Automatic Podcast Generation

*Saniya Zahoor[1], Shabir A.Sofi[2]
[1]Post Graduate Department of Computer Science, University of Kashmir, Hazratbal
[2]Department of Information Technology, National Institute of Technology Srinagar
[1]saniyazahoor.ku@gmail.com, [2]shabir@nitsri.ac.in

Abstract: A massive leap forward in the field of Human Computer Interaction in living memory has been achieved by the Google Duo system to sustain a natural sounding and coherent phone call with a human being without them being able to tell the difference. The computer system capitalized on recent developments in the field of Synthetic voice generation along with real time processing and response generation. The aim of this work is to replicate the success of that presentation as well as to build upon that body of work and generate useful content summaries which can be converted into high quality podcasts. In particular, our approach first comprises of extracting text data from web pages using various Natural Language Processing (NLP) tools as well as deep neural networks. After that it summarises text into byte sized chunks using extractive summarisation. Then, in the end it generates clear, high quality audio podcasts from the produced summaries using recently developed text to speech engines.

Keywords: Podcast generation, Artificial Intelligence, Machine Learning, Natural Language Processing.

1. Introduction

In the summer of 2019, Google demoed Google Duplex. It was an AI system that solved a long- standing problem of Human-Computer Interaction, enabling people to have a natural conversation with computers, as they would with each other [1]. Digital content creation has seen an explosion in recent years. Every year, digital content creation has increased year over year to the tune of 40%. As the profitability and interest in this market segment increases, a new generation of tools will be needed to help content creators of the digital age. In the midst of the COVID quarantine of 2020, digital content consumption has reached unprece- dented volumes. Digital media is being produced, distributed, and consumed at a breakneck pace because of a large captive audience [4]. The aim of this paper is to explore the recent developments in the field of NLP and develop a concrete understanding of its fundamentals. Using this knowledge, we aim to create a working Proof of Concept that summarizes text data and can create a purely synthetic corresponding audio file. This audio file can be consumed in the form of podcasts or radio broadcasts. Two passive mediums for which the bar of human attention is lower.

In order to generate high-quality podcasts, high-quality summaries need to be created as well. However, creating high-quality text summaries is not an easy challenge and requires a certain amount of contextual awareness to be instilled in the approach used. One of our goals is to answer the question of whether it is possible to instill contextual awareness in ML models. Our research shows that it is indeed possible to do so [12]. Once we obtain high-quality summaries of text data, we also need to produce realistic-sounding audio. A robotic voice would detract from the immersion and user experience which would be obtained by using automatic content generation tools. Such a decline can possibly lead to complete rejection by humans due to uncanny valley effects [13]. As part of this paper, we also aim to generate a high-quality synthetic audio in order to successfully cross the uncanny valley of human perception.

In this paper, we aim to develop a system that passes the Turing test for speech. Recent developments in the field of NLP have pushed this previously unattainable holy grail of computer science into the realm of feasibility for the undergraduate researcher. However, just the generation of high-quality speech is not enough, it must be used in a meaningful way in order to create content that can be consumed by the masses. In the scope of this text and paper code, we set out to solve this problem. This paper can be used as a tool for Automatic Content Generation. With improvements on the horizon in the field of TTS as well as document summarisation, realistic human audio is not a pipe dream anymore and our motivation is to solve this inherently complex problem.

The rest of the paper is organised as follows. In Section 2, we survey the relevant literature. After that, we discuss our approach along with the road map we followed in Section 3 which is followed by a presentation about our hardware and software requirements in Section 4. We discuss our code implementation in Section 5 and inform about the corresponding software libraries used. In Section 6 we discuss our results as well as visualise our generated audio.We present the paper contributions in Section 7 and discuss the applications in Section 8. We conclude and list the possible future work in Section 9.

2. Literature Survey

In this section, we have surveyed the literature related to this paper. Each of the relevant papers will be discussed in detail. Transfer Learning from Speaker Verification to Multi speaker Text-To-Speech Synthesis [2]: In this paper, the authors describe the architecture needed for authentic sounding voice cloning. The authors find a 1:1 mapping for a person's identity from the input voice. The unique mapping is then used as an input along with a sentence for which text has to be synthesized to an encoder model from which a sequence of embedding tokens is obtained. The embedding tokens have self-attention applied to them and are fed to a decoder model in order to create Mel Spectrograms. The MEL Spectrograms are further discussed in section 2.1.4 of this text. Using the Spectrograms, a WAV audio file is obtained which resembles the input voice but is speaking a completely synthetic sentence. The authors made use of developments in TTS en- gines such and based their model architecture around Tacotron 2. Using the generated waveforms they show that their speech model has learned a high-quality representation of individual speaker representation.

A Benchmark for Systematic Generalization in Grounded Language Understanding [5]: In this paper, the researchers at FAIR (Facebook AI Research) labs make a peculiar observation about NLP text synthesis. They demonstrate that advanced transformer architectures can show an under- standing of sentences encountered within the training data sets. These models also show contextual awareness within the observed data set. However, when generalising to unseen data, the models completely lose contextual awareness. To counter bias towards unseen data, the authors propose a novel benchmark that tests text generative models on contextual awareness. They believe that the proposed benchmark is a more rigorous evaluation criterion for testing the contextual awareness of the model being evaluated. The benchmark uses a map and grid system which talks to an AI agent and asks it a set of questions before allowing the agent to move. This test is a better evaluator of the contextual understanding of the NLP models.

Natural TTS Synthesis by conditioning Wavenet on MEL Spectrogram Predictions [3]: This paper marks a ground breaking achievement in the field of Human-Computer Interaction. The re- searchers at Google's AI labs have proposed an architecture that first converts text into a sequence of embeddings. The embeddings are then sent to a CONV+LSTM encoder model architecture to obtain intermediate embeddings for a sentence. Attention is then applied to the sentence along with another CONV+LSTM decoder model. The model outputs are fed to a wavenet described in the next section in order to produce high- quality Mel Spectrograms. The audio obtained from this architecture is nearly indistinguishable from professional audio recordings and has clear intonation and verbalisation. The presentation contains a demonstration of the capabilities of this model.

To evaluate the model outputs the author used MOS scores where the quality of the audio is rated on a scale of 1-5. The speech is evaluated by multiple independent speakers. The "Tacotron" model achieved a score of 4.53 which was a new State of The Art at the time of presentation. Using the knowledge gleaned from these papers, we developed a broad understanding of the recent developments in the field of NLP. Google's Tacotron is capable of generating near indistinguishable voice output [3], researchers have also made massive strides in generating sentence embeddings, question answering, and content summarisation. We have summarized the papers in Table 1.

Table 1. Literature Review

Name	Description	Year	Group
Transfer Learning from Speaker Verification to Multi Speaker Text-To-Speech Synthesis [2]	Uses voice embeddings plus a limited Tacotron network to produce Mel Spectrograms that clone a person's voice	2019	Google AI
A Benchmark for Systematic Generalization in Grounded Language Understanding [5]	A benchmark for evaluating the ability of NLP predictions to learn from small sentences and apply the results to large unseen sentence	2019	FAIR
Natural TTS Synthesis by Conditioning WAVENET on Mel Spectogram Predictions [3]	Uses a deep CNN plus LSTM network to produce Mel Spectrograms. Uses Wavenets to convert Mel Spectrograms to natural sounding audio	2018	Google AI
WAVENET: A Generative Model for Raw Audio [6]	Deep neural network that is used to produce Mel Spectrogram Structures and Indexes	2016	Google AI
Boosting Question Answering byDeep Entity Recognition [7]	Unstructured Data. Parses questions and uses extracted data for ranking all available documents. Uses Ranking results to answer questions	2016	Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
Generalized end-to-end loss for speaker verification [8]	This loss is used for extracting embeddings from voices to determine speaker identity. used for Spoof Voice, Audio generation	2019	Google AI, USA
Language Models are Unsupervised, Multitask Learners [9]	Structure of the GPT-2. This is now considered seminal work.	2019	OpenAI
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [10]	Structure of the Bidirectional Encoder, Representations from Transformers model. One of the most famous and benchmark beating NLP models	2019	Google AI, Language Group

3. Proposed Approach

The novel work in this paper is the process of summarising combined with a text to speech model to produce end to end audio data. In this section we discuss our end to end model and provide insights about the approach we took to build this paper.

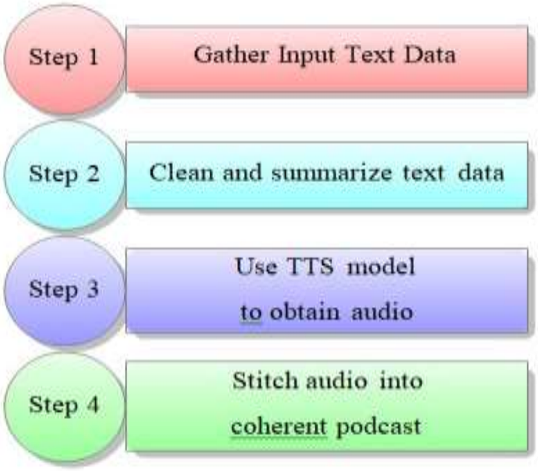
Step 1
Gather Input Text Data
Step 2
Clean and summarize text data
Step 3
Use TTS model to obtain audio
Step 4
Stitch audio into coherent podcast

Figure 1. Condensed Approach

3.1. End to End Model

The End to End model approach will involve 4 steps:

- Collection of Structured Input Text from Unstructured Data
- Summarization of Structured Text
- Sentence by Sentence voice generation
- Combination of generated sentences into a curated podcast

3.2. Obtaining Structured Text Input

The purpose of this section is to elucidate how we obtained structured data from web page dumps as inputs to our speech to text models. First, a raw dump of a large number of web pages is obtained and stored in a database using web crawling tools such as Scrapy. Once the web page links are gathered, they can be parsed individually in order to gather meaningful data from them. We explored Extractive, Abstractive and Generative Text Corpus Summarisation techniques. Of these 3 methods, we found that extractive text summarisation worked best for the purposes of Summary generation. Extractive summarisation techniques aim to find the sentence which has the same contextual infor- mation as a large number of sentences in the given text corpus. Multiple sentences that span the corpus and contain representative information are found and presented as a summary. Abstractive summarisation tries to generate sentences on the basis of the information present in the text corpus. These generative sentences are designed to be similar in style and representative of the substance of the given text data. Generative summarisation aims to fit data to a template which is provided and create summaries. However, these kinds of architectures do not generalise well to varied and unseen data. Once the text summaries have been obtained, the text needs to be converted into audio. For this, we have tested 2 different model architectures, Tacotron 2 and Voice Cloner, and obtained audio results.

We used Google's Tacotron model architecture along with pre-trained weights open-sourced by NVIDIA for synthesizing audio. The voice is natural and approaches the quality of professional audio recordings [3]. The model has a low latency and is capable of creating audio in near real- time.

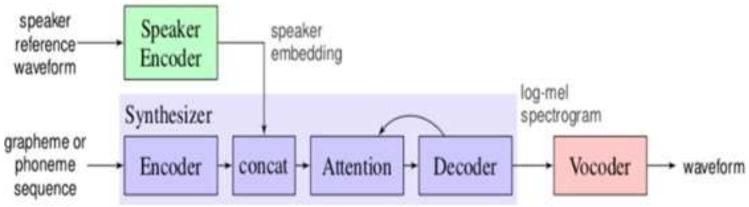

Figure 2. Voice Cloner Architecture

The Voice cloner architecture as described in [2] was also implemented by me to obtain cloned voices. The voices obtained were strikingly similar to the provided audio sample. However, because of inadequate training, the model is unable to pronounce all syllables and emits robotic sounds for certain words. I generated audio samples for the voice of Morgan Freeman. Once the individual sentences are generated, they can be stitched together using audio libraries such as librosa which use ffmpeg as a backend.

4. Results

In this section we discuss our code implementation results. We talk about our results obtained for text extraction using Spacy, Regex, Beautiful Soup + Regex and Dragnet. We present the extractive summaries generated by our BERT Extractive summarizer and we compare audio files generated by our two different TTS engines for the same text.

4.1 Web page Text Extraction

In this section we discuss the results we obtained while attempting to extract high quality text from webpages. The results obtained using Spacy contained a mixture of image captions and other extraneous text. A large amount of header text was included in the final processed text. Along with the header text, a large amount of titles, captions, and citations were also included. Spacy delivered all the text which would be visible to human eyes. However, all the visible text may not be relevant to creating coherent summaries. Because of these drawbacks, this approach was abandoned. The regex approach resulted in improvements over the results of spacy, however, a large amount of relevant content was also dropped as it usually contained brackets and punctuation which was caught by the regex rules. The improvements led to an improvement in the quality of the generated summaries. The rules-based approach was highly obtuse in an effort to prevent unwanted content from being present in the final processed text. However, a lighter set of rules made the processed text contain a garbled mixture of captions, headings, and citation texts.

A Combination of Beautiful soup for initial document extraction and regex rules was tried and the results were promising. However, a large part of the text containing contextual information was removed by the model. An entire web page was being distilled down to parts that were not representative of the whole. Because of this, we stopped using this approach. The obtained text was of high quality, which made the podcast generation process more streamlined and accurate. It did not contain extraneous header text, minimal formatting errors were present which could be caught with regex rules. The final text is clean legible and can be used for creating excellent summaries.

4.2 Voice Generation

In this section compare the generated audio between Tacotron 2 and the Voice Cloner architecture. When we compare the generate audio from Tacotron 2 and Voice Cloner, the generated audio for Tacotron is a lot more stable then the generated audio for Voice Cloner. When we compare the two audios, we can observe the amplitude spike which is caused by the imperfect RNN decoding. This spike sounds like a sharp whistle which detracts from immersion. When we compare the two audios, Voice Cloner audio is gibberish in between but Tacotron 2 is coherent throughout.

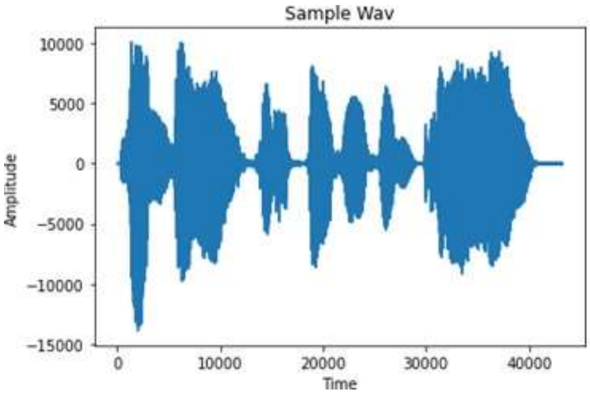

Figure 3. Audio wave form from Tacotron 2

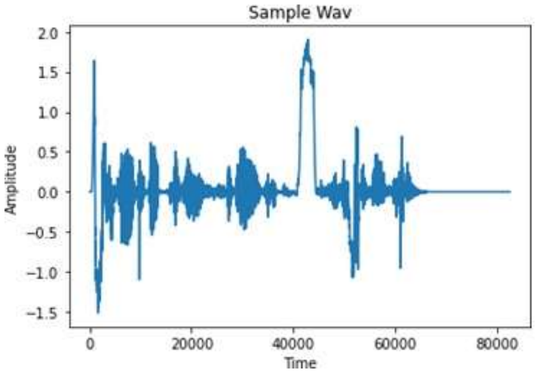

Figure 4. Audio wave form from Voice Cloner

5. Paper Contributions

This paper enables the creation of natural-sounding audio summaries from web articles. To the author's knowledge, no such tool exists. Due to the sequential approach developed by the authors, automatic podcast generation can be embarrassingly parallelised in order to obtain results faster and at a rate that surpasses human capability. This tool can be modified to any kind of text data. With the model trained to output the correct kind of embeddings, the paper can be used for:

- Summarizing agreement documents to produce a list of understandable
- Summarising sports Results
- Summarising political developments into byte-sized chunks
- Summarising financial data
- Summarising company balance sheets
- Summarising movie scripts and books for movie agents

6. Conclusion and Future Work

We explored the domain of structured data collection from unstructured text. We used rule based as well as deep learning approaches and found that deep learning approaches achieved better results. We surveyed various content summarisation techniques and found that extractive summarisation worked best and provided contextual information which was relevant to the end users. We also generated high quality, natural sounding audio using state of the art Tacotron 2 model while exploring the Voice Cloner architecture. For human ears, the Tacotron architecture proved superior to the Voice Cloner architecture. This paper enables the creation of natural-sounding audio summaries from web articles. However, this tool can be applied to any kind of text data. With the model trained to output the correct kind of embeddings, my paper can be used for summarizing legal documents, movie scripts, textbooks, and any other dense prose.

This paper can be further modified to obtain summaries of videos, songs, and plays using the text present in the audio stream. Using techniques similar to the one described in this text, video summarisation, and music summarisation can also be created and automated. For videos, we can trans-code dialogue to text and summarise the text. Clipping the relevant videos, we can create video summaries. For music, we can convert text into different tonal sections. Extracting appropriate embedding from the sections, we can split the song into its core melodies. This song summarisation can be used for creating remixes and music compilations.

REFERENCES

[1] Google Duplex: An AI System For Accomplishing Real-World Tasks Over The Phone. Avail- able at: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html [Accessed 28 May 2020].

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predic- tions," arXiv preprint arXiv:1712.05884, 2017.

[3] M. Peters, and D. Lecocq. Content Extraction Using Diverse Feature Sets. In Proceedings of WWW '13, pages 89-90, 2013.

[4] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. arXiv preprint arXiv:1912.08777, 2019.

[5] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I.

[6] Watson, Amy. "Coronavirus Impact: Global in-Home Media Consumption by Country 2020." Statista, 30 Apr. 2020

[7] L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to- speech synthe- sis," arXiv:1806.04558, 2018.

[8] Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt and Brenden M. Lake. A Benchmark for Systematic Generalization in Grounded Language Understand- ing.arXiv:2003.05161 2019.

[9] Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv:1609.03499, 2016.

[10] Przybyła, Piotr. Boosting Question Answering by Deep Entity Recognition. 2016.

[11] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In

Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018.

[12] AlecRadford,JeffreyWu,RewonChild,DavidLuan,Dario Amodei, and Ilya Sutskever. Lan- guage models are unsupervised multitask learners. OpenAI Blog, 2019.