Memories AI/ Resurrection AI

(Just temporary names)

Introduction:

The ache in your heart when you long to hear the voice of someone you've lost, or no more in your life is pain that many individuals go through in life at some point. The sound of a loved one has a profound healing power that matches none other.

To resurrect a voice, there are multiple available tools in contemporary advancements, however, unless the user is technically savvy and is upto date on the AI advancements its a nightmare to try to get a voice to life, all the way in the mounting sadness of not able to do it properly.

Firstly, the user would need a clean audio file of the missing person. The user might have an old video, with tons of background noise and others speaking in the video. The user must isolate, amplify and clean the audio before trying to use that voice for cloning. Once, this step is done, he must know all the latest AI advancements to understand what he could leverage to clone this voice.

Furthermore, once a voice is cloned, users naturally expect to engage in natural conversations, receive heartfelt birthday wishes, or wake up to a comforting good morning message. However, currently, there are no services available to facilitate these essential aspects of bringing a cloned voice to life in such a natural and personalized manner.

With all these issues in mind, we intend to create a well-defined, easy-to-use software to clone a person’s voice responsibly and use it.

Idea Validation:

1. There are many Quora posts, medium articles etc explaining their story on how they dealth with the loss of their loved ones. And in most cases what I observed is when they were in pain they mentioned they called their loved ones phone just to hear their voice mail message.

2. Reddit Link - This link explains how a person used various softwares to clone their dead fathers voice. But the comments below and having 350+ likes for the post kind of validates this idea might bring value to people looking for something similar.
3. My colleague at the company recently was enquiring with us (Data Scientists) if there was a tool to clone the voice of her dead father 10 years ago, which Ihad to say no, but kind of proves my idea on how this service app could provide a huge sentimental value to them.

Technical Process:

(Just Backend - Not Detailing User Management + Front End in this section)

Video or Audio file as input:

Firstly, The user may have an audio file of maybe a recorded phone conversation / or a recorded audio of them from years ago. This audio file could be of anytype WAV/ WMA/ MP3/ MPEG etc. We need to have an adapter to receive any kind of audio file.
This also could be a video of any sort, from which we would have to extract the audio to simplify the process.

Voice Diarization and Audio Enhancement:

In the audio file, there could be multiple people speaking - in fact, overlapping each other's conversation. In that case, we need to diarize the audio - meaning identify different speakers and isolate them.

Once the voice is isolated, we have to clean the isolated audio - background noise removal + equalization + dynamic range compression + Amplification + Removal of echo etc - Some of these are optional during the time of POC.

Tools - Pyannote / SpeechBrain - These are open-source models and Python packages that should theoretically enable us to achieve the above-mentioned process.

SpeechBrain - Takes over 40 mins on GPU to process a 5 min audio file and yet does not segregate voices well. Overlapped voices is an issue.

Pyannote- Works fast - takes 50 seconds to process 5 min audio file and breaks down individual speakers + overlapped timelines. However, this does not offer a direct functionality to separate

audio with multiple speakers into different audios, it transcribes the audio times for each speaker from which we could manually get the audio segments. As to the quality of results - it distinguishes broad audio differences well, if a male, female and child are speaking it very well distinguishes between them and lets us create 3 voice tracks, however, if it's an audio with very similar voices, it gets confused easily.

Voice Cloning:

API's

Eleven Labs
https://elevenlabs.io/ - This is a company that enables voice cloning on their website. But it is very inexpensive. 5$ a month for up to 30 mins of audio which in our case might be fine. And this has a ton of good reviews on cloning a voice especially if the audio length is significantly less too. They have a few API's too, but need to research to see if we could clone a voice in an API and use that voice without licensing issues.

- The Best results on instant audio cloning. Sounds almost like the same person when cloned even in 'Instant Cloning' option with as less as 40 seconds of clean audio by a single person.

  They even have the option to integrate phone calls with Twillio and handle incoming calls with for a given number.

  They seem to have an API to add/delete/edit voices.

  Biggest roadblock is their limits - Depending on the tier of subscription, we can save only 3-10 voices at a time.

  I thought of instantly cloning each voice to generate text, and then deleting it for another user however - they have limits on that too, for a given account - "Is there a limit on how many times I can edit / add / remove voices?

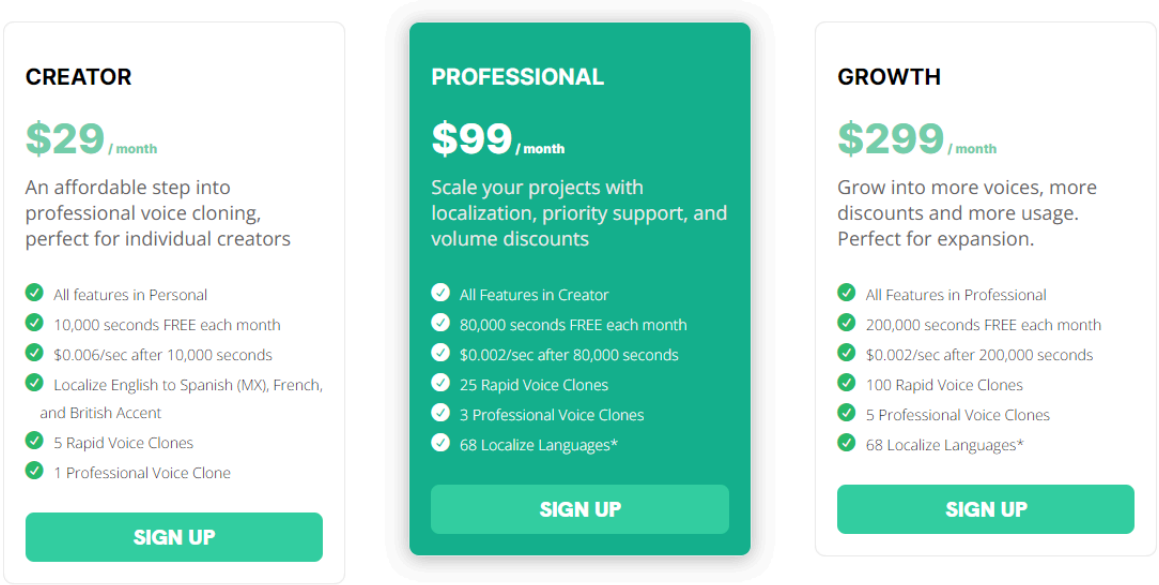
  Yes, depending on your subscription tier there are different monthly limits on how many voice operations you can make (adding a new voice or editing an existing one). Starter: 65 Creator: 95 Pro: 290 Scale: 1040"

Azure Voice Cloning - Needs 40 mins of audio in the minimum to create a cloned voice. Takes about 2 hours to do it, in their demo's.

Resemble AI

- They have 2 options.Instant and Professional Clone.
  To create a professional clone - in the audio samples we provide - The cloner should read a consent line provided by resemble AI and record their voice with it. Only then the clone would be successful and be available for further use.

  They too seem to have strict limits to how many voices can be added at a time.
  I haven't tested the quality of results yet.



Speechify

They have a UI Playground version - Tested in that.

Speechify is the only platform that has reasonable consent methodologies. We still have to get the consent but their API allows us to send an email to the customer using that voice to get their consent.

Also, Speechify does not have any restrictions on how many voice clone's we could do, unlike Resemble AI and Eleven Labs.

Results are decently good, For example, if we upload an indian accent female voice to clone and read a new statement - It does clone it and reads back in an Indian female accent - But if we just listen to the cloned voice - it dosen't exactly feel like that person. Since, we are cloning it, we can understand there is a hint of their actual voice in this cloned voice, but if we have to listen just to the cloned voice and guess who it is, might be hard to guess.
(This is slightly better than openVoice cloning but has the same problem, more often than not, it feels like it's not the same voice).

Open Source

MetaVoice-1B: https://huggingface.co/metavoiceio/metavoice-1B-v0.1 (OpenSource, Apache License)

- Tested in their web interface - results are not good.

OpenVoice: https://github.com/myshell-ai/OpenVoice (OpenSource) (MIT Licence)

- Implemented V2 version - With short audio's it's not good. There is a hint it might be the same person but doesn't really sound like that exact person

From my initial research -
https://github.com/CorentinJ/Real-Time-Voice-Cloning - is an old repo but demonstrates good results in YouTube demo.

- Watched YouTube demos where this worked great, but when tested locally it was terrible

https://github.com/coqui-ai/TTS - This is another open-source repo with instructions on how to clone a voice.

- The above models are built on the TTS Coqui backend. So, if openVoice and RTVC are underperforming, then this has to be worse than that. Tried testing it, but the package's backend models support only espeak-Ng backend, which is hard to install on a Windows machine.

Create Natural Conversations:

During the signup process, we could ask the user to provide a brief description of the person the voice is to be cloned. Were they funny, or smart, or very kind, how did they address the user when they were around etc.

We could use OpenAI, Anthropic, AWS bedrock to generate meaningful conversations from the details provided by the user.

Communicate with the user recurrently:

Once the voice is cloned, we could send birthday wishes, or good morning messages etc to user in whatsapp or sms or even in a call. All of these are generally facilitated from -

https://www.twilio.com/en-us - This is not expensive. Depending on the services we subscribe - for ~1000 clients it would be 20$ a month or lesser in my estimated guess, I could be wrong. This is used world wide for sending mass messages/ whatsapp or customer service robo calls. But would have to dwell into it to understand more.

https://www.plivo.com/ - Another Alternative.

There are a few indian based companies kind of doing the same - which are kind of cheaper and sometimes better in my opinion. Need to explore a bit more.

Functionality:

(My thoughts so far - Open to changing them as needed)

- User would come to our webpage.
- They could directly go through the technical process of using the audio they have, to try and clone a voice - and play a generic sentence with their loved ones voice so they could judge how the cloned voice sounds. Till this stage - no login/ no signup/ no credit card details etc. If they like the voice that's cloned - Then they can create an account and pay a subscription every month - to receive custom natural messages everyday in the morning or on their bdays etc.

Future Thoughts:

After POC - My initial thoughts on how this could go further:

- We could let them have a detailed conversation in whatsapp - where they type or speak - for which using chatgpt we generate an appropriate reply and create a voice over in the cloned voice and reply back to them.
- Microsoft Vasa - Was just released this week. With a single photo, we could create natural conversations. This is not open to public yet, and no other tool is close to this level of accuracy in video production. When this is opensourced or a similar alternative comes out, we could easily build a video call option to speak with the people with their face and audio naturally.
- If the product is working till here - Then there are a lot of things we could do - using cloned voices, we could enable users have an API to create voice clips with the cloned voice.
- The karaoke idea was an amazing touch too, tweaking the same technology.