Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-7 Year 2023** Page **419:427***

Voice Cloning Using Artificial Intelligence and Machine Learning: A Review

Fatima M Inamdar[1], Sateesh Ambesange[1], Renuka Mane[2], Hasan Hussain[3], Sahil Wagh[4], Prachi Lakhe[5]

[1]Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune.
[1]Co-founder & CEO, PragyanAI, Bangalore.
[2]School of Computer Engineering & Technology, Dr. Vishwanath Karad MIT World Peace University, Pune
[3,5]Department of Information Technology, Vishwakarma Institute of Information Technology, Pune;
[4]Department of Electronics and Telecommunication, Vishwakarma Institute of Information Technology, Pune;.
Email: fatima.inamdar@viit.ac.in, [1]sateesh.ambesange@gmail.com, Renuka.Suryawanshi@mitwpu.edu.in[2] hasanbaratopiwala53@gmail.com[3] , waghsahil196@gmail.com[4] , prachilakhe2@gmail.com[5]

**Corresponding author's E-mail: fatima.inamdar@viit.ac.in*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 13 Dec 2023

Abstract

This paper represents a thorough method for integrating emotions, texttospeech conversion, and state of the art voice cloning. The paper focuses on novel background noise adaptation, emotional voice synthesis, and multi-speaker voice cloning for better speech synthesis. The synthesis of emotive voices, multi-speaker voice cloning, and creative methods for modifying background noise to improve speech synthesis quality are among the topics covered in this study. Additionally, the study explores the domain of emotional artificial intelligence by adding a variety of emotions to artificial voices, improving user engagement through sympathetic reactions. The study also looks at how background noise can be altered to change it from a disturbing to a silent, non-disruptive state. The texttospeech systems usability in noisy conditions is greatly enhanced by this improvement. By integrating these components, the project makes a substantial contribution to text to speech, emotional AI, and voice cloning, creating new avenues for human-computer connection.

Keywords: *Voice cloning, Background voice, emotion*

1. Introduction

Artificial intelligence-powered voice cloning models are a cutting-edge development in machine learning. These models, which use a big collection of audio recordings of different people, have the astounding ability to reproduce a person's voice very perfectly. The advancement of these models lies in their capacity to replicate and simulate distinct vocal characteristics, ranging from intonation and pitch to regional accents.

Text-to-speech capabilities enable these models to convert written text into human-like speech, which has great potential for personalizing audio content and enhancing the viewing experience. With this function, you can create dynamic voiceovers and bespoke audiobooks, among many other things.

Furthermore, the creative inclusion of background noise synthesis enhances audio content that was previously difficult to create with authenticity and immersion. Artificial intelligence voice cloning models hold great promise to transform audio usage. We are working on this project because we want to use these models to give individuals in diverse places a more engaging and personalized audio experience.

Literature Survey

PAPER NAME & AUTHOR	YEAR OF PUBLICATION	PURPOSE	TECHNOLOGY USED	TOOLS	ACCURACY	ADVANTAGE & LIMITATION
Preserving Background Sound in Noice-Robust Voice Conversation via Muti-Task Learning Authors: - Jixun Yao, Yi Lei, Qing Wang, Pengcheng Guo, Ziqian Ning, Lei Xie, Hai Li, Junhui Liu, Danming Xie	2022	to close a major gap in the literature by presenting a solution to the issue of background sound preservation during voice conversion and demonstrating the enhanced performance attained with the suggested framework.	Modules for extraction of bottleneck features, SS, and VC Three steps to solve the VC background sound issue: 1) using the SS module to separate the vocal and background sound from the input signal; 2) using the VC module to convert the vocal with the bottleneck feature as an input; and 3) superimposing the converted voice on top of the background sound that the SS module retrieved.	Librosa ,pythorch	Proposed SISDR (11.11) PESQ (2.56) Scale-invariant-signal-to-distortion ratio (SISDR), and perceptual evaluation of speech quality (PESQ)	Benefits include: Create VC of the highest caliber while there is background noise. ability to adaptably control speech timbre, background noise, and language content. The suggested framework successfully bridges the gap between the upper bound and baseline in terms of ensemble quality thanks to its improved speech quality and similarity.
A Voice Cloning Method Based on the Improved HiFi-GAN ModelAuthors :- Zeyu Qiu, Jun Tang, Yaxin Zhang, Jiaxin Li, Xishan Bai	2022	Realistic-sounding artificial voices are produced using an enhanced HiFi-GAN model. These voices can be used for voice acting on computers, creating animated films, and even altering movie actors' voices. Having a	Advanced HiFi-GAN model with natural language processing and deep learning capabilities. It creates realistic artificial voices from text by utilizing big voice datasets.	HiFi-GAN Framework, TensorFlow, TTS libraries, NLP libraries	MOS(CI) 4.38 ± 0.06, PESQ 3.74 Mean opinion score (MOS), Perceptual Evaluation of Speech Quality (PESQ).	The development of high-fidelity and customisable voices is one benefit of the voice cloning project, which also improves user experiences in text-to-speech, dubbing, and personalized interactions. It helps with the preservation of historical voices, enhances

		machine speak in their own voice is also a useful tool for those who are nonverbal. This technology can even restore voice quality or allow people to speak again if they become speechless in the future. .				accessibility for people with speech impairments, and supports multilingual dubbing in entertainment. Notable constraints include, however, potential uncanny valley effects, resource-intensive training, ethical issues about misuse, and data quality. Furthermore, licensing and copyright issues may arise from voice cloning for commercial use, and this area of regulation is still developing.
VOICE CLONING Authors: - Saiesh Prabhu Verlekar, Saili Kulkarni, Varad Naik, Aaron Mendes, Saiesh Naik	2022	Natural-sounding synthetic voices are advantageous for a variety of applications, which is why simple voice cloning technology is useful. It improves content development, accessibility, customisation, and user engagement in a variety of businesses.	Speaker adaptation, Speaker encoding, Vector Quatization, Contrastive predictive coding	Scikit-learn, kaldi	After a careful analysis of MOS, it can be said that although the voice that the system cloned is quite close to the original human voice, it is lacking in naturalness and accent—two aspects that can be worked on. logy provides realistic-sounding synthetic voices for a variety of applications. It improves content development, accessibility, customisation, and user engagement in a variety of businesses.	Although it takes longer to see results, speaker cloning produces audio that is higher quality than speaker encoding.
Synthetic Speech		The development	Deep learning-		Deep learning-	A novel and promising

Detection Through Emotion Recognition: A Semantic Approach Authors:Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew Stamm, Stefano Tubaro	2022	of a novel synthetic speech detection system that makes use of emotion recognition elements is discussed in the study.	based emotion recognition model and a novel transfer-learning approach		based emotion recognition model and a novel transfer-learning approach	method for detecting synthetic speech is presented in this research. The suggested method is resilient to cross-dataset scenarios and can get high accuracy on a range of datasets. Nevertheless, there are certain drawbacks to the system, including its dependence on a SER system that has already been trained and its susceptibility to artificial speech produced by cutting-edge methods.
Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward Authors: Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza	2021	This paper offers a thorough examination and in-depth study of current deepfake generating tools and machine learning (ML)-based approaches, as well as the techniques for identifying these alterations in both audio and visual deepfakes.	Deep Learning, Generative Adversarial Networks (GAN), for generating deepfakes12	TensorFlow or Keras, open-source trained models, economical computing infrastructure, and the rapid evolution of deep-learning (DL) methods, especially.		The limits of deepfake production and detection methods as of right now are covered in the paper. But it doesn't offer precise restrictions for every approach that is covered.
Transfer Learning from Speech Synthesis to Voice Conversio	2021	introducing a novel voice conversion method that offers advantages in terms of	Encoder decoder architecture, Sequence to sequence models, Nonparallel	Gaussian mixture model (GMM),	Model – AutoVC 15.59%(Best), 27.24%(worst) PPG-VC 11.55%(best),59.78%(worst)	Benefits include: This study provides a unique TTS-VC

n with Non-Parallel Training Data Authors: - Mingyang Zhang, Member, IEEE, Yi Zhou, Student Member, IEEE, Li Zhao, and Haizhou L		flexibility, training effectiveness, and superior voice conversion performance. This method uses a text-to-speech system for training.	data processing (Variational autoencoder), Frequency warping, deep neural networks (DNN)		TTL-VC 43.05%(best),4.44 %(worst) MS-TTS 29.81%(best), 8.54%(worst)	transfer learning method with a more straightforward run-time inference system design. greater efficiency compared to other systems with comparable architecture. The suggested approach offers prosodic rendering in addition to excellent spectral mapping.
Voice Cloning: A Multi-Speaker Text-To-Speech Synthesis Approach Based on Transfer Learning Author: - Giuseppe Ruggiero, Enrico Zovato, Luigi Di Caro, Vincent Pollet	2021	to introduce a novel technique for text-to-speech (TTS) that seeks to allow a system to mimic the voices of several speakers without requiring intensive, individual training for each speaker. As a result, TTS technology might be greatly enhanced by being able to generate a variety of voices without having to record and retrain for every new speaker.	Speaker encoder, Synthesizer, Neural vocoder	Librosa, PyTorch	Baseline MSS 2.59 ± 1.03 Proposed MSS 3.17 ± 0.97 MSS – mean similarity score	An advantage The system consists of a neural vocoder model, a sequence-to-sequence with attention architecture, and an independently trained speech encoder network. a transfer learning method based on utterance embeddings rather than speaker embeddings from a speaker-discriminative encoder model Both the vocoder and the synthesizer can produce high-quality speech even from speakers who

						have never been seen before. Exception: - Comparing the suggested system to a single speaker, it falls short of human-level naturalness. The speaker prosody of the target audio cannot be replicated by the technology.
Real Time Voice Cloning Authors: - Kaushik Daspute, Hemang Pandit, Shweta shinde	2020	Voice synthesis or voice cloning technology, sometimes referred to as real-time voice cloning, has a variety of uses. It entails producing a computer-generated voice that imitates a certain person's tone and speaking patterns.	Coding and implementing Encoder module, Synthesizer Module and vocoder module.	TensorFlow GPU, Umap-Learn, Visdom, Webrtcvad, Librosa, Sounddevice, Unidecode, PyTorch, Inflect	Naturalness 4.10 ± 0.12 Similarity 2.19 ± 0.20	There was some odd prosody, and the framework's voice cloning ability was passably decent but not as good as techniques that used greater reference speech time.
Voice Converter Using DeepSpeech and Tacotron Author: Sree Nithy Chandran	2020	The goal of this project is to propose a voice conversion pipeline that uses Tacotron and DeepSrpeech models to transform audio signals from input speakers into text, which is then used to synthesize speech in the target speaker's voice.	DeepSpeech,Tacotron		Using a dataset of ten speakers, the suggested voice conversion pipeline was assessed and shown to be highly accurate in producing synthetic speech. For speaker recognition and voice similarity, the voice conversion pipeline's average accuracy was 95.2% and 92.5%, respectively.	A novel and promising method for voice conversion is the voice conversion pipeline that has been suggested. It doesn't require a lot of training data and is reasonably easy to deploy. It can also be used to instantly translate the voice of an arbitrary source speaker into the voice of an

						arbitrary target speaker.
Designing an Emotionally Realistic Chatbot Framework to Enhance Its Believability with AIML and Information States Authors: - RhioSutoyoa , Andry Chowandaa,, Agnes Kurniatia , Rini Wongso	2019	To make the chatbot more believable by adding feelings and highlighting the advantages of this	NLP techniques (e.g., pattern matching, rule-based, statistical methods, machine learning)	NLTK (Natural language toolkit), SpaCy, SpaCy Matcher, Tensorflow	63.33% of the respondents perceived Aero and Iris as two different individuals.	Benefits include: An emotive dialogue system is conceivable for a chatbot. Because chatbots are modular, they may be easily integrated with any current system, including chat apps, websites, mobile phones, virtual human systems, and more.

Existing System And Algorithm

The existing system for Voice Cloning uses following algorithm:

1. Advanced HiFi-GAN model: One kind of generative adversarial network that uses deep learning and natural language processing is the High-Fidelity Generative Adversarial Network. It makes use of big voice datasets to produce realistic-sounding synthetic voices.
2. Deep learning-based emotion recognition model: It is a machine learning system that can identify and comprehend human emotions from a variety of inputs, including text, audio, and photographs, by using deep learning techniques.
3. Sequence-to-Sequence Models: An input sequence is mapped to an output sequence in this class of models. They are trained to generate an output sequence given an input sequence and frequently employ an encoder-decoder architecture.
4. Bottleneck Feature Extraction Module: Neural networks, such as convolutional or recurrent neural networks, and deep learning techniques are commonly used in this module. Variational autoencoders can also be used to extract features.To extract important features from audio streams, utilize the Bottleneck Feature Extraction Module. These elements record significant aspects of the voice, like prosodic or phonetic information. This module converts unstructured audio data into a structured representation so that the AI system can process it further.
5. Sound Separation Module: Technology Used: Machine learning and signal processing techniques are used to separate sounds. For this, deep learning models like recurrent neural networks or convolutional neural networks may be used. In an audio transmission, the SS Module isolates vocal sounds from background noise or music. It filters out extraneous noise while identifying and isolating the voice component. This is essential for producing clean audio for voice conversion or additional processing.
6. Voice Conversion Module: Technology Used: Deep learning methods, including generative adversarial networks or deep neural networks, are frequently used in voice conversion. These models are trained to switch between voices. The vocal component split by the SS Module is taken by the VC Module, which then alters it based on the voice or emotion that is selected. It

keeps the voice natural while converting it. After that, the altered speech can be overlaid over the ambient noise to provide a fully customized audio output.

Deep Neural Networks: They are used in voice cloning to create or clone human speech or voice. DNNs are used in voice conversion models or text-to-speech synthesis systems in this context. These models translate text to speech or alter a given voice to mimic the voice of another speaker using DNN architectures.

2. Materials And Methods

Several steps are included in the suggested process for creating a voice cloning model to produce a flexible and interesting voice synthesis system. First, a diverse dataset of text and audio recordings, representing a range of emotions and speaking styles, are assembled as part of the data collection and preprocessing phase. To produce accurate transcriptions, this data is carefully cleaned and preprocessed to guarantee that it is of the highest quality and noise-free. Text-to-Voice Conversion is the following essential element. Here, we use sophisticated deep learning methods to convert text to realistic speech, like Tacotron and WaveNet. The gathered dataset is then used to fine-tune the TTS model, improving voice quality and naturalness. The AI cloning model is advanced by voice transformation and customized voice creation. To generate latent representations of the voices in the dataset, we use generative models such as the Variational Autoencoder. Custom voices can be created by altering latent variables, giving users the ability to create completely unique and personalized voices. A voice similarity metric is used to measure how close a user's preferred voice is to custom voices to guarantee accuracy. This process for creating custom voices is further improved by user feedback.

Emotion Infusion, a crucial component for expressing subtle emotional overtones in synthesized speech, is also included in the model. Sentiment analysis is one of the Natural Language Processing techniques used to interpret the intended emotions in the input text. To further detect emotional cues in the text, a deep learning-based emotion recognition model is applied. These observations are then utilized to modify the voice generation process, modifying prosody, intonation, and tempo to precisely correspond with the designated emotions.

The last topic discussed is Background Noise Generation, which improves audio quality overall. To separate background and vocal sounds from the input audio, a Sound Separation module is developed. The vocal component is altered according to the voice or emotion that is chosen using a Voice Conversion module. The transformed voice is placed on top of the background noise that the SS module has retrieved. Maintaining the intended audio quality while ensuring a smooth and organic audio output is the guarantee of fine-tuning.

With the help of this extensive methodology, a voice cloning model that excels in text-to-voice conversion, voice transformation, custom voice creation, emotion infusion, and background noise generation will be created, providing a voice synthesis experience that is both user-friendly and emotionally rich

3. Results and Discussion

The application of state-of-the-art technologies to voice cloning has produced remarkable outcomes. Profoundly high-fidelity synthetic voices are generated by the Advanced HiFi-GAN model, and emotional depth is added by deep learning-based emotion identification. Contextually correct voice production is ensured via sequence-to-sequence models. Various voice generation is facilitated by variational autoencoders. Deep Neural Networks are the best in speech synthesis and conversion; they create synthetic voices that sound realistic, are emotionally expressive, and are contextually accurate. The development of fusion technology One development in voice synthesis is speech cloning. Moral worries concerning consent requirements and misuse are significant. Research must carry on in order to make artificial sounds more robust and natural-sounding. It's critical to find ethical solutions and provide guidelines for responsible usage. Although the technology shows promise, more advancements and ethical behavior are need to complete the task.

4. Conclusion

A comprehensive analysis of numerous voice cloning research investigations has been conducted, with the results synthesized and evaluated critically. Several important conclusions and trends have been identified after a careful examination of the body of literature. Furthermore, the strengths and weaknesses of the current study have been emphasized in this review. While knowledge of voice cloning techniques has advanced significantly. For scholars and practitioners who wish to work in this topic, this review is a useful resource. This work advances our knowledge of voice cloning methods and provides a roadmap for deploying chatbots powered by artificial intelligence.

References:

Preserving Background Sound in Noice-Robust Voice Conversation via Muti-Task Learning ,Jixun Yao, Yi Lei, Qing Wang, Pengcheng Guo, Ziqian Ning, Lei Xie, Hai Li, Junhui Liu, Danming Xie

[2] A Voice Cloning Method Based on the Improved HiFi-GAN Model ,Zeyu Qiu, Jun Tang, Yaxin Zhang, Jiaxin Li, Xishan Bai

[3] VOICE CLONING ,Saiesh Prabhu Verlekar, Saili Kulkarni, Varad Naik, Aaron Mendes, Saiesh Naik

[4] Synthetic Speech Detection Through Emotion Recognition: A Semantic Approach , Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew Stamm, Stefano Tubaro

[5] Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward ,Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza

[6] Transfer Learning from Speech Synthesis to Voice Conversion with Non-Parallel Training Data, Mingyang Zhang, Member, IEEE, Yi Zhou, Student Member, IEEE, Li Zhao, and Haizhou L

[7] Voice Cloning: A Multi-Speaker Text-To-Speech Synthesis Approach Based on Transfer Learning, Giuseppe Ruggiero, Enrico Zovato, Luigi Di Caro, Vincent Pollet

[8] Real Time Voice Cloning, Kaushik Daspute, Hemang Pandit, Shweta shinde

[9] Voice Converter Using DeepSpeech and Tacotron, Sree Nithy Chandran

[10] Designing an Emotionally Realistic Chatbot Framework to Enhance Its Believability with AIML and Information States ,Rhio Sutoyoa , Andry Chowandaa,, Agnes Kurniatia , Rini Wongso