

Making AI Song Covers with RVC

- Google Colab or Local Install

These are the two main options for making AI song covers.

You can run RVC on your computer if you have a PC with a decent **NVIDIA graphics card** (GPU), or you can run it **for free** through the **Google Colab** web page.

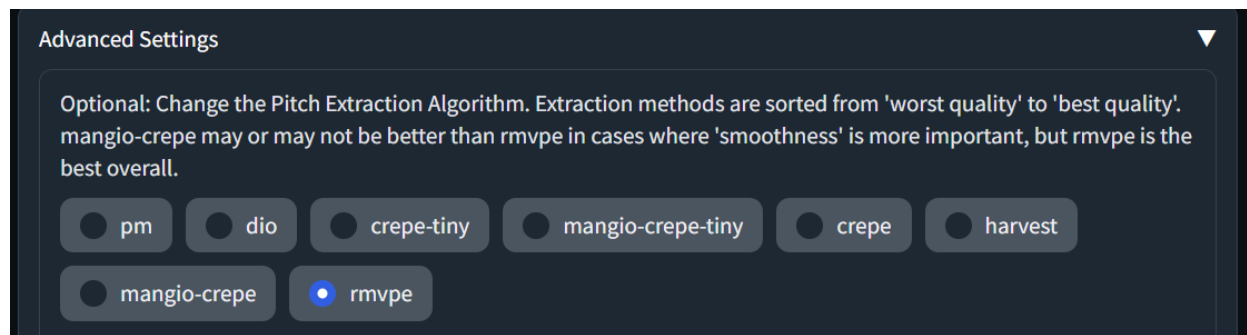
- Running Google Colab

This is the recommended Google Colab for using voice models:

https://colab.research.google.com/drive/1Gi6UTf2gicndUW_tVheVhTXIIYpFTYc7?usp=sharing

After enough time, Google limits your GPU usage and you have to wait to use the GPU again.

This will slow down your conversion speeds, but it will still be usable as long as you use 'rmvpe' mode (considered to be the general best mode, tied with mangio-crepe). ~3 minute song took 9 minutes for me without the GPU.



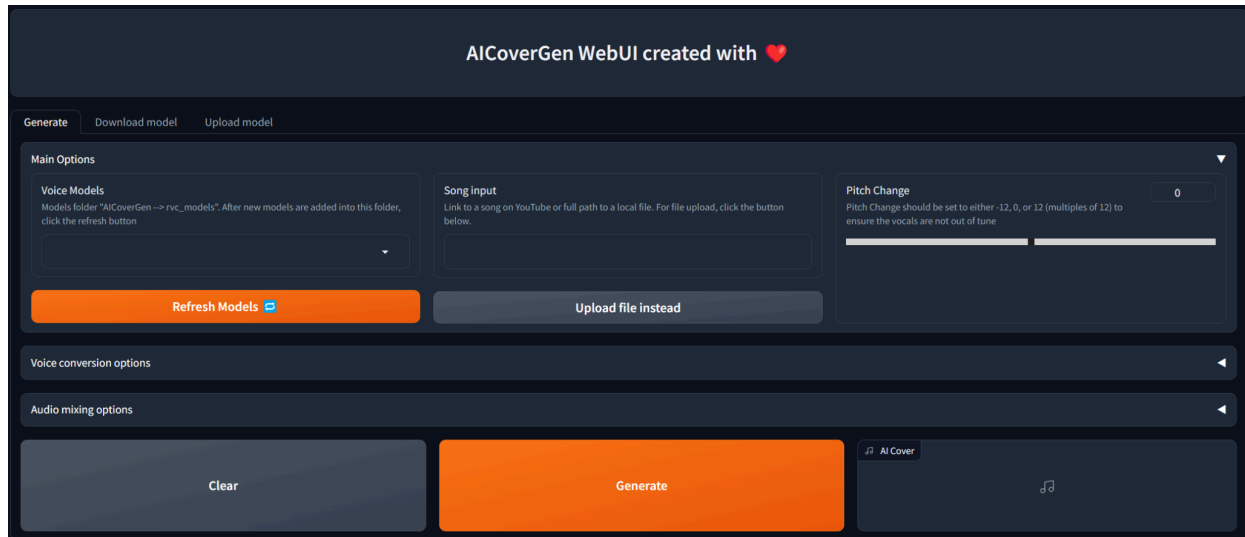
Some people make alternate Google accounts to get around the GPU limits, or they pay for Colab Pro. Most commonly happens for people training their own voices since that requires a lot of GPU power.

- Running Locally

Check out this local install guide:

<https://docs.google.com/document/d/1KKKE7hoyGXMw-Lq0JWx16R8xz3OfxADjwEYJTqzDO1k/edit?usp=sharing>

(NEW) Easy WebUI for AI Cover Generation



Tired of isolating vocals/instrumentals, conducting RVC inference for voice conversion and doing audio mixing manually? This WebUI does all of that for you automatically in a single click! All you have to do is download an already trained RVC v2 Voice model from a huggingface/pixeldrain link, and provide a link to a song on YouTube. The WebUI will take care of the rest! You can even make finer adjustments for RVC voice conversion, such as index rate, filter radius, rms mix rate, protect... and audio mixing options such as volume of vocals/instrumentals, or reverb settings.

Setup guide + Showcase on YouTube: <https://www.youtube.com/watch?v=pdlhk4vVHQk>

Local/Colab Install instructions in README.md:

<https://github.com/SociallyIneptWeeb/AICoverGen>

Google Colab:

https://colab.research.google.com/github/SociallyIneptWeeb/AICoverGen/blob/main/AICoverGen_colab.ipynb

This is under constant development, with new features coming soon!

- Local audio file browse button instead of copying and pasting full file path (DONE)
- Upload of locally trained RVC voice models (DONE)
- Faster Colab requirements installation via zip file

- Preparing Song Acapellas


See [this section](#) for more information on making song vocal isolations.

- Where to find voices?

AI Hub discord server, in #voice-models there is a large collection of different voices that you can search from:

<https://discord.gg/aihub>

Or, you can check out my RVC archive sheet, which automatically tracks colab download stats so you can see which voice models are the most popular. (Kanye, Dio, Weeknd, Mr Krabs, Gura, Jschlatt, etc...)

 RVC Archive Sheet

<https://huggingface.co/QuickWick/Music-AI-Voices/tree/main>

(AI Hub backup, has not been updated in a long time)

- How to actually use RVC?

 How to make AI covers (RVC, with crepe)

This guide ^ covers how to use voice models, what the settings do, and how to properly mix them later into full covers (using basic [Audacity](#) settings.)

Keep in mind the official Crepe seems to underperform compared to **Mangio-Crepe** or **RMVPE**, but you can still try it if you want. I recommend **rmvpe as a general option**, or mangio-crepe for 'smoother' results (but less pitch accuracy), generally.

The hop size option doesn't work for official Crepe. Also, I've heard poor results from training with the official Crepe option so far.

This is a mini guide explaining UVR vocal isolation in a bit more detail.

<https://youtu.be/ITNeuOarHHw>

Vocal Isolation

- Isolating acapellas

First, find your source material you want to use a voice model on. Preferably you get this in the highest quality possible (.flac preferred over mp3s or YouTube rips, but lower quality stuff will still function).

In order to isolate vocals from music you will need to use one of the following:

- [UltimateVocalRemover](#) (can be ran locally on good PCs or within the RVC Google Colab pages at the end). Kim vocal 1 or Inst HQ 1 is the best 'general' vocal model, Kim vocal 2 will sometimes isolate non-vocals but it can sound

better overall (you can run it and then a karaoke model after it to deal with this.. sometimes). Small UVR video explanation: <https://youtu.be/ITNeuOarHHw>

- MVSEP.com (totally free web app, but the queue can be long. I've been told MDX B is the general best option for vocal isolation here, but haven't used it myself).
- Vocalremover.org or X-minus.pro; these are not as high quality options but will get the job done quickly. Vocalremover.org has no option to remove reverb, and IIRC X-minus.pro doesn't either.

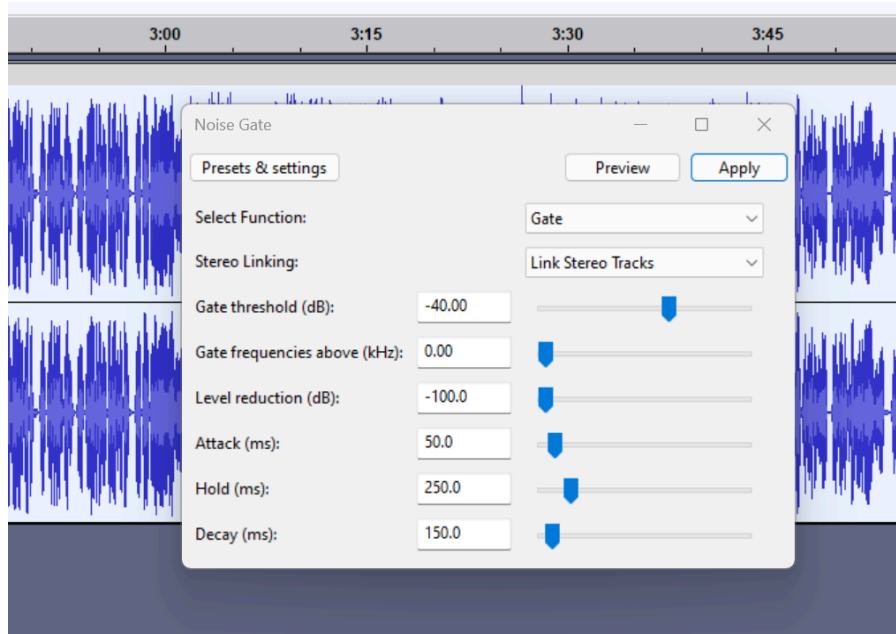
- Removing reverb / echo

It is necessary to remove reverb / echo from the song for the best results. Ideally you have as little there as possible in your original song in the first place, and isolating reverb can obviously reduce the quality of the vocal. But if you need to do this, under MDX-Net you can find Reverb HQ, which will export the reverbless audio as the 'No Other' option. Oftentimes, this isn't enough. If that did nothing, (or just didn't do enough), you can try to process the vocal output through the VR Architecture models in UVR to remove the echo and reverb that remains using De-Echo-DeReverb. If that still wasn't enough, somehow, you can use the De-Echo normal model on the output, which is the most aggressive echo removal model of them all.

There's also a [colab for the VR Arch models](#) if you don't want to run or can't run UVR locally. No clue how to use it though so *good luck*. The main RVC colab also has UVR's MDX-Net models (so Kim vocal) at the end. Without a good GPU on your PC, UVR will still run locally in most cases, but it will be quite slow, if you're okay with that. But if you have a super long dataset, be prepared to have it running overnight...

- Noise gating to remove silence

I like to noise gate my stuff in Audacity to remove noise at the super quiet parts of the audio. Usually -40db is a good threshold for this.



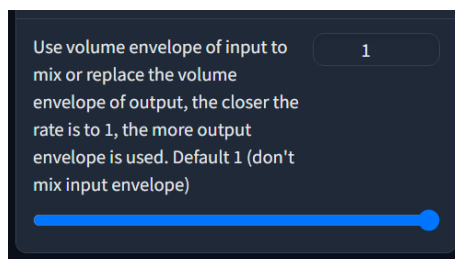
Adobe Audition probably has more advanced tools to do this automatically (idk how to use it), but this is a good preset to start off with for people using basic Audacity mixing. If it cuts off mid sentence, redo it with it turned up for the Hold ms. Maybe even turn down the gate threshold to -35db or lower if necessary.

- Isolating background harmonies / vocal doubling

In some cases, these are too hard to isolate without it sounding poor quality. But if you want to try anyways, the best UVR models for doing so would be 5HP Karaoke (VR Architecture model) or Karaoke 2 (MDX-Net). 6HP is supposed (?) to be a more aggressive 5HP I think? Dunno. YMMV so try out the other karaoke options unless it literally just isn't working no matter what.

Advanced Conversion Tips

Vocal Conversion Options, Explained:



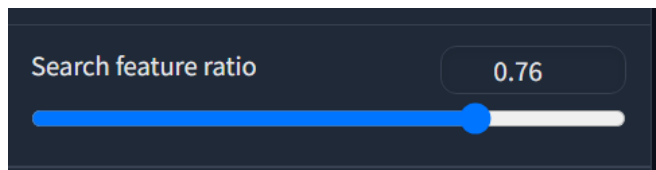
The lower you set this to, the more it will capture the original volume range of the original song.

A value of 1.0 will be equally loud throughout the whole conversion; 0 will make it mimic the volume range of the original as much as possible.

I would recommend you set this volume setting to a decently low value such as **0.25 or 0.2**.

Transpose / Pitch setting:

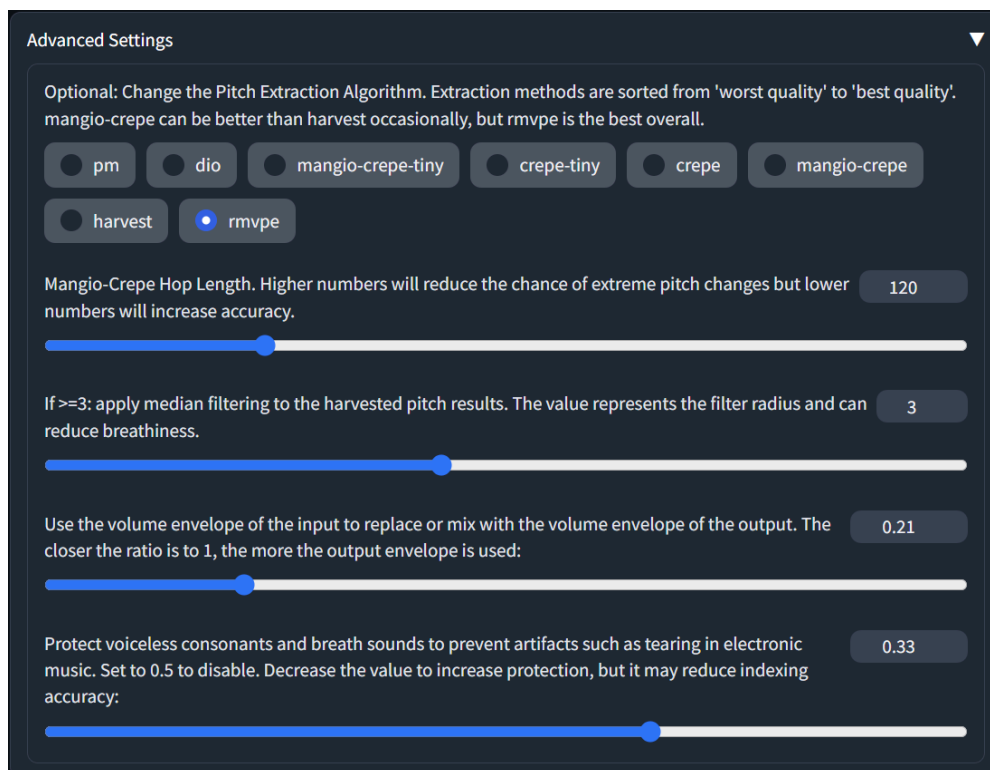
-12 or 12 will both stay in the original key but at different octaves. Good for extreme differences, like a male singer doing a song sang by a female. For more subtle differences, -5 and -7 are the least dissonant settings, as they represent perfect fourth and fifths respectively, but they may still feel 'off'.



This value ^ controls how much influence the .index file has on the voice model's output. (The index controls mainly the 'accent' for the model.)

If your model's dataset isn't very long or it's not very high quality, (or both), this should be lower, and if it's a high quality model, you can afford to go a bit higher.

Generally, my recommended value would be **0.6-0.75**, and reduce it if you think it's truly necessary.



The options on top here are for the pitch detection methods. The best option here is *generally* **rmvpe**, and I would recommend that, or **mangio-crepe** with different hop sizes between 64-192 for most cases. Mangio-crepe tends to be better for 'smoothed out' results, but higher hop length values will lead to less pitch precision. You also need a GPU for it to convert reasonably fast.

It seems like mangio-crepe is best for when you want 'smoother' vocals (which is most singing and some rapping), and rmvpe is better for when you need more 'raspiness' or 'clearness' in the vocal, e.g fast rapping (think [Andre 3k/JID/Eminem](#)).

You should experiment to see which sounds best for your specific song if you're unsure, but I would say mangio's crepe is still the best generally.

Harvest is a slower, 'worse' version of rmvpe. It might visually 'error out' on the colab, but it will eventually finish and the wav will be in the TEMP folder despite the visual error; keep that in mind.

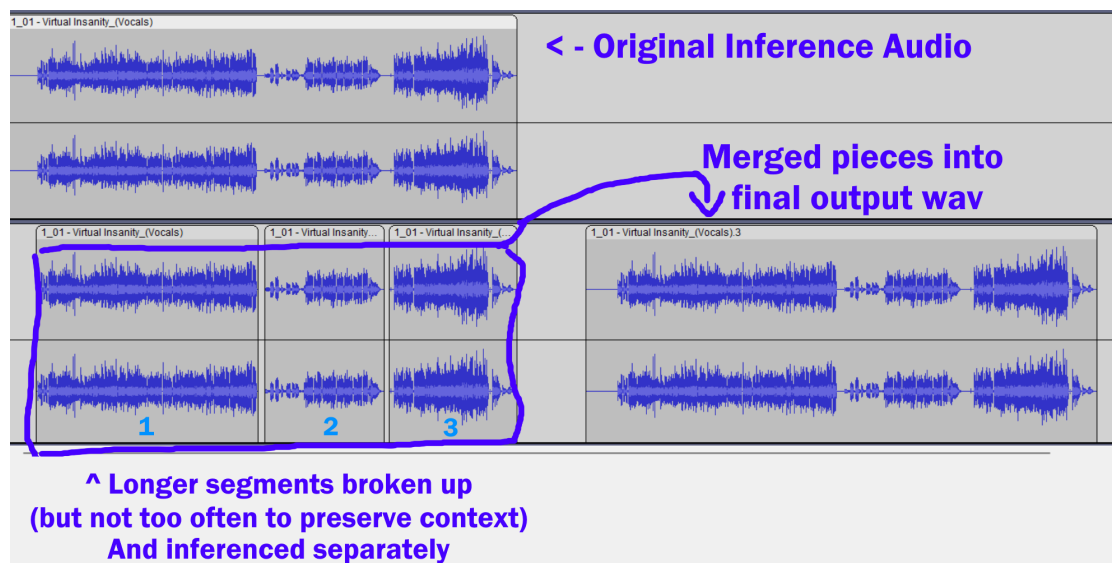
Crepe hop length controls how often it checks for pitch changes in milliseconds when using crepe specifically.

The higher the value, the faster the conversion and less risk of voice cracks, but there is less pitch accuracy.

The default value is 128, so that means it checks roughly 8 times a second for pitch changes. Anything lower than 64 is almost always pointless from my tests. **Start with 128**, and lower if you think you need to. Heighten it if you think being less pitch accurate might help the end result sound 'smoother' (yes, that can happen; I've noticed ~160-200 can help in some cases, some people prefer 192)

Crepe-tiny is just a faster, but worse sounding version of crepe.

Vocal Segmenting



If you have long silence periods, or a long song, exporting separately based on vocal pieces and then manually merging them later can help improve the output result, but it can be somewhat subtle of an improvement (depends on amount of silence). You can easily segment in Audacity with CTRL+I and then perfectly align each piece later.

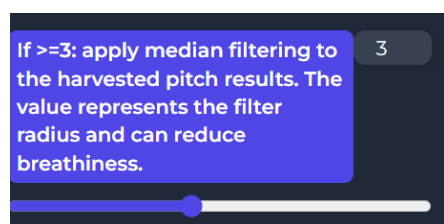
<https://cdn.discordapp.com/attachments/1089076875999072302/1117471825614618815/test-sample.mp4>

^ How including long periods of silence impacted quality negatively

<https://cdn.discordapp.com/attachments/1089076875999072302/1117472101583040662/test-sample-2.mp4>

^ How doing the pieces separately in large chunks helped quality (I added 1 min silence to test this theory)

This is the only way to properly do harvest conversions on a colab without running into visual gradio errors. But it helps both mangio-crepe and harvest of course.



Higher = more 'blurred', or smoothed out outputs. Might help slight cracking issues, but potentially makes the pronunciation worse.

- Training Guide

☰ Training RVC v2 models Guide (by kalomaze)

Consider subscribing to my Patreon!

Benefits include:

- Full on tech support for AI covers in general, including mixing and how to train your own models, with any tier, but priority given to the latter tier.

<https://patreon.com/kalomaze>

Your support would be greatly appreciated!