WGUiSW

OxO1 Say Hello to Security - Al & Voice Cloning

Krystian Bajno, 2024



[Table of Contents]

0x01 Say Hello To Security - AI & Voice Cloning

0x02 Whoami

0x03 Links section

0x04 Exploring online

0x05 Exploring the world of self-hosted

<u>0x06 Anatomy of a deepfake</u>

0x07 Practical steps to perform voice cloning

0x08 SO VITS SVC

<u>0x09 RVC</u>

0x0A Threat modelling

<u>0x0B What are possible Al uses – examples</u>

OxOC Let us create a song!

<u>0x0D Falling into a rabbit hole</u>

<u>0x0E Q&A</u>

0x0F Thank you

[Links are clickable]

OxO2 Whoami

Krystian Bajno

Cyber Security Specialist
Penetration Tester

Full-Stack Software Engineer **Backend, Frontend, Mobile**



Comp. Sci. I - Cloud Computing Technology

Comp. Sci. II - Cloud Computing Architecture and Security





WGUiSW _baycode.eu

0x03 Links section

Useful projects

https://github.com/voicepaw/so-vits-svc-fork https://github.com/WadRex/RVCompact https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI https://github.com/litagin02/rvc-tts-webui https://github.com/w-okada/voice-changer https://github.com/facebookresearch/demucs https://github.com/AUTOMATIC1111/stable-diffusion-webui https://github.com/comfyanonymous/ComfyUI https://github.com/Illyasviel/Fooocus https://github.com/s0md3v/roop https://github.com/guovww/AnimateDiff https://github.com/Illyasviel/ControlNet https://github.com/facebookresearch/llama https://huggingface.co/microsoft/phi-2 https://github.com/henrymaas/AudioSlicer https://github.com/flutydeer/audio-slicer https://github.com/cogui-ai/tts

Model repositories

https://huggingface.co https://civitai.com

Useful links

https://machine-learning.paperspace.com/wiki/machine-learning-models-explained https://www.hardware-corner.net/guides/computer-to-run-llama-ai-model/ https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-smalllanguage-models/ https://colab.google/ https://paperswithcode.com/task/speech-synthesis/ https://paperswithcode.com/ https://openart.ai/workflows

Recommended AI YouTube channels

https://www.youtube.com/@Fireship https://www.youtube.com/@NerdyRodent https://www.youtube.com/@sentdex https://www.youtube.com/@sedetweiler https://www.youtube.com/@houseofdim https://www.youtube.com/@sebastiankamph https://www.youtube.com/@enigmatic_e https://www.youtube.com/@OlivioSarikas https://www.youtube.com/@EndangeredAl

OxO4 Exploring online

The 2022/3 rise of Generative AI (GAN, LLM)

$\mathbf{\dot{\succeq}}$ Selected speech synthesis providers

ElevenLabs

https://elevenlabs.io

- Polish company! 💻
- Text to Speech
- Speech to Speech
- Dubbing (video translator)
- Make custom models (paid)

Pros:

- 1. Ease of use, cutting edge
- 2. No tech knowledge requirement
- 3. No hardware requirement
- 4. Support, big funding
- 5. API integrations

HeyGen

https://www.heygen.com

- Lip sync!
- Creating whole digital avatars
- Translating videos
- Speech to Speech

Cons:

- **1. Commercial,** limited free plans, pricy tier limitations, credit systems
- 2. Potential political correctness limitations
- **3.** No real-time voice streaming (at the current moment, all it takes is WebRTC)
- 4. Data collection

Voice

https://www.resemble.ai https://www.heygen.com https://elevenlabs.io

Imagery

https://www.bing.com/search?q=Bing+Al&showconv=1 https://openai.com/dall-e-3 https://www.midjourney.com/ https://www.krea.ai/home

LLM

https://www.bing.com/search?q=Bing+Al&showconv=1 https://chat.openai.com/ https://bard.google.com/chat

> NVIDIA Microsoft Google Meta Apple Open Source Community

The biggest players in **AI** as of 2024.

OxO5 Exploring the world of Self-Hosted

Let's do it offline and open source

Requirements

- For training, voice infering, images generation, and smaller LLMs (such as phi-2 and LLAMA-7B, approx 10B parameters), 8 GB of GPU VRAM is sufficient (eg. RTX 3060, or Apple M1 will do due to chip mixing RAM and VRAM, but M1 works 9x slower than RTX). For bigger LLM's, such as LLAMA-70B - 52 GB of GPU VRAM is a bare minimum.
- 2. Lots of free storage.

Alternatively deploy cloud environment for training

Deploy model training on selected **Cloud Computing Platform**, train it in minutes, download the checkpoints, and kill the instance, just like spinning up and destroying a hash cracking machine. You can also use **Google Colab**. Bigger LLM's however require that much resources to **run**. It does not mean you need bigger LLMs to get good results.

Pros:

- 1. Free
- 2. Huge model base
- 3. Huge amount of software
- 4. Full control
- 5. Limitless possibilities
- 6. Cutting edge

Cons:

- 1. Requires a bit of tech knowledge
- 2. Requires good GPU
- 3. Only community support

Voice

https://github.com/voicepaw/so-vits-svc-fork https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI https://github.com/litagin02/rvc-tts-webui https://github.com/w-okada/voice-changer https://github.com/facebookresearch/demucs https://github.com/flutydeer/audio-slicer https://github.com/coqui-ai/tts

WGUiSW

_baycode.eu

Imagery

https://github.com/comfyanonymous/ComfyUI (strongly recommended) https://github.com/AUTOMATIC1111/stable-diffusion-webui https://github.com/Illyasviel/Fooocus (love the simplicity) https://github.com/s0md3v/roop

LLM

https://github.com/facebookresearch/llama https://huggingface.co/microsoft/phi-2

Model repositories

https://huggingface.co https://civitai.com/ **Useful links**

https://paperswithcode.com/



OxO6 Anatomy of a deepfake

Unsupervised deep-learning nature of Generative Adversarial Networks

Picture:

- 1. Find **one**, good picture of the impersonated person.
- 2. Substitute the face on a selected picture with impersonated person face (eg. using **Roop**, **Fooocus**, **ComfyUI**), or generate a completely fake image with substituted face.
- 3. Upscale the picture (for example using **Stable Diffusion**, **R-ESRGAN** upscaler).

Video:

- 1. Find a picture of the impersonated person (one is all it takes).
- 2. Substitute the face on all frames of the video, or generate completely fake video using **ComfyUI.**
- 3. Upscale all frames of the video.
- 4. Voice clone the impersonated person and substitute your own voice in the video with a clone.

Mind your ethics. Everything is possible, but what for? Ask the person you clone for approval.

OxO7 Practical steps to perform voice cloning Make me sound like you OVERFITTING:

- **1.** Acquire a 10 or more minutes sample of the impersonated person voice.
- 2. Cut the sample into approx. 5 second pieces and remove noise, silence (using AudioSlicer, audacity), this sample will be later used for GAN unsupervised model training and **no data labeling** is needed.
- **3.** Train the model using RVC or SO-VITS-SVC for 200 epochs or more. (Can take a few hours). Increasing epochs too much will not improve the quality due to *overfitting*.
- 4. Use the created model for inference or real-time inference streaming.

The **RVC** and **SVC** models are not interchangeable and need to be trained separately.

Make sure to create a **loopback interface** or use a **mixer** in order to avoid audio feedback while streaming in **real-time.**

Streaming voice is easier than a complete deepfake video real-time streaming due to resources extensive use. Expect the voice streaming lag to be 0.3 of a second, whereas processing video takes way more time. In order to stream-infer voice + video, you will need to use cloud computing platform or a powerful computer. It is needed to match the audio/video latencies, yet it is perfectly doable in 2024.

> Mind your ethics. Everything is possible, but what for? Ask the person you clone for approval.



Ox08 SO VITS SVC Step by step

Activate Python environment

.\venv\Scripts\Activate.ps1 - Windows
source ./venv/Scripts/activate - Linux, MacOS

1. Place dataset files created with AudioSlicer into

dataset_raw/<speaker_id>/**/<files.wav>

2. Pre-process the dataset

svc pre-resample - converts your audio to mono 44.1khz files svc pre-config - downloads a few configuration files and puts them in the correct directory.

3. Modify number of epochs and batch size

in logs/44k/config.json to match **VRAM** or training will crash.

4. Continue pre-processing to optimize to "crepe" prediction method

svc pre-hubert -fm crepe - downloads and runs a speech model
pre-training.

5. Train the model (this could take a few hours)

Model checkpoints will be available in logs/44k - G_x.pth, D_x.pth svc train -t

6. Use the model

svc gui

WGUISW _baycode.eu

Usage:

Select the Model path, Config path, and the Speaker.

- 1. To infer file to file, select file input audio path, output audio path and press Infer.
- 2. To **change your voice in real-time** select input and output device, disable auto pitch prediction, adjust Pitch to your voice, press (Re)/Start Voice Changer, and use the output device as input in selected application (eg. Discord, Teams,

Telegram). Make sure to use a loopback interface and a set of headphones to prevent audio feedback.

0.000000											
Paths		50 - 11		File	-					a 1 1	
Model path	E;/mis/so-vits-svc-fork/logs/44k/G_1	58.pth	Browse	Input audio path	-]	Browse	Browse(Fold	per) Play
Config path	E:/mls/so-vits-svc-fork/configs/44k/ci	onfig.json	Browse	Output audio path							Save As
Cluster model path (Optional)			Browse	Auto play							
Common				Realtime							
Speaker				Crossfade seconds					0.	.050	
Silence threshold		-35.0		Block seconds						0.350	
Pitch (12 = 1 octave) ADJUST THIS based on your vo when Auto predict F0 is turned	pice I off.	-36 -24 -12 0	12	Additional Infer sec	conds (before)				0.	150	
Auto predict F0 (Pitch may	become unstable when turned on in rea	al-time inference.)		Additional Infer sec	onds (after)				0.1	100	
F0 prediction method		crepe	-	Realtime algorithm					1 (D	ivide constant	:ly) 두
Cluster infer ratio		0.00		Input device		Line (M	G-XU) (MME)				-
		0.40		Output device		Line (M	G-XU) (MME)				
Noise scale		0,40		Passthrough or	iginal audio (for	r latency ch	eck)			Refr	esh devices
Pad seconds		0.10		Notes-	ince:						
Chunk seconds		0.50		- Setting F0 pre - Auto Predict F If the audio sound	diction method 0 must be turne Is mumbly and c	to 'crepe' r ed off. choppy:	nay cause performa	ance degrada	tion.		
Max chunk seconds (set lower	if Out Of Memory, 0 to disable)	40		Case: Mic input	is low (Decreas	se Silence t	hreshold)	LK SELUIIUS)			
Absolute threshold (ignore	d (True) in realtime inference)			Presets							
				Presets		<u> </u>	efault VC (GPU, GT	TX 1060)			elete preset
				Preset name					Add cu	rrent settings	as a preset
				✔ Use GPU							
				Infer (Re)Start Vo	ice Changer	Stop Voice	Changer				

Ox09 RVC Step by step

Activate Python environment and launch the web-ui.

.\venv\Scripts\Activate.ps1 - Windows
source ./venv/Scripts/activate - Linux, MacOS
Web interface will be available at <u>http://localhost:7897</u>

1. Name the experiment (project)

2a. Load and pre-process the dataset in the user interface RVC can split and denoise audio, but you can specify a directory with pre-processed files.

2b. Extract the pitch and perform HuBERT pre-training technique

1. Training view

Step 11 m to experimental configuration. Experimental adda is stored in the logit foodor, who each experiment having a separate folder. Manually enter the experiment name path, which contains the experimental configuration files.					Model Inference Vocals/Accompaniment Separation & Reverberation Removal Train ckpt Processing Ex	port Onnx FAQ (Frequently Asked Questions)	😢 RVC - GUI				
Enter the experiment name: Tar mi-test	arget sample rate: • 40k 48k	Whether the model has pitch guidance (required for singing, optional for speech): Version true false	• v2	Number of CPU proce for pitch extraction an processing:	Inferencing voice: Refresh voik GADZIUMODEL.pth Uniload voice	to save GPU memory:	Load model E:/mls/wc/RVC/assets/weights/GADZIUM Select the .pth file				
Step 2a: Automatically traverse all files in the training folder that can be decoded into audio and perform slice normalization. Generates 2 way folders in the experiment directory. Currently, only single-single-speaker training is su				singer/speaker training is su	华次推理 批量推理		Filmlan (DV/Compart(DV/Class/and/integrated)				
Enter the path of the training folder: Please specify the 0 speaker/singer ID: Process data		Output information		Transpose (integer, number of semitones, raise by an octave: 12, lower by an octave: -12): Resample the output audio in post-processing to the final sample rate. Set to 0 for no resampling:		Lamianvarke compacting organged dumanization of the second states and states					
Level 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1					0		Audio device (please use the same type of driver)				
Step 2ts: Use CPU to extract pitch (if the model has pitch), use GPU to extract features (select GPU index):					Enter the path of the audio file to be processed (default is the correct format example):	Adjust the volume envelope scaling, Closer to 0, the more it mimicks the volume of the original voca Can help mask noise and make volume sound more natural when set relatively low. Closer to 1 will b	Input device Line (MG-XU) (MME)				
Enter the GPU index[es] separated by ¹ , e.g., 0-1-2 to use GPU 0, 1, and 2:	 ・	取算法输入或声可用pm提進流活量语音 用的逻辑。harves或量量好化增,mvpe效 gropuseu		Output information	C:\Users\Desktoplaudio_example.wav	more of a consistently loud volume:	Output device Yamaha Steinberg USB ASIO (ASIO)				
0	pm harvest	dio			Path to the feature index file. Leave blank to use the selected result from the dropdown:	Protect voiceless consonants and breath sounds to prevent artifacts such as tearing in electronic mu Set to 0.5 to disable. Decrease the value to increase protection, but it may reduce indexing accuracy	Reload device list				
GPU Information	💿 rmvpe 💽 rmvpe_gpu	Feature extra	tion		C:\Users\Desktop\model_example.index		Caparal asttings	Performance acttings			
0 NVIDIA GeForce RTX 3060 Laptop GPU	Enter the GPU index(es) separated use 2 processes in GPU0 and 1 proc	y '-', e.g. , 0-0-1 to sss in GPU1			Auto-detect index path and select from the dropdown:	If >=3: apply median filtering to the harvested pitch results. The value represents the filter radius an reduce breathiness.	General settings 45 Associate threshold	Sample length			
Step 3 Fill in the training settings and data training the model and index Same frequency 5 Tecal training rots 20 Batch size per GPU: 3 Same only the latest" cleref file to same Cache all training sets to GPU Same a small				U Save a small f	Select the pitch extraction algorithm ("pm" faster extraction but lower-quality speech; "harvest": better bass but extremely slow; "crepe: better quality but GPU intensive), "mmpe": best quality, and little GPU requirement	Search feature ratio (controls accent strength, too high has artifacting):	Pitch settings	Number of CPU processes used for harvest pitch algorithm			
Lana, veny, goodi; inau, goodi; inau, goodi; inau alaose i province da ana aporte: inau alaose i province da aporte: i province da a			than 10 minutes) can speed training, but caching large d will consume a lot of GPU m	up atasets emory	pm harvest crepe e mvpe		Index Rate	Fade length 0.15			
			ipeen	c	onvert	0.00	246				
			🔘 Yes 💽 No		Output information	# Export audio (click on the three dots in the lower right corner to download)	loudness factor	Extra inference time			
Lead pre-trained bias model (c part: asset)pretrained, v2, 856-66, pth Lead pre-trained bias model (c) part:				Output informat		69	pitch detection algorithm C pm C harvest C crepe @ rmvpe	☐ Input noise reduction ☐ Output noise reduction			
						Start audio conversion Stop audio conversion C Input voice monitor C Output converted voice Algorithmic delays(ms): 0 Inference time					
assets/pretrained_v2/f0D40k.pth	Train model	Train feature index	One-click training								

2. Inference view

WGUISW _baycode.eu

-

3. Train the model and create feature index

a) One-Click training (Recommended)

This will perform step 2b, model training, and feature index creation **automatically**.

b) Train model button - train the model manually after step 2b.

c) Train feature index button - create feature index manually.

Model checkpoints and index will be saved in logs directory. If you want to share the model, you'll need to extract and compile the .pth checkpoint from logs directory using **ckpt processing** tab, export should be about **60 MBs** in size. The model that is ready to use should reside in assets/weights directory.

4. Use the model

There are two user interfaces available

a) For **inference** stay in the web-ui

b) For **real-time streaming** launch gui_v1.py

3. Real-Time inference view

OxOA Threat modelling

The wide emerging threat of 2024

Due to making AI accessible to wide-public, the new threats have emerged.

What/Who could be the possible target?

- Anyone who gets his voice cloned
- Anyone who gets his face cloned
- Applications with **bio-authentication** (voice)
- Publicly speaking people
- YouTubers
- Influencers
- Politicians
- The internet

What is the threat?

- AI phishing, vishing, impersonation, identity theft
- Bypassing **bio-authentication** (voice)
- Blackmailing
- Disinformation and increase of trolling
- Automated false identity **scamming** bots, **catfishing**
- Fake, bot generated **influencers**
- Increase of AI generated **explicit** pictures
- Abuse of people voice and **copyrights**
- Overreliance on Artificial Intelligence could have potential

negative impact on cognitive abilities of specialists.

- Flood of LLM generated **spam** content over the internet
- Growing concerns about the **reliability** of digital data due to **AI hallucinations** and **impersonation. Anything digital can be generated.**



OxOB What are possible AI uses - examples

Do good, not evil

Automation

Creating bots that auto-respond to clients, leveraging LLMs, text-to-speech, and GAN networks for voice inference.

Parsing big sets of data

LLMs can parse big sets of data in order to give useful insights, or even provide code solutions.

Security Assessments

Useful in social-engineering, red-teaming, and comprehensive phishing assessments.

Art creation

Creating art - pictures, animations, songs.

Breaking the language barrier

Translating, dubbing the video/audio into various languages in order to reach wider audience.

Sock Puppets

Creating sock-puppet false identities in order to achieve anonymity and infiltrate criminal communities is easier than ever.



OxOC Let us create a song!

Let's do it now

- 1. Clone the voice of a friend using RVC or SO-VITS.
- 2. **Download** the song to exchange the voice in.
- 3. Demux the singer, bass, drums, guitars and pianos on origin using Demucs.
- 4. Infer the singer voice using the cloned model.
- 5. Reassemble the song in audacity or any other DAW.
- 6. Enjoy making your friend a rockstar.



OxOD Falling into a rabbit hole

Win a bunny

- 1. What cipher is unbreakable provided it is used correctly and only once, with the key length of plaintext?
- 2. What is steganography?
- **3. What web vulnerability** is number one in OWASP 2021?
- 4. **Imagine** there is a banking trojan that replaces the addresses during transfers in client browser memory. How can you secure the *client* web application?



OxOE Q&A

Ask me anything you want



OxOF Thank you

Presentation in PDF format is available on https://news.baycode.eu

Meet me again at https://wguisw.org

[BEC Poland]

WGUiSW | Jan 9th, 2024

[Copyright © 2024 Krystian Bajno]