Biggest Bird LLC. – Configurable Speech Censorship

Cole Bianchi, Dante Dodds, Kevin Dong, Nathan Litzinger, Andre Mitrik, Efe Sahin

Penn State Department of Electrical Engineering and Computer Science

CMPSC 442: Artificial Intelligence

Dr. Christopher L. Dancy

April 28, 2023

Introduction

AI & Policing in Society

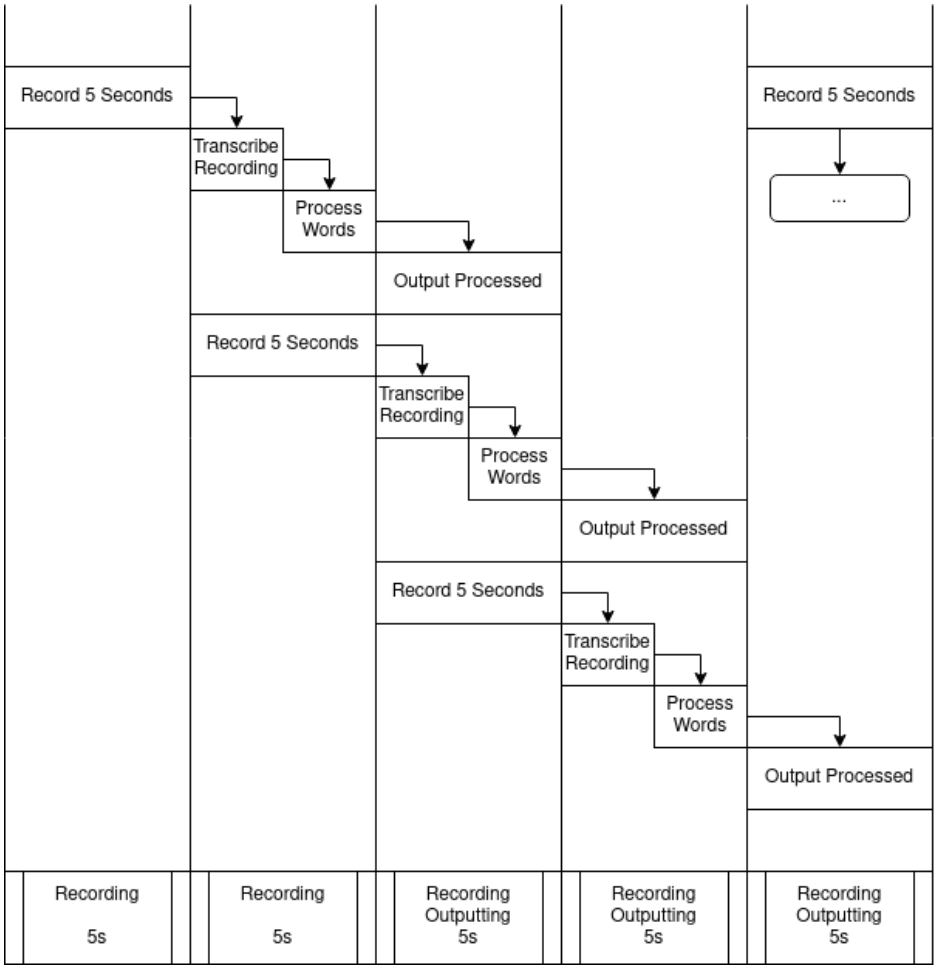
In a world dominated by interconnected technology outlets, the rapid growth of artificial intelligence (AI) systems and models in practical society continues to facilitate its evolution into a double-edged sword; these powerful systems have great capabilities requiring an equal amount of moral and ethical responsibility. In this work, we discuss the theme of AI and Policing in the context of filtering and masking data, particularly in the form of incoming audio streams (primarily human-speech based) that are processed and altered/censored real-time.

The scenarios for requirements of this type of system's success can range anywhere involving some kind of media outlet; television broadcasts, podcasts, live streams, etc. The usability question then also extends into both ethical and semantic dilemmas: *how* can we process/filter this data real-time, and *what* rationale do we have to support our censoring/filtering of the raw input data? The butterfly effect is obvious for the industries consumers see most; for instance, advertisers and commercial companies don't like swearing in programs, companies need advertisers to survive, and the question becomes simple: can we save even more money by automating a censorship process, or by hiring somebody to do it manually? From a business perspective, the choice seems obvious if the results are feasible.

Speech recognition (SR) systems also remain a widely discussed research topic – some outlets continually support ST's potential uses in fields including healthcare, education, and entertainment. We already see widespread transcription for most streaming/video sources, but the push to make SR more practical – including potentially automating data service pipelines, arbitrary diagnoses, and decreasing maintenance for existing services through training – is the innovative question of the future. This work delves into one possible solution for censorship, by using deep learning models to transcribe audio segments, and thereafter flagging audio segments that violate a certain list of defined words and reactively eliminating/censoring these segments as a proof-of-concept to be used in a larger framework with a more distinct purpose.

Implementation

At a high level, the system is implemented by combining three key modules: sound input, transcription, filtering, and sound output. These three key areas are run in parallel using Python's threading library. Shared queues are set up between the three parallel threads to transfer data in a pipe-lined fashion. By running these tasks in parallel real time performance can be achieved. The threaded process visualized in the figure below shows real time performance.



The first module, sound input, is accomplished using Python's sounddevice library. This library attaches to the computer's default microphone and generates audio segments of a predefined length in a predefined format. For this implementation we define this format to be raw PCM data recorded at 16,000hz in mono channel. This data is recorded in 4-second-long segments. At the end of each 4 second long recording the

raw PCM data as well as a unique sound frame identification number is added to the shared memory queue in the form of a tuple.

The second module is the transcription model. The chosen model was OpenAI's Whisper Tiny-en for its high level of accuracy in a range of audio scenarios. This model accepts raw PCM data in the form of NumPy arrays padded to be 30 seconds in length. This model is used within the software by first looping on 100ms intervals checking the shared queue for new data. When new data is found in the queue the raw PCM is converted to a float16 NumPy array and padded to 30 seconds to meet the model's requirements. The data is then fed into the model and a transcription of words and time stamps within the 30 second segment for the words are saved. This output is then sent to the next module using another shared queue along with the input audio and identification number associated with it.

The third and final module and step in our pipeline is the filtering and output module. This module, like the previous, loops on 100ms intervals scanning the shared queue for new data. Upon finding new data in the queue, the audio transcription is scanned for instances of words that appear in the provided banned-word list. If a banned word is found, the start and end timestamps of that word (provided by the whisper model) are recorded in a list of times to be censored out. Once the entire transcription has been scanned and banned-word instances identified, the module moves onto modifying the original audio stream to “bleep out” the banned-words. Because the audio is maintained as a float16 NumPy array, bleeping out the audio simply involves modifying those array entries in the audio stream array that correspond with banned word instances. This is accomplished by mapping the timestamps recorded for each banned word instance to an audio sample index within the audio stream NumPy array. This index corresponding to each timestamp is calculated as the product of the timestamp with the sample rate of the audio. When the indices corresponding to both the start and end timestamps for a word have been calculated, a 1KHz signal spanning those indices is generated as a float16 NumPy array and is spliced into the original audio stream, in place of the audio samples corresponding to the banned word. The modified (censored) audio stream is then placed

in a final shared queue where the playback module will pick it up and stream it to the output device.

Relevant Connections

Literary Review

A wide array of scholarly articles and journals discuss practical usages of speech recognition technology implementations like the one discussed in this work, with results in practice varying.

Article 1: Speech Recognition as a Transcription Aid: A Randomized Comparison with Standard Transcription

The 21st century provides extensive branched research into the idea of speech recognition, dating back to the beginning of the 2000's; one such work depicts SR as a transcription aid according to Mohr et al., 2003. Within their study, they acknowledged the potential cost-cutting on information entries for clinical information system databases. Referencing the idea of cutting costs through automated labor, the idea present here is creating computer-generated documents emulating the purpose of a physician and following up with the typical procedures in their institutional infrastructure.

As a backbone to their trial, two specialty divisions were randomized for standardized and speech-recognition based transcription processes, where the dictation of each separate dictation was a primary measurement for both human and computer. At the end of the experiments, the study interestingly yielded a result where humans were surveyed and deemed to have *not* improved productivity even with the help of the speech recognition system; however, it did prove to show some reasonable success as an accurate baseline model. In general, it served more as a proof-of-concept idea rather than a true practical improvement. It should also be noted that this study was conducted in 2003, meaning that the AI models between then and now have drastically improved.

Article 2: A systematic review of speech recognition technology in health care

In a similar medical context, but in a scenario much more modern (by roughly 10 years), the work done by Johnson et al. in 2014 connects extended existing literature related to SR technology in healthcare to the innovations that focus on clinical

information maintenance and patient record modifications. In a study like the one described by Mohr et al. in 2003, where a speech recognition system (composed of a microphone and conversion software) isolates words which are recognized through a minimum prediction residual, it was reported that their study conversely noted a reduction in turnaround times of reports for document processing procedures within endocrinology and psychiatry teams reduced from 15.7 hours to 4.7 hours. By using transcriptions for their daily interactions, in unison with more modernized forms of technology (better databases, more affluent users, larger cloud network usage, effective filtering of transcript) the effect of a successful voice recognition model becomes more evident in a high-pressure environment; this study bolstered the idea that automated transcriptions pipeline success may require reactive technologies that efficiently utilize the data.

Article 3: Possibilities of using Speech Recognition Systems of Smart Terminal Devices in Traffic Environment

Covering a different context, the work done by Husnjak et al. (2014) illustrates the benefits and complications of smartphone voice recognition in the automotive industry. It used to be extremely challenging to find simple and effective use cases for voice recognition with cars, but a prominent field of research focused on applications within cars that facilitate speech recognition pipelined from the phone to the car. For instance, a phone connected to the audio system of a car nowadays can integrate Apple CarPlay, who can be controlled by Siri to do commands like calling somebody, using speech-to-text, and other general use cases where the smartphone acts as a terminal; the key in justifying these advancements is orienting around the voice so that the focus remains primarily on the road ensuring safety. This idea has grown rapidly in popularity over the last decade, as many modern cars now incorporate these functionalities within Apple CarPlay, Android Auto, etc. Additionally, companies like Tesla have gone out of their way to make their own software to incorporate these growing changes, exhibiting the fact that these technologies are here to stay and will only improve with the Internet of Things (IoT) rapidly growing.

Futuristic Scenarios

The literary review discusses the scenarios of applications already discussed, but the future of these technologies matters equally; the pros and cons of speech recognition opens the possibility of worlds where there are benefits facilitating a utopian-style society and negatives that may promote a dystopia. Of course, there are a near-infinite number of outcomes in the future, so every "utopia" or "dystopia" is a slippery slope. However, the good & bad possibilities are what move humans the most, so the best and worst possible outcomes are the ones worth examining.

Utopian

In a Utopian world, viewers/individuals would have the ability to disable/enable any form of censorship at their own will (under the impression of the implications of our specific application). This alleviates the legal tension behind censorship, because if an individual doesn't like what they're hearing, they would quite literally have all the tools they need to not listen to it. This would imply that companies don't need to be restricted by ethics boards and/or policymakers that govern what should and shouldn't be censored.

This utopian idea would also be innovative for child-filtered content, because parents would be able to solely decide what their own child gets to hear. Although it is true that not all parents are good parents, the fault/blame can ultimately be attributed to a single person. Thus, the problem in these instances can be traced & resolved much more easily than if the fault were attributed to a large system/bureaucracy that makes vast, generalizing (and therefore possibly oppressive) decisions.

In summary, everyone would have everything they need to stop themselves from being exposed to vulgar or scary content made online, thereby also making the internet a safer place. Allowing each user to define their own list of banned/censored words to exclude from conversations would also remove the necessity of strict, systemic regulation, because the responsibility would be shifted down to the individual level.

Dystopian

To truly understand the dystopia, it first needs to be clear how this dystopia is even possible, given the results of this project. In the implementation, one of the sections of code performs censorship by adding a 1000Hz "bleep" noise into the audio file. In

theory, people could do anything at this stage; in other words, the decision to generate a "bleep" noise over the swear word is completely arbitrary. A possible expansion of the project would be to replace the "bleeping" segment of code with LM-and-deepfake-based code. The language model (LM) would be used to understand what words would make sense given the context; since LM's are heavily prone to bias from training datasets, given that there are open-source language models, anyone can train a "politically biased" LM to generate politically biased words/phrases/sentences that fit in any given context. These words from the LM would then be fed as input into a deepfake model that was trained on the speaker's live audio, perhaps through an initialization process. The end product would basically be an automated model that could twist people's words.

Now that the power of the LM-and-deepfake-based model is obvious, even the most honest and noble institution of regulation would not want this model to get into the hands of the public, as it would only serve to make the internet unreliable beyond salvation. However, a key cynical assumption of the systemic dystopia is that the systems of power would take the LM-and-deepfake-based model and use it for their own personal gain, instead of discontinuing/banning the model for the greater good.

Ever since COVID, societies have revised many systems so that the systems can continue in a completely online setting. Grade School, College, and Corporate America as we know it had begun to dig a hole towards a fully virtual world. The only thing the systems of power would need is another reason for the members of society to pick up their virtual shovels and keep digging. Given that the world does go virtual again for some reason, this would be the perfect opening for systems of power to force high-profile individuals to communicate through livestreams. For context, the original project was intended to be implemented into a livestream. Who says that the LM-and-deepfake-based model can't be used in real-time, also? So, there is nothing stopping the systems in power from forcing egregious agendas into well-intended, high-profile individuals' mouths. This would essentially give those in power the ability to compel speech. From this point on, "1984" by George Orwell perfectly explains the rest of the rabbit hole that is the systemic dystopia.

Model Card – Whisper Models

Model Details: ASR and Whisper Models were trained and released by OpenAI. These encoder-decoder transformer-based models were created to perform speech recognition and translation tasks capable of transcribing speech audio and translating them into English with recognizing the language they are spoken in. In total there are 9 models of different sizes and capabilities.

Size	Parameters	Layers	English-only model	Multilingual model
tiny	39 M	4	✓	✓
base	74 M	6	✓	✓
small	244 M	12	✓	✓
medium	769 M	24	✓	✓
large	1550 M	32		✓

Model Date: September 2002 (original series) and December 2022 (large-v2)

Model Version: There are 9 models in total.

Model type: Sequence-to-sequence ASR and transformer-based speech translation model

Citation: Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv: Electrical Engineering and Systems Science, Audio and Speech Processing, 2212.04356. https://doi.org/10.48550/arXiv.2212.04356

Intended Use:

- **Primary Intended Use:**
    - Studying robustness, generalization, capabilities, biases and constraints of the current models.
- **Primary Intended Users:**

    - AI researchers and developers who want to use ASR systems in their applications for either transcriptions or translation of speech.
- **Out-of-scope Users:**
    - Group of developers planning for malicious use cases that can be made possible with ASR systems. These models shouldn't be used for classification of people. These models shouldn't be used in high-risk domains either.

Factors:

- Despite age, gender and accent not being explicitly provided to the model, these factors may result in hallucination where the output includes words which were not present in the original audio.

Metrics:

- **Model Performance Measures:**
    - Measured Word Error Rate (WER) which is based on string edit distance that compares model output with expected output.
    - Robustness was tested by comparing accuracy of Whisper models to human accuracy with noisy or perturbed speech audio.
    - Transcription performance in languages correlates with the amount of training data used; models show near state-of-the-art accuracy.
    - Hallucination sometimes occurs when the model tries to guess the next word while processing the current word in the speech.
    - The model performs unevenly across different genders, races, ages, accents for the same language being processed.

Ethical Considerations:

- **Data:** The datasets used for training include audio recordings of possible sensitive topics, insults, slurs, etc. The datasets include a variety of words being spoken so that every word can be recognized.
- **Human Life:** The model is not created for identification of ethnic groups, genders, ages of people.

- **Risk & Harms:** If this model is used for classification of people or to use on people that don't give permission, the model can be harmful to society.

Training Data & Evaluation Data:

- **Datasets:**
  - Training dataset consists of 680 000 hours of audio data and their respective transcripts. 65% of the data (438 000 hours) is in English.
  - 18% of the data (126 000 hours) is non-English audio paired with English transcripts.
  - The last 17% of the data is non-English audio and corresponding transcript. There are 98 different languages in this category.
- **Motivation:**
  - Multiple languages were included in the training dataset with English audio as the majority. The focus was to train the model such that it would be able to transcribe or translate any non-English audio if specified to English.
- **Preprocessing:**
  - Datasets included human-written and ASR-generated transcripts; heuristics were used to remove machine-generated transcripts.
  - Complex punctuation, formatting, stylistic capitalizations and white spaces were removed from transcripts.
  - The dataset was divided into 30-second segments with relevant transcripts including silent segments.
  - After initial training, the audio-transcription pairs with high error rate and length of original audio were removed for better training.
  - All audio was re-sampled to 16000 Hz.

Caveats and recommendations: Equalize the number of samples for each language to potentially reduce hallucinations.

Conclusions

In this work, we explored the caveats and potential real-world applications (both presently and in the future) for speech recognition technologies and their ability to impact varying environments and fields. As shown through the literary review and for our implementation of a speech-recognition program for censorship, the practicality of these systems are typically found when they represent supplemental quality-of-life add-ons to existing features. For instances like high-pressure hospitals, or in scenarios where attention must be divided elsewhere (like in a car), speech recognition technologies can make ease of access and mobility of a system much more prominent while also sometimes adding features that aid in general safety/usability; the ability alone for our program to be executed on a mobile phone is a testament to the ability for this technology to be widespread.

The rise of additional advanced AI-based speech recognition systems like Siri and text-to-speech features on phones foreshadow the emergence of similar evolved technologies dominating the tech industry in the near future. As AI models become more refined and accurate, the butterfly effect of practical text-to-speech applications promises to breach into even more fields for the betterment of machine learning and human-computer interactive applications.

References

Husnjak, S., Perakovic, D., & Jovovic, I. (2014, March 25). *Possibilities of using speech recognition systems of smart terminal devices in traffic environment*. Procedia Engineering. Retrieved 2023, from https://www.sciencedirect.com/science/article/pii/S1877705814003002

Jain, N., & Rastogi, S. (2019, January). *Speech Recognition Systems - a Comprehensive Study of Concepts and Mechanism*. ResearchGate. Retrieved 2023, from https://www.researchgate.net/publication/331679755_SPEECH_RECOGNITION_SYSTEMS_-_A_COMPREHENSIVE_STUDY_OF_CONCEPTS_AND_MECHANISM

Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., & Dawson, L. (2014, October 28). *A systematic review of speech recognition technology in health care - BMC Medical Informatics and Decision making*. BioMed Central. Retrieved 2023, from https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-14-94

Mohr, D. N., Turner, D. W., Pond, G. R., Kamath, J. S., De Vos, C. B., & Carpenter, P. C. (2003). *Speech recognition as a transcription aid: A randomized comparison with standard transcription*. Journal of the American Medical Informatics Association : JAMIA. Retrieved 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC150361/

Ok, M. W., Rao, K., & Ulloa, P. R. (2020, December 17). *Speech recognition technology for writing: Usage ... - sage journals*. Sage Journals. Retrieved April 24, 2023, from https://journals.sagepub.com/doi/abs/10.1177/0162643420979929

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv: Electrical Engineering and Systems Science, Audio and Speech Processing, 2212.04356. https://doi.org/10.48550/arXiv.2212.04356