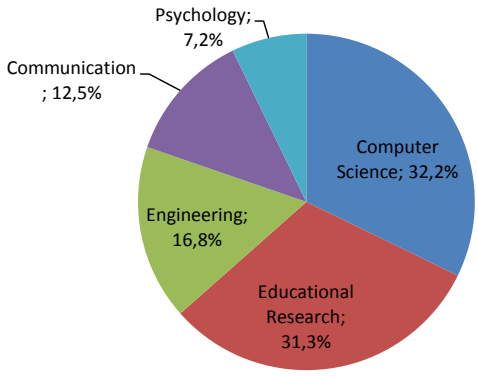

Figure 2. Percentage of research of text-to-speech aspects per areas.

It is worth to mention that although currently there are several text-to-speech applications available for educational purposes, this paper is focused on the open source applications due to the following reasons:

- From the trainers’ point of view, this advantage is not clearly seen, but in the open source applications, it is possible to carry out modifications in order to adapt such applications to specific educational needs. To do so, it is necessary to have the appropriate knowledge to modify the source code but if the end user of the application has the proper skills to do it, the text-to speech applications could be customized to fulfil specific educational requirements.
- The investment needed to use such applications is zero since with the open source community, users do not have to spend anything purchasing licenses. The only cost that the end users have to assume is the cost related to the learning curve (familiarization) of such applications.
- If the trainer also has the necessary knowledge, the maintenance tasks could not depend on a specific company, but the trainer could make that necessary maintenance and updating activities.

However, not all the trainers have the technological and informatics aptitudes to customize the open-source text-to-speech application. For this reason, the literature review is also focused on those applications that are free and moreover their user interfaces are user friendly. At this point, it is worth mentioning that there are many open-source text-to-speech applications that offer a wider range of possibilities from a scientific point of view such as linguistic, phonetic or phonological research that will not be analysed in this paper. Some examples of these applications are: Praat [5], FreeTTS [6], Festvox [7], and Flite [8]. Other text-to-speech applicatiosn such as: The MBROLA Project [9], The Festival Speech Synthesis System [10], SpeakRight [11], Kaldi - Speech Recognition Toolkit [12] are open source, however they have not been used for the development of educational materials.

For the above reasons, the main focus is based on the easy to install and user-friendly free and open source text-to-speech applications. The text-to-speech applications reviewed and deeply analyzed are the following ones: Balabolka [13], eSpeak [14], Natural Reader [15], MaryTTS [16], Text to Voice Internet Browsers Extensions (add-on) [17, 18], WordTalk [19] and YAKiToMe! [20] (alphabetically ordered).

Finally, it is also important to mention that there are open source applications that do the reverse process as voice recognition applications. They allow users to transform the speech into text. Some examples of these open sources application are: Simon Speech Recognition [21], CMU Sphinx [22], Wryte [23], among others. However, the study of these applications is out of the scope of this paper.

2.1 Balabolka

Balabolka is a text-to-speech program that transforms the text into voice in audible files with the following formats: .wav, .mp3, .mp4, .ogg or .wma file. The program offers different possibilities to read the text, it can read the clipboard content, view text from documents, customize font and background colour, control reading from the system tray or by the global hotkeys. Moreover, it supports a wide range of file formats such as: .azw, .azw3, .chm, .djvu, .doc, .docx, .eml, .epub, .fb2, .html, .lit, .mobi, .ods, .odt, .pdb, .prc, .pdf, .rtf, .tcr, .wpd, .xls, and .xlsx. It supports 31 different languages [13].

Balabolka is not open source. Although the licence of this application is freeware, it uses various versions of Microsoft Speech API (SAPI), that allows to alter the voice's parameters, including rate and pitch [13].

The user interface is very intuitive as it can be shown in Fig. 3.

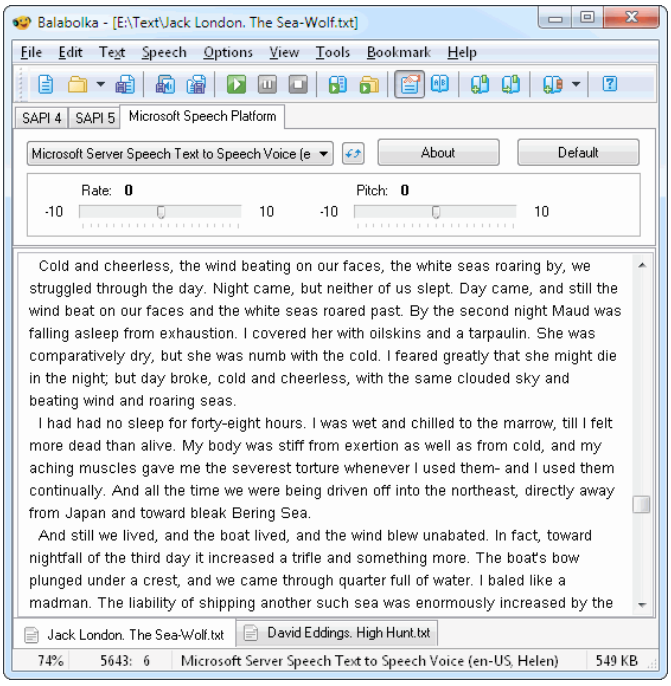

Figure 3. Balabolka User interface [13].

2.2 eSpeak

eSpeak is a compact, multi-language, open source text-to-speech application, for Linux and Windows, written in C [14]. It includes different voices, whose characteristics can be altered and furthermore it supports 51 languages. It produces speech output as a .wav file and it is downloadable as a desktop application. Moreover, it has a compact size as the program and its data, including the different languages, totals about 2 Mbytes. SSML (Speech Synthesis Markup Language) is supported (not complete), and also HTML. It can be used as a front-end to MBROLA diphone voices. eSpeak converts text to phonemes with pitch and length information. It can translate text into phoneme codes, so it could be adapted as a front end for another speech synthesis engine.

Fig. 4 shows the interface of the eSpeak. The text to be converted into sound, could be loaded through a .txt file or pasting directly the text in the screen [14].

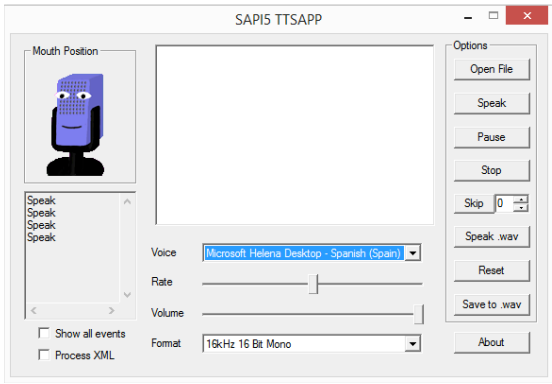

Figure 4. eSpeaker interface [14].

2.3 Natural Reader

It is a text-to-speech application that it is not open source but offers a free version that runs in different operating systems. The application works with .pdf, .docx, .txt and .epub. It offers the possibility to change the speed and the speaker. Natural Reader offers two different ways to read text. On one side, different documents can be loaded in its library to convert the text into voice as it is shown in Fig. 5. This option also offers the possibility to scan a document and Natural Reader is able to transform the scanned text into voice. On the other side, the application also reads text from different resources such as web browser, word processor… The user has to highlight text in these applications and Natural Reader, through a floating toolbar, controls the different options to transform the content into voice files [15].

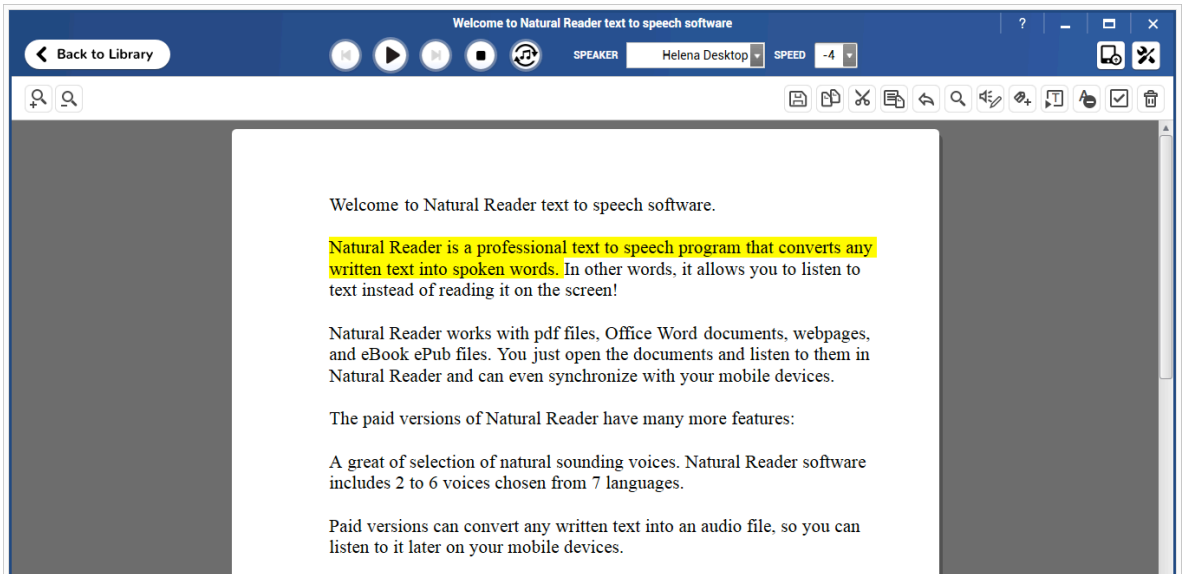

Figure 5. Natural Reader library interface [15]

2.4 MaryTTS

MaryTTS (**M**odular **A**rchitecture for **R**esearch on speech s**Y**nthesis) is an open-source, multilingual text-to-speech synthesis platform written in Java. It was originally developed as a collaborative project of DFKI's Language Technology Lab and the Institute of Phonetics at Saarland University. It is now maintained by the Multimodal Speech Processing Group in the Cluster of Excellence MMCI and DFKI [16].

The main features of MaryTTS are the following ones [24]:

– Several diphone sets for a number of male and female voices can be used.
– It supports 9 different languages.
– It can produce speech output as .wave, .au and .aiff files.
– MBROLA is used for synthesising the utterance.
– It is available online and as a downloadable version.

Fig. 6 shows the web client interface of the MaryTTS where users can paste the text to be recorded and chose different effects by changing the defaulting parameters.

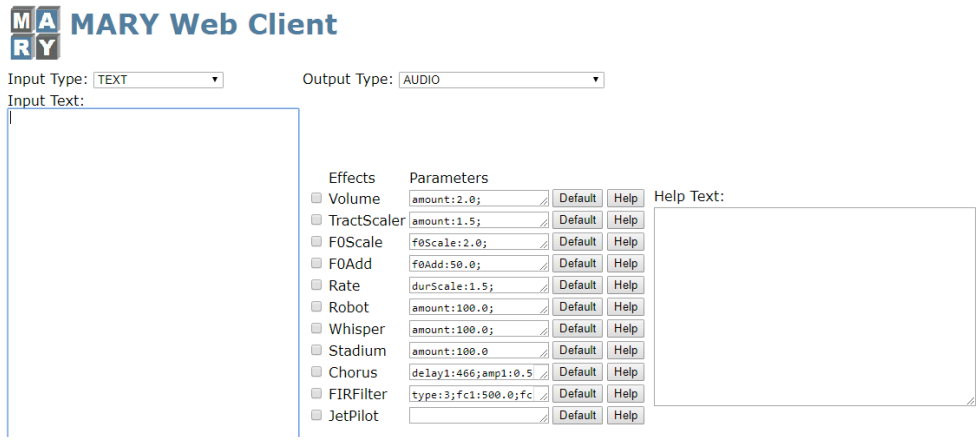

Figure 6. MaryTTS web client interface [16].

2.5 Text to Voice Internet Browsers Extensions (add-on)

Some Internet Browers such as Firefox or Chrome have extensions (add-on) that add text-to-speech functionalities to the browsers [17,18].

On installation a speaker icon is created on the bottom-right of the Internet Browsers (Fig. 7) [17]. The operation is very easy as the user only has to select the text he/she wishes to record and then select the speaker icon and this add-on speaks the selected text. The audio file can also be downloaded as .mp3 file (Fig. 8) [18].

The main disadvantages of this extension are that it is not compatible with some versions of the different Internet Browsers, for example the Text to Voice extension of Firefox does not work with Firefox Quantum. Moreover, it can only play and record the text selected in the different browsers what makes the customization of the text trainers need to include in their educational materials, difficult [17,18].

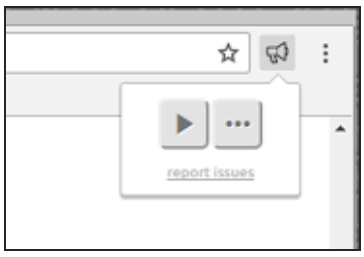

Figure 7. Speaker icon extension of Chrome [18].

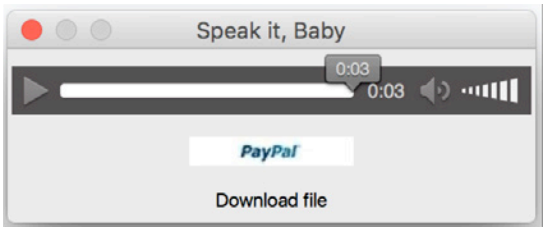

Figure 8. User interface to download the .mp3 text to voice file in Mozilla [17].

2.6 WordTalk

It is a free text reader, in form of a toolbar add-on that was developed by the University of Edinburgh. WordTalk creates a spoken sound version of the text in a Microsoft Word document. Therefore, the application can read an entire document, paragraph or word. Moreover, it highlights the text as it is reading. It offers the possibility to change the voice and the speed, and the output voice files are .wav

or .mp3 [19]. Fig. 9 shows the different controls and options of the WordTalk toolbar in the 2007 Word version [25].

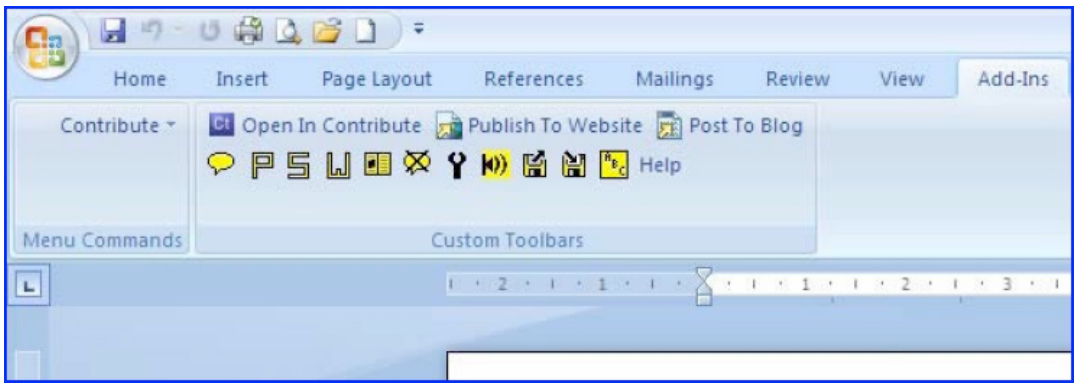

Figure 9. The WordTalk toolbar in Word 2007 from the 'Add-Ins' tab [25].

2.7 YAKiToMe!

YAKiToMe! is an online service that reads aloud electronic text. It was conceived to help readers with visually impairments, dyslexia, and others learning and reading disabilities. However it is also applicable to record the text to be included in the education slides presentation materials [20].

It is compatible with texts in different formats (.txt, .doc, .pdf,…) that can be uploaded, and read by such a service. It can produce speech output as .mp3 or .ogg files. It allows to download and podcast the outputs files [20].

Fig. 10 shows the online user interface of YAKiToMe! by which trainers could convert the text into audible files.

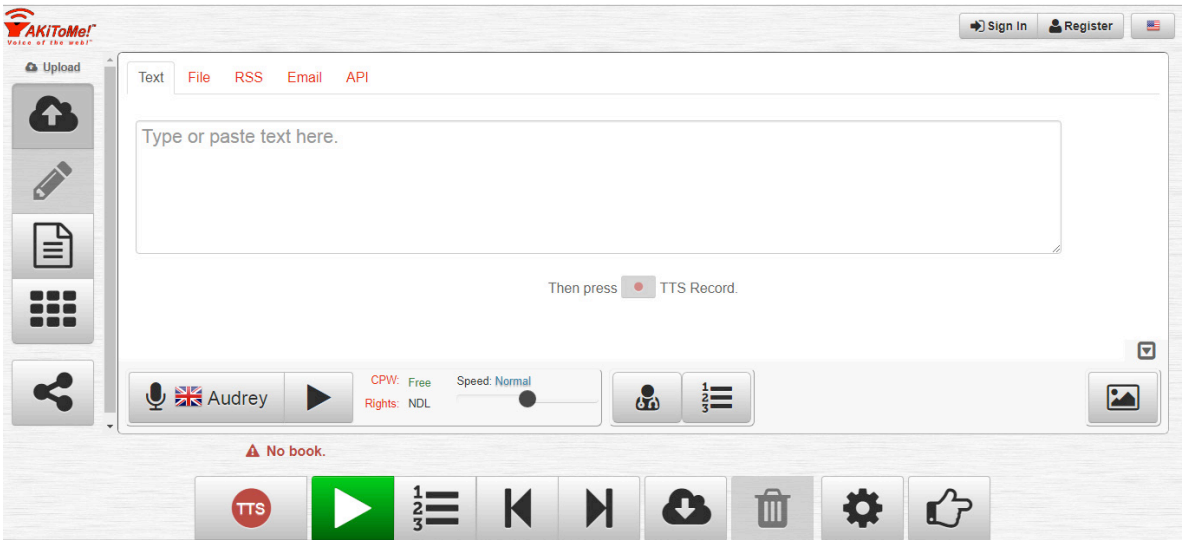

Figure 10. YAKiToMe online user interface [20].

3 COMPARISON AMONG THE DIFFERENT TEXT-TO-SPEECH APPLICATIONS

The choice of the most appropriate application for the development of educational material with read-aloud' or spoken version of typed text depends on the requirements of each trainer. It is generally recognized by trainers that one of the most important features is related to the voices. Aspects such as naturalness (natural voices in a number of languages), fluency (speed and volume, control of the reading by inserting pauses), comprehension (pitch control, pronunciation…) are key for the selection of the most suitable application. These factors depend to a large degree on the subjective judgment of the different trainers. In further research lines it is expected to survey a representative sample of

trainers using the different applications analysed in this paper to rank the most proper text-to-speech applications based on a collection of general educational requirements.

Moreover, trainers also give importance to the ease with which they can obtain the voice files. Therefore, the applications' ergonomics, simplicity, user-friendly factors are also vital for the selection of the application. But not only, the use of the application itself but the process to download and install it, in case that the application is a desktop one.

From a technical point of view, trainers with informatics and programming background are more interested in open source applications while those focused on only the development of educational materials with the spoken version of the presentation slides are more interested in being able to use the application for free than in the fact that it is open source. Furthermore, it seems that trainers prefer desktop applications than online ones because they can use them without Internet connection. Finally, it also should be noted that operating systems have a great influence when choosing the text-to-speech application.

To support trainers in the selection of the most suitable text-to-speech application, Table 1 offers an overview of the main functionalities of the analysed applications.

Table 1. *Characterization and comparison among the different text-to-speech applications.*

	Balabolka	**eSpeak**	**Natural Reader**	**MaryTTS**	**Text to Voice Browsers Extensions**	**WordTalk**	**YAKiToMe!**
Open-source	No	Yes	No	Yes	Yes	No	Yes
Sound file generation	.wav, .mp3, .mp4, .ogg,.wma	.waw	.mp3	.au, .aif, .waw	.mp3 (Mozilla)	.wav, .mp3	.mp3 .ogg
Adjustments	Pitch, volume and rate	Rate and volume	Speed	Volumne, rate, robot, whisper, stadium, chorus	-	Voice and the speed	Speed, pronun-ciation, and chapter chopping
Number of languages	31	51	4	9	All languages supported by the browsers	All languages supported by Microsoft Word	23
Online / Desktop application	No/Yes	No/Yes	Yes/Yes	Yes/Yes	Yes (add-on)/ No	No/ Yes (add-on)	Yes/No
Operating systems	Windows	Linux, Windows	Mac OS, Windows	Linux (strongly recommende d). Mac OS, Windows with dependent tools	Linux, Mac OS, Windows	Windows	Linux, Mac OS, Windows
Format texts	.azw, .azw3, .chm, .djvu, .doc, .docx, .eml, .epub, .fb2, .html, .lit, .mobi, .ods, .odt, .pdb, .prc, .pdf, .rtf, .tcr, .wpd, .xls, and .xlsx.	.txt, .xml	.pdf, .docx, .txt and .epub	-	Browser selected text	.doc, .docx	txt, .doc, .pdf

As aforementioned, the selection of the most convenient text-to-speech application depends on many factors (subjective, technological, educational…). Nevertheless, it is universally known that these applications have a positive impact in education. They are mainly used for the development of education e-learning materials in which it is opportune to add the spoken version of the presentation to facilitate to the audience the following of the educational session. With this spoken version, the audience fatigue is diminished as they do not have to be reading all the time. Moreover, the pronunciation is always the same and a wide range of different languages can be used.

When comparing online and desktop applications, the first ones present the advantage that do not require any installation and for this reason users have not to be worried about their operating systems or software requirements. However, by counterpart, online versions require internet connection while the desktop applications are always available once installed.

Conversely, the text-to-speech applications also have detractors since some trainers state that the face-to-face interaction cannot be substituted for such applications. One solution would be to combine both through blended learning and use, in the case of e-learning sessions, the text-to-speech applications to make the comprehension of the educational materials' contents, easier.

4 CONCLUSIONS

The paper reviews the main open source and free text-to-speech applications in order to offer to trainers, appropriate applications for developing educational materials that need spoken versions. Through the use of spoken versions of educational materials, the trainees will find the training process easier and simpler. The review performed mainly covers two types of applications, ones that are offered online and the others that provide an installable file as a desktop application. The choice between both types will depend on the preferences of the trainers. Moreover, the selection of the most appropriate text-to-speech application will depend also on other factors such as:

- Technological aspects, such as if the text-to-speech application is open source or only free. If the trainer wishes to customize the application, he/she will need to get access to the programming code and his/her selection will be open source. If the trainer only wishes to use the application, his/her selection could be a freeware application and it is not necessary an open-source one. From the technological point of view, another important aspect is the operating system as the analysed text-to-speech applications have different technological requirements (as it is specified in Table1).
- Trainers' opinion aspects, such as the naturalness and pronunciation of the voices used in each of the text-to-speech applications. The choice will depend on subjective motives of the selector.
- Ergonomics aspects, such as the ease, simplicity and comfort to use the text-to-speech application.

The overview about the main text-to-speech applications offered in this paper tries to support the trainers' decision making process.

REFERENCES

[1] A. H. Gardiner, The theory of speech and language. 1932.

[2] N. Swetha, K. Anuradha. "Text-to-speech conversion", International Journal of Advanced Trends in Computer Science and Engineering, vol. 2, no. 6, pp. 269-278, 2013.

[3] M. Pacella "Eight types of content to include in your training program", 2017. Retrieved from URL. https://www.litmos.com/blog/course-design/8-types-of-content-to-include-in-your-elearning-training

[4] Web of Science, 2017. Retrieved from URL. wos.fecyt.es

[5] Praat: doing phonetics by computer, 2017. Retrieved from URL. http://www.fon.hum.uva.nl/praat/

[6] FreeTTS 1.2.3 - A speech synthesizer written entirely in the JavaTM programming language, 2005. Retrieved from URL. https://freetts.sourceforge.io/docs/index.php

[7] Festvox, 2017. Retrieved from URL. http://festvox.org/

[8] Flite: a small, fast run time synthesis engine, 2017. Retrieved from URL. http://www.speech.cs.cmu.edu/flite/

[9] The MBROLA Project, 2006. Retrieved from URL. http://tcts.fpms.ac.be/synthesis/mbrola.html

[10] The Festival Speech Synthesis System, 2017. Retrieved from URL. http://www.cstr.ed.ac.uk/projects/festival/

[11] SpeakRight, Building VoiceXML apps faster using reusable Java components. 2008. Retrieved from URL. http://speakrightframework.blogspot.com.es/

[12] Kaldi - Speech Recognition Toolkit, 2017. Retrieved from URL. http://kaldi-asr.org/

[13] Balabolka, 2018. Retrieved from URL. http://www.cross-plus-a.com/balabolka.htm

[14] eSpeak text to speech, 2017. Retrieved from URL. http://espeak.sourceforge.net/

[15] NaturalReader, The Most Powerful Text to Speech Reader, 2018. Retrieved from URL. https://www.naturalreaders.com/

[16] The MARY Text-to-Speech System (MaryTTS), 2016. Retrieved from URL. http://mary.dfki.de/

[17] Text to Voice. Firefox Adds-on, 2017. Retrieved from URL. https://addons.mozilla.org/en-US/firefox/addon/text-to-voice/

[18] Texto a Voz, 2017. Retrieved from URL. https://chrome.google.com/webstore/detail/read-aloud-a-text-to-spee/hdhinadidafjejdhmfkjgnolgimiaplp

[19] WordTalk, Free text-to-speech plugin for Microsoft Word, 2018. Retrieved from URL. http://www.wordtalk.org.uk/download/

[20] YAKiToMe!, Voice of the web, 2018. Retrieved from URL. https://www.yakitome.com

[21] Simon Speech Recognition, 2017. Retrieved from URL. https://alternativeto.net/software/simon-speech-recognition/

[22] CMU Sphinx, 2007. Retrieved from URL. http://www.speech.cs.cmu.edu/

[23] Wryte, 2017. Retrieved from URL. https://alternativeto.net/software/wryte/

[24] M. Schröder, J. Trouvain. “The German text-to-speech synthesis system MARY: A tool for research, development and teaching”, International Journal of Speech Technology, vol. 6, no.4, pp. 365-37, 2003.

[25] Text to Speech, 2017. Retrieved from URL. http://web2specialeducation.pbworks.com/w/page/22883536/Text%20to%20Speech