Preliminary Analysis of an AI-powered Transcription Bot for FM Transmissions

Zhemin Zhang KD2TAI[1], Brian Robert Callahan AD2BA[1]
1: Rensselaer Polytechnic Institute

Abstract

Amateur Radio communications mediated by Artificial Intelligence (AI) and Machine Learning (ML) provide a wide variety of experimentation. One potential experiment is developing a bot using AI/ML that can read FM voice transmissions, transcribe the audio heard into text, and finally post the transcriptions on social media, all without human intervention [1]. In this paper, we consider some factors that may influence transcription accuracy. While accuracy is important, there are additional considerations for transcriptions, especially those that may be archived on the Internet: these include not posting gibberish and indecent words. We consider these points as improvements to our initial design.

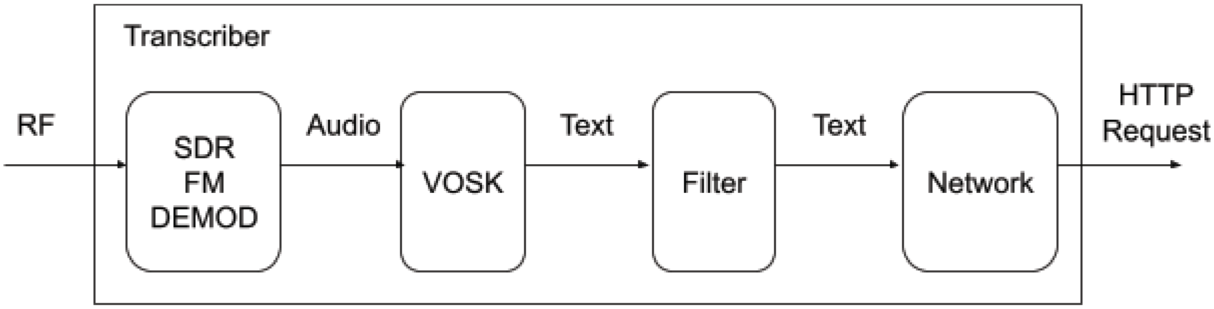


Figure 1. System Block Diagram

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) is an exciting new focus for amateur radio experimentation. In a forthcoming article, we demonstrated a prototype of an AI-powered transcription bot for FM transmissions, named FMBot [1]. The system uses AI/ML technology to transcribe text from audio demodulated from an FM signal using an SDR, then push the transcript to a social media account on the Internet. As our initial prototype was solely a minimum viable product, more questions must be addressed in order to produce a more full-fledged system. Some of the most important questions identified center around accuracy. How accurately can the system transcribe words, and how does it select the messages to be broadcasted? Additionally, as we came to refine our understanding of the problems the AI/ML bot intends to solve, we have come to adopt a broad view of the concept of accuracy to incorporate ideas of unintelligible and inappropriate transmissions.

In this article, we conduct experiments to explore possible factors that may influence the transcription accuracy of FMBot, and measure the degree of such impacts. We also explore methods to filter out meaningless (also known as “gibberish”) or inappropriate messages.

On accuracy

A key takeaway from the minimum viable product was that different models yielded different levels of accuracy. We used VOSK [2] as the open source off-the-shelf AI/ML speech recognition software for transcription. For its English-language models, VOSK offers three: the largest model we learned would only run on quite powerful desktop machines; a moderately powerful laptop was unable to run the largest model. There is a midrange model for such laptops and other similarly midrange-specced machines. Finally, there is a small model for lightweight machines such as the Raspberry Pi.

The size of the model made a noticeable difference in accuracy of the transcribed text. While we didn’t measure the exact accuracy of each model, as conditions between radio and computer setups might skew exact numbers, we felt anecdotally confident that the improved accuracy was noticeable by comparing outputs between different models.

Accuracy in transcription is a common problem for many disciplines. To take just two disparate examples, psychotherapists have argued that increased automatic transcription would increase effectiveness, training, and monitoring [3], and musicians are interested in automatic transcription to aid in capturing music heard accurately into other forms more suitable for analysis [4]. Depending on context, accuracy for our bot might be more or less needed: amateur operators looking for amusement might not require much more accuracy than is needed to enjoy their bot. However, we also envision a bot like the one we are developing acting in part as a response to Covid-19: as clubs turned to Zoom and other Internet services for meeting, our bot can be used to put those meetings squarely back into the amateur radio space over the airwaves without needing to record audio or have someone furiously take notes. For this activity, clubs likely need a higher level of accuracy than those looking for amusement. Moreso, we also envision the bot enabling voice activities such as voice contesting for amateur operators who might not otherwise be able to participate in such voice-based activities.

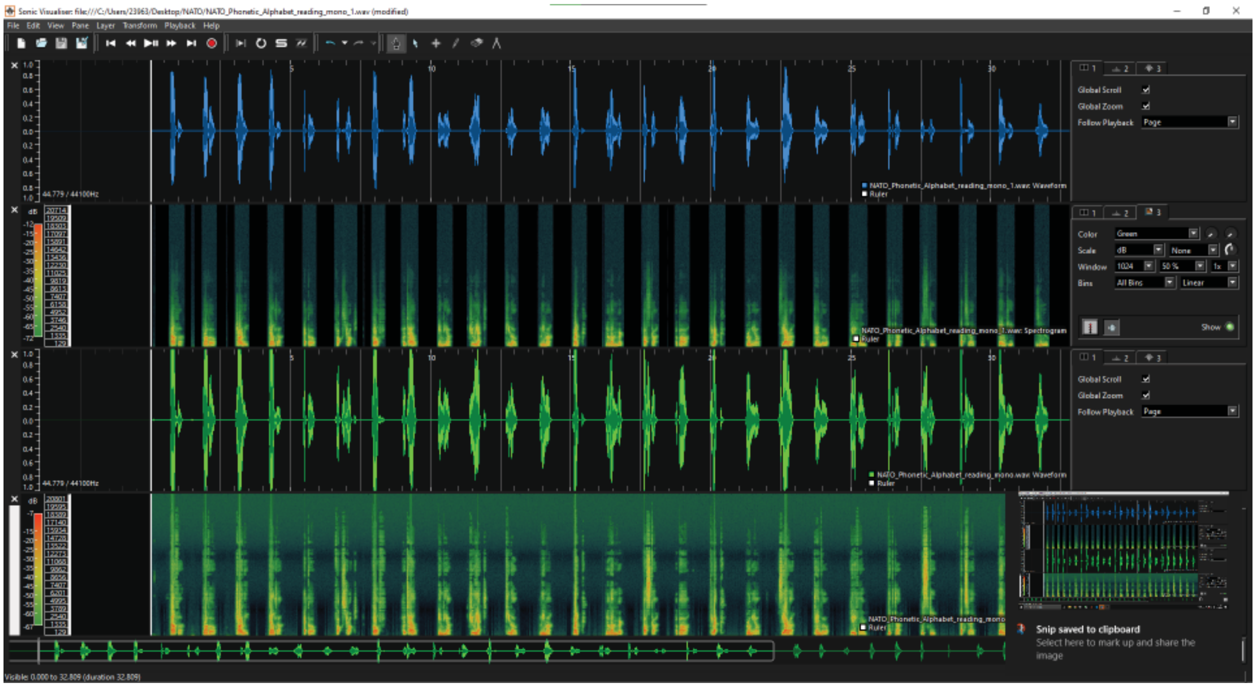
Not only does accuracy mean the literal English-words-to-English-text that we often think, in our preliminary experiments we noticed oftentimes the bot would post transcripts of gibberish, effectively nonsense and nonsense words. We also noticed that the bot would interpret pockets of silence as the word “the” which led us to eventually implement a filter that would reject all single-word messages. We also think of accuracy in terms of not posting inappropriate transmissions. Such transmissions could include hate speech, offensive words, and other transmissions that may be deemed “not safe for work.” While the FCC does have rules prohibiting the transmission of “obscene or indecent words or language” [5], it does not necessarily follow that all amateur transmissions are free of such language. Amateurs with transcription bots of their own may not want such language transcribed into text and then posted on social media on the open Internet on their accounts. A bot ought to be able to identify inappropriate transmissions and prevent the posting of transmissions including such words to the best of its ability.

Experiments

We have identified two factors that are important to transcription accuracy: signal-to-noise ratio (SNR) and bandwidth. Intuitively, it is harder to understand another person's words with noise in the background or talking over the telephone, which demonstrates the effect of the two aforementioned factors. We previously discussed the impact of SNR on accuracy quantitatively [1], so we will focus solely on bandwidth here.

These are the specifications for the hardware and software we used to conduct our experiments.

a. Hardware specifications
    i. Processor: AMD Ryzen 7 3800X 8-Core Processor 3.90GHz
    ii. Installed RAM: 32 GB
    iii. Software-Defined Radio: RSP1 kit connected via USB
b. Windows specifications
    i. Edition: Windows 10 Home
    ii. Version: 2004
c. Software specifications
    i. Python 3.9.12
    i. VOSK 0.3.42
    ii. PyAudio 0.2.11


Sonic Visualiser: file:///C:/Users/23963/Desktop/NATO/NATO_Phonetic_Alphabet_reading_mono_1.wav (modified)

Global Scroll ☑
Global Zoom ☑
Follow Playback: Page

Color: Green
Scale: dB, None
Window: 1024, 50 %, 1x
Bins: All Bins, Linear

Snip saved to clipboard
Select here to mark up and share the image

Visible: 0.000 to 32.809 (duration 32.809)

File
Edit
View
Pane
Layer
Transform
Playback
Help
44.779 / 44100Hz
NATO_Phonetic_Alphabet_reading_mono_1.wav: Waveform
Ruler
NATO_Phonetic_Alphabet_reading_mono_1.wav: Spectrogram
NATO_Phonetic_Alphabet_reading_mono.wav: Waveform
NATO_Phonetic_Alphabet_reading_mono
dB
Show
5
10
15
20
25
30

Figure 2. Visualization of the filtered spectrogram (top) vs. the original one (bottom). 300 Hz low cutoff and 2700 Hz high cutoff. Note the intensity difference at higher frequencies.

The bandwidth contains two variables: the low and high cutoffs. First, we test the recognition accuracy of a clear recording (SNR > 64 dB) by adjusting two boundaries of a 4th-order Butterworth filter at 100 Hz or 200 Hz intervals, respectively, for low and high cutoffs. Since the way the filter was implemented prohibits low-cutoff to be zero, we set it to 1 Hz instead.

We then iterate over filtered audio files to perform transcription. Punctuation is included in VOSK output, however we chose to drop all punctuation and then format the output one word per line. Some recognized words are replaced using sed (e.g., “nine” to “niner”), as some words in NATO phonetic alphabets are read differently than in common English. Then we use the command

```
$ diff  -y --suppress-common-lines standard.txt result.txt | wc -l
```

to report the differences between the transcript and ground truth. Another script collects the result and plots a 2D heatmap, as shown in Figure 3, also known as the contour map for discrete values, to demonstrate how two variables together affect the accuracy.

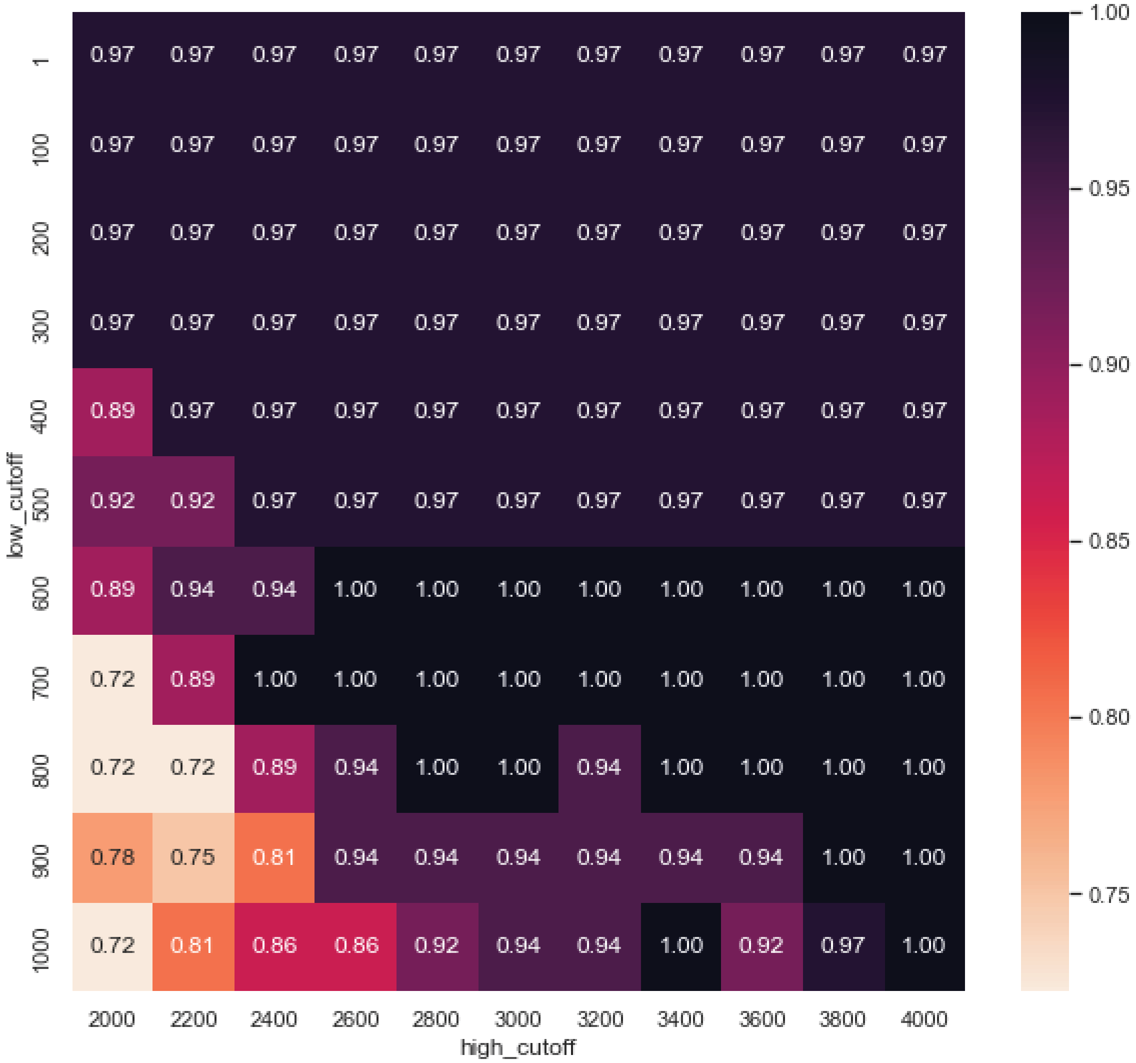


Figure 3. Heatmap for NATO phonetic alphabet recognition accuracy with various low and high cutoffs in Hz.

The accuracy deteriorates as the bandwidth gets narrower, in other words, towards the bottom left of the map. For example, to yield an accuracy greater than 90%, the minimum bandwidth needs to be greater than 1700 Hz, whereas it takes 2800 Hz on average to reach near 100% accuracy.

Note that this does not mean that lower lower-bound and higher upper-bound are always desirable. For certain words, we expect cutoffs to be at the appropriate location to improve recognition. The error most likely happening in this test setup is the pair “kilo” and “kino,” and this is fixed as the low cut-off rises, which is visualized by the bottom right corner being darker than the top.

Like the previous SNR test, human perception outperforms the AI transcriber. Even at the narrowest bandwidth (1000, 2000), it is still understandable. This is probably due to the context that humans can understand. For example, in this NATO alphabet testing, human operators specially trained for it will subconsciously correlate the word heard with one of 36 possibilities in the set, while to achieve the same effect, the programmer needs to specify such information to a machine.

In the above bandwidth test and the previous SNR test, we used NATO phonetic alphabet recording as audio input. Despite being a good starting point as it is a standard format to exchange call signs, it is only a small portion of ham radio traffic. Furthermore, the recording pronounces a single word at a time with extra long spacing; this speech pattern is uncommon in real conversation. As such, we chose an excerpt from the Wikipedia page on artificial intelligence [6], recorded by Zhemin himself in a young male voice, to test the system's performance in a more practical situation. Below is an annotated excerpt; the contents in parentheses were unintelligible to the bot due to their unstressed nature, and the parts in square brackets indicate a human error made when reading. These factors are compensated for in our accuracy calculations.

> *Artificial intelligence is intelligence demonstrated by machines, as opposed to a natural intelligence displayed by animals including humans. AI research has been defined as the field of study of intelligent agents, which refers to any system that perceives its environment and takes actions that maximize (its) chance of achieving its goals.*
>
> *The term "artificial intelligence" had previously been used to describe machine(s) that mimic and display "human" cognitive skills that are associated with (the) human mind[s], such as "learning" and "problem-solving." This definition has since been rejected by major AI researchers who now describe AI in terms of rationality and acting rationally, which does not limit how intelligence can be articulated.*

The microphone, an RØDE NT-USB Mini, was placed 40 centimeters away, with Windows Voice Recorder running on the PC with a default sampling rate of 48000 Hz. The audio is then noise reduced using Audacity 3.0.2 with reference to silence before talking starts. After that, we applied the compressor with default settings to yield the reference audio for this test. Finally, we repeat the above experiment steps to generate the result in Figure 4.

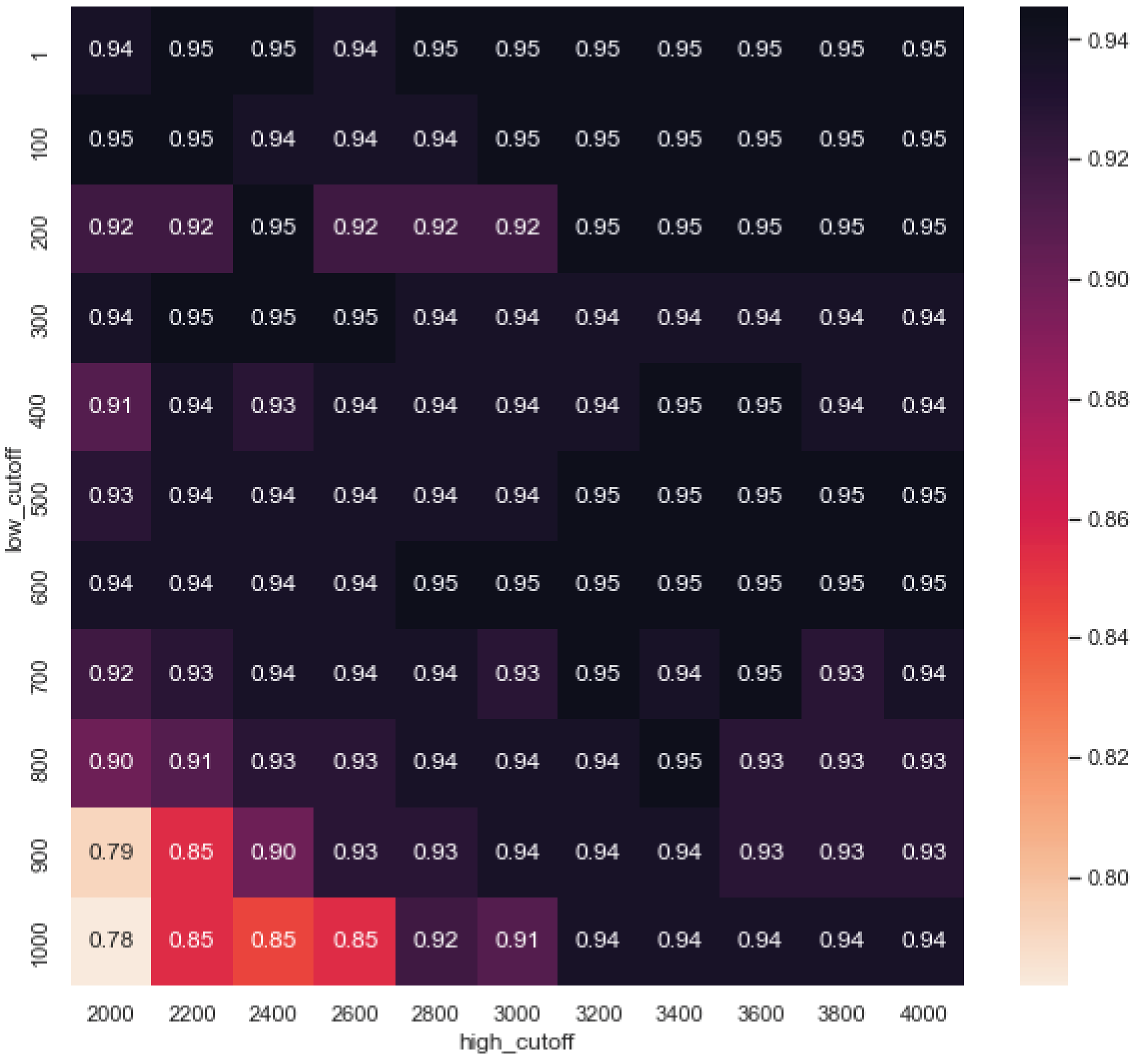


Figure 4. Heatmap for AI Wikipedia article recognition accuracy with various low and high cutoffs in Hz.

As expected, the overall accuracy decreases in this test set. We believe the most plausible explanation lies in the limited vocabularies of the model; accuracy would likely be improved if the model could be trained on a large corpus of actual amateur radio transmissions and conversations. The stressed and unstressed syllables and words found in natural speech patterns are a challenge to the model. The countermeasure for the vocabularies issue is to train the model with an even larger pool of words, although improvement will be limited given the law of diminishing returns; the size of the model grows rapidly as words rarely used in everyday conversation get included. To the stress and unstress issue, a professional announcer may overcome the obscurity, but as the system targets the public, there may be little that can be done.

What is unexpected, however, is that in our experiments the male voice allows a higher low-cutoff frequency before the accuracy deteriorates compared to the female voice in the NATO test set. Intuition suggests that the male voice in general has a lower fundamental frequency. With a higher low-cutoff, we might expect a more severe loss of information, yet the measurement suggests opposing evidence. One possible explanation is that the model extracts information from the harmonics more than the fundamental frequency, but without solid evidence, this phenomenon remains unresolved and is worth further study.

Recently, the editor-in-chief of *QEX* called for more rag chewing [7, 8], and we could not agree more–as that would significantly aid in training an amateur radio-specific model to increase the accuracy of the bot! CQ CQ CQ any audio you would like to share for improving the accuracy of the bot.

Message Filtering

The filtering process involves multiple filters for different purposes, all cascaded together. Each unit specialized in specific topic detections is joined by “OR” logic to assemble a complete filter module. Figure 5 shows an example configuration and its flowchart. An inaccurate RF analogy to this is a series of bandstop filters that reject undesirable information. This flexible design allows more user-defined filters to be added later on.

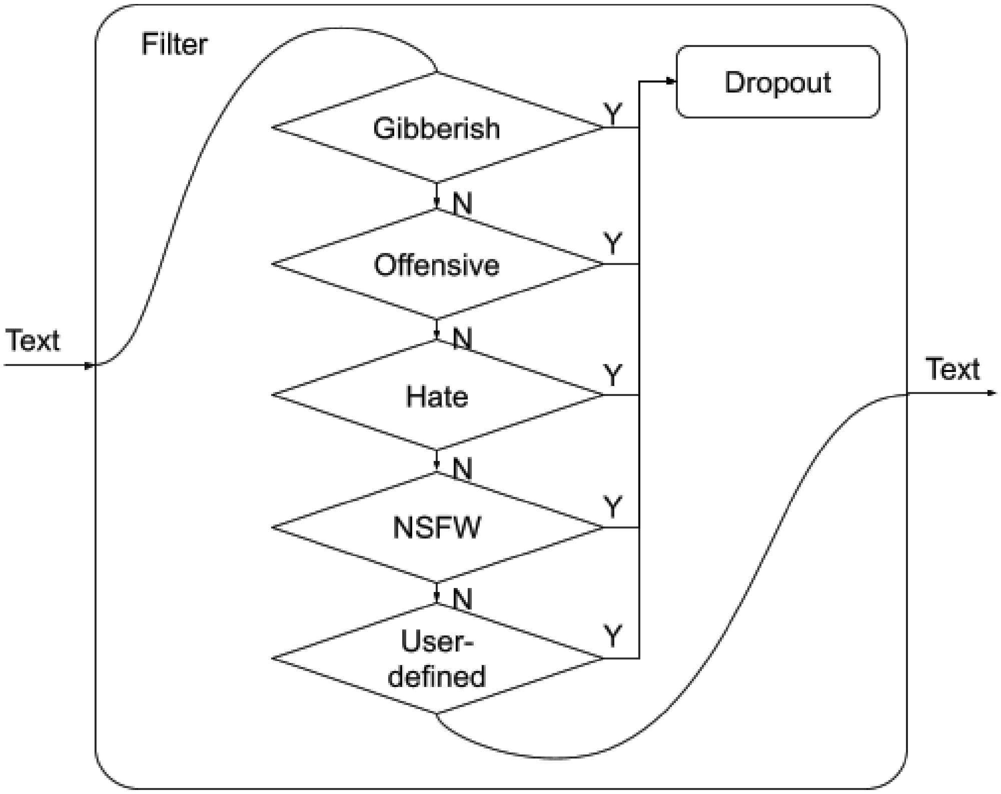


Figure 5. A Closer Look at Filtering Mechanism

There may be some concerns with a filtering mechanism that an amateur operator ought to be aware of. First is the idea of censorship. While it is of course true in the United States that a private person has the right to determine what speech makes it out of the filter, one operator's peace of mind can sometimes be felt as another operator's censorship. Nonetheless, censorship may be an important consideration in other locations.

Additionally, different taboos exist across different cultures; what constitutes an offensive word to one person may be a totally innocuous word to another, even within the same language. Consider the differences in slang among American, British, and Australian English for some notable examples. This means there will never be a one-size-fits-all filter. The filter will always have some degree of cultural context implicit in its design and operation. We consider this to be perfectly acceptable and perhaps even desirable, as it helps to imagine the flexibility of the filter design overall while being able to be highly adaptable to any situation.

Future filtering work

We have identified two important areas of further work that could be added to enhance the power of the filtering mechanism. First, we might consider the informativeness of a message. A message that was not sufficiently informative by any number of metrics might be a good candidate to drop. Second, we might consider messages containing sensitive information, such as passwords to online accounts, and it would be important to drop those messages as well lest an operator's secrets be divulged to the world.

Conclusion

In designing the original AI/ML transcription bot, we discovered several issues with the accuracy of transcription that required further exploration. In this paper, we conducted a number of experiments designed to better understand how we could refine the accuracy, defined broadly, of the bot. Our experiments examined the effects of bandwidth on transcription accuracy. We then experimented with filters for unintelligible and inappropriate transmissions, designing a system that is flexible enough to deal with cultural contexts.

Future work may identify additional variables that affect transcription accuracy. Future work may also explore concepts such as informativeness of a message in determining if it should be posted to social media. The filter may also be extended to not post sensitive information, such as passwords.

The continued development and experimentation of the AI/ML transcription bot demonstrates the viability of research and experimentation at the intersection of amateur radio and AI/ML. It is our hope that this article inspires future work by other amateur operators in this space.

Works Cited

[1] Callahan, B. R., and Z. Zhang. Forthcoming. “An AI-powered transcription bot for FM transmissions.” *QEX*.

[2] https://alphacephei.com/vosk/

[3] Miner, A. S, A. Haque, J. A. Fries, S. L. Fleming, D. E. Wilfley, G. T. Wilson, A. Milstein, D. Jurafsky, B. A. Arnow, W. S. Agras, L. Fei-Fei, and N. H. Shah. 2020. “Accessing the accuracy of automatic speech recognition for psychotherapy.” npj Digital Medicine 3, 82.

[4] Mauch, M., C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon. 2015. “Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency.” Available online: https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/7247/tony-paper_preprint.pdf

[5] § 97.113 Prohibited transmissions. Available online: https://www.ecfr.gov/current/title-47/chapter-I/subchapter-D/part-97#97.113

[6] https://en.wikipedia.org/wiki/Artificial_intelligence

[7] Siwiak, K. 2022. “Perspectives.” *QEX* July/August 2022: 2.

[8] Siwiak, K. 2022. “Perspectives.” *QEX* September/October 2022: 2.