New life for old media:

Investigations into Speech Synthesis and Deep Learning-based colorization for audiovisual archives

Rudy Marsman[1,2], Victor de Boer[1,2], Themistoklis Karavellas[1], Johan Oomen[1]

1. Netherlands Institute for Sound and Vision, Hilversum, the Netherlands
2. Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

The Netherlands Institute for Sound and Vision (NISV) is the national audiovisual archive and media museum of the Netherlands. The collections comprise of over one million hours of audiovisual material. One of the collections is that of the so-called “Polygoon” newsreels from the 20th century, published under open licenses on the Open Images platform[1]. This paper outlines recent explorations where AI technologies are used to enrich this archival material to allow for new types of engagement.

Firstly, we investigated leveraging an existing, limited corpus of broadcast narration by a single person to build a working text-to-speech (TTS) system. In our case, we used the voice of Philip Bloemendal, a well known narrator, almost instantly recognisable for many people in the Netherlands. His voice announcing the Polygoon newsreel from 1946 to 1986, led to being nicknamed "the voice of the Netherlands". For this type of Limited Domain Speech Synthesis [1], we developed a slot-and-filler type TTS system which uses several heuristics to optimize the available words. One heuristic is *word decompounding,* as many Dutch words -which might not appear in the corpus- can be split up into compounds which have a higher probability of occurring in the corpus (ie. “sinterklaasoptocht” -> “sinterklaas” and “optocht”). A second heuristic uses a background thesaurus of *synsets* (Open Dutch Wordnet[2]) to replace words that do not occur in the corpus with synonyms or hypernyms. The heuristics were evaluated on four different corpora: contemporary news articles, 1970s news articles, contemporary e-books and Twitter messages (see Figure 1). Combining the heuristics produces the best results, with performance ranging from 49% word coverage for Twitter messages to 89% for contemporary news articles. In a user-test (*n*=8), 100% of participants reported understanding generated sentences and being able to correctly identify the speaker.

Secondly, we investigated the possibility of colorization of old black-and-white video footage from the Polygoon newsreel collection using Deep Learning approaches. We used a Convolutional Neural Network, pre-trained on 1 Million web images to colorize black and white still images [2]. The conversion pipeline developed firstly extracts the frames of a video on 24fps rate and a 200x200 pixel resolution. The images were then one-by-one presented to the neural network and colorized. Finally, the images are stitched back together to produce colorized videos. The pipeline was tested using six videos from the Open Images collection. The results

[1] http://openimages.eu
[2] http://wordpress.let.vupr.nl/odwn/

were published on the Open Images platform[3]. An example is shown in Figure 2. One of the colorized videos -after being shared on a social media platform- received over 61,000 views, 1,700 likes and was shared 521 times, illustrating the potential to engage new audiences.

Collection-specific TTS systems, can be used for audio-enrichments of archive material or for multimedia applications. The colorization of old media allows for a new view on existing images. NISV, following its mission to keep their audiovisual material *alive*[4], will continue using these emerging technologies and expand on them to enrich collections, enable new types of interaction and to further engage new audiences with archival material in unexpected ways. In the future, these emerging technologies will be used in the media museum operated by Sound and Vision, and on its public-facing online channels.

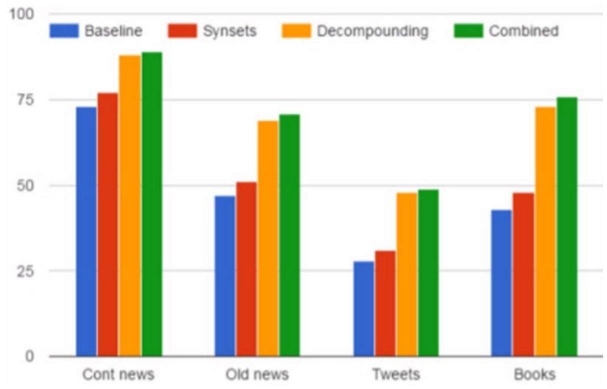

Figure 1: Percentages of words covered in each of the four corpora for the baseline algorithm, with each of the heuristics and the combination of heuristics.

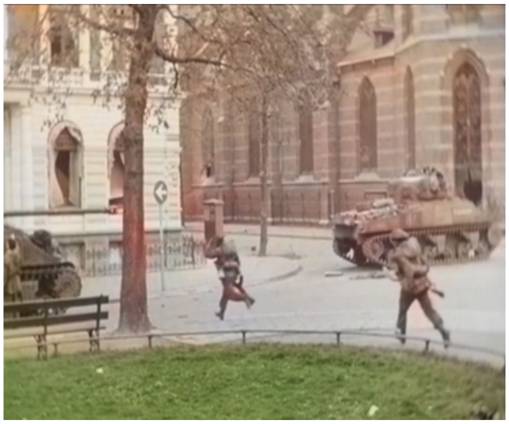

Figure 2: Screenshot of the colorized video “Bevrijding stad Groningen

References

[1] Marketa Jzova and Daniel Tihelka. Minimum text corpus selection for limited domain speech synthesis. In Text, Speech and Dialogue, pages 398–407. Springer, 2014.

[2] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. *arXiv preprint arXiv:1603.08511, 2016*

[3] https://www.openbeelden.nl/tags/ingekleurd

[4] http://www.beeldengeluid.nl/missie-visie-en-beleid

About the authors

Rudy Marsman holds an MSc. in Information Sciences from Vrije Universiteit Amsterdam. He conducted this research as part of his degree while embedded at NISV. He currently is working as a Health Insurance developer at Oracle inc.

Victor de Boer (v.de.boer@vu.nl) is an Assistant Professor at the Web and Media group of the Computer Science department of Vrije Universiteit Amsterdam and Senior Research Fellow at the Netherlands Institute for Sound and Vision. His research focuses on Artificial Intelligene applications for Cultural Heritage and Digital Humanities. He has been involved in a number of European and Dutch projects in this field focusing on (semi-)automatic content enrichment, and using Semantic Web technologies for data integration.

Themistoklis Karavellas (tkaravellas@beeldengeluid.nl) is a Scientific Software Engineer in the R&D department of the Netherlands Institute for Sound and Vision. His research interests lie in the field of Computer Vision and Machine Intelligence. He holds a BSc. in Computer Science and an MSc. in Information Science.

Johan Oomen (joomen@beeldengeluid.nl) is head of the R&D department at NISV and researcher at Vrije Universiteit Amsterdam. He is board member of the Europeana Foundation and board member of CLICKNL. His research at the VUA in on how active user engagement helps to establish more open, smart and connected cultural heritage. Oomen holds a BA in Information Science and an MA in Media Studies. He has given talks at leading conferences (SXSW, JTS), published numerous articles in journals and is lecturer at the ICCROM training course 'Sound and Image Collections Conservation'.