Say It Out Loud

Author: **Sarah Schembri**

Our everyday lives depend on reading things off screens, paper, or notices. Text-to-speech technology can help the visually impaired, but first, we need a robust technology that can handle the complexity of the task.

The power of technology can be harnessed to help daily life become more equitable for all. One example of this is text-to-speech technology, sometimes called read-aloud technology. Text-to-speech technology is available on most electronic devices and within applications such as MSWord and Adobe, and it speaks aloud the text written on the screen. Additionally, apps and devices such as ReaderPens go a step further. Instead of simply reading text on screens, they can read text on digital images or paper. ReaderPens and similar apps have the additional task of recognising the text on an image before converting it to audio. Text-to-speech tools help the visually impaired and anyone who struggles with reading, allowing them more independence and lowering the hurdles faced in everyday life.

IT STARTED WITH A DETOUR

Dr Alexandra Bonnici (Head of the Department of Systems & Control Engineering) and Dr Ing. Stefania Cristina (Senior Lecturer with the Department of Systems & Control Engineering) were originally inspired to start the Doc2Speech project when they noticed there are no ReaderPens which convert Maltese text to speech. However, before they could start development, they encountered some bigger hurdles with the software that recognises text.

When trying to read from a children's book for example, where the text is typed on top of designs or pictures, most text-to-speech software fails at correctly recognising the text. 'The software assumes that the background is a single, flat colour and mistakes the background patterns with symbols,' explains Bonnici while showing me text written on top of an illustration of a garden fence with its vertical lines running through the text. This is a problem even in established technology that reads text in English. Some advances have been made at recognising text on old documents. However, the mild distractions from yellowing stains that could confuse the text-to-speech software are nothing next to the bright colours, bold designs, and cartoon animals on unlikely adventures found

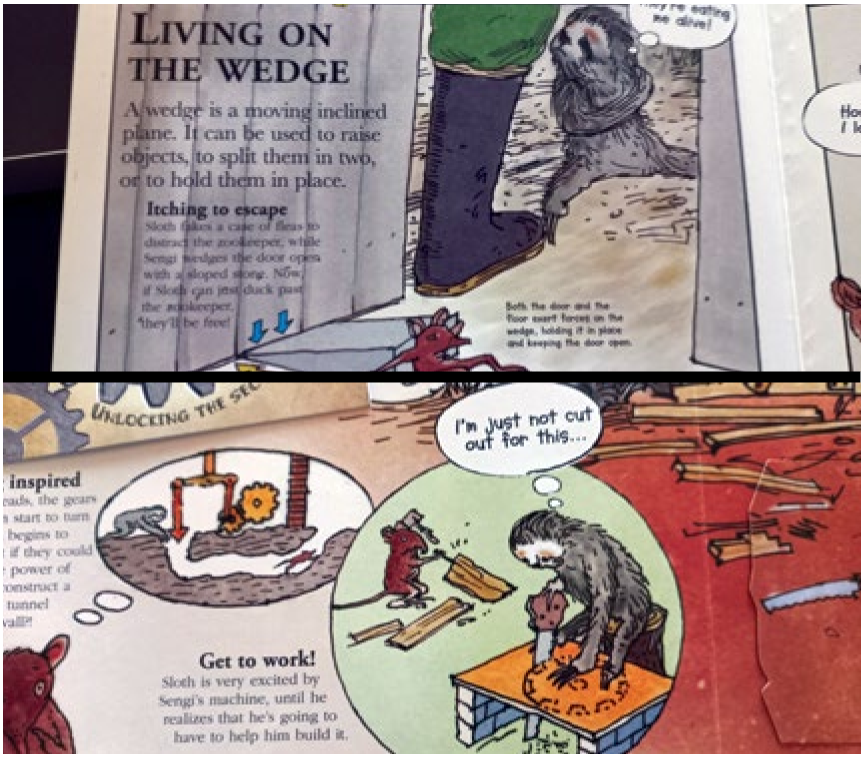
LIVING ON THE WEDGE
A wedge is a moving inclined plane. It can be used to raise objects, to split them in two, or to hold them in place.
Itching to escape
Sloth fakes a case of fleas to distract the zookeeper, while Sengi wedges the door open with a sloped stone. Now, if Sloth can just duck past the zookeeper, they'll be free!
...they're eating me alive!
Both the door and the floor exert forces on the wedge, holding it in place and keeping the door open.
UNLOCKING THE SE
inspired
I'm just not cut out for this...
Get to work!
Sloth is very excited by Sengi's machine, until he realizes that he's going to have to help him build it.

Figure 1: Text-to-speech software tends to fail at correctly recognising the text on top of colourful designs and images.

Illustrations from* How Machines Work *by David Macaulay

in children's books. Bonnici and Cristina recognised this as an important flaw, especially since text-to-speech software could be of immense help to children with varying abilities when learning how to read.

DEVELOPING THE TECHNOLOGY

The engine behind text-to-speech software is called a binarisation algorithm. The algorithm converts a greyscale image into a black-and-white one. In its simplest form, it goes through each pixel one by one and if the pixel has a value below a certain threshold, the value is set to 0 (black) while a pixel above the threshold is set to 255 (white). The result is a picture that, instead of a range of greys, has either black or white pixels, hence why it's called *binarisation*. The text on the binary, black-and-white image is easily recognisable as text to the software and is then converted to speech.

However, using one threshold for the whole picture does not work when trying to read from children's books with multi-coloured backgrounds. Bonnici and Cristina, with their team of research support officers, developed a more robust algorithm for the Doc2Speech project. Before the binarisation process, the algorithm has several other steps. First, it recognises and separates each line of text. Then, it segments each line into several sections or text windows. After these two steps, the binarisation is performed on each text window using a specific threshold automatically calculated for that particular text window. This way, the threshold is different for each window or section. This makes it less likely that the value of pixels that represent text is above the threshold or that

Dr Alexandra Bonnici
Photo by James Moffett

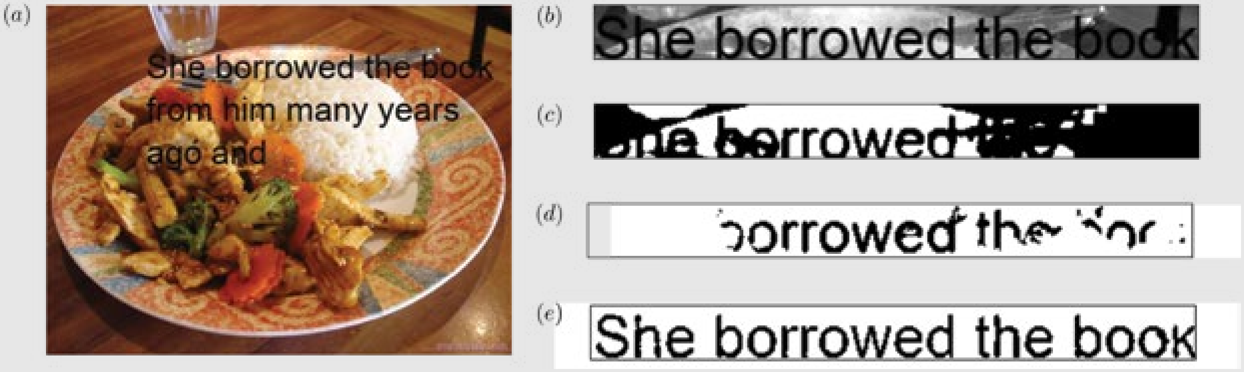

Figure 2: An example of how the Doc2Speech algorithm extracts text compared to other methods. (a) An example image from which text was extracted. The image was taken from a pre-established dataset, and the text was superimposed by the researchers. (b) A line of extracted text in grayscale. (c) and (d) The text extracted by other binarisation methods – Otsu and ROBIN respectively. (e) The text extracted by the algorithm developed in the Doc2Speech project.

pixels that represent background details have a value below the threshold. As shown in figure 2, binarisation algorithms that do not utilise Doc2Speech's extra steps fail to extract text when there is a large range of hues in the background. In the meantime, the algorithm developed for Doc2Speech manages to adapt to different backgrounds and still successfully extract the text.

Dr Ing. Stefania Cristina
Photo by James Moffett

The new algorithm can be applied to text written in any language. Naturally, it can be used to read Maltese; however, more work is needed to develop the speech part of text-to-speech. Text-to-speech systems use an AI voice to read the text aloud. A computer voicing one word at a time is one thing. Reading out complete sentences, including punctuation, with the correct tonality is a complex, manifold, and essentially human skill. Most people have heard the stilted, unnatural inflexion of bus announcements or GPS directions. While the robotic voice suffices for these services, a better reader is needed to realistically help people who will use text-to-speech technology in their everyday lives.

FUTURE AVENUES OF RESEARCH

As with any scientific progress, solving one challenge leads to others. In children's books, the text and the illustrations work hand in hand to tell a story, so simply reading the text is not enough for a complete understanding. Some software already exists where you can point a phone camera at something, take a picture, and the software extracts an explanation of what the picture is. This software could be used in combination with text-to-speech to give a holistic description of what is on the page.

Top to bottom: Erica Spiteri Bailey, Luke Abela, and Andre Tabone

Photos courtesy of the Docs2Speech Team

One problem is that the software that describes photos has been trained on real pictures, not on illustrations and cartoons. The software is completely baffled when encountering an illustration of a sloth sawing a plank of wood, for example, which is a fairly pedestrian thing to find in a children's book. Furthermore, a large dataset on which the algorithm could be trained does not exist yet.

A training dataset is a large number of photos or pictures with descriptions made by humans on which the algorithm could be trained so that it 'learns' to make descriptions of new photos by itself. To create a training dataset of illustrations with descriptions, the researchers took a DIY approach. They used a style-transfer technique to convert regular photos in the training dataset (which already have descriptions attached) into illustrations in different styles. Bonnici and Cristina do not expect this to be the only solution because the variety of styles of illustration in books is unlimited, while the style-transfer techniques are limited in how they can transform a photo.

NOT THE TYPICAL KIND OF ENGINEERING

A project like Doc2Speech is the work of a team. Research support officer Ms Erica Spiteri Bailey worked on the first step of the algorithm – extracting the lines of text individually. Mr Luke Abela worked on the binarisation of the windows of text, while Mr Andre Tabone worked on the segmentation of illustration images into their individual objects.

Projects like Doc2Speech might not be the first to come to mind when thinking about engineering research, but Bonnici and Cristina encourage students to join projects like theirs. 'Students who are interested in this work will have the satisfaction of advancing the field,' explains Bonnici, while Cristina points out, 'It could help people in need, and the images are fun to work with. It's a different type of engineering!' Their research combines programming and AI with art, linguistics, and social service. The technology developed helps people live better lives, and although conceived for a specific purpose, text-to-speech technology could have many other, as yet undiscovered, applications. T

This project is funded by the MCST Research Excellence Programme with grant agreement number REP-2022-006.