

Automatic Speech Writer

May 2020

Rahul Malavalli
rahul.malavalli@gmail.com

Scott Bauersfeld
sbauersfeld@ucla.edu

Maggie Xiao
zxiao2015@gmail.com

Abstract

We present a pipeline combining state-of-the-art natural language processing transformers to produce speech transcripts. To generate the speech, only a short summary of the intended speech goals and the target tone are required. We web-scraped publicly available speech transcripts and generated abstractive summaries for each speech using BART, a denoising autoencoder for pretraining sequence-to-sequence model. We used IBM Watson Tone analyzer to generate the tone tags for the speeches. A GPT-2 model is then fine-tuned on the dataset to generate speeches when given summary and tone as the input. Compared to the baseline approach using pre-trained GPT-2 with a perplexity score of 22, the fine-tuned model reaches a perplexity score of 18. The generated speeches from the proposed pipeline are fluent, coherent, follow the suggested tones, and successfully capture the key information in the given summary.

1 Introduction

In this paper, we aim to emulate a human speech writer who creates a speech based on their client’s requested objectives and the emotions they wish to convey. Speakers and leaders in all fields, ranging from government to private industry, often utilize these types of expert writing services in order to articulate their thoughts in the most effective manner. By automating such a process, we hope to equip those lacking access to traditional speech writing resources with a tool that allows them to more powerfully convey their thoughts.

Here, we build a model that can produce the first draft of a speech transcript, given a short summary of the intended speech goals and desired tone. To do so, we take inspiration from the Grover model [1] for neural fake news generation by fine-tuning a GPT-2 model on a series of distinct fields as required for the speech transcript writing task. Since

we do not have access to a pre-existing dataset of speech transcripts with corresponding summaries and tone labels, we generate and augment our own training data set. Specifically, we utilize publicly available speech transcripts and apply a pre-trained Bart summarization model to generate the summaries. The IBM Watson Tone Analyzer is applied to extract the emotional tone from the speech transcript. The speech data is then organized into a sequence of summary, tone, and transcript fields, and GPT2 is fine-tuned on this dataset.

Figure 1 shows the schematic of the pipeline starting from data acquisition, summary generation, tone analysis, to GPT-2 medium model fine-tuning and speech generation.

2 Data acquisition

To create a model that outputs a transcript from a desired summary and other metadata, a training dataset is needed that contains reliable pairs of inputs (metadata) and outputs (transcripts). For this project, we focused on producing potential first drafts of speech transcripts as our output. The input metadata contained the speech’s summary and tone analysis, both of which are covered in section 3.

The training data was obtained through an online speech bank from [American Rhetoric](#) [2], containing various “public speeches, sermons, legal proceedings, lectures, debates, interviews, other recorded media events, and a declaration or two”. No easily packaged version of the online speech bank was available, so speech data was scraped via the [webscraper.io](#) tool. The obtained dataset contains information on speech title, speaker name, speech delivery year, and speech transcript. The original speech data contains 1,202 speeches from the 1600s to 2020.

Note that the speeches extracted from American-Rhetoric do not contain a ground truth of human-

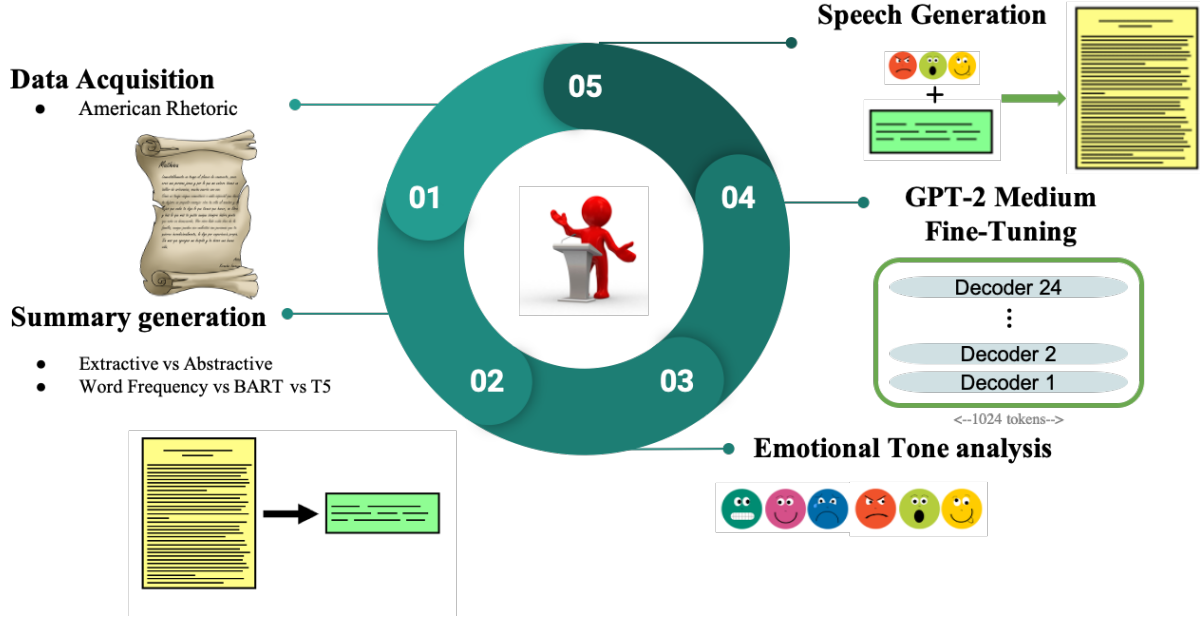


Figure 1: Overview of the automatic speech writer pipeline with five steps.

created summaries, sentiment, or other synoptic features. Therefore, we performed data augmentation via summarization and tone analysis, as described in section 3 below.

3 Data augmentation

To create a complete speech dataset, we augment the parsed speech transcripts by incorporating summarization and emotional tone analysis. Both features were generated programmatically for efficiency.

3.1 Summaries

To summarize the speeches acquired in the data acquisition step, we implemented and compared three different methods that are either extractive or abstractive. Extractive summaries are formed by reproducing exact words or sentences from the original text thought to represent the text’s meaning; abstractive summaries, on the other hand, are created by generating new words and sentences intended to capture the original text’s meaning.

The extractive method used is based on the weighted word frequency, obtained by dividing the frequency of word occurrence for all the words by the frequency of the most common word in each speech. Next, “sentence scores” are calculated as the sum of the weighted frequencies of the words in each respective sentence. The ten sentences

with the highest sentence scores are taken as the extracted summary for the speech.

In terms of the abstractive summary, we experimented with two recent transformers, T5 [3] developed by Google, and BART [4] proposed by Facebook AI. T5 stands for “Text-to-Text Transfer Transformer” model. It uses the same model, loss function, and hyperparameters on a diverse range of tasks. The main highlight of this model is that it can reframe all NLP tasks into a unified text-to-text-format. It claims to have the best of existing transfer learning techniques, and can be applied flexibly to a diverse set of NLP tasks including summarization, as shown in Figure 1(a). BART is a denoising autoencoder that essentially generalizes BERT and GPT training schemes, combining bidirectional encoder and left-to-right decoder, along with other recent pretraining schemes (Figure 1(b)) [4]. It is fine tuned particularly for text summarization. We applied on a single GPU and compared the outputs of 3 T5 models and 2 Bart models of different sizes (T5 small (60 million parameters), T5 base (220 million parameters), T5 large, bart-large-xsum, and bart-large trained on CNN/DailyMail Dataset.) The pre-trained models of different sizes are obtained from [Hugging Face Transformers](#) library. All of the generated summaries are truncated to have 100 words max and have the incomplete sentences removed as a result of text generation and max word limit. The average summary lengths are 50 words.

Method	Generated Summaries
Bart-xsum	Former US President Ronald Reagan and his wife, Nancy, arrived in Canada for a state visit in 1987, and it was a visit that will be remembered for many years to come, writes Prime Minister Brian Mulroney, who was then Prime Minister of Canada and now US President.
T5-large	bob greene: Ronald Reagan was a president who inspired his nation and transformed the world. greene says he embodied the unusual alchemy of history, tradition, achievement. he says Reagan’s vision of a united u.s. was based on a sense of the nation’s majesty.
Word Freq	And here we were waiting for our wives. At home and on the world stage, his were not the pallid etchings of a timorous politician. They did. And we know now who was right. I have been truly blessed to have been a friend of Ronald Reagan.

Table 1: Generated Summary on the same speech by three methods.

Comparing the generated summaries from the three methods, we observe that word frequency method extracting sentences with highest frequency may miss information from the transcript, hence abstractive summary serves the purpose better. T5 summaries using T5-base or T5-large occasionally append random speaker names to the beginning of the summary. This phenomena is less prevalent in the T5-small generated summary for this dataset, so we attribute the strange behavior to the model development. On the contrary, Bart-large-xsum and Bart-large-cnn that are pretrained on XSum and CNN/DailyMail dataset both consistently yield coherent summaries. According to Lewis et al. [4], XSum is highly abstractive and CNN/DailyMail resembles the source sentences more. Based on the qualitative analysis of all generated summaries, we decide to use BART generated summary using bart-large-xsum model for downstream tasks.

Table 1 shows an example of the sample summaries generated for the same speech. BART generated summary outperforms the other two.

3.2 Tone analysis

Although the summarization techniques explored in section 3.1 seem to adequately encapsulate the content of a speech, they tend to omit its emotional style. This is evident through a qualitative review of the chosen bart-xsum model’s output in table 1, which seems more akin to statements of fact. To allow the user of the model to influence the generated speech’s style, we decided to augment the dataset further by explicitly encoding emotional and style.

Initially, we experimented with sentiment anal-

ysis through BERT models. However, we realized that most sentiment analysis techniques only provided a polarity (positive, negative, or neutral), without much more useful context.

So, we decided to explore tone analysis techniques, which assigns a list of ”tones” to each transcript in an attempt to describe the emotional characteristics of the speech. To do so, we utilized IBM Watson’s Tone Analyzer, which selects from 7 possible tones listed in table 2. To perform tone analysis, IBM trained a support vector machine (SVM) using, at least, the following features [5]:

- n-grams, i.e. sets of n contiguous words, used extensively in various NLP models.
- Lexical features (from dictionaries), such as words or groups of words with known meanings and sentiments.
- Existence of second-person references, such as the phrase ”You did...”, that may indicate more active speech or more pronounced tones.
- Dialogue-specific features that carry known connotations, such as ”thank you” or ”sorry”.
- Other distinguishing features, such as consecutive exclamation points or question marks (ex: ”!!!” or ”????”), that convey meaning unconventionally.

In addition to a list of tones for each transcript, the IBM Watson Tone Analyzer also returns a score for each tone, indicating how prevalent the analyzer thinks the tone is in the transcript. To utilize this information, we sort the list of tones per transcript

Tone	Frequency
Anger	1%
Fear	4%
Joy	35%
Sadness	23%
Analytical	26%
Confident	5%
Tentative	5%

Table 2: The frequency of each tone across all speeches in the dataset.

in decreasing order of their respective scores; therefore, the first tone in the list is considered the most prevalent, the second tone is considered the second most prevalent, and so on. Adopting this structure for the tones feature allows the user of the model to input multiple desired tones with different levels of importance.

4 Model building and training

In order to train a model to generate a speech from a given title and summary, we organize our speech documents into three sections per speech:

1. Speech summary
2. Speech tone
3. Speech body (transcript)

The summary and tone sections are respectively identified by the special tokens $\langle |summary| \rangle$ and $\langle |tone| \rangle$. The speech body directly follows the tone section after a $\langle |cls| \rangle$ token. We then train a language model capable of generating the body of the speech given the summary and tone sections. An example data sample is shown in Figure 2.

	$\langle summary \rangle$	At the Commonwealth Club of New York on 7 May, 1776, Thomas Jefferson addressed a group of leaders of the nation who were concerned with the development of government through non-partisan means.
	$\langle tone \rangle$	analytical, sadness
	$\langle cls \rangle$	
• Summary		My friends:
• Tone		I count it a privilege to be invited to address the Commonwealth Club. It has stood in the life of this city and state, and it is perhaps accurate to add, the nation, as a group of citizen leaders interested in fundamental problems of government...
• Speech		

Figure 2: Each data sample is divided into summary, tone, and speech transcript fields.

Given our limited computational resources, we utilize transfer learning with the medium-sized GPT-2 architecture [6] to efficiently fine-tune the 345 million parameter language model in a reasonable amount of time. The medium-sized GPT-2

transformer is an auto-regressive language model with 24 decoder layers, each of which consists of a masked self-attention and feed forward layer. The self-attention layers enforce causal language modeling by masking all future tokens so that each decoded token depends only on previously seen tokens. Mathematically, GPT-2 models the probability $p(x) = \prod_{i=1}^n p(s_i | s_1, \dots, s_{i-1})$ where x is a sequence of tokens and s_i is a token in x . As a result, the GPT-2 transformer performs very well for text generation tasks because there is no need to incorporate information from future tokens.

We use a GPT-2 model provided by the [Hugging Face Transformers](#) library that has been pre-trained on 40GB of text. The library also provides a GPT-2 tokenizer that converts raw text into identifiers that can be understood by the model. We add the special tokens $\langle |summary| \rangle$, $\langle |tone| \rangle$, and $\langle |cls| \rangle$ to the tokenizer so that the model can learn to distinguish between each section of the speech. We can then fine-tune our model by applying the tokenizer to each speech and forcing the model to predict sequential tokens. After training on about 1200 speech documents, our GPT-2 model successfully learns the structure of each document as defined above and can generate reasonable speech transcripts when given the speech title and summary. Figure 3 presents an overview of the fine-tuning pipeline.

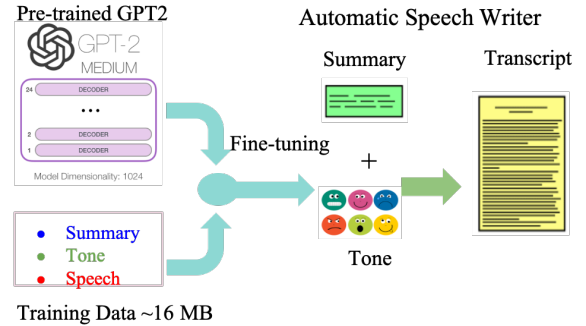


Figure 3: GPT-2 fine-tuning pipeline.

4.1 Results and discussion

Due to the nature of the generation task, most model evaluation was done qualitatively. However, perplexity scores were also used as a basic form of quantitative comparison. Perplexity measures how well probabilistic models predict test samples by calculating the exponentiation of the cross entropy loss, $P(m) = 2^{-\sum_x p(x) \log_2 p(x)}$ where x is a sample in the test data set. Low perplexity scores

indicate that the model is able to accurately predict the test data. The pre-trained GPT-2 model achieved a perplexity score of 22 while our fine-tuned GPT-2 model reduced the perplexity score to 18.

To demonstrate the final model’s performance, examples of generated speeches are provided in tables 3 and 4. In the speeches generated in table 3 on a summary-tone pair from the testing dataset, which was acquired from real speeches, the model effectively expands on the content in the summary. At times, the model also spins historical information into the narrative while preserving the prevailing tone.

Table 4, on the other hand, displays an artificially constructed summary-tone pair that was not found in the training and testing sets. The summary was created to pose multiple seemingly unrelated objectives to the model in a single speech; namely, a combination of the usually disparate topics of public health and economic stability. While maintaining a conversational style, the model was able to cover both objectives with appropriate transitions. In one of the speeches, the model was even able to offer basic diet and exercise recommendations in response to the health-related prompt.

5 Conclusions

Especially with the advent of larger and larger models like GPT-2, and most recently the 175 billion parameter GPT-3 model [7], computers have become better and better at understanding language and performing a plethora of NLP tasks. Alongside seemingly harmless applications like machine translation, however, these models have also become particularly proficient at generating realistic text. We believe that our automatic speech writer, albeit a basic implementation, has the ability to be a useful tool for those wishing to improve their speeches and the efficiency with which they write them. As exciting and awe-inspiring as language modeling can be, however, we should keep in mind that these models suffer from many potentially crippling drawbacks; at the very least, they do not necessarily have a grasp of truth, may inherit the biases of their training datasets, and could be used by malicious actors to simulate real humans. Even in some of the example speech generations we include in this paper, for example, the model attempts to offer basic health-related advice. Although this guidance may be general and either correct or harmless, ac-

tually heeding or delivering those suggestions without consulting the qualified resources very much has the potential to be dangerous.

So, we encourage continued research in language modeling because of the immense possibilities for technological advancement. But, we also need to be conscious of the implications of progress in this field and careful about how we handle them.

References

- [1] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9054–9065. Curran Associates, Inc., 2019. URL <http://papers.nips.cc/paper/9106-defending-against-neural-fake-news.pdf>.
- [2] Michael E. Eidenmuller Ph.D. American rhetoric: The power of oratory in the united states. URL <https://www.americanrhetoric.com/>.
- [3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019. URL <https://arxiv.org/abs/1910.10683>.
- [4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. URL <https://arxiv.org/abs/1910.13461>.
- [5] IBM Watson. The science behind the service. URL <https://cloud.ibm.com/docs/tone-analyzer?topic=tone-analyzer-ssbts>.
- [6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. URL <https://openai.com/blog/better-language-models/>.
- [7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020. URL <https://arxiv.org/abs/2005.14165>.

Table 3: Generated Speeches using summary from test data

Speech Generated based on Summary from Test Set	
Summary	The United States Department of Education was created to protect the rights of all students, not just those who are victims of discrimination or who are accused of discrimination.
Tone	analytical, joy, sadness
Generated Speech Option 1 (Abbrev.)	It is an honor for me, as the Chairman of the Commission for Civil Rights at the Department of Education, to be here today. In the past, this Commission has held hearings and hearings, sometimes in secret, in order to bring to light issues of discrimination. But today we feel compelled to present ourselves before the world for the first time. We do so only to discuss the latest issues facing the education of our children. The past four years have seen some of the most dramatic incidents of racial discrimination in our country's history. These include the racially motivated murders of two young white women in Oklahoma City, the shooting of four young black men by police in New York City, and the racist attack on the federal building in Birmingham, Alabama, and the beating of the civil rights lawyer Kenneth Starr. These horrific crimes, along with the vicious harassment, the death threats, and the assaults on other civil rights leaders, prompted my predecessor, Attorney General Tom Corney, to resign. This Administration has taken steps to root out these racist practices. The Attorney General has launched investigations. And I have directed the federal civil rights office to launch an investigation to find out how these incidents were allowed to fester for so long. We want to be clear: We do not intend to condone the actions of our law enforcement officers, but we also do not seek to punish – we do not seek to punish any individual – but we want to find out where – where these problems and abuses came from, where they've been going on and to put our policies in place to stop them.
Generated Speech Option 2 (Abbrev.)	Mr. Bossert, Vice Chancellor of the University of Missouri, Chancellor Thompson, Mayor Thomas Brown, Dr. Mabry, Secretary Stauffer, Dr. King, many, many others: I am grateful to you for taking the time to talk with us today. I want to address an issue that has troubled many of you: What happens when a student is stopped or even detained for what the police believe to be minor violations of federal law or campus procedures? This is an issue on which we can make a difference. As I have indicated to you, this has long been a Department issue – a problem of police accountability. It is a problem which requires action by law enforcement. It also needs to be dealt with by education. In the past 10 years, we have made significant progress, and we are seeing a rapid improvement in the training of police across the country. For example, we have made tremendous strides in developing training programs with a focus on non-racial profiling by police, a practice which has been a major source of distrust, mistrust, and bias in the past. As President Obama recently told a crowd in Houston, "We have to be honest. We have to get real with our communities if we want to reform policing." This training is critical. And it can only happen in partnership with law enforcement. As you know, at the University of Missouri and throughout the country, the Department of Education maintains the department's Civil Rights Office. Our Civil Rights Office is dedicated to enforcing the equal rights laws with respect to all people regardless of race, color, and national origin. We are committed to advancing an open dialogue with our communities to address the needs of all Americans.

Table 4: Generated Speeches using an arbitrary summary and suggested tone

Speech Generated based on user-provided summary	
Summary	Hospitals will need more personal protective equipment. The stock market is volatile. Everyone should adhere to social distancing and general hygiene to stay healthy.
Tone	fear, confident
Generated Speech Option 1 (Abbrev.)	<p>The last month, the stock market of our country has gone through a complete change, from a stable base to a bear market. From October the value of our stock market reached 20 billion dollars. It went into negative territory. We can't be caught without money. We should have more funds to spend. We're in a slump right now. The only thing we need is more capital. This is not just a business issue; it's also a social issue. We all feel the impact of these stock market crashes. I had one colleague who was in the hospital for 3 weeks. He was in a lot of pain. And then he was discharged with pain. Everyone else here is suffering the same thing. But we need to do better. We need to do a better job. We must change the way we think. This means change in the way we think about our family. How to raise the children and look after their health? How to help other families? And how to provide for ourselves? We want to live up to our heritage as a nation. . . . Thank you very much.</p>
Generated Speech Option 2 (Abbrev.)	<p>I've been thinking and thinking and thinking and thinking about this whole "you'll be fine" business for about 18 months. And I've been thinking that I was pretty darn fine the first few weeks and months here. And then I lost 30 pounds and my life changed. And that's when I took the plunge. That's when I decided I needed to go to the doctor. And I'm on that. I'm not going to lie to you and tell you that you're gonna be fine. And the fact of the matter is that as soon as you start to feel sick, it's not good for anybody. It causes your body to overheat. It causes you to contract infections. It causes you to become more susceptible to the infections you contract. The body loses control over its own body temperature, and in turn, you start losing body hair and becoming increasingly susceptible to infections, and as a result, people lose all the weight that they otherwise would have lost and the body starts losing a large amount of blood. But there are ways to control your temperature, and that's why you need to exercise. And those include:</p> <ul style="list-style-type: none"> • Regular exercise, both in terms of weight training and weight management, and exercise in general. • Eating nutritious foods in a diet free of refined sugar, artificial sweeteners, and fatty foods. • Exercise in short bursts; and being in an indoor setting at night. • A good hydration system. • A healthy diet. • A good sleep schedule. <p>That's why I'm so sorry to all the doctors and nurses out there, all the medical staff and all the patients out there, but they're – your lives will change in very short order if they are not careful. And they don't really know what kind of changes they're going to be making. And the consequences will be pretty devastating, and that's not their fault.</p>