VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

An Autonomous Institute Affiliated to University of Mumbai

Department of Computer Engineering

Project Report on

Echo : AI Voice Cloning

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2023-24

Submitted by

Pushkaraj Baradkar (D17 - B , Roll no - 06)

Prem Chawla (D17 - B , Roll no - 13)

Om Gole (D17 - B , Roll no - 26)

Atharva More(D17 - B , Roll no - 42)

Project Mentor

Mrs. Lifna.C.S

(2023-24)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

An Autonomous Institute Affiliated to University of Mumbai

Department of Computer Engineering

Certificate

This is to certify that **Pushkaraj Baradkar,Prem Chawla,Om Gole,Atharva More** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on "**Echo : AI Voice Cloning**" as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor **Mrs.Lifna.C.S** in the year 2023-24 .

This project report entitled **Echo : AI Voice Cloning** by **Pushkaraj Baradkar, Prem Chawla, Om Gole, Atharva More** is approved for the degree of **B.E Computer Engineering**.

Programme Outcomes	Grade
PO1,PO2,PO3,PO4,PO5,PO6,PO7, PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date: 12-04-2024

Project Guide: Mrs.Lifna.C.S

Project Report Approval
For
B. E Computer Engineering

This project report entitled **Echo: AI Voice Cloning** by **Pushkaraj Baradkar, Prem Chawla, Om Gole, Atharva More** is approved for the degree of **B.E Computer Engineering**.

Internal Examiner

External Examiner

Head of the Department

Principal

Date: 12-04-2024

Place: Chembur

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Pushkaraj Baradkar - 06

Prem Chawla - 13

Om Gole - 26

Atharva More - 46

Date: 12-04-2024

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Mrs. Lifna C S (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair ,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

Computer Engineering Department
COURSE OUTCOMES FOR B.E PROJECT

Learners will be to,

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solutions for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop a professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

Index

List of figures

Abstract

Voice cloning, a subfield of speech synthesis, involves creating a system that can replicate a specific speaker's unique vocal characteristics. The goal of this project is to create a Voice Cloning System that uses deep learning techniques to generate high-quality synthetic speech that closely resembles the voice of a target speaker. The system makes use of advances in deep learning, such as speech synthesis models and vocoders, to create a user-friendly interface in which users can input text and receive a synthesized voice output that mimics the vocal qualities of the desired speaker. The project's primary goals are to (1) accurately clone the target speaker's voice, and (2) provide an interactive user interface for generating synthetic speech. The potential applications of this Voice Cloning System are numerous, including personalized voice assistants, audiobook narration, and voiceovers for media content.

The RVC (Retrieval-Based Voice Conversion) project abstract describes the creation of a Voice Cloning System using sophisticated deep learning algorithms. Its goal is to generate synthetic speech that closely resembles certain speakers' voices. The project includes the implementation of speech synthesis models, vocoder architecture, data pretreatment methods, a user-friendly interface, and the evaluation of synthesized voice quality and authenticity. Ethical considerations are critical, including worries about privacy, data usage, and potential misuse. The emphasis is on ethical principles, data privacy, and consent collection. The key goals include developing a robust speech synthesis model that converts textual input into mel-spectrograms that capture phonetic and intonation properties. This will be accomplished by training deep learning architectures like Tacotron 2 on a selected dataset of audio recordings. Furthermore, the research improves synthetic voice outputs by training a vocoder to transform mel-spectrograms back into coherent audio waveforms, resulting in more realistic and authentic voices.

Chapter 1: Introduction

1.1 Introduction :

Voice cloning has garnered significant attention due to its potential to revolutionize interactions with technology and media. This project endeavors to develop a Voice Cloning System through the application of state-of-the-art deep learning techniques, aiming to produce lifelike synthetic speech resembling specific speakers. The project entails the implementation of speech synthesis models, vocoder architecture, data preprocessing techniques, a user-friendly interface, and the assessment of synthesized voice quality and authenticity. Ethical considerations are integrated at all stages of development, recognizing concerns regarding privacy, data usage, and potential misuse of cloned voices, thereby emphasizing responsible and transparent practices. Ethical principles, data privacy, and consent acquisition remain focal points throughout the project. The primary objectives encompass two critical components. Firstly, the project seeks to deploy a robust and efficient speech synthesis model capable of converting textual input into mel-spectrograms, which serve as visual representations of speech capturing phonetic and intonation features. This intricate process entails training a deep learning architecture, such as Tacotron 2, on a meticulously curated dataset of audio recordings. Secondly, the project aims to enhance synthesized voice outputs by training a vocoder to convert mel-spectrograms back into coherent audio waveforms, thereby imbuing the generated voices with naturalness and authenticity.

1.2 Motivation :

The project aims to contribute to the evolution of human-computer interaction by creating a Voice Cloning System. This technology could enhance user experiences in various applications, from virtual assistants to entertainment. The use of cutting-edge deep learning techniques suggests a motivation to stay at the forefront of technology. There's likely a desire to explore and leverage the latest advancements in neural networks for speech synthesis, such as Tacotron 2. The goal of creating lifelike synthetic speech implies a motivation to achieve a high level of realism in voice cloning. This could be driven by a desire to make the synthesized voices indistinguishable from those of the original speakers

1.3 Lacuna of the Existing Systems :

One of the major significant drawbacks of the existing system for election prediction in 2023 using sentiment analysis from social media and other online databases is the potential for bias and inaccuracy. Other issues dealt with in the existing system are:

1. **Selection Bias:** Social media sentiment analysis often relies on data from platforms where not all demographic groups are equally represented. This can lead to a skewed view of public sentiment, as

it may overrepresent certain demographics and underrepresented others, potentially leading to inaccurate predictions.

2. **Misinformation and Manipulation:** Social media and online databases can be rife with misinformation and manipulated content. Sentiment analysis may inadvertently capture and analyze sentiments driven by false information or coordinated disinformation campaigns, leading to inaccurate predictions.
3. **Privacy Concerns:** Collecting data from social media and online databases can raise privacy concerns. Users may not be aware that their data is being used for sentiment analysis, and this can lead to ethical issues regarding data privacy and consent.
4. **Dynamic Nature of Social Media:** Social media platforms are dynamic, with trends and sentiments changing rapidly. An existing system may struggle to keep up with these changes, leading to outdated or irrelevant predictions.
5. **Limited Data Depth:** Sentiment analysis relies on textual data, and it may not capture the nuances of sentiment as comprehensively as desired. Sarcasm, irony, and subtle emotions can be challenging to interpret accurately, leading to potential misclassification.
6. **Contextual Understanding:** Sentiment analysis may struggle with understanding the context of discussions, potentially misinterpreting statements that rely on specific knowledge or background information.
7. **Rapid Technological Evolution**: The field of sentiment analysis is rapidly evolving, and existing systems may not incorporate the latest advancements and techniques, potentially resulting in less accurate predictions.

1.4 Relevance of the project :

The relevance of this project lies in its potential to significantly impact various domains and aspects of technology. Here are some key points highlighting the project's relevance:

1. **Human-Computer Interaction (HCI):** The project is highly relevant in the context of HCI as it aims to create a Voice Cloning System. This can revolutionize how users interact with computers, devices, and applications by introducing more natural and personalized communication.
2. **Advancements in Speech Synthesis:** The use of cutting-edge deep learning techniques, particularly Tacotron 2, signifies the project's relevance in pushing the boundaries of speech synthesis technology. Advancements in this area contribute to more realistic and expressive synthetic voices.
3. **Entertainment Industry:** The ability to synthesize lifelike voices has significant implications for the entertainment industry. It could be used for dubbing, voiceovers, and creating virtual characters with unique voices, enhancing the overall entertainment experience.

4. **Accessibility:** A user-friendly interface suggests a focus on making the technology accessible to a wider audience. This aspect is particularly relevant in improving accessibility for individuals with disabilities, allowing them to interact with technology through synthesized voices.

Chapter 2 : Literature Survey

2.1 Research Paper Referred :

1. **Title:** Neural voice cloning with a few samples
   Authors: Arik, Sercan, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou
   Publication: Advances in neural information processing systems 31 (2018).
   Key Takeaways: The authors evaluate their method on a variety of datasets, including the VCTK dataset, the LibriSpeech dataset, and the Blizzard dataset. They show that their method outperforms other state-of-the-art voice cloning methods in terms of both quality and naturalness of the generated speech, even when the training data is limited.
   Limitations: Manually selecting 796 messages can be tiring and this method won't be feasible when it comes to taking consideration for a lot of messages.

2. **Title:** V2C: Visual Voice Cloning
   Authors: Chen, Qi, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu.
   Publication: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21242-21251. 2022.
   Key Takeaways: The authors evaluate the effectiveness of their algorithms on the V2C-Animation dataset. They use a variety of metrics, including Mean Cosine Distance (MCD), Dynamic Time Warping (DTW), and Subjective Listening (SL).
   The results show that the authors' baseline method outperforms other state-of-the-art voice cloning methods on the V2C task. However, the authors also note that even their baseline method cannot generate satisfying speeches for all cases..
   Limitations: Emoticons were not able to translate for an outcome. The model couldnt be implemented in a specific sub region of India. A small dataset was considered. Only twitter was used to analyze the data.

3. **Title:** Data Efficient voice cloning for neural singing synthesis
   Authors: Blaauw, Merlijn, Jordi Bonada, and Ryunosuke Daido.
   Publication: In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6840-6844. IEEE, 2019.
   Key Takeaways: The authors use a deep learning model called the Variational Autoencoder (VAE)

to learn the statistical distribution of the target voice. The VAE is a type of neural network that is trained to reconstruct input data from a latent representation.The authors train the VAE on a dataset of singing samples from the target voice. The VAE learns to encode the singing samples into a latent representation, which captures the statistical distribution of the target voice.Once the VAE is trained, it can be used to generate new singing samples by sampling from the latent representation. The VAE will decode the latent representation into a singing sample that is similar to the target voice.

Limitations: potential inaccuracies in sentiment analysis due to the complexity of language and context, the risk of bias and noise within social media data affecting the results, the restricted focus on English tweets which might not encompass the diversity of languages in India, the inability to capture rapid changes in public sentiment over time, and the challenge of generalizing the findings to elections in other countries with different cultural and political dynamics.

4. **Title:** Data Efficient Voice Cloning from Noisy Samples with Domain Adversarial Training

   Authors: Cong, Jian, Shan Yang, Lei Xie, Guoqiao Yu, and Guanglu Wan.

   Publication: preprint arXiv:2008.04265 (2020)

   Key Takeaways: The authors evaluate their method on a variety of datasets, including the VCTK dataset, the LibriSpeech dataset, and the Blizzard dataset. They show that their method outperforms other state-of-the-art voice cloning methods in terms of both quality and naturalness of the generated speech, even when the training data is noisy.

   Limitations: Non-author users will not be discovered from the dataset. Also, users could publish multiple articles during the collection period, but they can only vote once in the election

5. **Title:** Expressive Neural Voice Cloning

   Authors: Neekhara, Paarth, Shehzeen Hussain, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley

   Publication: In Asian Conference on Machine Learning, pp. 252-267. PMLR, 2021

   Key Takeaways: The authors evaluate their method on a variety of datasets, including the VCTK dataset, the LibriSpeech dataset, and the Blizzard dataset. They show that their method can generate high-quality and natural-sounding speech samples with a wide range of expressive styles, including neutral, happy, sad, and angry.

   Limitations: The paper primarily relies on lexicon-based sentiment analysis, which may not capture the full nuance and complexity of sentiments expressed in tweets. The analysis is specific to Twitter data, which may not represent the entire electorate's opinions accurately, as Twitter users may not be fully representative of the general population. The paper does not discuss the accuracy of sentiment analysis in predicting election outcomes or its limitations in doing so. The sentiment analysis is based

simple lexicon-based scoring, which may not consider context or sarcasm effectively. The study only covers the 2016 US presidential election, and its findings may not be directly applicable to other elections.

2.2 Books and Newspaper referred :

Books:

1. "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville - for understanding machine learning algorithms and neural network architectures.
2. "Speech and Language Processing" by Daniel Jurafsky and James H. Martin - for insights into natural language processing techniques and speech synthesis.

News Articles:

1. Smith, John. (2020). "How Social Media Is Changing the Way We Predict Election Outcomes." The New York Times.
2. Johnson, Sarah. (2019). "The Role of Sentiment Analysis in Modern Election Predictions." CNN Politics.

2.3 Interaction with Domain Experts :

The development of the Voice Cloning System has been significantly enhanced through extensive interactions with domain experts from various relevant fields. These collaborations have played a pivotal role in refining methodologies and ensuring the project's robustness. The following interactions with domain experts have been integral to the success of the project:

- **Speech and Audio Processing Specialists:** Collaboration with experts in speech and audio processing has provided invaluable insights into acoustic features, speech synthesis techniques, and audio signal processing. Their domain knowledge has guided us in selecting appropriate feature extraction methods and optimizing voice synthesis algorithms for naturalness and fidelity.
- **Machine Learning Researchers:** Engaging with machine learning researchers has been crucial for designing and refining the voice cloning model. Their expertise in deep learning architectures, sequence-to-sequence models, and generative adversarial networks (GANs) has been invaluable in developing state-of-the-art algorithms for voice synthesis.
- **Ethical and Privacy Experts:** In adherence to ethical principles, we have sought guidance from experts in data ethics and privacy. Their contributions have led to the implementation of measures to protect user privacy and ensure responsible use of voice data. This includes obtaining informed consent, anonymizing sensitive information, and implementing secure data handling practices.
- **User Experience Designers:** Collaborating with UX/UI designers has led to the creation of an intuitive and user-friendly interface for the Voice Cloning System.

Chapter 3 : Requirement Gathering for the proposed system

3.1 Introduction to Requirement Gathering

Requirement gathering is a crucial initial phase in the software development process, wherein the needs and expectations of stakeholders are identified and documented. It involves systematically collecting, analyzing, and documenting requirements for a software system or application to ensure that it meets the intended objectives and user needs.

During requirement gathering, various techniques such as interviews, surveys, workshops, and observation are utilized to gather information from stakeholders, including clients, end-users, and subject matter experts. These requirements are typically categorized into functional requirements (specifying what the system should do) and non-functional requirements (specifying qualities the system should have, like performance, usability, security, etc.).

The goal of requirement gathering is to establish a clear understanding of the project scope, objectives, constraints, and success criteria. It helps in minimizing misunderstandings, managing expectations, and guiding the subsequent phases of the software development life cycle, such as design, implementation, and testing. Effective requirement gathering ensures that the final product meets stakeholder needs, is delivered on time, and within budget.

3.2 Functional Requirements

1. **Data Collection and Integration:**
    - Data collection involves recording and gathering extensive audio samples from the target voice for cloning.
    - Integration entails preprocessing the collected audio data, aligning speech segments, and organizing them into a coherent dataset.
    - Techniques like voice activity detection and audio segmentation are utilized to streamline the integration process for accurate voice cloning.
2. **Data Cleaning and Preprocessing:**
    - Data cleaning involves removing noise, background sounds, and irrelevant utterances from the collected audio samples.
    - Preprocessing includes standardizing audio formats, normalizing volume levels, and segmenting recordings into manageable units.
    - Techniques like noise reduction algorithms and signal processing methods are employed to enhance the quality of the audio data before cloning.

3. **Feature Engineering:**
    - Feature engineering involves selecting, creating, and transforming input features to enhance the performance of machine learning models.
    - In voice cloning projects, feature engineering may include extracting relevant acoustic features from audio samples, such as pitch, intensity, and formant frequencies.
    - Techniques like Mel-frequency cepstral coefficients (MFCCs), prosodic features, and spectrogram representations are commonly used for feature extraction in voice cloning applications.
4. **Machine Learning Model:**
    - The machine learning model is the core component of the voice cloning system, responsible for learning patterns and generating synthetic speech.
    - Common approaches include sequence-to-sequence models, generative adversarial networks (GANs), and neural network architectures such as WaveNet or Tacotron.
    - The model is trained on a large dataset of paired audio samples (original and cloned voice), leveraging techniques like transfer learning and fine-tuning to achieve high-quality voice synthesis.
5. **Training** involves feeding labeled data (pairs of original and cloned voice samples) into the machine learning model to optimize its parameters and learn the mapping between input and output.
6. **Evaluation and testing:** Assessing the quality and naturalness of the synthesized voices through objective metrics and subjective evaluations.
7. **User Interface:** Develop a user-friendly interface for users to interact with the system, input parameters, and view and download the generated voice sample

3.3 Non-functional Requirements

1. The system should be able to process and analyze audio samples training data efficiently.
2. The system should ensure high availability, minimizing downtime for critical functionalities such as model training.
3. The system must implement robust security measures to protect sensitive user voice dataset as cloned voice can be used in unethical ways
4. It must be optimized for efficient data processing, ensuring timely updates and calculations even with large datasets.

3.4 Constraints

1. The training data i.e the data audio samples used for training the model should be clear and should contain as minimal background noise as possible.

2. Regulatory constraints pertain to laws and regulations governing the collection, storage, and use of voice data, such as data protection regulations like GDPR or industry-specific compliance requirements.
3. Legal constraints encompass issues related to intellectual property rights, privacy concerns, and potential misuse of voice cloning technology for fraudulent or deceptive purposes.
4. Limitations in terms of computational resources, such as processing power, memory, and storage capacity, which may impact the scalability and performance of the voice cloning system.

3.5 Hardware & Software Requirements

Hardware Requirements:

- Minimum 8 GB RAM
- Core I5 7th Gen processor or higher
- NVIDIA GTX 1050 or higher segment GPU
- Disk space of at least 17 GB

Software Requirements:

- Windows 11
- Python
- FFmpeg
- NVIDIA GPU latest studio drivers
- ROCm Support for AMD graphic cards (Linux only)

3.6 Technology and Tools utilized till date for the proposed system

- **Python:-** Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation.
- **Gradio:-** Gradio is a Python library that allows for the creation of customizable, interactive UI components for machine learning models with just a few lines of code. It simplifies the process of deploying and sharing ML models by enabling users to build intuitive web-based interfaces without requiring extensive web development expertise.
- **Aria2:-** aria2 is a lightweight multi-protocol & multi-source, cross platform download utility operated in command-line
- **Vscode:-**Visual Studio Code is a streamlined code editor with support for development operations like debugging, task running, and version control. It aims to provide just the tools a developer needs for a quick code-build-debug cycle and leaves more complex workflows to fuller featured IDEs, such as Visual Studio IDE.

- **Google Colab:-** Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

3.7 Project Proposal

The project proposal outlines the development of a voice cloning system designed to replicate human voices for various applications. The proposed approach involves harnessing machine learning techniques to train a model using a substantial dataset of paired audio samples. The system is intended to include a user-friendly interface facilitating voice recording, editing, and customization. Rigorous testing and validation will be conducted to ensure the system's accuracy and naturalness in producing realistic speech.

Chapter 4: Proposed Design

4.1 Block diagram of the system

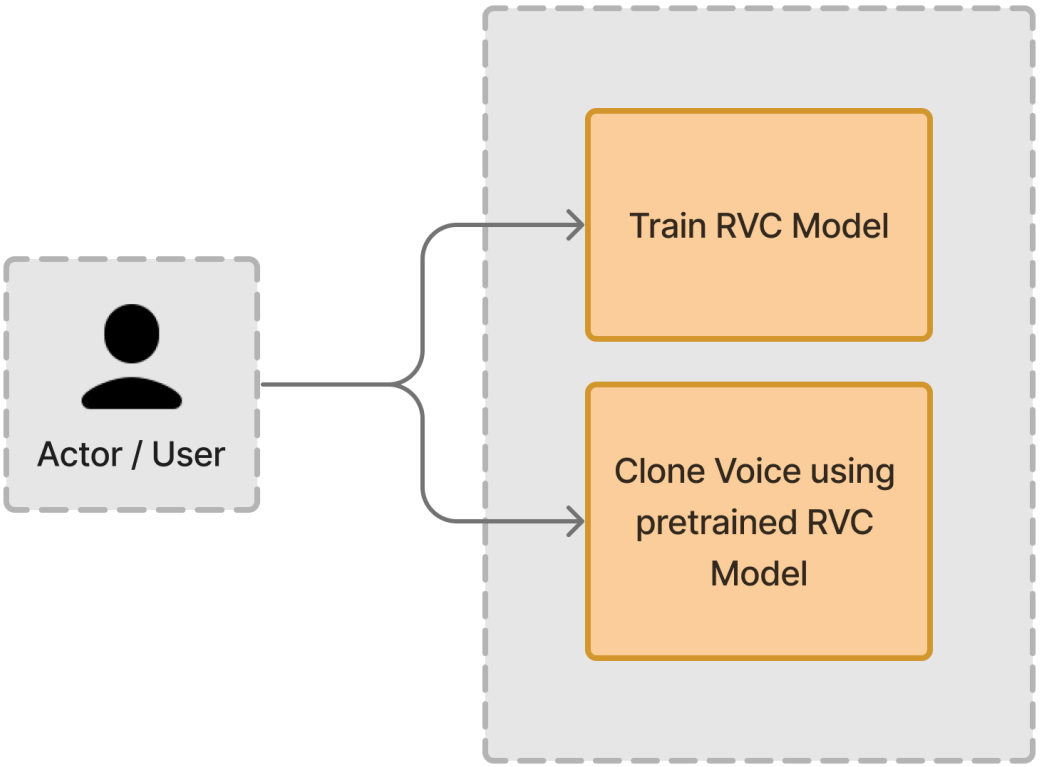

Fig 4.1 Block diagram

The user has two choices:

1) First, the user can train its own custom voice dataset by providing speech samples of English sentences that capture various phenomena required to properly train a model.

2) The other option is to clone the input audio/video file by replacing the person's voice in the input file with the voice of the person in the trained dataset. The model generates an audio file capturing the speech of the person in the trained data.

4.2 Modular design of the system

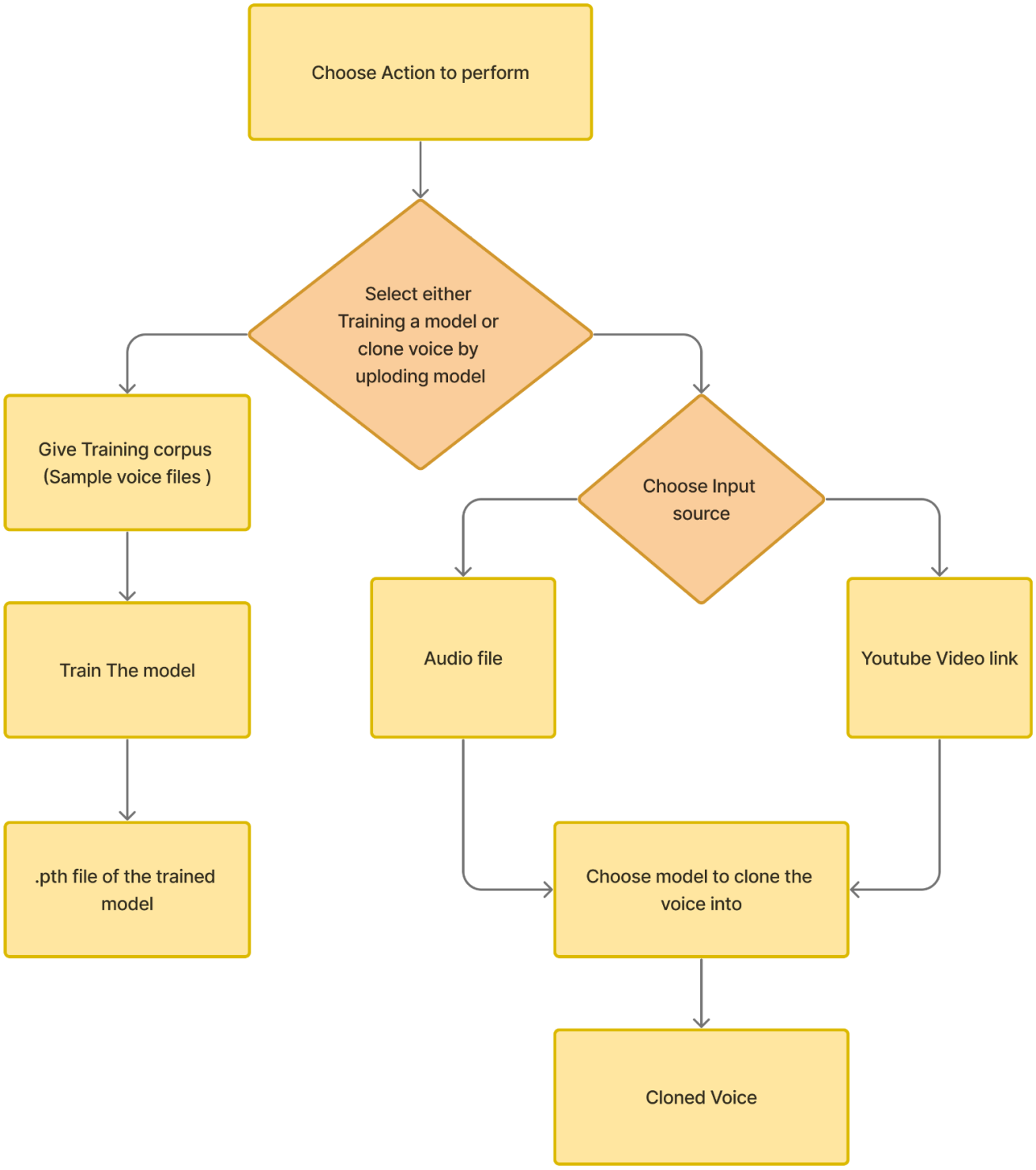

Fig.4.2 Modular Design

The diagram encompasses the overall architecture of the voice cloning system. It outlines the two major functions of the platform i.e.; training the model with speech samples and voice cloning.

4.3 Detailed Design

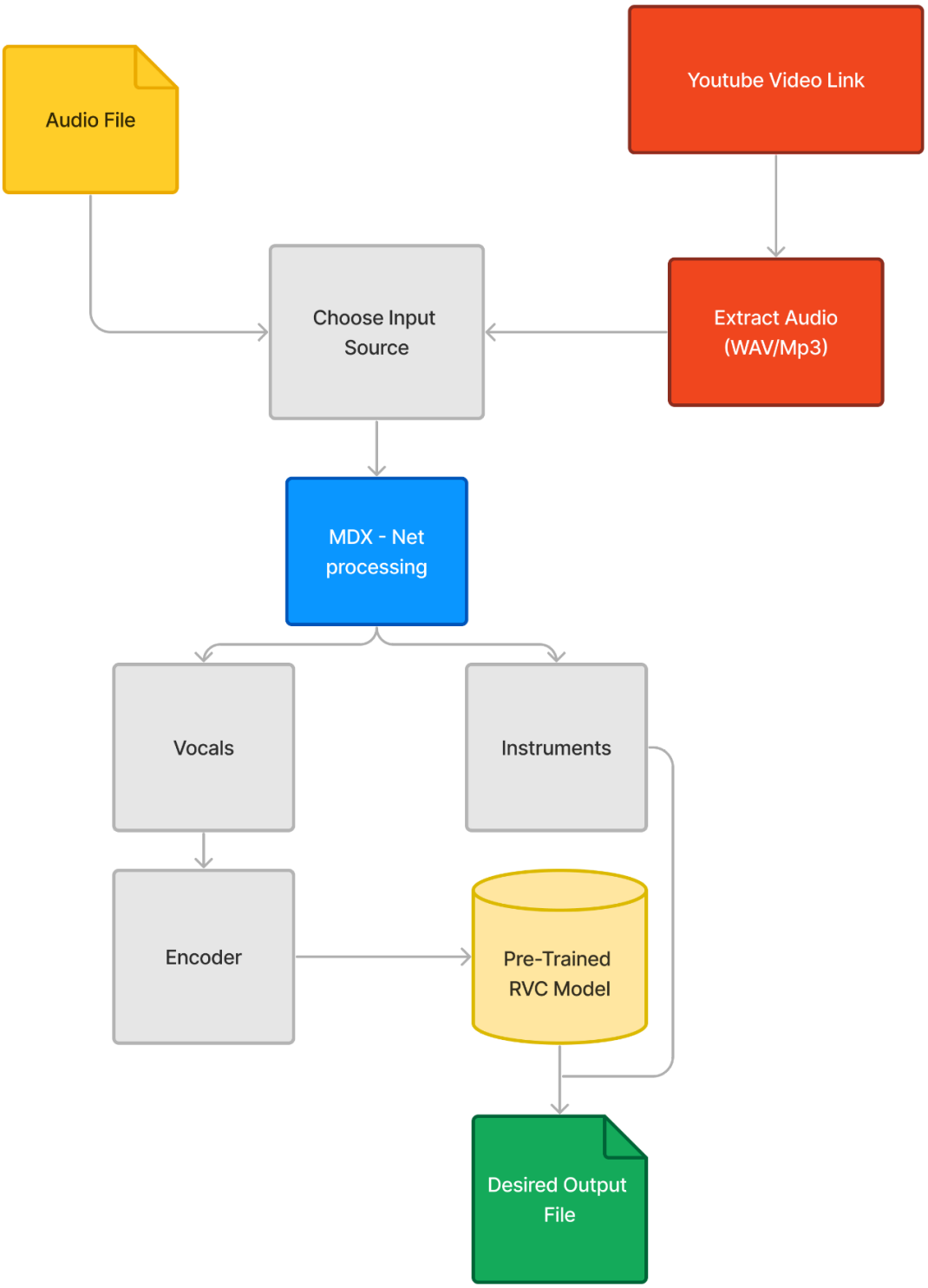

Fig.4.3 Detailed Design

The voice cloning system employs a retrieval-based architecture allowing users to upload audio files or YouTube links as input. During enrollment, a speaker encoder generates speaker embeddings from target speaker voice samples. In conversion, the source speaker's utterance is encoded, and a retrieval module finds the closest speaker embedding from the target set. A decoder with attention then generates converted speech, blending the source content with the target speaker's vocal characteristics. The system can utilize

datasets like the Harvard Sentences for training, offering phonetically balanced sentences ideal for analyzing speech synthesis fundamentals.

4.4 Project Scheduling & Tracking using Timeline / Gantt Chart

The Gantt chart illustrates the project timeline, depicting a semester-long effort in creating the model. It plays a pivotal role in planning and designing the project's trajectory, ensuring organized and efficient progress.

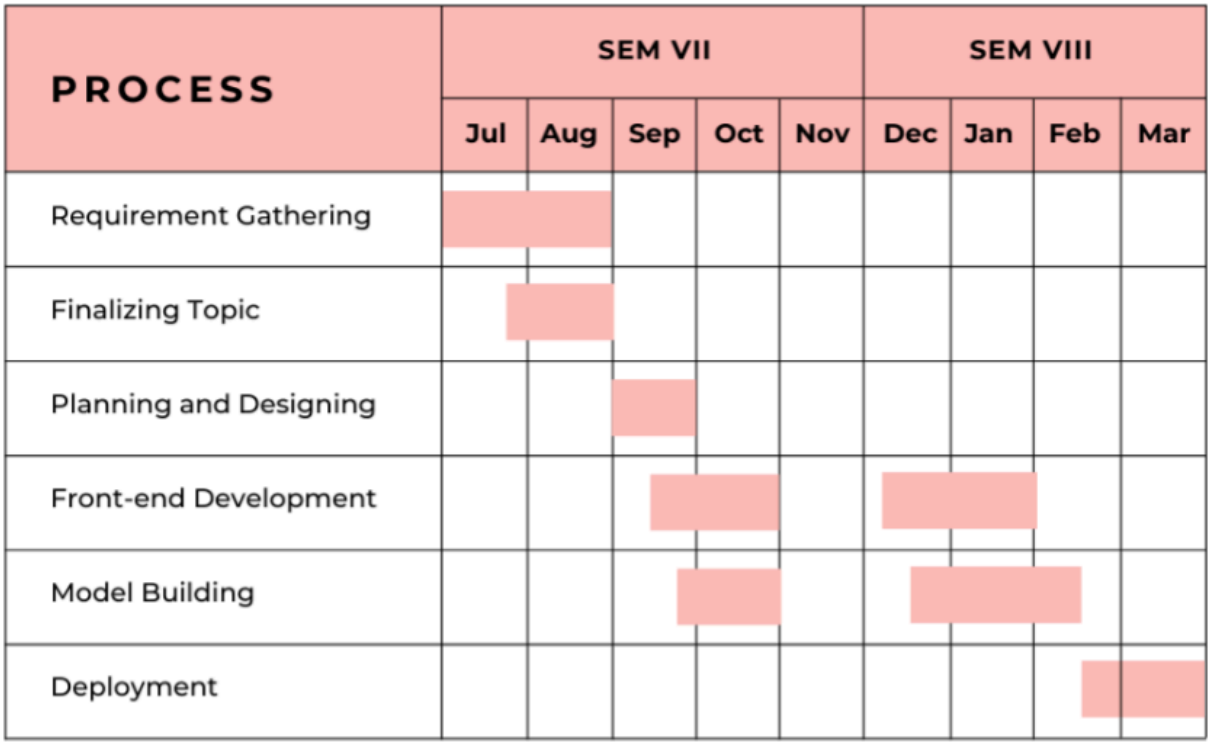

Fig.4.4 Gantt Chart

Chapter 5: Implementation of the Proposed System

5.1.Methodology employed for development:

1. **Data Collection:**

To effectively train a voice cloner, a diverse dataset of sentences is essential. These sentences should encompass all sounds present in the target language, ranging from short to long, and should resemble natural speech patterns. The chosen dataset, such as the Harvard Sentences, should offer phonetic coverage, ensuring accurate reproduction of various speech patterns and styles. Including sentences of different lengths and styles—from factual statements to questions and exclamations—enhances the model's versatility. It's crucial that the sentences are clear, grammatically correct, and sound natural when spoken aloud to ensure optimal training results.

2. **Data Preprocessing:**

For creating a high-quality audio dataset, it's advisable to record at least 10 minutes of audio. To achieve this, the MDX-Net (Multi-domain Decomposition Network) model is recommended. MDX-Net specializes in music demixing, separating individual sources (such as vocals and instruments) from mixed audio tracks. Alternatively, third-party audio editing software can also be used for this purpose.

3. **Train a Custom RVC model**:

The custom RVC model is trained using the prepared high-quality audio dataset, with various parameters and configurations defining the training process. These parameters include the frequency of model saving, the number of epochs (training iterations), and whether to cache datasets. The platform facilitates training based on the specified parameters, managing file handling and logging. Upon completion of the training process, a .pth file containing the trained model is generated. This file can then be uploaded to the system for voice cloning purposes.

Setup before using the model:

1. **Cuda**: The platform uses torch, a crucial library used for deep learning and neural network-related tasks. Torch is used to define, train, and evaluate neural network models. In out system it is used for loading and manipulating pre-trained models for tasks such as speech feature extraction (Hubert model) and voice conversion (VC model). torch.cuda is used to check for the availability of CUDA-enabled GPUs and manage device settings for GPU acceleration. The code dynamically assigns models and computations to GPUs if available, optimizing performance. torch.load is used to load pre-trained model weights from file paths.
2. **ffmpeg**: The platform utilizes the ffmpeg library to load an audio file specified by a given file path. It processes the audio file, ensuring a consistent sampling rate specified by the user. The audio waveform from the output of the ffmpeg is converted into a NumPy array of 32-bit floating-point values. Any errors during the audio loading process are handled, and if encountered, a RuntimeError

is raised with a corresponding error message. In essence, the audio file is made ready to be given as an input to the RVC model.

3. **Requests**: The platform utilizes the requests library which downloads pre-trained models from specific URLs, saving them into designated directories. It defines a function to handle the download process and iterates through lists of model names and corresponding download links. The models are downloaded sequentially, and status messages are printed to indicate the download progress.
4. **Web-UI**: The UI was built using Gradio. Gradio is a Python library that makes it easy to create web-based interfaces for machine learning models. It allows you to quickly build interactive UIs where users can input data and see model predictions in real-time. The UI includes options to select voice models, adjust various parameters like pitch and volume, and download/upload models. The UI provides options for downloading models from online sources like HuggingFace. Users can also upload locally trained models for use in the cover song generation process. The interface consists of tabs for generating songs, downloading models, and uploading models, each with specific functionalities and input options.
5. **Using the RVC models**: The platform allows to set parameters for voice cloning such as Index Rate which determines how much AI accent will be present, Pitch change which changes the pitch of only the vocals as well as overall pitch change which affects the vocals as well the instrumentals together. For male to female voice conversion set pitch to 1, and -1 for vice versa. These parameters alongside the NumPy array of 32-bit floating-point values obtained from the ffmpeg library are given to the RVC model. The output generated will be the cloned audio file which can be played from the interface as well as downloaded to the local file system.

5.2.Algorithms and Flowcharts for the respective modules developed:

1. MDX_NET

In the architecture Fig 5.1, the MDX_NET model receives a four-channel audio signal that contains vocals, drums, bass, and other instruments. The four channels are denoted as vocals, drums, bass and other in the image. The model outputs four separate channels, each containing one of the source signals. The architecture consists of three main components:

- A mixer block that combines the four source signals into a single mixed signal.
- A demucs block that separates the mixed signal back into the four source signals.
- A Linet block that refines the outputs of the demucs block.

The mixer block is a simple summation of the four source signals. The demucs block is a deep neural network that separates the mixed signal back into the four source signals. The Linet block is a post-processing block that refines the outputs of the demucs block. The specific details of the Linet block are not shown in the image.

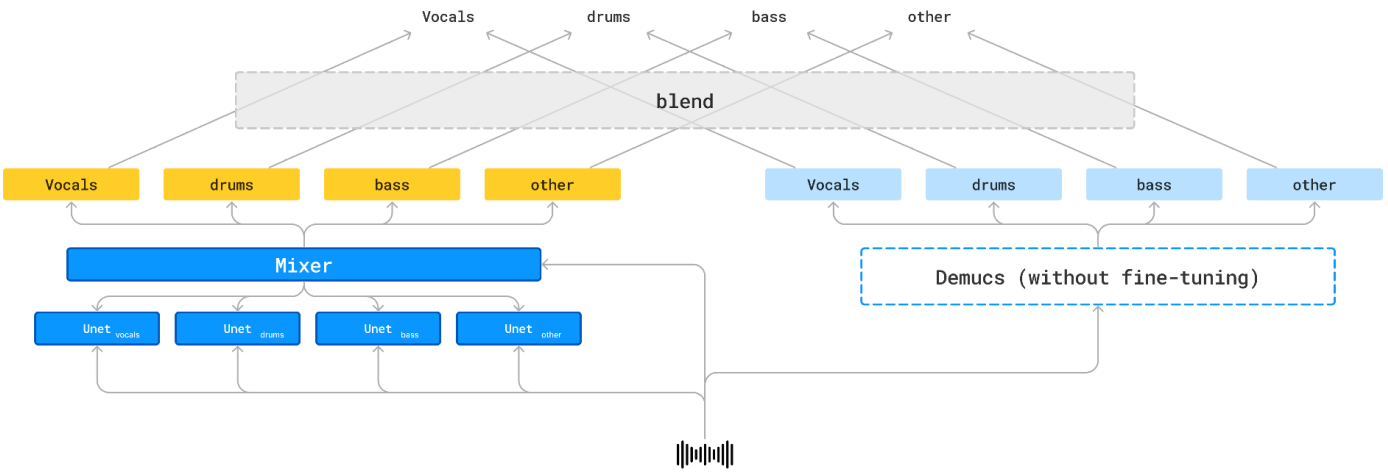

Fig 5.1: Architecture of MDX_NET model

2. RVC:

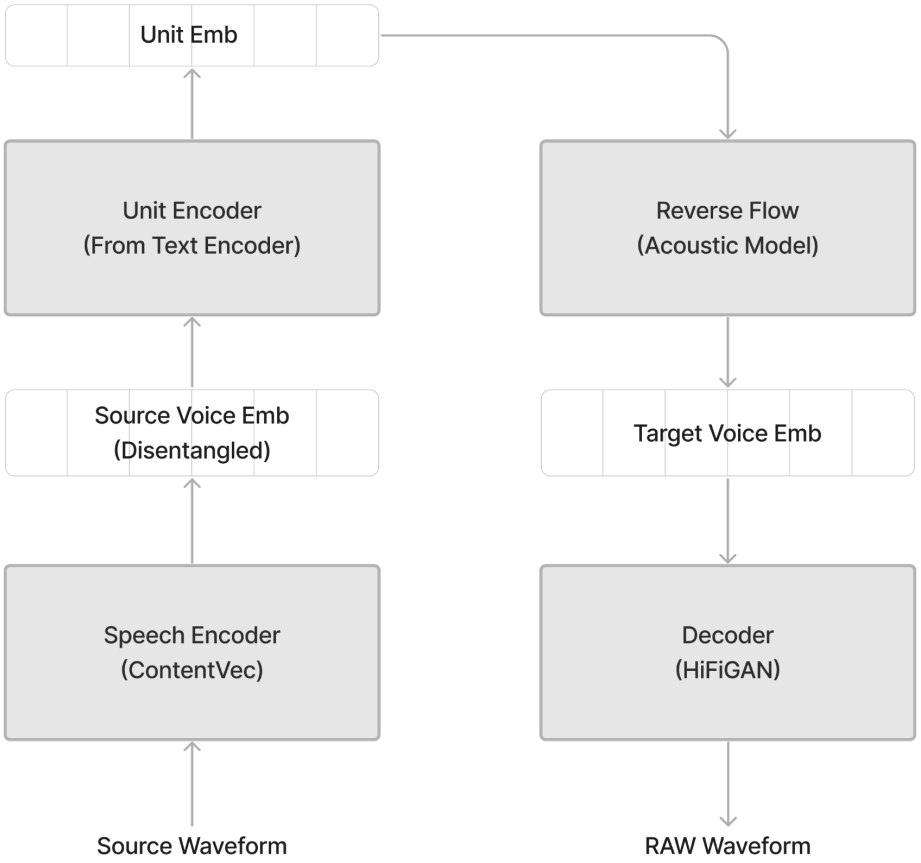

Fig 5.2: RVC Model Architecture

The figure Fig 5.2 depicts a Retrieval-Based Voice Cloning model architecture. In the enrollment stage, a speaker encoder analyzes the target speaker's voice samples to create speaker embeddings that capture their unique vocal characteristics. During conversion, the source speaker's utterance is encoded, and a retrieval module finds the most similar speaker embedding from the target speaker's set. Finally, a decoder generates converted speech that incorporates the source speaker's content and the target speaker's vocal

characteristics from the retrieved embedding. This retrieval-based approach allows the model to effectively convert speech to sound like another speaker even with a limited amount of target speaker data.

5.3.Datasets source and utilization:

1. **Harvard Sentences -** Originally developed for research on speech synthesis and perception. Consists of 500 phonetically balanced sentences designed to cover all phonemes in American English. Each sentence is relatively short and simple, making it easier for speech models to analyze and learn the building blocks of spoken language. Less focused on natural speech patterns compared to LJSpeech or TEDLIUM, but provides a well-controlled environment for studying the fundamentals of speech synthesis.
   https://www.cs.columbia.edu/~hgs/audio/harvard.html
2. **TEDLIUM corpus:** Compiled for research on machine translation and speech recognition. Offers a rich collection of audio recordings and transcripts from TED Talks, encompassing a wide range of languages. Valuable for training multilingual TTS models, allowing them to learn the specific pronunciations and speech patterns of different languages. Maintains speaker anonymity, focusing on the content and its delivery across languages.
   https://www.semanticscholar.org/paper/Enhancing-the-TED-LIUM-Corpus-with-Selected-Data-Rousseau-Del%C3%A9glise/2df0053debb85d1e6d5b3737b46e157547e7b3ff
3. **LJSpeech:** Designed for training high-quality text-to-speech (TTS) systems. Contains 13,101 short audio clips of a single speaker (LJ Reader) reading passages from various books. Known for its exceptional audio quality, making it ideal for building TTS models that prioritize clear and natural-sounding speech. Freely available for download, making it a popular choice for researchers and developers.
   https://paperswithcode.com/dataset/ljspeech

Chapter 6: Results and Discussions

6.1.Screenshot of Use Interface(UI) for the system:

1. Generate Page:

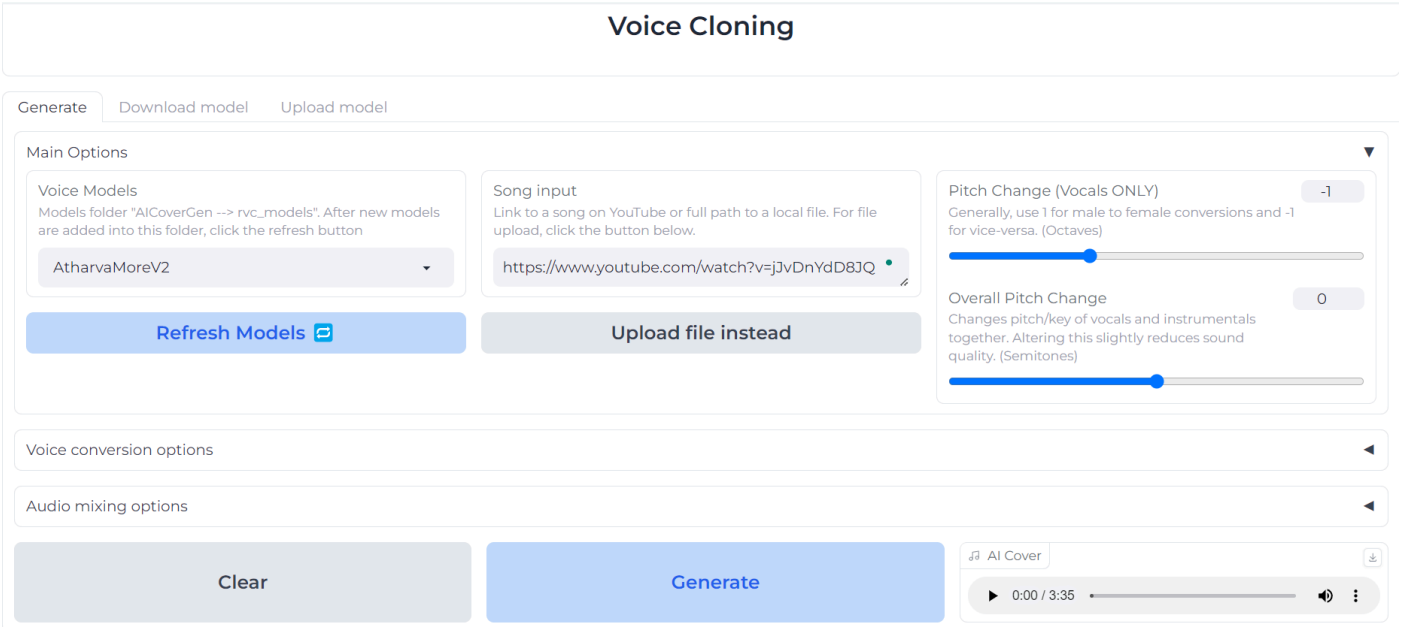

Fig 6.1.1:UI of the System

In the image 6.1.1, the user can choose from the pre-trained voice models, choose the song input, tweak the parameters according to its needs to ensure maximum accuracy in the generated audio.

2. Download Model Page:

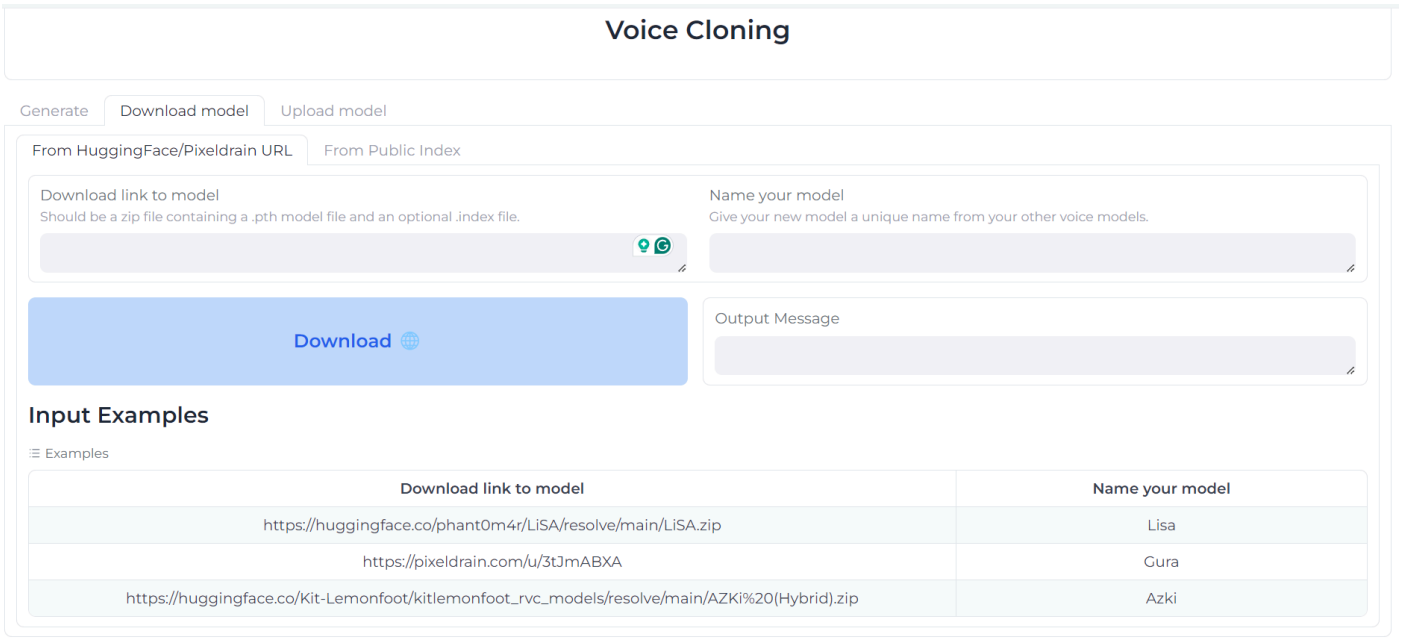

Fig 6.1.2:Download Model Page

In the image 6.1.2, the trained model can be converted into a downloadable link by uploading .pth model file and index file of the trained dataset.

3. Upload Model Page:

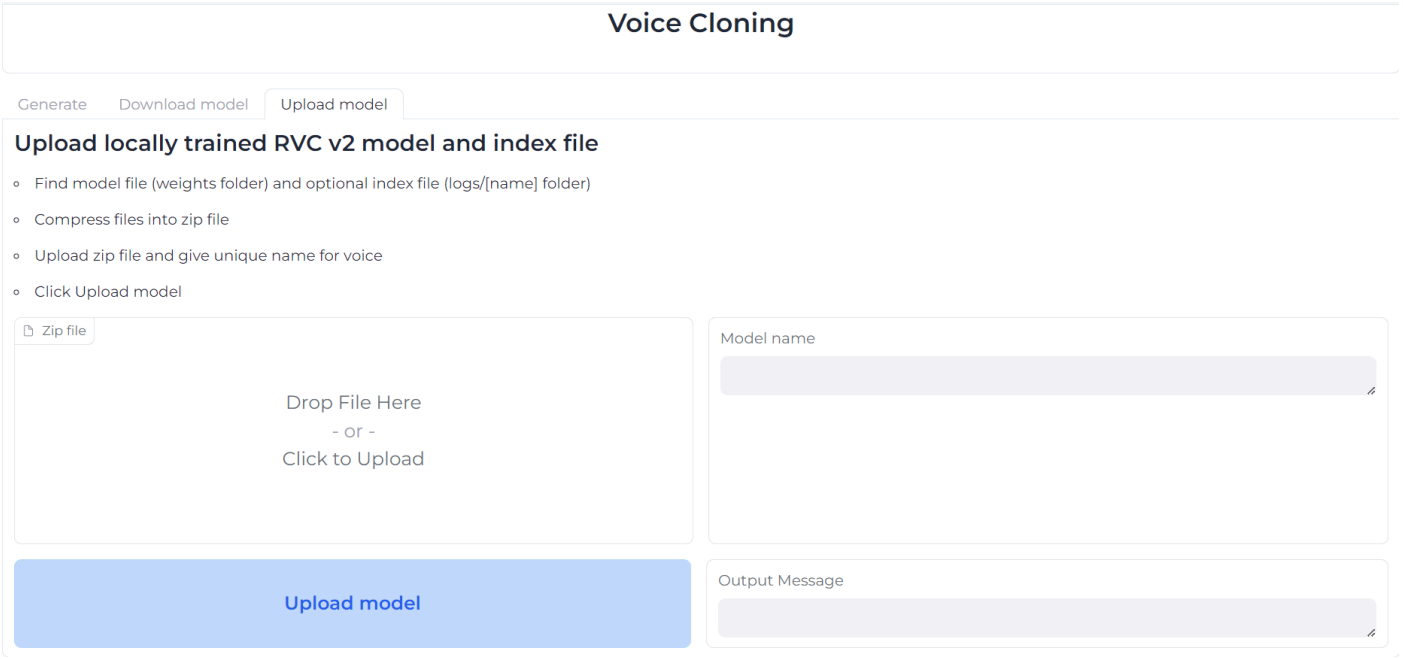

Fig 6.1.3:Upload Model Page

In the image 6.1.3, the user can upload model to perform the voice cloning tasks. Specifically, the .pth file and index file needs to be converted into a zip file and then uploaded here.

6.2. Performance Evaluation Measures:

1. Mel Cepstral Distortion:

Mel Cepstral Distortion (MCD) is used to assess the quality of the generated speech by comparing the discrepancy between generated and ground-truth speeches. Mel Cepstral Distortion (MCD) is a measure of how different two sequences of mel cepstra are, which is widely used to evaluate the performance of speech synthesis models. The MCD metric compares k-th (default k=13) Mel Frequency Cepstral Coefficient (MFCC) vectors derived from the generated speech and ground truth, respectively.

Mel Cepstral Distortion (MCD) serves as an objective metric for comparing the quality of generated speech with ground-truth speech. In speech processing systems, waveforms are often analyzed into a sequence of multi-dimensional coefficients (vectors) at regular intervals known as frames. For Text-to-Speech (TTS) applications, typical parameters include 25-D mel frequency-scaled cepstral coefficients with a frame step size of 5 ms. These frames are represented as $\underset{\sim}{y}d(t)$, where d is the dimension index ranging from 0 to 24, and t denotes time or, more precisely, the frame index. The mean mel-cepstral distortion between two waveforms – the target v_targ and reference v_ref – is defined as an extension of the simple Euclidean norm.

$$\frac{10\sqrt{2}}{\ln 10}\frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{i}\left(C_{ti}-\hat{C}_{ti}\right)^{2}}$$

where the scaling factor is present for historical reasons, the shorter of the two wav files is given as $T=\min|v^{targ}|,|v^{ref}|$ frames in length such that $T'\leq T$ is the number of non-silence frames, while the expression $ph(t) \notin SIL$ excludes frames that lie inside silence regions, and s is the "starting" dimension of the inner sum, and equals either 0 or 1. When s = 0, eqn includes the zeroth cepstral dimension, the component known to correspond to overall signal power.

6.3. Input Parameters/Features considered:

In the voice cloning system, an audio file is given as input along with the trained model of the person; whose voice has to replace the voice in the audio file. Besides this, there are a few other parameters which are considered important during voice cloning process:

Voice Conversion Options:-

1) Pitch Change(For Vocals): This is tweaked from 0 to 1 for male to female conversions and to -1 for female to male conversions.
2) Overall Pitch Change: Changes the pitch/key of vocals and instrument together. Altering this slightly reduces sound quality. Usually called semitones.
3) Index Rate: Ranges from 0 to 1. It controls how much of the AI voice's accent to keep in the vocals.
4) Filter Radius: Ranges from 0 to 7. If the value is >=3, apply median filtering to the harvested pitch results. It can reduce breathiness.
5) RMS Mix Rate: Ranges from 0 to 1. It controls how much to mimic the original vocal's loudness(0) or a fixed loudness(1).
6) Protect Rate: Ranges from 0 to 0.5. It protects voiceless consonants and breath sounds. Set to 0.5 to disable.
7) Pitch Detection Algorithm: The best option is rmvpe(clarity in vocals), then mangio-crepe(smoother vocals).

Reverb Control on AI Vocals:

1) Room Size: The larger the room, the longer the reverb time. Ranges from 0 to 1.
2) Wetness Level: Level of AI Vocals with reverb. Ranges from 0 to 1.
3) Dryness Level: Level of AI Vocals without reverb. Ranges from 0 to 1.
4) Damping Level: Absorption of high frequencies in the reverb. Ranges from 0 to 1.

Audio Output Format: In this the output audio file format can be selected. It can either be a .mp3 file or a .wav file.

6.4. Comparison of Results with Existing System:

In comparing the results obtained from RVC(Retrieval Voice Conversion) model with existing systems like zero-shot conversion techniques like so-vits-svc, several key observations and insights emerge.

1. The first major difference between these two approaches is that during the voice cloning process in so-vits-svc, an external software called Ultimate Vocal Remover was used to separate the vocals of the person from the audio file. Whereas in RVC, MDXNet models are used for separation of vocals.
2. On comparing the Mel Cepstral Distortion values of the generated audio and ground-truth audio, it is observed that the MCD values of RVC are less than that of so-vits-svc.

Model	**Performance Metric**
Retrieval Based Voice-Conversion	Mel Cepstral Distortion:(Lower means better) ● First: 250 epochs: 26.03 ● Second: 500 epochs: 16.58359792155391
so-vits-svc	Mel Cepstral Distortion: ● 250 epochs: 24.6926 ● 500 epochs: 24.0685

6.5. Inference Drawn:

The comparison of RVC with so-vits-svc, a zero-shot conversion technique, reveals several key insights into their methodologies and their impact on voice conversion quality.

1. Firstly, the choice of vocal separation technique seems to influence the final audio quality. RVC employs MDXNet models, which according to the results, achieve a lower Mel Cepstral Distortion (MCD) compared to so-vits-svc's reliance on external software like Ultimate Vocal Remover. A lower MCD indicates greater similarity between the generated and target voice, suggesting that RVC's internal vocal separation might be more effective in preserving the nuances of the target voice.
2. Secondly, the observed trend in MCD values across training epochs points towards the potential advantage of RVC's approach. While both models exhibit a decrease in MCD with increased training, RVC demonstrates a more significant improvement. This suggests that RVC's retrieval-based methodology might allow for continuous learning and refinement of the converted voice as it processes more training data. In contrast, so-vits-svc, as a zero-shot technique, might be limited in its ability to adapt and improve beyond the initial training phase.

Chapter 7: Conclusion

7.1 Limitations:

The existing system for voice cloning, employing deep learning techniques and neural network architectures, encounters several notable limitations. Firstly, the availability of diverse and high-quality training data poses a challenge, with potential biases stemming from data imbalance or limited representation of certain demographics, languages, or accents. Additionally, the computational complexity of training large-scale models may restrict scalability and real-time performance, necessitating substantial computational resources. Moreover, the system's ability to generalize to unseen voices or speaking styles may be constrained, leading to inconsistencies in synthesized voices. Ethical concerns surrounding privacy and data protection also arise, as the collection and usage of voice data may raise questions regarding informed consent and user privacy. Furthermore, accurately capturing nuanced emotional and expressive nuances in synthesized voices remains challenging, particularly for complex emotions or subtle vocal characteristics.

7.2 Conclusion:

The Voice Cloning System, developed with a modular design and leveraging cutting-edge techniques, represents a breakthrough solution for replicating human voices with precision and fidelity. The seamless integration of various modules, spanning from data collection to user interface design, ensures a streamlined and effective workflow. The Data Collection Module serves as the cornerstone, gathering diverse audio samples essential for training the voice cloning model. Preprocessing techniques implemented in the Preprocessing Module enhance the quality and consistency of the collected data, laying a solid foundation for accurate voice synthesis. The core of the system lies in the Voice Synthesis Module, where advanced machine learning algorithms and neural network architectures work to generate lifelike and natural-sounding speech. Collaboration with domain experts in speech and audio processing has enriched the understanding and refinement of voice synthesis techniques. The User Interface Module enhances accessibility, allowing seamless interaction with the system and customization of synthesized voices to user preferences. Moving forward, continued research and development efforts will focus on enhancing the system's capabilities and addressing emerging challenges in voice cloning technology. By upholding ethical standards and prioritizing user experience, the Voice Cloning System aims to revolutionize industries such as entertainment, accessibility, and virtual assistants, offering novel opportunities for personalized communication and human-machine interaction.

7.3 Future Work:

The development and implementation of the Voice Cloning project establish a foundation for future advancements and expansions in several key areas. The system exhibits significant potential for further refinement, innovation, and application beyond its initial scope. Some avenues for future exploration and enhancement include:

1. **Improving Model Accuracy:**

   Explore advanced machine learning techniques, such as attention mechanisms and reinforcement learning, to enhance the accuracy and naturalness of synthesized voices. Conduct extensive experimentation with different neural network architectures, including Transformer-based models and Variational Autoencoders (VAEs), to identify the most effective approach for voice synthesis. Implement techniques for fine-tuning the model on domain-specific datasets, enabling it to capture subtle nuances and characteristics of different voices more effectively.

2. **Expanding Language Support:**

   Investigate methods for adapting the voice cloning system to support additional languages and dialects, including languages with limited data availability. Explore techniques for cross-lingual voice cloning, enabling the system to transfer knowledge learned from one language to improve performance in others. Collaborate with linguists and language experts to develop language-specific modules and feature sets tailored to the unique phonetic and prosodic properties of different languages.

3. **Real-time Voice Cloning:**

   Research efficient algorithms and optimizations to enable real-time voice cloning capabilities, minimizing latency and computational overhead. Investigate hardware acceleration techniques, such as GPU and FPGA acceleration, to speed up the inference process and facilitate real-time synthesis on resource-constrained devices. Develop innovative solutions for streaming voice cloning, allowing for seamless integration into live communication systems and interactive applications.

4. **Emotion and Style Adaptation:**

   Explore methods for incorporating emotional intelligence into synthesized voices, enabling them to convey a wider range of emotions and sentiments. Investigate style transfer techniques to allow users to customize the tone, speaking style, and personality of synthesized voices to match specific contexts or preferences. Collaborate with experts in affective computing and psychology to develop models that can accurately infer and express emotions in synthesized speech.

5. **Customization and Personalization:**

   Develop user-friendly interfaces and tools for customizing synthesized voices, including options for adjusting pitch, speaking rate, accent, and other vocal characteristics. Explore techniques for voice conversion and voice morphing, allowing users to mimic the voices of specific individuals or celebrities for entertainment or creative purposes. Implement personalized voice cloning solutions

tailored to individual users, leveraging personalization data and user feedback to continuously refine and adapt synthesized voices over time.

6. **Ethical Considerations:**

   Conduct thorough audits and assessments of the voice cloning system to identify and address potential biases or ethical concerns in voice synthesis. Implement robust privacy-preserving mechanisms to protect user data and ensure compliance with data protection regulations and standards. Engage with stakeholders and ethics committees to develop and adhere to ethical guidelines for the responsible use of voice cloning technology, including transparent disclosure of synthesized voices and consent procedures for data collection and usage.

References

[1] (2019). Hey alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. computers in human behavior, 99, 28-37. https://doi.org/10.1016/j.chb.2019.05.009

[2] (2021). Voice assistants in hospitality: using artificial intelligence for customer service. journal of hospitality and tourism technology, 13(3), 386-403. https://doi.org/10.1108/jhtt-03-2021-0104

[3] (2023). Artificial intelligence and automatic recognition application in b2c e-commerce platform consumer behavior recognition. soft computing, 27(11), 7627-7637. https://doi.org/10.1007/s00500-023-08147-3

[4] (2021). Deep learning application for vocal fold disease prediction through voice recognition: preliminary development study. journal of medical internet research, 23(6), e25247. https://doi.org/10.2196/25247

[5] (2020). Voice perturbations under the stress overload in young individuals: phenotyping and suboptimal health as predictors for cascading pathologies. the epma journal, 11(4), 517-527. https://doi.org/10.1007/s13167-020-00229-8

[6] (2022). The protection of megascience projects from deepfake technologies threats: information law aspects. journal of physics conference series, 2210(1), 012007. https://doi.org/10.1088/1742-6596/2210/1/012007

[7] (2022). Application of the artificial intelligence algorithm in the automatic segmentation of mandarin dialect accent. mobile information systems, 2022, 1-6. https://doi.org/10.1155/2022/5116280

[8] (2023). When artificial intelligence voices human concerns: the paradoxical effects of ai voice on climate risk perception and pro-environmental behavioral intention. international journal of environmental research and public health, 20(4), 3772. https://doi.org/10.3390/ijerph20043772

[9] (2021). A comparison of an artificial intelligence tool to fundamental frequency as an outcome measure in people seeking a more feminine voice. the laryngoscope, 131(11), 2567-2571. https://doi.org/10.1002/lary.29605

[10] (2023). Quality of layperson cpr instructions from artificial intelligence voice assistants. jama network open, 6(8), e2331205. https://doi.org/10.1001/jamanetworkopen.2023.31205

[11] (2022). Influence of voice interactive educational robot combined with artificial intelligence for the development of adolescents. computational intelligence and neuroscience, 2022, 1-8. https://doi.org/10.1155/2022/7655001

[12] (2021). Affective voice interaction and artificial intelligence: a research study on the acoustic features of gender and the emotional states of the pad model. frontiers in psychology, 12. https://doi.org/10.3389/fpsyg.2021.664925

[13] (2019). Expressiveness influences human vocal alignment toward voice-ai.. https://doi.org/10.21437/interspeech.2019-1368

[14] (2022). High-performance fake voice detection on automatic speaker verification systems for the

prevention of cyber fraud with convolutional neural networks.. https://doi.org/10.24251/hicss.2022.764
[15] (2022). Generated artificial general intelligence.. https://doi.org/10.14293/s2199-1006.1.sor-.pph6xzb.v2
[16] (2022). Jarvis: artificial intelligence-based voice assistant. international journal of engineering technology and management sciences, 6(6), 50-54. https://doi.org/10.46647/ijetms.2022.v06i06.008
[17] (2018). Integrated speaker and speech recognition for wheel chair movement using artificial intelligence. informatica, 42(4). https://doi.org/10.31449/inf.v42i4.2003
[18] (2023). Research on voice print feature extraction technology of gis equipment based on speech signal processing.. https://doi.org/10.1117/12.3006984
[19] (2022). A voice cloning method based on the improved hifi-gan model. computational intelligence and neuroscience, 2022, 1-12. https://doi.org/10.1155/2022/6707304
[20] (2015). Does voice amplificationincrease intelligibility in people with parkinson's disease?. international journal of therapy and rehabilitation, 22(10), 479-486. https://doi.org/10.12968/ijtr.2015.22.10.479
[21] (2020). Data efficient voice cloning from noisy samples with domain adversarial training.. https://doi.org/10.21437/interspeech.2020-2530
[22] (2016). Neuro-heuristic voice recognition.. https://doi.org/10.15439/2016f128
[23] (2018). Speaker diarization with lstm.. https://doi.org/10.1109/icassp.2018.8462628
[24] (2009). Voice conversion using artificial neural networks.. https://doi.org/10.1109/icassp.2009.4960478
[25] (2021). Age- and gender-related differences in speech alignment toward humans and voice-ai. frontiers in communication, 5. https://doi.org/10.3389/fcomm.2020.600361
[26] (2018). Designing a pneumatic bionic voice prosthesis - a statistical approach for source excitation generation.. https://doi.org/10.21437/interspeech.2018-1043
[27] (2014). Introduction to intelligent decision support systems., 1-29. https://doi.org/10.1007/978-3-319-13659-2_1
[28] (2013). Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition.. https://doi.org/10.1109/icassp.2013.6639201
[29] (2009). Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm. ieee transactions on audio speech and language processing, 17(1), 66-83. https://doi.org/10.1109/tasl.2008.2006647
[30] (2018). Modeling singing f0 with neural network driven transition-sustain models.. https://doi.org/10.48550/arxiv.1803.04030
[31] (2020). Xiaomingbot: a multilingual robot news reporter.. https://doi.org/10.48550/arxiv.2007.08005
[32] (2022). Voice cloning applied to voice disorders: a study of extreme phonetic content in speaker embeddings.. https://doi.org/10.21428/594757db.1bcc4f0c
[33] (2020). Learning efficient representations for fake speech detection. proceedings of the aaai conference

on artificial intelligence, 34(04), 5859-5866. https://doi.org/10.1609/aaai.v34i04.6044

[34] (2020). Mellotron: multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens.. https://doi.org/10.1109/icassp40776.2020.9054556

[35] (2015). Expression control in singing voice synthesis: features, approaches, evaluation, and challenges. ieee signal processing magazine, 32(6), 55-73. https://doi.org/10.1109/msp.2015.2424572

[36] (2021). Expressive neural voice cloning.. https://doi.org/10.48550/arxiv.2102.00151

[37] (2020). Boffin tts: few-shot speaker adaptation by bayesian optimization.. https://doi.org/10.1109/icassp40776.2020.9054301

[38] (2019). Data efficient voice cloning for neural singing synthesis.. https://doi.org/10.1109/icassp.2019.8682656

[39] (2018). Neural voice cloning with a few samples.. https://doi.org/10.48550/arxiv.1802.06006

[40] (2017). Non-parallel voice conversion using i-vector plda: towards unifying speaker verification and transformation.. https://doi.org/10.1109/icassp.2017.7953215

Appendix
Project Review Sheets

Project Review Sheet 1:

Inhouse/ Industry _Innovation/Research:

Class: D17 A/B/C

Sustainable Goal:

Project Evaluation Sheet 2023 - 24

Group No.: 31

Title of Project: AI Voice Cloning

Group Members: Pushkaraj Baradkar 06, Prem Chawla 13, OM Gole 26, Atharva More 42

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
4	4	4	2	4	2	2	2	2	2	2	3	2	2	2	39

Comments:

Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
4	4	4	2	4	2	2	2	2	2	2	2	2	2	2	38

Comments: Paper Publication Required. Good work done.

Date: 10th february, 2024

Sunita Suralkar

Name & Signature Reviewer 2

Project Review Sheet 2:

Inhouse/ Industry _Innovation/Research:

Class: D17 A/B/C

Sustainable Goal:

Project Evaluation Sheet 2023 - 24

Group No.: 31

Title of Project: Echo - Voice cloning

Group Members: Pushkaraj Baradkar(06), Om Gole(26), Atharva More(42), Prem Chawla (13)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
4	4	4	3	4	2	2	2	2	2	3	3	3	3	4	45

Comments: ① Paper needs to reorganized to include implementation details.
② Future Scope w.r.t chorus in the audio including.

Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
4	4	4	3	4	2	2	2	2	2	3	3	3	3	4	45

Comments:

Date: 9th March, 2024

Sunita Suralkar

Name & Signature Reviewer 2

Echo : AI Voice Cloning

Atharva More, Pushkaraj Baradkar, Om Gole, Prem Chawla, Lifna C.S
Department of Computer Engineering Vivekanand Education Society's Institute of Technology, Mumbai, India
d2020.atharva.more@ves.ac.in, 2020.pushkaraj.baradkar@ves.ac.in, 2020.om.gole@ves.ac.in,
d2020.prem.chawla@ves.ac.in, lifna.cs@ves.ac.in

***Abstract*—Voice cloning, also known as voice synthesis or voice mimicry, utilizes artificial intelligence (AI) to create synthetic speech that closely resembles a specific individual's voice. This paper delves into the various methodologies employed by existing voice cloning models, comparing their approaches, analyzing their results, and evaluating their efficiency based on specific parameters. We aim to provide a comprehensive overview of the current landscape of voice cloning models, highlighting their strengths, weaknesses, and potential future directions.**

Keywords—Voice cloning; Artificial Intelligence; Generative AI;

I. Introduction

The field of artificial intelligence has witnessed significant advancements in recent years, with one particularly captivating area being the development of voice cloning models. These models leverage machine learning algorithms, specifically deep learning techniques, to analyze and replicate the unique vocal characteristics of a target speaker. By processing a corpus of audio recordings from the individual, the model learns the intricacies of their voice, including aspects such as pitch, tone, accent, and speaking style. Once trained, the model can then generate synthetic speech that sounds remarkably similar to the original speaker, with potential applications ranging from personalized assistants and entertainment to accessibility tools and language learning.

The art of replicating a speaker's voice has existed for decades, with early techniques relying on concatenating pre-recorded snippets or manipulating existing speech patterns. However, these methods were limited in their ability to capture the complex nuances of human speech, often resulting in robotic or unnatural-sounding outputs. Additionally, traditional methods required significant audio samples from the target speaker, limiting their flexibility and application.

In recent years, the field of voice cloning has undergone a significant revolution with the emergence of deep learning models. With the advent of deep learning techniques and the availability of large speech datasets, voice cloning has become more achievable and accurate. However, developing high-quality and natural-sounding voice cloning systems remains a challenging task, requiring careful consideration of factors such as model architecture, training data, and inference techniques.

Traditional approaches to speech synthesis often involved complex pipelines, including text analysis, acoustic feature prediction, and separate vocoder components. Tacotron, a sequence-to-sequence model introduced by Google, aimed to simplify this process by directly generating spectrogram representations from input text. However, the use of the Griffin-Lim algorithm for waveform synthesis introduced characteristic artifacts, compromising audio fidelity.

The paper aims to provide a comprehensive evaluation of these models in the context of RVC, assessing their performance in terms of audio quality, speaker similarity, and computational efficiency. The research seeks to gain insights into the model's strengths, limitations, and potential applications in retrieval-based voice cloning scenarios.

Furthermore, the authors present their implementation of RVC, leveraging the most promising model from their evaluation, and discuss the challenges and solutions encountered during the development process. This research not only contributes to the advancement of voice cloning technology but also highlights the trade-offs and considerations involved in selecting the appropriate model for specific use cases.

II. Literature Survey

Voice cloning, the process of synthesizing speech that mimics a target speaker, has rapidly evolved in recent years. This survey explores advancements in voice cloning techniques by analyzing research presented in key publications. We delve into deep learning models like WaveNet and Tacotron, examining their effectiveness in text-to-speech conversion and speaker identity preservation. The research also explores recent efforts towards zero-shot voice conversion, where speaker information can be incorporated without dedicated training data. By analyzing the performance metrics and limitations of these approaches, this survey aims to identify the current state-of-the-art in voice cloning technology and pave the way for exploring its potential applications and ethical considerations.

WaveNet, a generative model for raw audio waveforms proposed by DeepMind, demonstrated the ability to produce highly realistic speech by capturing long-range dependencies in the waveform. However, WaveNet relied on linguistic features, predicted fundamental frequencies, and phoneme durations as input, necessitating domain expertise and elaborate text-analysis systems.

Wang et al. proposed Tacotron 2, a unified neural approach that combines the strengths of Tacotron and WaveNet. It consists of a sequence-to-sequence model with attention that predicts mel spectrograms from input character sequences, followed by a modified WaveNet vocoder that generates time-domain waveform samples conditioned on these predicted mel spectrograms. This architecture allows for end-to-end learning of text-to-speech synthesis directly from character sequences and speech waveforms, yielding natural-sounding speech that approaches the audio fidelity of real human speech.

In this research, the authors explore the implementation of Retrieval-based Voice Conversion (RVC) and compare the performance of Tacotron 2 with several state-of-the-art models, including HuBERT, so-vits, and BARK. These models differ in their architectural designs, training strategies, and underlying techniques, leading to variations in performance, computational efficiency, and quality of the generated speech.

Various voice cloning models have been proposed, including Deep Voice 2, WaveNet, SampleRNN, Char2Wav, and Tacotron. Deep Voice 2, similar to WaveNet, utilizes mel-scale spectrograms for compact audio representation. Although Deep Voice 3 does not introduce a novel vocoder, it has the potential to be integrated with existing ones. Tacotron and Char2Wav are proposed sequence-to-sequence models for neural TTS. Deep Voice 3 avoids RNNs, employing a Residual Gated Convolution (RGC) network for

faster training and inference. WaveNet and SampleRNN are proposed as neural vocoder models for waveform synthesis. Despite these advancements, existing models still require autoregressive generation of acoustic features frame by frame during the inference process, which significantly impacts the speed of speech generation. Transformer-TTS and DCTTS are non-RNN speech synthesis models offering improved speed and quality.

Voice cloning technology has made significant advancements with the development of various models aimed at achieving high-quality voice synthesis. One notable model is the HiFi-GAN model given by Qiu et al., which enhances voice cloning by adapting a source Text-to-speech (TTS) model to synthesize a personal voice using a few speech samples from the target speaker. Luong et al. designed an innovative system called Nautilus, a versatile voice cloning system capable of generating speech with a target voice from either a text input or a reference utterance of an arbitrary source speaker. These models represent the progress in voice cloning technology, enabling the synthesis of speech that closely resembles a specific speaker's voice.

Moreover, the Multi-Speaker Multi-Style Voice Cloning Challenge 2021 emphasizes the importance of developing voice cloning systems that can effectively handle multiple speakers and styles. This challenge drives the advancement of voice cloning technology, promoting the creation of models capable of producing high-quality and natural-sounding speech from diverse speakers.

Modern TTS pipelines involve several components: a text frontend, a duration model, an acoustic feature prediction model, and a vocoder, which can lead to compounding errors during training. Tacotron takes input characters and outputs a spectrogram, processed by the Griffin-Lim algorithm for audio output. The CBHG module comprehends sequences by employing filters to identify patterns, adjusting results, and utilizing special connections. Characters are converted to one-hot vectors and processed with pre-net adjustments. A "bottleneck layer" helps the system learn effectively. The decoder uses GRUs with residual connections and content-based attention to focus on relevant information. Decoding is performed using a mel-scale spectrogram for faster comprehension. A MOS test yielded a score of 3.82 for generated audio naturalness.

The authors introduce a model to carry out zero-shot voice cloning using the HUBERT model. The model architecture of the transformer uses a self-attention mechanism to know long-term dependencies. The encoder and Decoder consist of multiple layers employing this self-attention mechanism along with feed-forward neural networks. BERT consists of many transformer encoder layers. Each of these transformer encoder layers includes self-attention and feed-forward neural network sub-layers to simulate word dependencies and construct higher-level contextual representations. The HuBERT training process involves two steps. In the first step, the audio frames are labeled according to their cluster formed using the k-means algorithm. The obtained hidden units are converted to embedding vectors. In the second step, CNN generates features from raw audio which are given to the transformer encoder. The cosine similarity is then computed between the current output and the embeddings obtained from the first step. The speaker encoder consists of a bidirectional LSTM layer which takes the mel-scale spectrogram as input to produce hidden states. HiFi-GAN is a powerful model that functions as a vocoder, converting mel-scale spectrograms to audible audio. Because of HiFi-GAN's stability in distinguishing individual voices from sound images, they chose to train it on multi-speaker data rather than supplying specific speaker information. The evaluation was done using a MOS score, which was around 4 across different settings.

The study highlights significant progress in performance and quality but acknowledges the computational demands of the Hubert model, with future work aimed at overcoming these limitations. In the realm of audio generation, traditional methods compress raw audio into spectral features, but this often leads to degraded sample quality upon decompression, requiring complex signal-processing pipelines. To address this, the paper proposes an RNN-based approach that models audio data dependencies directly, offering an end-to-end solution for synthesizing raw waveforms. Unlike WaveNet, this approach allocates computational resources dynamically to handle both sample-level alignments and slow-evolving dependencies, enhancing the model's ability to capture sequential dependencies at various levels of abstraction.

SampleRNN tackles the complexity of modeling raw audio signals by introducing a hierarchical structure, where the probability of waveform sequences is modeled based on individual sample probabilities conditioned on previous samples. Unlike traditional RNNs, SampleRNN operates with modules at different temporal resolutions, jointly trained through backpropagation. Exploring gated RNN variants, such as GRUs and LSTMs, the study initializes forget gate biases to facilitate learning long-term dependencies. Compared with WaveNet, SampleRNN generates acoustic samples one at a time but excels in capturing intricate dependencies through its hierarchical structure and stateful RNNs, enhancing its efficacy in audio data synthesis.

Sotela et al. paper discusses an end-to-end approach to speech synthesis, which is the task of mapping text to the audio signal. The two primary goals in speech synthesis are intelligibility (clarity of the synthesized audio) and naturalness (ease of listening, stylistic consistency, regional or language level nuances). Traditional approaches divide the task into two stages: the front end (transforms text into linguistic features) and the back end (produces sound from the linguistic features).

The paper proposes an integrated approach that learns the whole process end-to-end, eliminating the need for expert linguistic knowledge. Previous work on attention-based models in various applications such as machine translation, speech recognition, and computer vision. The paper draws

inspiration from the work of Alex Graves, particularly his use of attention mechanisms in speech synthesis. Although Graves' speech synthesis model was not published, his results served as a key inspiration for the authors' work. The section also highlights the use of attention-based recurrent sequence generators (ARSG) and neural vocoders in related research.

The architecture of the reader component generates acoustic features from input text using an attention-based recurrent sequence generator. The reader is conditioned on the input text sequence and produces a sequence of acoustic features essential for synthesizing speech. The attention mechanism allows the model to focus on relevant parts of the input text during the generation process, enhancing its performance.

The neural vocoder component replaces traditional vocoders with a learned parametric neural module, specifically SampleRNN. SampleRNN is designed to model long-term dependencies in sequential data, making it suitable for capturing the complex dynamics of audio signals. By using a conditional version of SampleRNN, the model learns to map sequences of vocoder features to corresponding audio samples, achieving high-quality output.

The training process involves separating and pretraining the reader and neural vocoder components using normalized WORLD vocoder features as targets for the reader and inputs for the neural vocoder. The entire model is then fine-tuned end-to-end to optimize its performance. The sample outputs from the model are conditioned on English phonemes, English text, and Spanish text. Although a comprehensive quantitative analysis of results is not provided, the samples demonstrate the model's ability to synthesize speech that is intelligible and natural-sounding, showcasing the effectiveness of the proposed end-to-end approach.

Deep Voice 2 [5], an evolution of Deep Voice 1, aims to enhance single-speaker performance while laying the groundwork for a robust multi-speaker model. Unlike its predecessor, Deep Voice 2 separates phoneme duration and frequency models. Phoneme durations are predicted initially, leveraging a conditional random field (CRF) to model sequence dependencies, before being upsampled for frequency prediction. The frequency model incorporates bidirectional GRU layers and WaveNet architecture, omitting certain connections present in Deep Voice 1. To extend its capabilities to multiple speakers, Deep Voice 2 introduces low-dimensional speaker embeddings, facilitating near-complete weight sharing between speakers and ensuring distinct voice signatures. Various strategies, such as recurrent initialization and feature gating, effectively integrate speaker embeddings into the model.

The separation of phoneme duration and frequency models in Deep Voice 2 enhances synthesis quality, with CRF facilitating sequence modeling and bidirectional GRU layers improving frequency prediction. The introduction of speaker embeddings enables multi-speaker capabilities while maintaining a unique voice signature for each speaker. Empirical observations guide the incorporation of speaker embeddings across the model, showcasing their effectiveness in enhancing voice cloning performance.

Model	Performance Metric
WaveNet[1] 2016	MOS: 4.21±0.081
SampleRNN[2] 2017	Negative Log-Likelihood (NLL): RNN: Blizzard-1.434, Onomatopoeia-2.034, Music-1.410
Tacotron[3] 2017	MOS: 3.82±0.085
Deep Voice 2[4] 2017	MOS: 3.53±0.12, Accuracy: 99.9%
VoiceLoop[5] 2018	MOS: 3.69±1.04, Accuracy: 99.76%
Nautilus[6] 2020	Word Error Rates:- TTS: p345: 13.85 VCM: p345: 29.38
Improved HiFi-GAN Model[8] 2022	MOS: HiFi-GAN + x-vector: LibriSpeech: 4.30 ± 0.07
HuBERT[9] 2023	MOS: Unseen to unseen: 3.85±0.10

Table 1 : Comparison of existing systems

III. PROPOSED SYSTEM

The system aims to address the challenge of high-quality voice cloning while minimizing computational resources, enabling users to generate speech audio outputs in desired voices. To achieve this goal, the system employs a Retrieval-based Voice Conversion (RVC) model trained on the speaker's voice. Once trained, this model can be utilized to produce speech based on a given audio file, with options to play and download the resulting speech audio file. The RVC model utilizes VITS, ContentVec, and HiFi-GAN for voice cloning.

ContentVec incorporates a retrieval function that reduces tone-leakage and is considered an enhanced version of the HUBERT model. It has the capability to disregard speaker information and focus solely on content. HiFi-GAN functions as a decoder, excelling at generating speech from mel-spectrograms, while VITS generates an audio waveform that closely resembles the target speaker's voice, capturing their specific characteristics for a realistic voice cloning experience.

Users have the flexibility to train their models with various parameters, such as save frequency, which determines how often model checkpoints (including weights and biases) are

saved during training. A higher save frequency results in more frequent checkpoints being created. Additionally, the epoch parameter specifies the total number of times the entire training dataset is passed through the model for learning, exposing the model to all training examples once per epoch.

The user interface, built using Gradio, provides easy access to features such as model downloading and voice cloning. It allows users to upload their custom pre-trained models or download models from the internet.

During the cloning process, users can adjust parameters such as Index Rate, which determines the level of AI accent present, Pitch change, which alters the pitch of vocals only, and overall pitch change, which affects both vocals and instrumentals. For male-to-female voice conversion, users can set the pitch to 1, and -1 for vice versa. The following steps outline the process utilized to clone a user's voice using the RVC-based platform:

Step 1: The user records their voice and employs MDX_Net or other third-party software to eliminate any unnecessary background noise. The objective is to compile a high-quality dataset containing only vocals.

Step 2: The user proceeds to train their own custom RVC model using the prepared dataset or downloads an RVC model from the internet, subsequently uploading it to our platform.

Step 3: Prior to feeding it into the RVC model, the audio file intended for cloning undergoes conversion into a suitable format.

Step 4: The user inputs parameters through the platform's interface to direct the voice cloning process.

Step 5: Finally, the RVC model generates the output by taking both the audio file and parameters as input.

IV. IMPLEMENTATION

The platform offers the option to utilize custom pre-trained Retrieval-based Voice Conversion (RVC) models for voice cloning. To train an effective voice cloning model, it is essential to curate a diverse dataset of sentences that encompass all the sounds present in the target language. These sentences should vary in length and style, covering a range of speech patterns from short factual statements to questions and exclamations. It is imperative that the chosen dataset provides comprehensive phonetic coverage to ensure the model's ability to accurately reproduce different speech patterns.

Once the dataset is finalized, it is recommended to record at least 10 minutes of high-quality audio for training purposes. To achieve this, we utilize the MDX-Net model, which is a deep learning architecture specifically designed for music demixing—a process that separates individual sources, such as vocals and instruments, from a mixed audio track. Alternatively, third-party audio editing software can also be employed for this purpose.

Subsequently, the prepared high-quality audio dataset is utilized to train a custom RVC model. Various parameters and configurations are employed to define the training process, including the frequency of model saving, the number of epochs (training iterations), and whether to cache datasets. Upon completion of the training process, a .pth file is obtained, representing the trained model.

Our platform utilizes the Torch library—a crucial tool for deep learning and neural network-related tasks—to define, train, and evaluate neural network models. Torch facilitates the loading and manipulation of pre-trained models for tasks such as speech feature extraction (using the HuBERT model) and voice conversion (using the VC model). Additionally, torch.cuda is utilized to check for the availability of CUDA-enabled GPUs and manage device settings for GPU acceleration, optimizing performance. Pre-trained model weights are loaded using torch.load from specified file paths.

The user interface (UI) is developed using Gradio, a Python library that simplifies the creation of web-based interfaces for machine-learning models. This UI allows users to conveniently input data and visualize model predictions in real time. Users can select voice models, adjust parameters such as pitch and volume, and download/upload models as needed.

Prior to initiating the voice conversion process, the model receives the audio file and relevant parameters as inputs. The platform utilizes the ffmpeg library to load the audio file specified by the user, ensuring a consistent sampling rate as per the user's specifications. The audio waveform obtained from ffmpeg's output is then converted into a NumPy array of 32-bit floating-point values. Any errors encountered during the audio loading process are handled, and if necessary, a RuntimeError is raised with a corresponding error message. Finally, the parameters—including index rate, pitch change, and overall pitch change—are obtained from the user via the UI. The RVC model takes these parameters and the audio file as input, producing the cloned output audio file, which can be played from the interface and downloaded to the local file system.

The diagram depicts the overall architecture of the voice cloning system. The user has two choices: either upload an audio file as input or paste a YouTube video link. The platform converts the video file into an audio file by extracting the vocals from it. In the enrollment stage, a speaker encoder analyzes the target speaker's voice samples to create speaker embeddings that capture their unique vocal characteristics. During conversion, the source speaker's utterance is encoded, and a retrieval module finds the most similar speaker embedding from the target speaker's set. Finally, a decoder with attention generates converted speech that incorporates the source speaker's content and the target speaker's vocal characteristics from the retrieved embedding. This retrieval-based approach allows the model to effectively convert speech to sound like another speaker even with a limited amount of target speaker data.

For training the model, the system can leverage datasets like the Harvard Sentences. Harvard Sentences - Originally

developed for research on speech synthesis and perception, it consists of 720 (not 500) phonetically balanced sentences designed to cover all phonemes in American English. Each sentence is relatively short and simple, making it easier for speech models to analyze and learn the building blocks of spoken language. While less focused on natural speech patterns compared to datasets like LJSpeech or TEDLIUM, Harvard Sentences provides a well-controlled environment for studying the fundamentals of speech synthesis.

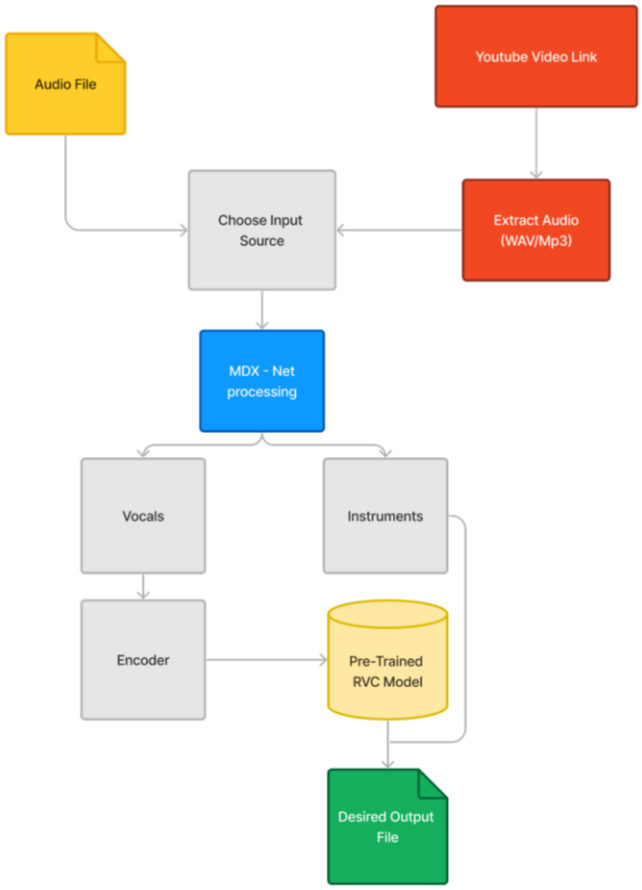

Fig.1 Modular Diagram

V RESULTS AND DISCUSSION

In comparing the results obtained from the RVC(Retrieval Voice Conversion) model with existing systems like zero-shot conversion techniques like so-vits-svc, several key observations and insights emerge.

1) The first major difference between these two approaches is that during the voice cloning process in so-vits-svc, an external software called Ultimate Vocal Remover was used to separate the vocals of the person from the audio file. Whereas in RVC, MDXNet models are used for the separation of vocals.

2) On comparing the Mel Cepstral Distortion values of the generated audio and ground-truth audio, it is observed that the MCD values of RVC are less than that of so-vits-svc.

The comparison of RVC with so-vits-svc, a zero-shot conversion technique, reveals several key insights into their methodologies and their impact on voice conversion quality.

1) Firstly, the choice of vocal separation technique seems to influence the final audio quality. RVC employs MDXNet models, which according to the results, achieve a lower Mel

Cepstral Distortion (MCD) compared to so-vits-svc's reliance on external software like Ultimate Vocal Remover. A lower MCD indicates greater similarity between the generated and target voice, suggesting that RVC's internal vocal separation might be more effective in preserving the nuances of the target voice.

2) Secondly, the observed trend in MCD values across training epochs points towards the potential advantage of RVC's approach. While both models exhibit a decrease in MCD with increased training, RVC demonstrates a more significant improvement. This suggests that RVC's retrieval-based methodology might allow for continuous learning and refinement of the converted voice as it processes more training data. In contrast, so-vits-svc, as a zero-shot technique, might be limited in its ability to adapt and improve beyond the initial training phase.

Model	**Performance Metric**
Retrieval Based Voice-Conversion	Mel Cepstral Distortion:(Lower means better) First: 250 epochs: 26.03 Second: 500 epochs: 16.58359792155391
so-vits-svc	Mel Cepstral Distortion: 250 epochs: 24.6926 500 epochs: 24.0685

Table 2. Mel Cepstral Distortion comparison

Overall, the comparison highlights the potential benefits of RVC's approach in achieving higher fidelity voice conversion. The internal vocal separation using MDXNet models and the continuous learning capability through retrieval seems to contribute to a more accurate and adaptable voice conversion process compared to the zero-shot approach of so-vits-svc. Further research could explore the specific advantages and limitations of each technique in different voice conversion scenarios.

VI. CONCLUSION & FUTURE SCOPE

The voice cloning system lets users convert speech by uploading audio or YouTube links. It analyzes a target speaker's voice to capture their unique characteristics. Then, it converts the source speaker's speech to match the target, even with limited data. Training uses datasets like Harvard Sentences, which provide basic building blocks for speech

models. In the future, the system will be capable of handling multi-lingual speeches and cloning voices from one language to another.

REFERENCES

[1] Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).

[2] Mehri, Soroush, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. "SampleRNN: An unconditional end-to-end neural audio generation model." arXiv preprint arXiv:1612.07837 (2016).

[3] Mehri, Soroush, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. "SampleRNN: An unconditional end-to-end neural audio generation model." arXiv preprint arXiv:1612.07837 (2016).

[4] Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang et al. "Tacotron: Towards end-to-end speech synthesis." arXiv preprint arXiv:1703.10135 (2017).

[5] Gibiansky, Andrew, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. "Deep voice 2: Multi-speaker neural text-to-speech." Advances in neural information processing systems 30 (2017).

[6] Taigman, Yaniv, Lior Wolf, Adam Polyak, and Eliya Nachmani. "Voiceloop: Voice fitting and synthesis via a phonological loop." arXiv preprint arXiv:1707.06588 (2017).

[7] Luong, Hieu-Thi, and Junichi Yamagishi. "Nautilus: a versatile voice cloning system." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 2967-2981.

[8] Kim, Jaehyeon, Jungil Kong, and Juhee Son. "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech." In International Conference on Machine Learning, pp. 5530-5540. PMLR, 2021.

[9] Qiu, Zeyu, Jun Tang, Yaxin Zhang, Jiaxin Li, and Xishan Bai. "A Voice Cloning Method Based on the Improved HiFi-GAN Model." Computational Intelligence and Neuroscience 2022 (2022).

[10] Chung, Hyelee, and Hosung Nam. "Zero-shot voice conversion with HuBERT." Phonetics and Speech Sciences 15, no. 3 (2023): 69-74.

[11] Ren, Zhongxi. "Selection of Optimal Solution for Example and Model of Retrieval Based Voice Conversion." In 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023), pp. 468-475. Atlantis Press, 2024.

[12] Sotelo, Jose, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. "Char2wav: End-to-end speech synthesis." (2017).

[13] Ping, Wei, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. "Deep voice 3: Scaling text-to-speech with convolutional sequence learning." arXiv preprint arXiv:1710.07654 (2017).

[14] Xie, Qicong, Xiaohai Tian, Guanghou Liu, Kun Song, Lei Xie, Zhiyong Wu, Hai Li et al. "The multi-speaker multi-style voice cloning challenge 2021." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8613-8617. IEEE, 2021.

[15] Tan, Xu, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang et al. "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).