Realtime AI-Supported Voice Conversion (Deepfake) and its applications on Vishing and Social Engineering exercises

Pablo Alobera, Pablo López, Víctor Lasa

Tags: vishing, deepfake, ai, machine learning, social engineering

Foreword

Recent and ongoing advancements in AI technology mean that the dangers of voice cloning attacks are more real than ever before. An attacker could create realistic voice clones from as little as five minutes of recorded audio. This hugely increases the risk that organisations face from attacks that exploit the trust that employees place in senior staff.

We've already seen multiple successful attacks carried out against companies using AI voice cloning.[1] Combined with phone number spoofing,[2] these attacks are extremely hard to detect - and the use of AI means that they can be launched by a much wider range of people.

In a recording linked below, you can listen to a voice clone created and recorded in real-time, to demonstrate the power of AI voice cloning.

Introduction

The purpose of this article is to demonstrate and publicise the very real risks of voice cloning as applied to vishing exercises. For security reasons, we do not share any specific design or implementation details that could be of use to real-world attackers. That said, it should be expected some threat actors have already developed these techniques themselves.

This article will not provide all the information required to conduct an AI-powered vishing exercise. It would not be responsible to disclose all the details of NCC Group's internal research. Nevertheless, we feel we have included enough information for the risks and opportunities arising from these technologies to be fully understood.

NCC Group's research team have been exploring the potential capabilities and impact of deepfakes on cyber security. One of the ways in which these potential risks have been realised over the last few years is voice impersonation by means of AI. The possibilities offered by this technology could allow some classic and very well-established social engineering attacks (such as vishing) to be refined and extended to previously unthought of levels.

---

[1] AI Voice Cloning Scams: https://edition.cnn.com/2024/09/18/tech/ai-voice-cloning-scam-warning/index.html

[2] Caller ID Spoofing: https://en.wikipedia.org/wiki/Caller_ID_spoofing

These techniques make it possible to impersonate the voices of people in key positions inside a company, and to do so in such a convincing manner that it is possible to gain access to privileged information and even issue rogue directives.

Existing Research and Known Limitations

When we started our research, we quickly identified several limitations on these kinds of attacks. At that time, the vast majority of the state-of-the-art deepfake technologies and architectures were focused on offline inferences. These were capable of achieving good results when cloning a pre-recorded extract of someone talking, but they couldn't change a user's voice in real-time – a key requirement for vishing attacks. In addition to this, many of them were strongly dependant on TTS (text-to-speech) models.

This placed real and definite limitations on the possibility of conducting a realistic AI-powered vishing exercise. An attacker would really only have two alternatives, both unsatisfactory. Firstly, they could use a set of pre-recorded sentences, with the obvious problems this would present in terms of having a real-time conversation. Or secondly, they could input sentences into the TTS model on-the-fly, to create realistic response to the victim's speech, although this would introduce an unnatural delay in the cloned responses.

The majority of research at that time focused on pre-recording voice samples and then applying a voice model to them. There was little available research into how any real-time voice cloning could be delivered in a way that would not arouse suspicion in the victim.

The main objective of our research was therefore to try to overcome these limitations by developing a framework capable of real-time voice cloning. That is, a framework which took the words spoken by the consultant in their own natural voice as input, so that those words could then be delivered in the desired voice, in a way that would not raise suspicion with the victim.

Capability and Accessibility

It is true that, even relatively recently, the technology was just not mature enough to support this approach. Also, the equipment or compute costs were simply not affordable for most. However, the reality today is that the tools and infrastructure needed for real-time voice cloning are accessible to those with even limited technical and financial means.

During our research, we learnt how to train a model in a few hours using only minutes of publicly available voice samples of a target voice. This was used to successfully clone the target's voice and make calls with real-time voice processing. This was combined with NCC Group's existing phone number spoofing capability and social engineering team to create a multi-dimensional attack. Not only that, but we have since proved in multiple applications of this technology in real security assessments, that it would just not be reasonable to expect the victims to detect the subterfuge.

It is worth pointing out that this was all possible using hardware, audio sources and audio processing software that were all 'good enough', rather than being exceptional. That is, the financial outlay to achieve our results would certainly be within the reach of many individuals and certainly of a small organisation.

Voice Cloning with AI

Voice cloning with machine learning and AI involves training a deep learning model to replicate a person's voice by analysing recordings of their speech. The process starts by converting audio samples into spectrograms, which visually represent sound frequencies over time. These are used to teach the model the unique vocal features of the speaker, such as pitch, tone, and rhythm. These are often referred to as identity content features.

Once trained on a suitable reference voice, the model can take any voice and generate audio that mimics that speaker. This system typically includes a neural network that predicts how the voice should sound for a given sound and a vocoder that turns this prediction into actual audio. The result is a synthetic voice that can say new things while sounding convincingly like the original speaker.

Several models, architectures and approaches were tried during the research. These included:

- the use of different pipelines involving both TTS (text-to-speech) and STT (speech-to-text)
- the analysis of several pitch extraction algorithms
- a number of code reviews to determine key elements and components that could be improved or changed
- audio processing and dataset preparation techniques, etc.

This enabled NCC Group to internally develop a real-time voice spoofing framework that achieved the quality standards defined in the research program.

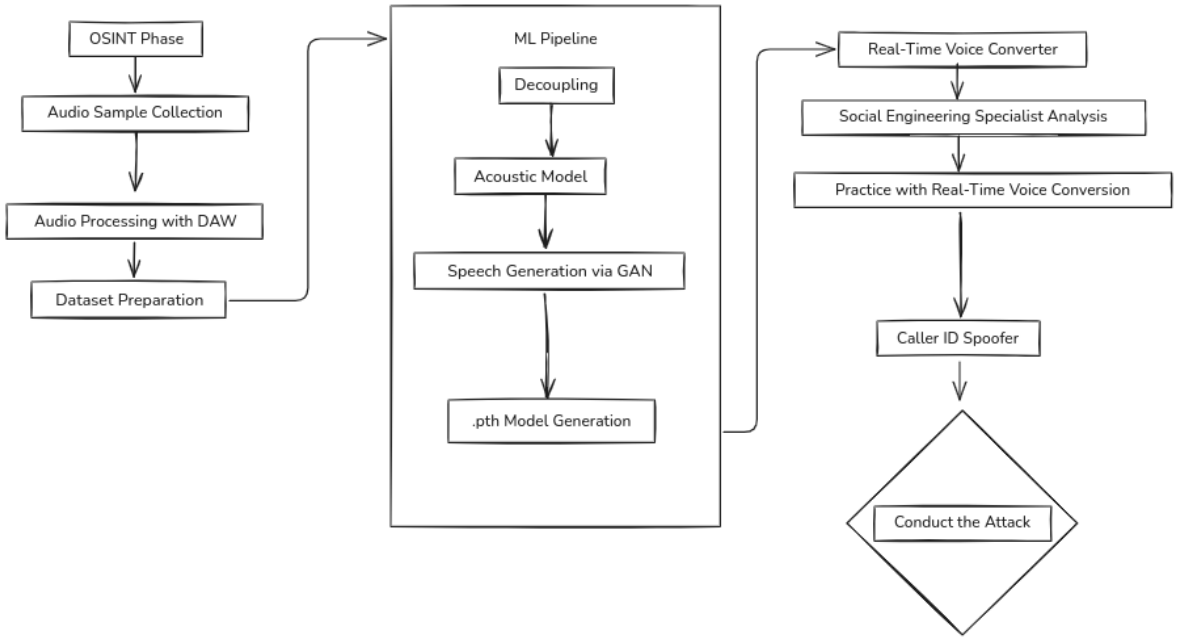

Figure 1 – Deepfake Vishing Assessment Workflow

Audio Source Gathering, Processing and Dataset Preparation

The first stage in creating the voice models used to create a voice clone required reasonable quality voice samples of the target voice. In order to mimic a realistic attack path, OSINT was used to obtain video and audio samples of the target's voice. Public sources on the Internet (such as corporate blogs and social media) proved to be a practical way to obtain samples of senior and executive committee staff members. The raw samples obtained for the voice cloning were either low quality or included background noise and other voices in addition to the target's. It was necessary to remove these extraneous voices to provide input of suitable quality to the machine learning models.

Parts of this audio process pipeline were automated, such as speaker identification (based on ML models). Other parts, such as noise-gating, equalisation, dynamic range compression or background removal, were performed manually using digital audio workstations (DAWs).

Once the audio processing was finished, the resulting audio samples were processed by several scripts that prepared the samples to be ingested by the ML model. This gave us our first version of the dataset that could be ingested by the feature extraction machine learning model. Thresholds and settings for the audio processing step were then modified in an iterative way until the results were sufficient for the final voice mask.

Model Training and Finetuning

The technical specifications for the selected AI/ML model involved several modules that were used in specific tasks. These tasks included *disentanglement* (decoupling linguistic, speaker-agnostic content from identity, speaker-dependent content), *pitch extraction* (the determination of a speaker's fundamental frequency) and *speech reconstruction* (synthesis of new linguistic content plus identity content).

For the disentanglement step, a self-supervised learning model was used, based on convolutional feature extraction blocks[3] and several layers of transformers.[4] This model could take an audio sample and extract separately identity, speaker dependent content (intrinsic features from the speaker) and pure linguistic content (what was said, independently from who said it; this is referred to as speaker-agnostic content).

For the pitch extraction, the RMVPE model was used. This model has proven its capability to extract vocal pitch, even with backgrounds as complex as polyphonic music. Another important element in this architecture was the acoustic model. This took information from previous processes on the ML pipeline and tried to reconstruct a suitable input for the final element in the

[3] Convolutional Neural Networks on Feature Extraction: https://towardsdatascience.com/exploring-feature-extraction-with-cnns-345125cefc9a

[4] Transformers: https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)

pipeline, the vocoder. The vocoder was based on a generative adversarial network (GAN[5]) model and crafted the final audio signal.

To optimise the generation of the voice model, a pre-existing model was fine-tuned with the voice samples of the target. We used the machine learning open-source framework PyTorch for this. The output of this process was a PTH file that contained the state of the model, weights, and other relevant information. This allowed the model to be trained on dedicated hardware and transported to any other environment for the real-time inference phase, once the training was complete. In addition to this, the setup supported a client-server architecture, which meant that the heaviest workload could be allocated to a dedicated hardware rig, specifically designed for this purpose.

That said, we were still able to create a successful proof of concept using only a laptop with a GPU NVIDIA RTX A1000 with 4096 MiB of memory and 2048 CUDA cores. Using this barebones hardware, the most resource-intensive process (such as training the model) only took a few hours, even when using computationally expensive pitch extraction algorithms. It was also possible to deploy this solution on cloud-based environments for the real-time inference phase. This had two distinct practical advantages for us. We could use the most capable GPU-optimised instances from anywhere in the world. And this approach removed the need to pay for new AI capable GPUs up-front, instead allowing us to rent them by the hour, substantially reducing our immediate costs.

Audio Devices

Once the fine-tuned model of the target's voice was ready, virtual audio devices were created to redirect audio signals in and out of the deepfake pipeline. First, the audio signal from the attacker's physical microphone was routed to a real-time voice modulator, which took advantage of the ML models and architecture mentioned above. The cloned voice was then delivered over a telephone call, making use of a caller ID spoofing solution which has been used in more traditional vishing exercises for decades.

In this proof of concept, the impersonated individual's phone number was spoofed with their consent. Like traditional vishing, spoofing is used to exploit the trust the victim will have when they are presented with a familiar caller ID on their device. Alternatively, the audio signal could also be routed directly to applications like Microsoft Teams or Google Meets, thus allowing the use of AI-impersonated voices in popular messaging applications during a voice call in real time.

[5] Generative Adversarial Networks: https://en.wikipedia.org/wiki/Generative_adversarial_network

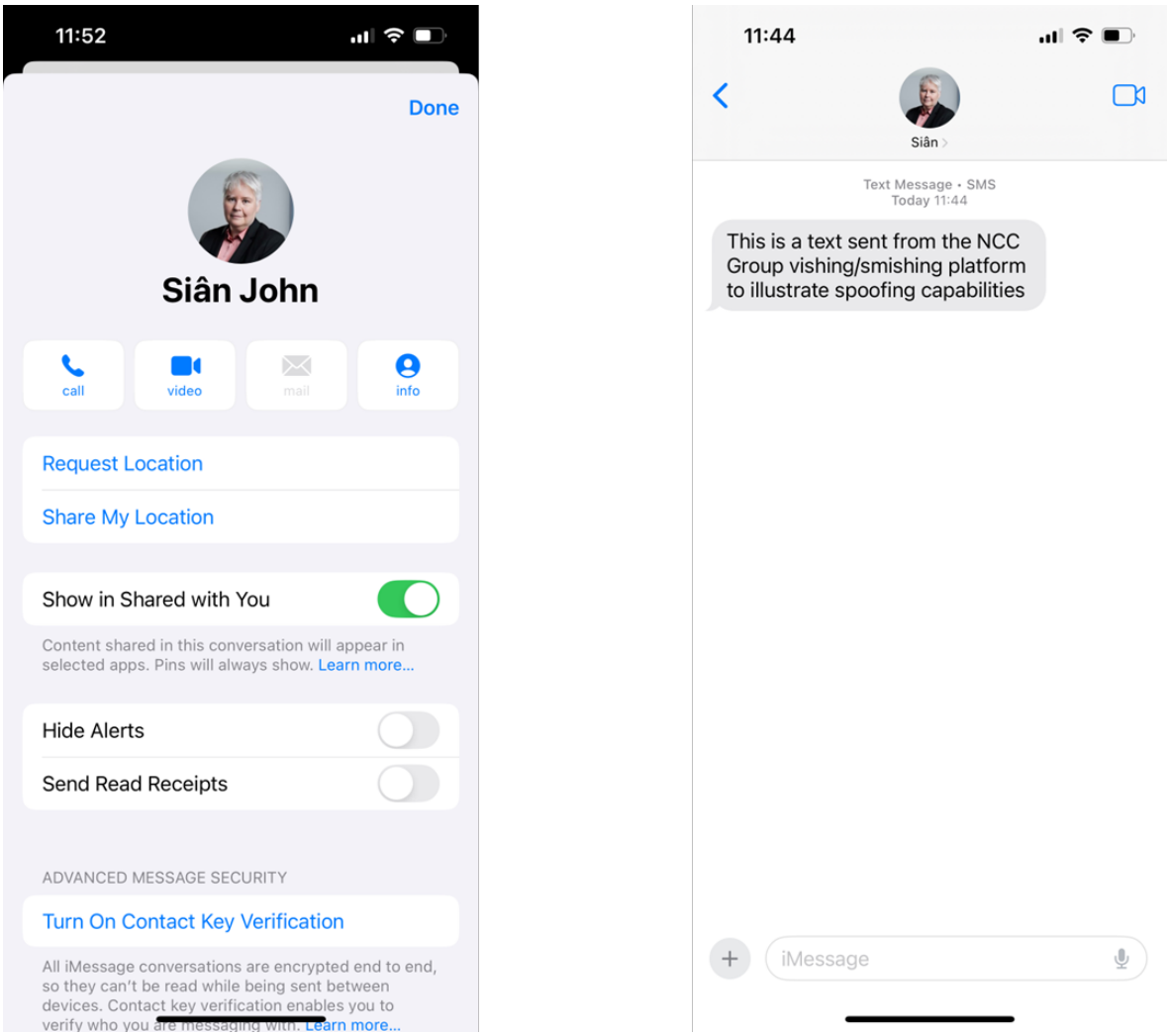

Figure 2 – Spoofed caller ID demo.

For this blog post, we cloned the voice of Siân John, who kindly agreed to lend her voice for this exercise. The samples were obtained from public sources and processed in a digital audio workstation. You can hear the resulting audio clip at:

https://www.nccgroup.com/media/5wohajk3/sian_df.wav

Or linked from our social engineering prevention page at:

https://www.nccgroup.com/technical-assurance/social-engineering-prevention/

Social Engineering

After an iterative process of audio processing, the model was optimised, and the dataset was ready. We now felt prepared to try the solution on a live client project and so we started working with one of our social engineering colleagues, a vishing specialist. Their responsibilities ranged from studying and analysing the prosody (that is, the rhythms and patterns of speech) of the person to be impersonated, reviewing and understanding the internal processes of the company to

be attacked and generating scripts and pretexts that could be used to keep the conversation on track during the attack.

Once our vishing specialist felt comfortable with both the materials and with using the real-time voice converter, the audio signal was plugged into the caller ID spoofing program to start the exercise.

Conclusions

NCC Group has already delivered many jobs which successfully used these deepfake vishing techniques. And by 'successful', we mean that we have run practical versions of the attacks described above, against real organisations, and these attacks recovered sensitive and confidential information. Not only that, but we have also shown how these techniques can convince people in positions of *key* operational responsibility to carry out actions on behalf of the attacker. In security assessments that simulated real-world attack conditions, we have been able to carry out actions such as email address changes, password resets, and so on.

The upward trend in the demand for this type of security exercise shows no signs of abating. Which is no surprise, given that the tools and resources required to create AI deepfake voice clones continue to get cheaper and easier to access.

Limitations and next steps

Having achieved the stated aims for this project, the next step would be to investigate video deepfakes, trying to replicate objectives analogous to the voice cloning. Preliminary investigations so far have mainly succeeded only in identifying new limitations and problems in this field. For example, there are difficulties in synchronising the modified audio and video signals. At the time of writing, a real-time solution or model that can offer video and voice cloning (and that mirrors the quality and operational criteria achieved by NCC Group in the audio only field) has not yet been found. However, given the unprecedented speed at which this technology is moving forward, it is expected that a deepfake solution which realistically simulates both audio and video is feasible, it is just a matter of time.