音声クローン 検討まとめ

自分の声でナレーションを作るための方法の整理(2026年10月19日時点)

目的

プレゼンやYouTube動画のナレーションを、**自分の声で読み上げさせたい**。原理は「自分の声の特徴をAIに覚えさせ、その声で任意の文章を喋らせる」技術(ボイスクローン)。

方法の全体像

- **クラウド型**(ElevenLabsなど):手軽で高品質。ただし有料で、声のデータを外部に預ける。
- **ローカル型**:自宅PC内で完結。声のデータが外に出ない。
  - **タイプA(ゼロショット型)**:訓練済みモデルに数十秒〜数分の声を「参照」として渡すだけ。学習不要、CPUで動く。
  - **タイプB(ファインチューニング型)**:自分の声で追加学習させ、専用モデルを作る。**今回の第一候補。**

クラウド型(参考)

ElevenLabsのStarter(月約6ドル=約900円、月約30分)から簡易クローンと商用ライセンスが使え、YouTubeにも使える。高品質クローンはCreator(月約22ドル)から。手軽だが、声データを預ける点と月額料金がかかる点から、今回は見送り。

※料金は確認時点の目安(1ドル=150円換算)。申し込み時に最新の金額を確認すること。

ローカル型 タイプB(詳しく)

選んだ理由

声のデータが手元から出ず、月額料金もかからない。YouTubeで長く使うなら理想的。手間はかかるが、クロコさん(Claude Code)と相談しながら進める。

仕組み

訓練済みの音声合成モデルに、自分の声のデータを追加で学習させ(ファインチューニング)、自分専用の声のモデルを作る。タイプAより再現度が上がりやすい。

学習と読み上げの環境

- **学習**:GPU(特にNVIDIA製)があると現実的な時間で終わる。自宅PCにない場合は、**Google Colab**の無料GPUで学習する方法がある。ただし無料版は使用時間の制限や、混雑時にGPUが使えないことがあり、数時間で終わる規模向き。

- **読み上げ(推論)**:CPUだけでも動く。遅くても、音声は事前にmp3として作っておけばよいので実用上は問題ない。
- **分担の例**:学習だけColabで行い、読み上げは自宅PCのCPUで行う。

事前準備

1. **自宅PCのGPUを確認する**:有無と種類で、自宅で学習するかColabを使うかが決まる。
2. **自分の声を録音する**:目安は30分以上、できれば1時間前後。
    - 静かな部屋で録る
    - 同じマイク、同じ距離で録る
    - いろいろな文章を自然なトーンで読む
    - 普段のYouTubeのナレーションに近い話し方で録ると、仕上がりが動画に合う

日本語に強いモデル

- **Style-Bert-VITS2**、**GPT-SoVITS**などのオープンソースモデルが、日本語で自分の声を学習させる用途でよく使われている。
- この分野は入れ替わりが早いので、始めるときに**クロコさんに最新の状況を調べてもらってから選ぶ**のが確実。
- 日本語は英語ほど完成度が高くない場合がある。

次の一歩

自宅PCのGPUを確認し、その結果に合わせてクロコさんに渡すプロンプトを作る。

共通の注意点

- クローンしてよいのは**自分の声だけ**。他人の声を許可なく使うのは、なりすましなどに悪用されうるため避ける。
- YouTubeでAI音声を使う場合、投稿時に「改変または合成されたコンテンツ」として申告する設定があり、迷ったら申告しておくのが安全。