New Life for Old Media: Investigations into Speech Synthesis and Deep Learning-based Colorization for Audiovisual Archives

Rudy Marsman
Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands / Netherlands Institute for Sound and Vision, Hilversum, the Netherlands
rudymarss@gmail.com

Victor de Boer
Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands / Netherlands Institute for Sound and Vision, Hilversum, the Netherlands
v.de.boer@vu.nl

Themistoklis Karavellas
Netherlands Institute for Sound and Vision, Hilversum, the Netherlands
tkaravellas@beeldengeluid.nl

Johan Oomen
Netherlands Institute for Sound and Vision, Hilversum, the Netherlands
joomen@beeldengeluid.nl

Abstract - 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. This paper outlines recent explorations where Artificial Intelligence 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. Secondly, we investigated the possibility of colorization of old black-and-white video footage from the Polygoon newsreel collection using Deep Learning approaches.

Index Terms – Artificial Intelligence, audiovisual archive, colorization, deep learning, speech synthesis

INTRODUCTION

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 (http://openimages.eu). This paper outlines recent explorations where AI technologies are used to enrich this archival material to allow for new types of engagement. As it is the responsibility of NISV to ensure that all archives are available for re-use, by extension the institute is also interested in exploring ways to make interaction with the material easier and to increase exposure to their archives. To do so, this paper explores two options. The first of which is the research to using a famous anchorman's voice in a modern text-to-speech engine. Here, we focus on natural language processing and the determination to what extent the language used in the 1970s is still comparable enough to con-temporary Dutch. Another part of the research was the autonomous colorization of old black-and-white video footage using Neural Networks. We describe these two explorations in the following sections.

LIMITED CORPUS SPEECH-SYNTHESIS

I. Introduction

Famous Dutch anchorman Philip Bloemendal is praised for his iconic voice and characteristic way of pronouncing his news reports [1]. He was 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". Even though he passed away in 1999, NISV is interested in the use of the voice of Bloemendal for various speech applications.

On a more generic level, institutions might be interested in investigating leveraging an existing, limited corpus of broadcast narration by a single person to build a working text-to-speech (TTS) system. For this type of Limited Domain Speech Synthesis [2], we developed a slot-and-filler type TTS system which uses several heuristics to optimize the available words.

Traditionally, TTS systems focus on using diphones to construct audio. This means that as a basis, every possible diphone in a language has to be built manually either

by recording or by dissecting recorded phrases. An additional benefit of using limited domain speech synthesis is that it may be easier to construct a natural sounding sentence. Moreover, LDSS is also useful in applications where there is a small or limited corpus available. There is a relation between the number of slots in a sentence and how well the sentence is perceived. Adding too many slots may lead to unnatural sounding pauses within a sentences, whereas if all sentences have to be prerecorded the corpus would grow too large. A method of finding an optimal solution and an algorithm to determine from what sentences in the corpus a slot is to be taken is described in [2].

The limitation of phoneme based voice synthesis is further-more described by Habib et al.[3]. The main issue described is that phone level coverage of a corpus fail to cater to co-articulatory effects between adjoining phonemes. They also suggest that a corpus should ideally be selected from a large number of domains to ensure diversity. This may prove to be difficult in the context of Bloemendal, where most if not all of the corpus stems from news recordings. This may lead to a rather formal or limited application, but it may also strengthen the iconic voice of Bloemendal who is mainly known as an anchorman.

Current state of the art text-to-speech engines can make use of Deep Neural networks. An example of such research has been made by Lu et al. [4]. In their research in 2013 they demonstrated how Neural Networks can be used to map sequences of phonemes to acoustic descriptions from which speech waveforms can be generated. Although this approach differs from what we intend to do, this research should be noted to demonstrate the difference between current state text-to-speech engine and our research, which focuses more on corpus expansion.

Natural language processing and translating may prove to be challenging, as the meaning and interpretation of words are largely dependent on the sentence and context in which they are used. A statistical approach by Jiang et al. may be used to determine the intended meaning of a word and to find a proper synonym in its absence in the corpus [5].

II.Audio Corpus from Open Images

NISV has an extensive archive of media, including many episodes of the Polygoonjournaal, which was presented by Bloemendal. These videos are available for the general public and can be accessed using the framework set up by the Open Archives Initiative. The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient accessibility of content [6].

At present, roughly 3,300 videos originating from Polygoon are present in the collection (http://openimages.eu). We retrieved these, using a custom written script in Python filtering out footage that did not fall under the license of Beeld en Geluid and footage that did not contain Polygoonjournaals. After these videos were scraped, the video was split from the audio to reduce the size of the files for storage. The audio files were put through speech recognition software to determine what was said. This resulted in an xml file for each individual audio fragment containing every spoken word, labeled with an exact begin and end time.

III. Speech index and baseline system

Based on the text-to-speech results, an index was created listing the words available in the speech corpus. The resulting list contained roughly 35,000 unique words. This index is then used by a simple slot-and-filler text to speech system that looks up words in a target sentence in this index. This constitutes our baseline with respect to checking how much of a given target sentence is available to be spoken by the TTS system. Of course many words will not appear in this limited corpus. To expand the coverage of this baseline we investigate the effectiveness of two heuristics and the combination of these two.

IV. Synonym expansion

One heuristic is to use background thesaurus of synonyms in the form of Open Dutch Wordnet [7]. The thesaurus is used to replace words that do not occur in the corpus with synonyms or hypernyms. For each of these words, a synset is generated containing various or no synonyms. For each of these words, we checked whether this specific word could be found in the index. If this is the case, the original word was removed from the list of unpronounceable words. Loading the OpenDutchWordnet library into memory is relatively slow, taking up to 10 seconds on a Core I3 5005u.

V. Word decompounding

As the Dutch language is known for containing many compounded words, a first heuristic to increase coverage is decompound these words. For instance the word "Regenwater" (rain water) might not be found, but the words Regen and Water of which it is compounded may be found in the index.

However, to construct bigrams or trigrams of each of the 35,000 words in the corpus would take too much time if done via a traditional brute force approach. To counter this, for each candidate word we made a list of candidate compounds. We assume that each part of the trigram can have a length of no more than N-1, N being the number of letters in the candidate word. Iterating through the index and removing all items with a length greater than a certain value can be done fairly quickly in Python. Furthermore, we had to make bigrams and trigrams of all the candidate words and check whether they matched the original word when concatenated. An additional step to increase

coverage was to check whether the words would match if they were concatenated with an "s" in between (e.g. staat + s + hoofd = staatshoofd). We constructed two nested for-loops, one checking for bigrams and one checking for tri-grams. If the bigram did not match, then the entire nested trigram loop would not be executed as well. This drastically improved performance time-wise beyond the brute-force method. If either a bigram or trigram would match with the original word, that word would be removed from the list of words that could not be pronounced. Both the Python scripts as well as the XML corpus are available online at [8].

VI. Evaluation

The heuristics and the combination of the two were evaluated on four different corpora:

- **Contemporary news articles.** This corpus consists of 1,022 sentences with 2,743 unique words, retrieved from articles scraped from the news sites of online news websites (http://www.nu.nl and http://www.nos.nl). The intuition behind the use of this data set is that the language used in these articles closely resembles actual contemporary natural language and the data is relatively clean and easily accessible.
- **1970s news articles,** The second corpus we selected to test our approach on was the set of 50 news articles out of newspapers in the 1970s (the same period as the Polygoonjournaals). These were retrieved from the National Library open newspaper collection (http://delpher.nl). The data set is not clean either as all the articles are scanned and have been processed using OCR algorithms which may not have yielded clean results. This may explain why there are so many more unique words in the corpus. The corpus consists of 16,191 unique words and 2,626 sentences.
- **Twitter messages.** A data set of Tweets has been selected as the use of social media has become more and more widespread in the past years. This popularity means there might be demand for a renewal in human computer interaction when it comes to these message. We expect many unfound words in this corpus (including smileys, typos and words like 'hahaha'). The twitter corpus was scraped using the Twitter API in Dutch language and consists of 27,180 words and 8,937 sentences.
- **E-Books.** The last corpus is made up of sentences from six public domain e-books from 1800-2010. Some of these e-books are the result of scanning and OCR (which introduces many incorrect words) The corpus consists of 2,657 unique words and 5,610 sentences.

For each of the evaluation corpora, we evaluate our methods both on the number of unique words found in the articles as well as the number of sentences that can be pronounced. Figure 1 and 2 show the percentages of covered words and sentences respectively for each of the four corpora and each of the four settings (baseline, with synset heuristic , with de-compounding heuristic and with both heuristics).

Combining the heuristics produces the best results, with performance ranging from 49% word coverage for Twitter messages to 89% for contemporary news articles. When looking at sentences, the coverage is considerably lower. Especially tweets have low coverage, originally only 2 percent of all tweets can be pronounced. Books perform best with a score of 16 percent. However it should be noted that a single unpronounceable word in a sentence renders the entire sentence as not found. In the case of twitter, many tweets contain URLs or emoticons rendering the entire sentence unfound.

Another interesting fact is that none of the corpora seem to benefit from looking for the synonyms heuristic except e-books. The comparison to the unique words should not be made to easily however, as not each word is used as frequently in natural language. Most likely synonyms can be found for most words in the sentence, but the unfound words are distributed to most of the sentences.

In a very preliminary user-test (n=8, all participants are students aged 18-24), 100% of participants reported understanding generated sentences and being able to correctly identify the speaker. The background music played in the Polygoonjournaals, which remained in the audio snippets did not seem to be explicitly decrease the comprehensibility of the produced audio, although the audio quality was not perceived as perfect and far from natural.

VII. Discussion

The limited corpus of words pronounced by Philip Bloemendal appears to be sufficient to cover most sentences currently used in the Dutch language. Considering contemporary news articles, close to 90% of the distinct words in all articles were pronounceable. Although other corpora such as tweets, e-books, and older news articles achieved lower coverage, this may be due to other factors such as OCR or grammatical errors. The search for synonyms to increase coverage was successful but the most successful technique implemented was the de-compounding of words. The combination of both techniques yielded the highest result in both the coverage in sentences and words.

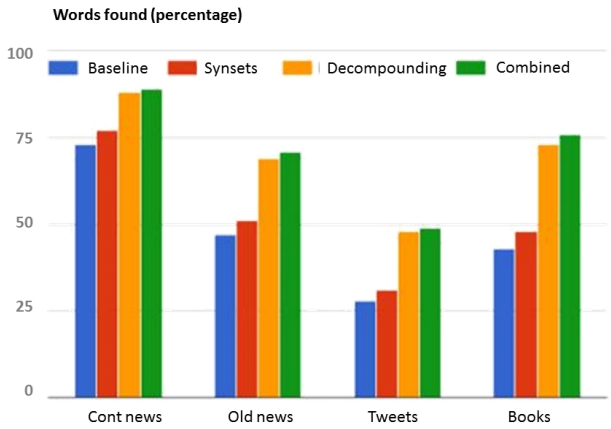

FIGURE 1: PERCENTAGES OF WORDS COVERED IN EACH OF THE FOUR CORPORA FOR THE BASELINE ALGORITHM, WITH EACH OF THE HEURISTICS AND THE COMBINATION OF HEURISTICS.

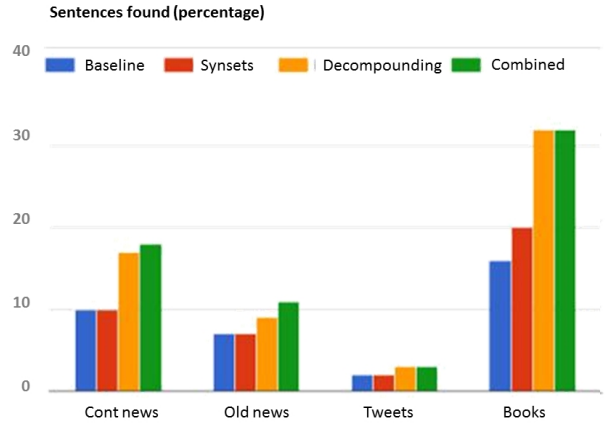

FIGURE 2: PERCENTAGES OF SENTENCES COVERED IN EACH OF THE FOUR CORPORA FOR THE BASELINE ALGORITHM, WITH EACH OF THE HEURISTICS AND THE COMBINATION OF HEURISTICS.

THE USE OF DEEP NEURAL NETWORK IN THE COLORIZATION OF VIDEOS

I. Introduction

Neural Networks have advanced enough to provide reasonable results in colorizing black-and-white photos. This can be extended to videos as well, although this introduces minor artifacts. Computational time is an issue when dedicated hardware is absent and when videos are colorized.

Many videos in the archives of NISV are currently stored in black and white. Colorizing these videos may draw attention and revitalize interest in these videos. The mission of NISV contains, among others, the statement that the re-use of the cultural heritage is to be promoted. In the context of this goal, implementing the cutting edge of current artificial intelligence technology to breathe new life in the old footage.

We investigated the possibility of colorization of old black-and-white video footage from the Polygoon newsreel collection using Deep Learning approaches. In recent years computing power has become sufficiently inexpensive to implement more and more complex neural networks, now finally being able to make accurate predictions in real world situations such as colorizing images. Especially the use of graphics card, capable of doing many computations in parallel, can be used to train these networks. The Convolutional Neural Network we will use for our research has been designed and rained by Zhang et al. on 1 Million web images to colorize black and white still images [9]. In training, all these images were converted from RGB to Grayscale, resized to 200 by 200 pixels (to the input layer size of the network) and fed into the neural network. The network then had to estimate for each pixel what color it should be. Correct assumptions were rewarded and remembered, wrong assumptions were discarded. A flaw in the network trained by Zhang et al., as is also described in their research, is that the network some-times estimates bright colors in areas with low contrast.

II.The pipeline

Building on this pre-trained network, our conversion pipeline firstly extracts the frames of a video on 24fps rate and a 200x200 pixel resolution. This is achieved with off-the-shelf FFMPEG tool (https://www.ffmpeg.org/). The frame images were then one-by-one presented to the neural network and colorized. The neural network was implemented in the TensorFlow framework (http://tensorflow.org). Colorizing an single frame using an Intel Core i5 6500 running at 3300mHz took roughly 7 seconds although total running time increased because memory errors forced us to re-colorize failed images often. Finally, the images are stitched back together to produce colorized videos using the FFMPEG tool.

III. Results

The pipeline was tested using six videos from the Open Images collection. The results were published on the Open Images platform using the open Creative Commons CC-by-SA license. These can be found at https://www.openbeelden.nl/tags/ingekleurd. Example screenshot of two videos are shown in Figure 3 and Figure 4.

Looking at Figure 4, it is clear that some artifacts appear. Part of the background behind the person of Wiesenthal has been colorized in the same tint as his skin. This might be due to the low resolution of the Neural Net - the image was scaled down to 200x200 pixels,

colorized, and the colorized image was laid over the original black and white image. Using linear extrapolation to colorize pixels in between artifacts such as these may appear. Furthermore, it seems that the color of the background it not equal on the left and right sides of his face. The left side accurately colorized the background a blue tint, as it is supposed to be a lake, but the right side has a brown tint. In practice, adjacent frames should be colorized similarly. This is addressed by Ruder et al. who designed a method to stylize videos whilst focusing on consistency between frames [10].

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.

FIGURE 3: SCREENSHOT OF THE COLORIZED VIDEO "BEVRIJDING STAD GRONINGEN "

FIGURE 4:SCREENSHOT OF THE COLORIZED VIDEO "ACTIE VOOR WIESENTHAL"

CONCLUSIONS

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 , 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.

REFERENCES

[1] Philip Bloemendal beeld en geluid. http://www.beeldengeluidwiki.nl/index.php/PhilipBloemendal accessed 2016-08-22

[2] M. Jzova and D. Tihelka. Minimum text corpus selection for limited domain speech synthesis. In Text, Speech and Dialogue, pages 398-407. Springer, 2014.

[3] W. Habib., R. H. Basit, S. Hussain & F. Adeeba Design of speech corpus for open domain Urdu text to speech system using greedy algorithm. In Conference on Language and Technology (CLT14), 2014.

[4] H. Lu, S. King, and O. Watts. Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis. Proc. ISCA SSW8, pages 281-285, 2013.

[5] J. J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008, 1997.

[6] Open Archives Initiative. https://www.openarchives.org/. Accessed: 2016-06-23.

[7] M. Postma, E. van Miltenburg, R. Segers, A. Schoen, and P. Vossen. t "Open Dutch WordNet." Proceedings of the Eight Global Wordnet Conference, Bucharest, Romania. 2016.

[8] R. Marsman Github Code depository. https://github.com/rudymars/general. Accessed: 2016-08-22

[9] R. Zhang, P. Isola, and A.A. Efros. Colorful image colorization. arXiv preprint arXiv:1603.08511, 2016

[10] M. Ruder, A. Dosovitskiy, and T. Brox. Artistic style transfer for videos. arXiv preprint arXiv:1604.08610, 2016.

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 is an Assistant Professor at 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 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 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'.