**VoiceCraftAI**

**Project Overview**

**Abstract**

Voice Craft AI is a cutting-edge dubbing software powered by artificial intelligence (AI), designed to seamlessly translate English audio into various Indian regional languages. Our innovative platform offers a simplified yet robust process to dub videos in multiple languages while retaining the nuances of the original speaker's voice.

**The Dubbing Process**

Voice Craft AI follows a systematic approach to transform English audio into regional languages with a touch of naturalness.

We follow a 8-step pipeline in order to achieve this end product.

**Input:** Youtube video / User uploaded video

**Step 1: Audio Extraction**

* The first step involves taking the user’s video content and extracting the audio from it. This serves as the foundation for the dubbing process.
* We do this with the help of ffmpeg and extract a “.wav” file from the user video. This is our original English audio.

**Step 2: Speech Transcription**

* The next step involves transcribing this audio into text with appropriate punctuations, to get the best possible text for translation.
* We use OpenAI’s Whisper to transcribe the speech from the extracted audio, as our research has concluded that this is the best way to transcribe our English speech.

**Step 3: Translation**

* In this step, we convert our English text into 20+ regional languages to get the transcription for our voice synthesis is various languages.
* Since the problem statement deals with non-colloquial translations, we can use Google translate to achieve this.

**Step 4: Voice Synthesis**

* Our next step involves utilizing Edge-TTS, a powerful AI-based text-to-speech technology.
* We choose between male and female native voices to produce a synthetic voice rendition of the transcribed script in the desired regional language.
* While this voice may sound natural, this initial synthesis is a bit robotic in nature, so we aim to fix this problem as well to make it sound natural.
* We plan to clone the user’s voice in order to achieve this.

**Step 5: Voice Model Creation**

* We can select an already trained voice model, load a pretrained model, or train our own voice model and proceed further with the cloning process.
* **Training a Voice Model:**
  + VoiceCraftAI makes it very easy for it’s users to train their own voice models using a Retrieval-Based Voice Conversion technology.
  + A voice model is simply an AI model of the user’s voice, which can be used to clone their voice into existing audio clips.
  + Creating a Dataset:
    - In order to train a voice model, the user must prepare a data set consisting of roughly 10-20 minutes of the user’s voice as a “.wav” file.
    - These can be separate files with duration of 20 seconds each or one single file.
    - These files must be named “file<number>.wav”, stored into a folder and that folder must be zipped.
    - This zipped file is now the dataset which needs to be imported to VoiceCraftAI
  + Once the dataset is created, we can import it and start creating a model. The number of epochs, the sampling rate, the batch size has already been determined with intense experimentation, so those need not be tinkered by the user.
  + Once the model is trained, the user can download the model as a “.zip” file.
  + The user can also choose to save this voice on the software, so that it is easier to choose the same voice for future conversions.

**Step 6: Final Audio Generation**

* Now that we have our voice model, we can proceed to clone the user’s voice over our existing edge-tts voices we generated for different languages.
* This will make the voices for different languages sound natural and more personal to the user, thereby making it more human.
* We use Retrieval-Based Voice Conversion for this step.

**Step 7: Video Integration and Subtitle Generation**

* The next step is to integrate our audios for over 20+ regional languages with our video to form the basis of the output videos.
* This is done with the help of ffmpeg.
* We also embed subtitles in each regional languages along with their videos in order to provide more clarity for the consumers.

**Step 8: Lip-sync Integration**

* The final step is to integrate lip-syncing with the audio of each regional language, to create the final end product.
* We achieve this using Wav2lip.
* Once this is done, the final set of videos are rendered out and neatly compiled into a “.zip” file with each language’s name in the title of that specific video.

This marks the end of our pipeline.

**Output:** 20+ voice cloned and lip-synced videos in various regional languages.

**The End Result**

Voice Craft AI's unique pipeline takes a single video as input and transforms it into multiple output videos. These output videos are not only professionally dubbed in various Indian regional languages but also feature the speaker's own voice and are synchronized with lip movements. The addition of subtitles further enhances the user experience, making the content more accessible to a wider audience.

With Voice Craft AI, the power of AI is harnessed to break language barriers, making content more inclusive, engaging, and relatable to diverse audiences across India. Whether for entertainment, education, or information sharing, our solution ensures that your message is heard and understood by all.

**Voice Craft AI - Where the essence of your voice meets the diversity of languages.**