The implementation of Artificial Intelligence in the re-use of old media corpora

Rudy Marsman
VU University
De Boelelaan 1105, 1081 HV Amsterdam
rudymarss@gmail.com

Preface
I have conducted this research partially on site at the Netherlands Institute for Sound and Vision (NISV).NISV is a cultural-historical organization of national interest. Sound and an vision has one of the largest audiovisual archives in Europe and ensures that the material is optimally preserved for (re) use. The collection of sound an vision contains the complete television archive of the Dutch public broadcasters and includes footage from old newsreels, to which the broadcasts of Philip Bloemendal belong. As it is the responsibility of NISV to ensure that all archives are available for re-use, by extension they are also interested in exploring ways to make interaction easier and to increase exposure to their archives. To do so, I have explored two options. The first of which is the research to using Philip Bloemendals, a famous anchorman, voice in a modern text-to-speech engine. To do so I have mainly focused on natural language processing and the determination to what extent the language used by Bloemendal in the 1970s is still comparable enough to contemporary Dutch. Another part of the research was the autonomous colorization of old black-and-white video footage using Neural Networks. The following research is split up into two parts for each part of the research respectively. The first part, the text-to-speech research, starts below.

Abstract
A useful way to transfer information may be to use spoken voice, such as the messages broadcast on train stations, trains, buses and airports. Those voices do not have to originate from humans but can also artificially be synthesized. The goal of this research is to establish whether a small, limited corpus can be used to construct a functioning text-to-speech function based on the voice of the Dutch news reader Philip Bloemendal. Whenever the corpus is not sufficient, we intend to use the Open Dutch Wordnet Thesaurus to construct a sentence of similar semantic meaning which can be synthesized from the corpus. This text-to-speech engine may improve user interaction and familiarity with the media collection NISV has gathered.

Keywords
Text-to-speech,Thesaurus, voice synthesis, Natural Language Processing,Neural Networks

1. INTRODUCTION

Famous Dutch anchorman Philip Bloemendal is praised for his iconic voice and characteristic way of pronouncing his news reports [1]. Public familiarity with his voice has lead him record messages announcing train stations in Amsterdam. However, as he passed away in 1999 and new train stations continued to be built, eventually his recordings had to be replaced with modern ones which did include newly built stations. The corpus of words spoken by Bloemendal is large in size and already structured, but as it is no longer possible to expand this by recording new messages it may be necessary to incorporate ways to circumvent this problem, for instance by looking for sentences with similar meanings but that are constructed of words that are present in the corpus. The Netherlands institute for Sound and Vision (NISV) is interested in the use of the voice of Bloemendal for various applications. The problem posed by the limited corpus may be interesting to solve in general because the research may prove that a smaller corpus than previously thought is required to construct a functioning TTS application.

A way to supplement missing words from the corpus is to implement a thesaurus such as Open Dutch Wordnet to find synonyms for missing words. As specific words such as smartphone or laptop may not be present in the dated corpus, words with similar meaning such as phone or mobile computer may be present.

Another solution to do this is to use the current corpus as the building blocks for a text-to-speech engine. Various methodologies exists to achieve this and we shall mainly focus on the so called Limited Domain Speech Synthesis (LTDS) [2]. Various types of LTDS exist, but the most straightforward example is the use of a limited set of fixed sentence structures with variable items (called slots in the research by Juuzova et al.), and sets of words to be placed in the slots.

2. RELATED WORK

Traditionally, text to speech 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. Many forms of LD text representations are available depending on the size of the domain. A trivial example is the application within a train station where sentences such as ”the [—-] bound train departs from platform [—-]”, where the variable entries (called slots) can be replaced with a set of words. 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 by Juzoka and Tihelka (2014)[2].

The limitation of phoneme based voice synthesis is furthermore 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. Auditive behaviour of a phoneme is partially influenced by its neighbouring phonemes, leading to a quickly growing set of phonemes required to construct text. Ideally, to combat this, tri-phone coverage can be applied. However, the number of combinations of 3 phonemes of a language is expensive to construct.

Furthermore, Habib et al. suggest that a corpus should ideally be selected from a large number of domains to ensure diversity. This may proof 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 expension.

Natural language processing and translating may prove to be challenging, as the meaning and interpretation of words are largely dependand 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].

Extensive research has been done on what words and types of words are used most often in the Dutch language [6]. Moreover, the research conducted by Zijlstra et al. describes in what context certain words are used most often which may be beneficial in determining the domain the text-to-speech algorithm will perform best.

In order to find proper synonyms and to use those to construct sentences of similar meaning, we intend to use the Open Dutch Wordnet Thesaurus [7]. Open Dutch Wordnet is constructed based on the Cornetto database, the Princeton Wordnet and various open sources and contains 51588 synsets. Synsets are groups of data that are considered to be semantically equivalent. Open Dutch Wordnet is especially beneficial for our research as it is especially designed for the Dutch language.

3. INPUT DATA

The Netherlands Institute for Sound and Vision 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 [8].

At present, roughly 3300 videos originating from Polygoon are present in the collection. We have gathered 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. Afterwards these audio files were put through speech recognition software to determine what was said. This resulted in an xml file for each individual audiofragment containing every spoken word, labeled with an exact begin and endtime.

4. RESEARCH QUESTIONS

The problem is that Philip Bloemendal can no longer expand the corpus of his voice. Additionally, parts of the corpus may be lost or unavailable for research. We wish to expand the number of sentences we can construct out of this limited corpus. The main research question for this thesis will therefore be:

How can the limited corpus of audio recordings of Bloemendal be used to construct a TTS engine? To help answer these questions, several sub questions have to be answered.

What percentage of the domain specific sentences can be constructed with the current corpus? The usage of words in the Dutch language is not homogeneous; some words and diphones are used more than others. It may be that if the most frequently used phrases can be constructed and some rarely used ones miss out the TTS engine still is perceived as complete. We intend to determine the most frequently used words by a literature research and incorporate the results to answer this question. This may furthermore be interesting for research as the domain of the corpus is strictly news messages from the 1960s and 1970s. To what extent natural language has evolved since then and to what extent the news domain applies to everyday or other specific domains is to be determined.

Can a thesaurus be used to find alternatives for missing words and phrases Natural language is constantly evolving and new words emerge. A clear example may be the word Smartphone. Although telephones have been common good for almost a century, mobile phones in general and smartphones in particular have only gained popularity in the last two decades. As the corpus predates this period, it may be that words as mobile phone and smartphone miss from the corpus. This may also apply to other ways of saying as well; the language used in the 1970s is viewed by many contemporary people as old fashioned. We intend to apply a thesaurus to find synonyms of missing words and phrases to still be able to construct sentences.

How well is the text-to-speech engine recognizable as Philip Bloemendal? Philip Bloemendals voice can be considered iconic. However, exactly what aspects of his voice make his so iconic may be abstract and subjective. We intend to carry out a user study to evaluate the performance of the voice generator. Specific applications may lie in a web application where users can create their own sentences to be spoken out by Bloemendal, or a more survay based approach where users are given sentences which are actually spoken out by Bloemendal compared to sentences constructed from the corpus and users have to label the sentences.

5. APPROACH

In order to carry out this research, two parts of the project had to be completed. First of all a tool to construct acoustic sentences had to be created. Secondly, we had to revise a way to replace missing words in the corpus by words that could be pronounced.

5.1 Text to speech

To convert text to speech we have constructed a Python script which makes use of FFMPEG to cut and concatenate parts of audio fragments. The script used a list of XML files, each for every individual Polygoonjournaal, and checked single words and listed them. The XML files contained all the words spoken in the Polygoonjournaal but also noted a language model and an acoustic model. The language model is the probability that the listed work is correct based on other words neighbouring it. The acoustic model is the probability that a word is correctly recognized based on the distance between the vector of the audio fragment itself and the vector where the speech recognition software is trained on. Generally, words with more syllables gave the software more data to work with and resulted in better acoustic model scores. Whenever a word was spoken multiple times in one or more Polygoonjournaals, the word with the highest acoustic model was selected. To test the usability of the proposed, various articles were gathered and we attempted to have Bloemendal pronounce those words. However, no articles were fully pronounceable. All articles contained sentences which could not be pronounced and many sentences contained words that could not be pronounced. A collection of all words, spoken by Bloemendal, can by found on github [9]. These words are split into XML files, each file represents one Polygoonjournaal. Within these XML files each word has a label, along with an ID (the word that is spoken), the begintime, the endtime, and either confidence or AM/LM labels to signify confidence that the measurement is correct.

5.2 Word replacement

As it is to be expected that not all words in the contemporary Dutch language are present in the corpus, a research has to be conducted to determine what words are missing, how many words are missing and what measures can be implemented to replace missing words.

5.2.1 *Baseline*

First of all a baseline method had to be constructed. We did so by extracting all the words from the Polygoonjournaals and deleted duplicate words. The resulting list contained roughly 35000 unique words. Secondly, we compared these words with all the unique words in various corpora (see section 6). The baseline was measured as-is, with no advanced methods applied. The code for generating a list of unique words can be found on GitHub [10] in woordenlijst.py.

5.2.2 *Synsets*

The first step we took to increase coverage was to find synsets of words that could not be pronounced. The intuition behind this was that a word could be replaced by any synonym of that word without drastically altering the meaning of the sentence. To do so we made use of OpenDutchWordnet. After the baseline test was set a list of unpronouncable words remained. For each of these words, a synset was generated containing various or no synonyms. For each of these words, we checked whether this specific word could be pronounced and if so, 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.

5.2.3 *Decompounding*

As the Dutch language is known for containing many compounded words, the next step to increase coverage was logically to decompound these words. For instance the word ”Regenwater” might not be found, but the words Regen and Water of which it is compounded may be found. However, to construct bigrams or trigrams of each of the 35000 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. The first heuristic we applied is 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 a list of 35000 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 2 for loops nested in each other, one checking for bigrams and one checking for trigrams. If the bigram did not match, then the entire nested trigram loop would not be executed as well. This drastically improved performance time-wise. 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.

5.2.4 *Combined method*

The combined method was the combination of looking for synonyms and decompounding words, although decompounding synonyms was not taken into account due to the quickly rising computational cost of such a method. We chose to look for synonyms first as this was quicker to do than decompounding. Additionally, words that have actually been pronounced by Bloemendal in contrast to the concatenation of two or three other words may also sound more natural, although this was not tested. A graphical representation of the algorithm can be found on the next page in figure 1.

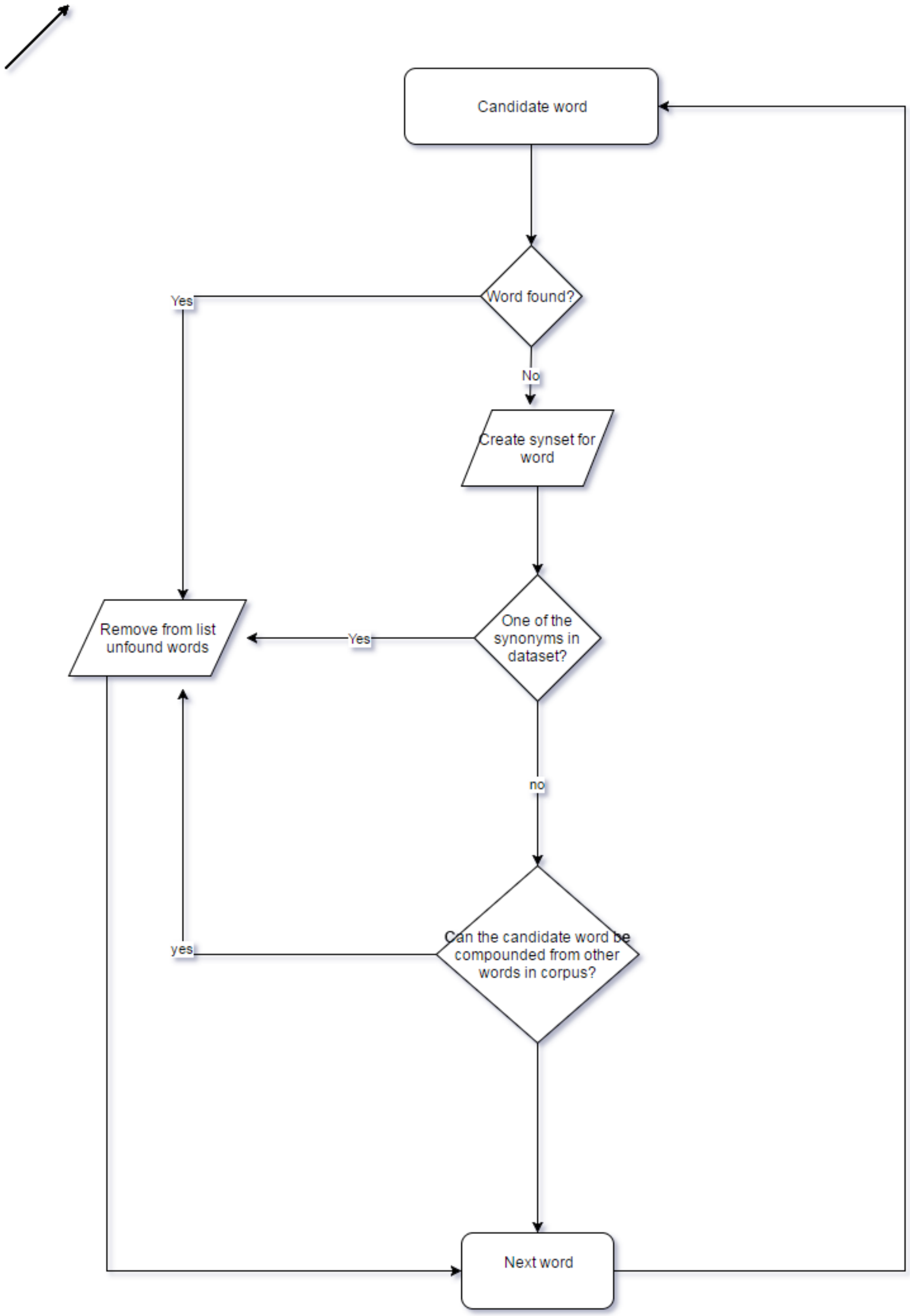
Candidate word
Word found?
Yes
No
Create synset for word
Remove from list unfound words
Yes
One of the synonyms in dataset?
no
yes
Can the candidate word be compounded from other words in corpus?
Next word

Figure 1: Graphical representation of algorithm

6. EVALUATION

6.1 Coverage experiment

Design

In order to test the effectiveness of the word replacement techniques implemented, we have gathered 4 different data sets from different domains. These are 50 contemporary news articles, 50 news articles written in the 1970s, 1000 tweets (messages from the social medium Twitter) and lastly 3 free e-books. For each corpus, we extracted all the unique words. We matched this with the unique words in our data set. Our data set, consisting of 3300 audio fragments, consists of roughly 35000 unique words.

6.1.1 Contemporary articles

The first corpus we selected to test our approach on was the data set of contemporary news articles. These articles are scrapes from the news sites of www.nu.nl and www.nos.nl. The intuition behind the use of this data set is that the language used in these articles closely resembles actual natural language and the data is relatively clean and easily accessible. The news articles consist of 2743 unique words and 1022 sentences.

6.1.2 News articles from the 1970s

The second corpus we selected to test our approach on was the set of 50 news articles out of newspapers in the 1970s. The intuition behind this is that language may evolve, but the language used in this corpus stems from the same domain and time of the corpus we constructed out of the audio fragments. However, as newspapers from this time were actually printed on paper there were no original machine readable files available. The corpus we constructed is not entirely clean with not just news appearing, but also other information displayed on the page such as weather reports and advertisements. Advertisements, for instance, are particularly sensitive to be pronounced in text-to-speech engines as they are prone to consist of abbreviations. Newspapers typically charge for each printed letter which leads to users abbreviating their advertisements. Rather than printing "zo goed als nieuw", "z.g.a.n." (as good as new) would be printed for instance. 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 16191 unique words and 2626 sentences.

6.1.3 Tweets

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 messages and a 1970s anchorman pronouncing tweets might be a refreshing change. Additionally the Twitter API allows for easy scraping of tweets of a specific language. This way we could gather a corpus of 1000 tweets. The language of Tweets may differ greatly from what Philip Bloemendal has used in his broadcasts. Tweets are limited to 140 characters and the use of so called URL shorteners, to share links without exceeding this limit, is widespread among twitter users. Additionally the lack of auditive information, facial expressions or other nonverbal communications may lead users to exploit other ways to clarify the meaning of their tweets. Rather than actually laughing, Twitter users can actually write out "hahaha", a word which is absent in the corpus of Bloemendal. The twitter corpus consists of 27180 words and 8937 sentences.

6.1.4 Books

The last corpus we have used to test our approach on is E-books. 6 e-books have been gathered from the public domain, half of them are originally digitally written and stem from the ere of the early 2000s through 2010 and the other half are books which are scanned in and converted into a more machine friendly format using OCR software. These books are written from the late 1800s through the early 1960s. Especially this part of of the corpus is prone to both errors in scanning, damage in the original books themselves or changes of language. The rationale behind the use of this corpus is that although the total coverage may be low, there are many unique words present and this may prove a good insight in how well the approach performs relatively to the baseline score. The corpus consists of 2657 unique words and 5610 sentences.

6.2 Results

We have tested our method both on the number of unique words found in the articles as well as the number of sentences that can be pronounced. The absolute tables shows the absolute number of sentences and words found, the percentage table shows the percentage of unique words and sentences that are found. The number in parenthesis shows the growth compared to the base number. It is clearly visible that contemporary news articles perform best. This is most likely due to the fact that these articles are natively digital meaning there are no OCR errors involved. Tweets score by far the worst, and this is most likely due to the abundance of URLS, emoticons and generally the bad grammar of twitter users. Old news articles perform relatively well especially considering the fact that these articles may contain scanning errors. When looking at words that are not found, most of them are words such as 'verantwoordelvjkheid' (which should be verantwoordelijkheid) or 'driehondex" (which should be driehonderd). Additionally, many words in the list of words not found generally are gibberish and clearly noise (yse, iven, rttaan). The same applies to books, of which the corpus partly consists of scanned books.

The relative improvement for each corpus seems to follow a general trend though. Each corpus' words that can be pronounced increases by roughly 4 percent when synonyms of words are implemented. The most growth stems from decompounding. Books benefit the most from decompounding, the pronounceable corpus growing by 27 percent. Contemporary news articles grow the least with 12 percent but this may be due the fact that the coverage was all ready high. In this case, many of the words not found are names or foreign words.

When looking at sentences, the coverage is 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 sentences renders the entire sentence as not found. In the case of twitter, many tweets contain urls or emoticons rendering the entire sentence unfound. In the case of Old news or Books a single OCR error removes the sentences from the list of pronoun cable sentences. Finally, when a name is mentioned in a sentence the odds are high that the sentences cannot be pronounced. Many news articles and sentences in books contain the names of people or characters.

Another interesting fact is that none of the corpora seem to benefit from looking for synonyms except 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. Another explanation may be that compounded words make up the majority of the language. Therefore decompounding the unfound words greatly increases coverage.

Dataset	Unique words	Unique words found	after synsets	after decompounding	combined method
Cont. News	2743 (100%)	2019 (73%)	2106 (77%)	2414 (88%)	2448 (89%)
Old news	16191 (100%)	7703 (47%)	8261 (51%)	11305 (69%)	11541 (71%)
Tweets	27180 (100%)	7692 (28%)	8446 (31%)	13209 (48%)	13440 (49%)
Books	26575 (100%)	11440 (43%)	12922 (48%)	19578 (73%)	20207 (76%)

Dataset	Unique sentences	Unique sentences found	after synsets	after decompounding	Combined method
Cont. News	1022 (100%)	106 (10%)	110 (10%)	176 (17%)	186 (18%)
Old news	2626 (100%)	183 (7%)	190 (7%)	230 (9%)	301 (11%)
Tweets	8937 (100%)	174 (2%)	181 (2%)	276 (3%)	296 (3%)
Books	56100 (100%)	9380 (16%)	11380 (20%)	18154 (32%)	18270 (32%)

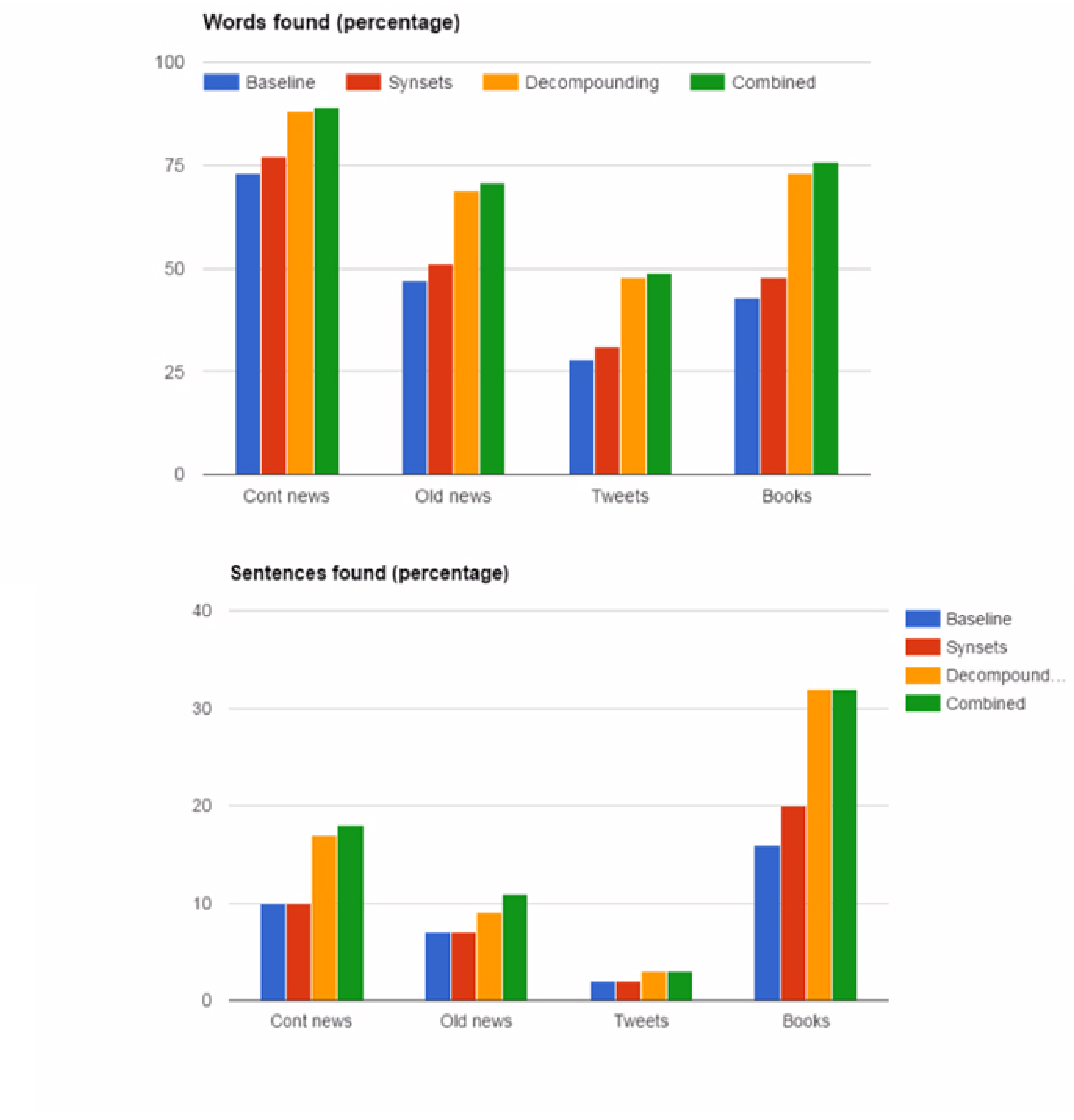

Figure 2: Coverage results

6.3 Understandability

6.3.1 Design

The text to speech evaluation took place among 8 users. These users, all students with ages between 18 and 24, were given audio files concatenated by our text to speech engine. First of all we constructed various sentences and words and asked the users what the words or sentences were. We did not measure time, but only whether they understood the words or sentences spoken by asking them to repeat the message. We also asked participants to fill in some words to be pronounced, chosen by them. The users where then given the chance to reflect on the audio with regards to understandability. The research took place face-to-face with a laptop running a python script. During the first phase of the experiment users did not see the screen of the laptop and could only hear the audio from the speakers. During the second phase, where users entered their own phrases, they could see the screen and thus 'read along' with the spoken words. The experiment took a maximum of 10 minutes per participant.

6.3.2 Results

In all cases, the audio was understandable and the users could tell what the spoken sentence was. However, users sometimes had difficulties doing so. It may be the case that the users extracted information from words they did understood and used this to fill in missing words. When users could enter their own sentences to pronounce the reactions were more positive when asked how well understandable the produced sentences were. Words with few syllables tended to be wrongly classified. Especially words with only single syllables were completely mislabeled by the speech analysis software. However, this could easily be solved by manually labeling all the one syllable words in a language. The background music played in the Polygoonjournaals did not seem to be explicitly decrease the comprehensibility of the produced audio files, although the audio quality was not perceived as perfect and far from natural. The newsreader was always recognized as either Philip Bloemendal or otherwise participants correctly stated that the voice was from 'that iconic newsreader'. However, it should be noted that the participants already knew the goal and title of the research and could derive the name of the newsreader from there. We did not observe users trying to have Philip pronounce one syllable words.

6.4 Conclusion

The corpus of words pronounced by Philip Bloemendal appears to be sufficient to cover most sentences currently used in our 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 errors or errors in grammar. In a natural setting users may be able to reformulate their sentences based on the missing words. The search for synonyms to increase coverage was successful but the most successful technique implemented was the decompounding of words. The combination of both techniques yielded te highest result in both the coverage in sentences and words. Although this is computationally expensive, proper programming techniques may greatly reduced computational time. A lot of time is spent in initializing the Open Dutch Wordnet and the opening of the corpus taking up to 10 seconds on a Core i3 5005u. The audio produced was good enough to be recognized as Philip Bloemendal.

The audio quality was good enough to be recognized as Philip Bloemendal. However, the quality of the audio was far from perfect. A rule based approach to filter out relatively rare occurring events, such as one syllable words and numbers, manually and only use the speech recognition software on the rest of the words may be beneficial to improve this. Additionally, the music which plays on the background of all the Polygoonjournaals may be at least partially filtered out by subtracting the original audio files from the Polygoonjournaals. To to so however the original audio track would need to be found.

7. REFERENCES

[1] Philip Bloemendal beeld en geluid. http://www.beeldengeluidwiki.nl/index.php/Philip$_{B}$*loemendal.A* 2016 – 01 – 24.

[2] Markéta Jzová and Daniel Tihelka. Minimum text corpus selection for limited domain speech synthesis. In *Text, Speech and Dialogue*, pages 398–407. Springer, 2014.

[3] Wajiha Habib, Rida Hijab Basit, Sarmad Hussain, and Farah Adeeba. Design of speech corpus for open domain urdu text to speech system using greedy algorithm.

[4] Heng Lu, Simon King, and Oliver 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] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[6] Hanna Zijlstra, Tanja Van Meerveld, Henriët Van Middendorp, James W Pennebaker, and RD Geenen. De nederlandse versie van de âĂŸlinguistic inquiry and word countâĂŹ(liwc). *Gedrag Gezond*, 32:271–281, 2004.

[7] Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. Open dutch wordnet.

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

[9] XML files - word dataset. https://github.com/rudymars/general/blob/master/xml%20files Accessed: 2016-08-22.

[10] Code depository. https://github.com/rudymars/general. Accessed: 2016-08-22.

The use of Deep Neural Network in the colorization of videos

Part 2

Rudy Marsman
VU University
De Boelelaan 1105, 1081 HV Amsterdam
rudymarss@gmail.com

1. ABSTRACT

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.

Keywords

Text-to-speech,Thesaurus, voice synthesis, Natural Language Processing,Neural Networks

2. INTRODUCTION

Although many forms for Moore's law circulate the internet, the general sentiment of each version is that the number of transistors on a computer chip doubles roughly after two years and with it, the computations said chip can perform per second whilst maintaining its price. Current progressions in technology have made it possible to train computationally expensive Deep Neural Networks, which are an advanced form of Artificial Intelligence. We investigate the feasibility of implementing a Neural Network to colorize old footage of Open Beelden, which falls under the jurisdiction of the Netherlands Institute of Sound and Vision.

3. PROBLEM STATEMENT

Many videos in the archives of NISV are currently stored in black and white. Although this is not inherently flawed, colorizing these videos may draw attention and revitelize 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.

4. RELATED WORK

Artificial Neural Networks have been described as early as 1970, although computing power in those days was expensive and computer use was not widespread [1]. However the potential use of computer models designed akin patterns found in nature was already noticed. 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 Convulutional Neural Network we will use for our research has been designed and trained by Zhang et al. [2]. Zhang et al. trained a Neural Network on over a million images. In short, all these images were converted from RGB to Grayscale, resized to 200 by 200 pixels (to fit the input layer size of the network) and fed into the neural network. The computer 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 sometimes estimates bright colors in areas with low contrast.

Examples of colorized images are shown in figures 1 and 2. It seems to be the case that the network is particularly well fit to predict colors of the sky, of foliage, of hair and of skin. It should be noted however that these figures may be cherry picked by the author to only show the cases of success.

Research has also been done to implement Neural Networks to stylize videos. Initially, the easiest approach to colorize videos using the method described above, would be to extract each frame and colorize it individually. Although this approach would be easy to implement but the flaw in it would be the assumption that each each frame exists completely individually from each other. In practice, adjacent frames should be colorized similarly. This flaw is addressed by Ruder et al. who designed a method to stylize videos whilst focusing on consistency between frames [3]. Their approach however is even more dependant on raw computing power than the approach Zhang et al. Ruder et al. go as far as to say using a CPU to stylize video is impractical and recommend a GPU with at least 4GB of memory for a video with a resolution of 450x350 pixels.

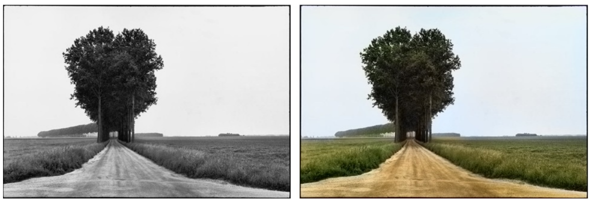

Figure 1: Colorized photo by Zhang et al.

5. INPUT DATA

The Netherlands Institute for Sound and Vision has an extensive archive of old black-and-white video footage. These videos are available for the general public and can be accesed

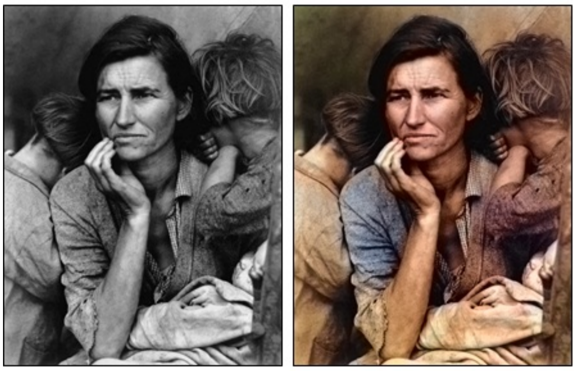

Figure 2: Colorized photo by Zhang et al.

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 accesibility of content [4].

At present, roughly 3300 videos originating from Polygoon are present in the collection. We have gathered 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. For our research we shall only colorize two videos. The videos we used are shot at the campus of the VU University, Uilenstede, in the 1970s. Another video we shall colorize is of a meeting with Simon Wiesenthal, famous for hunting down war criminals. These videos were stripped of audio and all their frames were extracted at a frame rate of 24 frames per second. The frames were then resized to 200 by 200 pixels to fit the input layer size of the colorization neural network.

6. RESEARCH QUESTIONS

We shall focus on the feasability of colorizing black and white footage. To determine this, we shall focus on the main question:

How can Neural Networks be used to autonomously colorize black and white video footage? To help answer these questions, several sub questions have to be answered.

How well do the colorized videos match real life colors? Black and white photos contain no information on color and although by looking at what is shown on the footagge a reasonable assumption can be made, whether a color is correct remains subjective. For instance a car can be blue in real life but if done properly, if colored red the car would be incorrectly colored but still seem correct. The same applies for instance on hair color, skin color and the color of cloth.

How well is the Neural Network by Zhang suited for video? As previously mentioned, our approach considers each frame of the video completely individual and does not act on temporal information or distance to other frames. As such, it may be the case that inter-frames artifacts appear. To what extent this is the case and how this affects performance is to be determined.

7. APPROACH

In order to carry out this research, we extracted the frames of each video, colorized them, merged them back and analyzid the videos. In order to do so we set up a virtual machine running Linux and ran the software there. The frame extraction and merging all was done on a windows machine.

7.1 Colorization

Zhang et al have made their research and the Python code used to implement their Neural Network open source. However, the TensorFlow framework used in their code only works on Linux. In order to colorize videos we have used FFMPEG to extract individual frames from Polygoonjournaals at 24 frames per second. These individual frames were then copied to a Virtual Machine running Ubuntu, which was capable of colorizing these frames. Colorizing an indivudal frame using an Intel Core i5 6500 running at 3300mHz took roughly 7 seconds althoug total running time increased because memory errors forced us to recolorize failed images often. After all the frames of a video were colorized, the frames were zipped and copied back to a windows machine. On the windows machine we merged the now colorized frames back into a video and we merged this video with the audio from the original video.

8. RESEARCH RESULTS

Two videos have been colorized and are uploaded to Youtube [5][6].

8.1 Wiesenthal

The first video we colorized is a meeting with Simon Wiesenthal. The color of his skin and tie appear to be properly colorized.

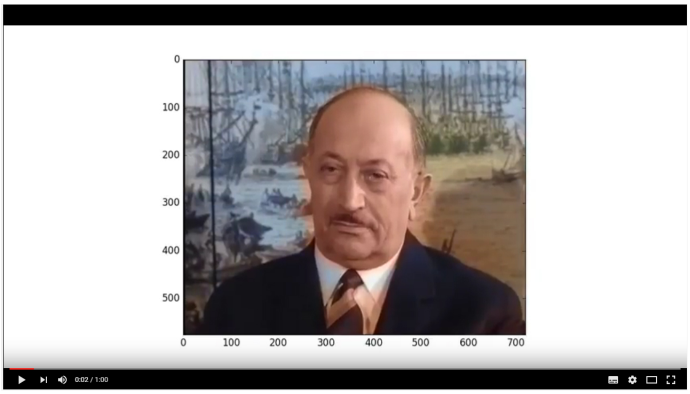

Figure 3: Colorized frame.

However, it is clear that some artifacts appear. Part of the background behing 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 blackgrounda blue tint, as it is supposed to be a lake, but the right side has a brown tint.

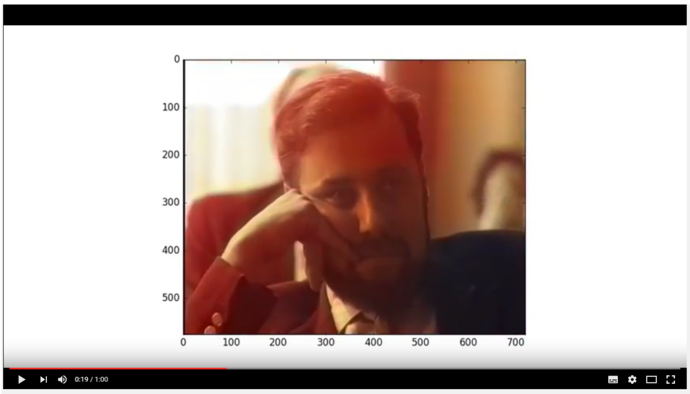

Figure 4: Colorized frame.

Not all frames are correctly colored though. As seen above, there is a reddish tint all over the image and the color of the persons' skin is hardly recognizable. Due to the nature of Neural Networks and our implementation, it is hard to exactl point to where the mistake was made and what can be done to prevent it.

8.2 Uilenstede

This video of Uilenstede is popular among its inhabitants, mostly students studying at Vrije Universiteit Amsterdam. Thusly we chose to colorize this video as it might spark interest among its inhabitants and drawing attention to NISV when the video is published. This video seems to suffer greatly from artifacts in between frames - rapid changes in color of objects when the camera pans or moves. Additionally, at areas of low contrast bright colors appear. This is a problem described by Zhang et al. The tiles in this bathroom are all white in real life, yet the neural network colorized many of them a bright cyan color.

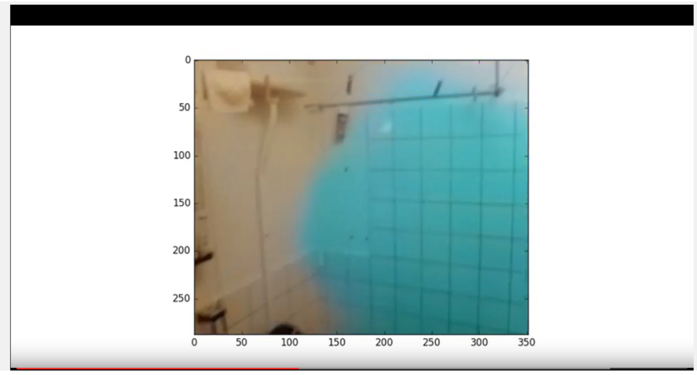

Figure 5: Colorized frame.

This frame is a still from a shot in a student's room. The colors seem to be accurate - the brownish tint of the wood is correctly predicted and the white of the walls has not been colorized. Although the correct color of different objects in the room cannot be correctly determined, the cololrs of the shot do not seem off.

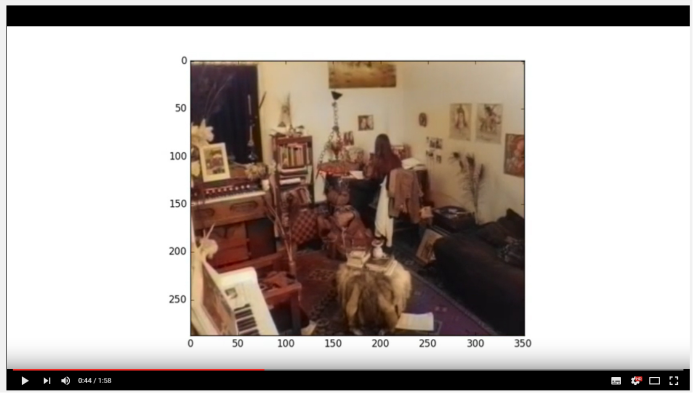

Figure 6: Colorized frame.

9. REFERENCES

[1] M French and F Recknagel. Modeling of algal blooms in freshwaters using artificial neural networks. *WIT Transactions on Ecology and the Environment*, 6, 1970.

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

[3] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. *arXiv preprint arXiv:1604.08610*, 2016.

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

[5] R Marsman. Colorized video 2 - Wiesenthal. https://www.youtube.com/watch?v=olsO2rOy$_i$4/, 2016.[*Online*; *a*

[6] R Marsman. Colorized video 1 - Uilenstede. https://www.youtube.com/watch?v=13SaYVnmnM8/, 2016. [Online; accessed July 2016].