# Robocall Audio from the FTC's Project Point of No Entry

Sathvik Prasad and Bradley Reaves

{snprasad,bgreaves}@ncsu.edu

North Carolina State University

## 1 Dataset Summary

This document describes a collection of over a thousand audio recordings of automated or semi-automated phone calls. Such calls are commonly called robocalls. These recordings were made available by the FTC through the Project Point of No Entry (PPoNE) initiative [2, 3, 4]. The dataset consists of 1101 robocall audio recording used in the real-world. Most of these robocalls are suspected illegal calls. Malicious actors used a majority of these recordings to defraud people. The dataset also includes the cease and desist letters sent by the FTC to the suspected call-originating entity (telephone carrier or the robocaller).

## 2 Data Collection

Each audio recording was collected using the links embedded within the Cease and Desist letters sent by the FTC to the suspected call-originating entity (telephone carrier or the robocaller). The webpage and the PDF files published on the PPoNE website were collected using automated crawlers. Links to audio files embedded within the PDF were extracted using *pdfgrep* and then downloaded using *wget*.

## 3 Audio Recording Setup

Although this dataset does not contain granular information about where or how these audio example were collected, most example robocall audio recordings are collected using telephony honeypots [5], voicemails, or reports from phone users who may have recorded the call using their own devices. These calls were likely generated by a robocalling system, and the audio traversed the phone network (over a logical channel) before being recorded by the recipient.

## 4 Curating and Normalizing the Dataset

Since these recordings are sourced from various honeypots and voicemails, the original audio format included wav, amr, and mp3. Some recordings were in stereo and others in mono.

All the recordings were converted to *wav* format (*pcm_s16le*) and re-sampled to 16kHz using *ffmpeg*. When the source audio was in stereo, it was converted into two mono streams (filenames *_left.wav* and *_right.wav*). The *_left.wav* contains the audio stream originated by the remote party (robocaller), and the *_right.wav* contains the audio stream originated by the local party (honeypot or voicemail). Only the *_left.wav* files were transcribed and included in the dataset. However, the respective *_right.wav* audio files are also included in the dataset for completeness.

## 5 Dataset Format

The *metadata.csv* format contains the filename and the transcription of the audio recording. It also includes the language used within the call and was detected automatically using Whisper [6]. The dataset consists of 1101 calls out of which 97.5% (1073) calls are in English and 2.5% (28) are in Mandarin/Chinese. The medium (multilingual) model was used to transcribe the audio. The specific cease and desist letter or the warning letter is also included for each audio recording.

## 6 Cease and Desist Letters and Warning Letters

The cease and desist letters and warning letters are included in the *pdf_files* directory. The *case_pdf* column in the *metadata.csv* file contains the link to the specific letter for each audio recording.

## 7 How to access and use the dataset

The dataset is hosted on GitHub [1] and can be easily accessed using Pandas and HuggingFace datasets.

```
import pandas as pd
df = pd.read_csv('metadata.csv')
df.columns
#Output: ['file_name', 'language', 'transcript', 'case_details', 'case_pdf']
```

The dataset can also be loaded using Huggingface's datasets library which can process raw audio files.

```
from datasets import Dataset, Audio
import pandas as pd
df = pd.read_csv('metadata.csv')
audio_dataset = Dataset.from_dict({
    "audio": df['file_name'].to_list(),
    "transcript": df['transcript'].to_list(),
    "language" : df['language'].to_list(),
    "case_pdf" : df['case_pdf'].to_list(),
}).cast_column("audio", Audio(sampling_rate=16000))
audio_dataset.head()
```

After loading the dataset using Huggingface Audio library, individual samples can be inspected as follows:

```
audio_dataset[0]

'''
#Output
{'audio': {'path': 'audio-wav-16khz/.._normalized.wav',
  'array': array([0.03210449, 0.03390503, 0.03796387, ..., 0.00616455, 0.00695801,
         0.0072937 ]),
  'sampling_rate': 16000},
 'transcript': 'We would like to inform you that there is an order placed for Apple iPhone 11
 ↪  Pro using your Amazon account. If you do not authorize this order, press 1 or press 2 to
 ↪  authorize this order. ',
 'language': 'en',
 'case_pdf': 'pdf_files/..inaljms.pdf'}
'''
```

[1] https://github.com/wspr-ncsu/robocall-audio-dataset.git

## 8 License and Contact Information

This document describing the data is released under the Creative Commons BY-ND [1] license. The data itself is in the public domain. If you find this structured data useful, we would appreciate (but do not require) an acknowledgement in any publications.

## References


[1] CC BY-ND 4.0 Deed | Attribution-NoDerivs 4.0 International | Creative Commons — creativecommons.org. https://creativecommons.org/licenses/by-nd/4.0/. [Accessed 14-11-2023].

[2] FTC, Law Enforcers Nationwide Announce Enforcement Sweep to Stem the Tide of Illegal Telemarketing Calls to U.S. Consumers — ftc.gov. https://www.ftc.gov/news-events/news/press-releases/2023/07/ftc-law-enforcers-nationwide-announce-enforcement-sweep-stem-tide-illegal-telemarketing-calls-us. [Accessed 13-11-2023].

[3] Project Point of No Entry Letters — ftc.gov. https://www.ftc.gov/legal-library/browse/project-point-no-entry-letters. [Accessed 13-11-2023].

[4] Project Point of No Entry Letters — web.archive.org. https://web.archive.org/web/20230418192421/https://www.ftc.gov/legal-library/browse/project-point-no-entry-letters. [Accessed 13-11-2023].

[5] Sathvik Prasad, Elijah Bouma-Sims, Athishay Kiran Mylappan, and Bradley Reaves. Who's Calling? Characterizing Robocalls through Audio and Metadata Analysis. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 397–414. USENIX Association, August 2020. https://www.usenix.org/conference/usenixsecurity20/presentation/prasad.

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. https://github.com/openai/whisper.