International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 3s, 2025
https://www.theaspd.com/ijes.php

Advanced Voice Cloning And Transcription Using Deep Learning: Implementation For High-Fidelity Speech Synthesis

Sreenivasa Rao Kakumanu[1*], Jupalli Pushpakumari[2], Kolli Veena[3], RamaRao Tandu[4], Padma TNS[5]

[1]Assistant Professor Department of CSE-AIML & IoT, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering & Technology Hyderabad, Telengana-500090,India*
Ph.D Scholar Department of IT, Andhra University Visakhapatnam, Email: cnu.kakumanu@gmail.com
[2]Assistant Professor, Department of Cse-Aiml & Iot Vnr Vignana Jyothi Institute Of Engineering And Technology, Email: pushpasahithi@gmail.com
[3]Assistant Professor, Department of Computer Science and Engineering, Hyderabad Institute of Technology and Management, Hyderabad, Telangana, India, Email: veenachowdary93@gmail.com
[4]CVR College of Engineering, Sr. Asst. Professor Dept: Computer Science and Engineering,
Email: ramaraotandu@cvr.ac.in, Pursuing Ph.D. in Cse, Gitam Deemed to Be University, Visakhapatnam.
[5]Assistant Professor Department of Cse- Data Science Sreenidhi Institute of Science and Technology Hyderabad Email: padmathandu@sreenidhi.edu.in

***Corresponding Author:** Sreenivasa Rao Kakumanu*
** Email: cnu.kakumanu@gmail.com*

Abstract
State of the art voice cloning methodologies employed conventional concatenative and parametric synthesis techniques, effective and efficient though they have remained, producing mechanical, if somewhat constrained speech. Advancements in deep learning however have allowed modern TTS systems to employ neural nets for the generation of richer, more natural and communicative speech. The approach presented here is for creating a sophisticated voice-cloning technique that integrates model found in NVIDIA's Tacotron2 and HiFi-GAN to achieve very natural-like speech synthesis. Tacotron2 uses a sequence-to-sequence architecture with an attention mechanism that converts the text into mel-spectrograms. Subsequently, the conversion of such spectrograms to real waveforms is done through a GAN-based vocoder called HiFi-GAN. It also incorporates other techniques such as denoising and super-resolution to enhance the quality of audio output in terms of clarity and naturalness. This work has further evaluation concerning its performance based on the RMS Loss during text-to-speech conversions. The resulting system shows striking improvements over state-of-the-art methods that achieve much better quality and efficiency toward practical applications for voice reproduction and digital communications.

Keywords : *Speech Synthesis, Tacotron2, HiFi-GAN, Mel-Spectrogram, Text-to-Speech (TTS)*

1. Introduction

The domain of voice cloning and speech synthesis has advanced considerably, transitioning from conventional concatenative and parametric synthesis methods to highly sophisticated neural network-driven techniques. Initial text-to-speech systems were based on concatenative approaches, which used the concatenation of pre-recorded speech segments, as well as parametric models, which generate speech according to predetermined rules. However, these approaches often suffered from a lack of natural intonation and expressiveness, making the voices sound rather mechanical. Recognizing these limitations, researchers began to explore machine learning techniques in voice cloning research and made contributions toward the development of noise-robust voice conversion systems. An example is as follows: Yao et al. [12] designed a multi-task learning approach to maintain interfering ambient noises during voice conversion without degrading audio quality.

With the advent of deep learning, TTS systems have furthered their development to yield high-fidelity, adaptive speech. Some of the more key innovations include models such as Tacotron2, a sequence-to-sequence model with an attention mechanism generating mel-spectrograms directly from text input, and HiFi-GAN, a GAN designed to synthesize high-quality waveform. This work uses the architecture of Tacotron2 with deep learning techniques

to create realistic and expressive audio output. Qiu et al. [13] also advanced this method by developing additional features for HiFi-GAN, which enhance the voices with high authenticity and realistic nature. Verlekar et al. [10] showed the real-time voice cloning of possibilities to generate content and also improve accessibility.

In the current work, we propose a new voice-cloning method by integrating NVIDIA's Tacotron2 with HiFi-GAN to achieve better effectiveness and quality of speech synthesis systems. Traditionally, as Shen et al. [1] mention, the vocoder of Tacotron2 used a WaveNet-based mixture of logistic distributions (MoL) vocoder to map the mel-spectrogram produced by the model to speech close to natural human voices. This could be computationally expensive and costly in terms of both training and inference time.

We, therefore replaced WaveNet MoL vocoder with state-of-the-art GAN called HiFi-GAN. It is renowned for outstanding performance in audio quality and real-time synthesis. This uses a combination of adversarial training with a multi-scale discriminator to generate high-fidelity audio signals from mel-spectrograms. Thus, speech synthesis becomes more natural and clearer at significantly lower computational costs.

The attention mechanism used by Tacotron2 enables accurate text-to-speech mapping by transforming textual sequences into mel-spectrogram representations, which HiFi-GAN subsequently synthesizes into high-quality waveforms. In addition, our methodology includes additional post-processing techniques, including denoising and super-resolution, which are designed to enhance audio clarity and reduce artifacts. We show that our system is usable in the production of realistic and expressive speech by measuring the model effectiveness using RMS Loss, MOS, Encoder-Decoder alignment. This further has massive applications in voice cloning and digital communications. We describe an all-inclusive methodology for high-fidelity voice cloning by integrating NVIDIA's Tacotron2 for generating mel-spectrograms and HiFi-GAN for synthesizing waveforms.

In the following sections we delve into Related Work, Section 2: The recent works and the innovation that have developed the voice cloning and speech synthesis bring a context to how our model is chosen. Proposed Methods, Section 3: It describes our model architecture, training procedures, post-processing techniques, including denoising and super-resolution methods for enhancing output quality. In Section 4, Results, we present a benchmarking of our model by applying metrics such as Root Mean Square (RMS) Loss to confirm both the alignment accuracy and audio quality. In Section 5, Conclusion, we summarize our findings, discussing the impact of the model and future work possibilities, ending with References in Section 6.

2. Related Work:

Voice cloning and multi-speaker text-to-speech synthesis. A total breakthrough in the field was made in 2018 by Arik et al. [2], who correctly approached the challenge of synthesizing high-quality voice clones using minimal audio data. Their work lay in speaker adaptation and encoding that was to be the root of efficient voice cloning techniques. Similarly, Jia et al. [3] have added by using the power of transfer learning from speaker verification to develop TTS systems that are capable of adaptation with unseen speakers, thereby leading to widespread applications of TTS technology.

More recently, in 2019 Qian et al. [4] introduced the groundbreaking AUTOVC framework, using an autoencoder to simplify the training process for zero-shot voice style transfer. One further advance toward many-to-many voice conversion applications, thereby making more accessible. Chen et al. [5] addressed the multilingual aspect of speech synthesis, suggesting the possibility of generalising voice characteristics across languages and speaker identity and naturalness.

It also saw Sutoyoa et al. [6] progress into a belief-realistic dialogue framework for the chatbot and advanced emotionally realistic dialogue, which further established the direction of voice synthesis technology.

At the time, this field had matured in 2021, where new techniques address limitations in earlier models. Ruggiero et al. [7] applied transfer learning for multi-speaker TTS synthesis: demonstrating how systems could generate synthetic speech even for voices unheard during training. Zhang et al. [8] refined it further by applying transfer

learning from speech synthesis to voice conversion using non-parallel training data, improving flexibility and training efficiency. The years 2022 observed a rise in research focused over the developing and discovering of synthetic speech technologies. Masood et al. [9] presented a comprehensive review about deepfake generation and detection techniques, identifying challenges and ethical implications involved through this technology. Prabhu Verlekar et al. [10] showed prospects about real-time voice cloning techniques and their applications focusing on user interaction and practical industry use cases. Conti et al. [11] addressed the crucial need of synthesizing speech detection using emotion recognition, a very new approach to this problem, which performed very satisfactorily in cross-dataset testing. Yao et al. [12] proposed a multi-task learning framework to address the background sound preservation in the voice conversion problem and achieved high quality speech even in quite noisy environments.

Qiu et al. [13] took this one step further by developing an advanced HiFi-GAN voice-cloning model that delivers realistic artificial voices for applications including dubbing and personalized TTS.
Until 2023, voice cloning and TTS synthesis continued advancing detection and anonymization techniques. Barrington et al. [14] presented methods for detecting cloned voices of single and multi-speaker, pointing out the use of perceptual features and learned representations for improving the accuracy of methods similar to these. It developed a framework for voice cloning, using traditional signal processing together with contemporary machine learning to distinguish between original and synthesized speech. Champion [15] advanced the field in 2023, evaluating speaker anonymization techniques intended for protecting identity while maintaining intelligibility and quality. Innovative strategies for protecting voices from cloning or misuse are essential for ethical applications of TTS.

In 2024, Roman et al. [16] proposed an approach to voice cloning abuse through localized watermarking by embedding imperceptible markers in synthetic speech for live detection of cloning, thus strengthening audio authentication and security in TTS systems. The watermarking results indicate the potential of detecting and ensuring accountability in voice synthesis-the tracing of synthetic speech to its origin and prevention of unauthorized use.

The field in the years evolved from basic voice cloning and speaker adaptation to complex multi-speaker and multilingual TTS models. Early approaches discussed the potential to address the underlying issues of few data requirements and speaker adaptation with the unseen speaker. Once the field matured, researchers started introducing complex models and architectures with capabilities for handling non-parallel data, preservation of background sounds, and synthesizing multilingual voices. The emphasis on real-time performance, combined with the need for robust detection mechanisms to protect against misuse of synthetic voice, marked a comprehensive evolution toward a more sophisticated and responsible application of voice synthesis technology.

3. Proposed Method

This voice cloning system with high-fidelity capability is based on NVIDIA's Tacotron2 and HiFi-GAN models. The text converted by Tacotron2 into the mel-spectrograms would be used by the model HiFi-GAN for generating high-quality waveforms. The following chapters cover architecture, dataset, training process, and various post-processing techniques to produce an acceptable realistic speech output.

3.1 Dataset Preparation

To achieve high-quality voice cloning, our model was trained on both a general and a custom dataset:
3.1.1 LJ Speech Dataset: We began with the LJ Speech Dataset, which is about 24 hours of speech recorded by a single speaker. It is a good basis for pre-training the Tacotron2 model because it contains rich audio-text pairs.
3.1.2 Custom Dataset: Next, a custom dataset was formed based on 500 samples of audio from a target speaker. These samples are transcribed using OpenAI's Whisper, and, hence, text-audio pairs are generated for the model to learn specific voice characteristics.

3.2 Model Architecture

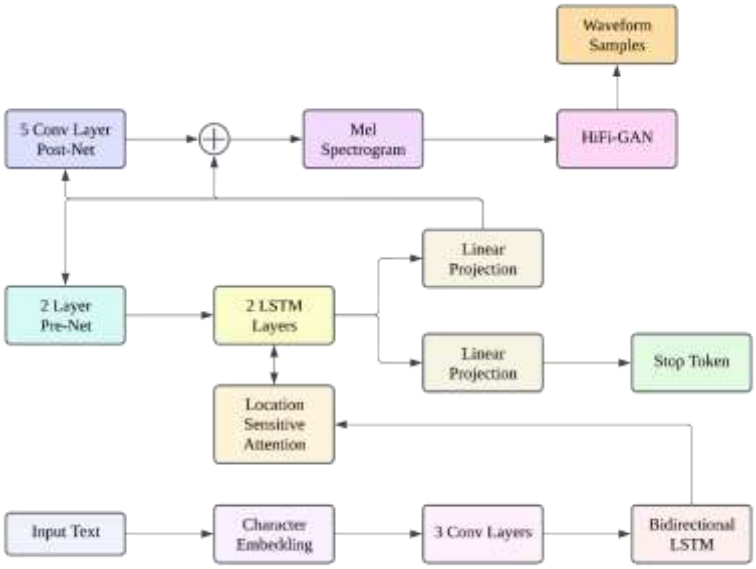

Fig 1 NVIDIA Tacotron2 with HiFi-GAN

The *Fig 1* above shows the architecture of the proposed system, where:

3.2.1 Text Input and Character Representation

The procedure starts with input text depicted as a series of character embeddings $x_1, x_2, .., x_T$, where T indicates the length of the input sequence. Every character is incorporated into a space of fixed dimensions to represent its linguistic characteristics.

3.2.2 Three Convolutional Layers

The character embeddings go through three convolutional layers, aiding in the capture of local dependencies and sequential information. This convolution process is described as:

$$\mathrm{Conv}(\mathrm{x_i}) = \mathrm{W_c} * \mathrm{x_i} + \mathrm{b_c} \tag{1}$$

Where in Equation (1), W_c and b_c represent the weights and biases of the convolution, respectively, and $*$ signifies the convolution operation.

3.2.3 Bidirectional Long Short-Term Memory

The extracted features are subsequently inputted into a Bidirectional Long Short-Term Memory ($BiLSTM$) network, which gathers contextual information from both the forward and backward directions of the input sequence. The encoder's hidden state h_i at step i is denoted as:

$$h_i = \mathrm{BiLSTM}\big(\mathrm{Conv}(x_i)\big) \tag{2}$$

The result from the $BiLSTM$ layer, $H = \{h_1, h_2, \dots, h_T\}$, creates the series of hidden states that serve as input for the attention mechanism.

3.2.4 Attention Sensitive to Location

The attention mechanism matches every decoder step with the corresponding segment of the input sequence. The attention weights α_{ij} linking the encoder output to the decoder hidden state are computed as:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})} \tag{3}$$

Where in Equation (3), e_{ij} represents a score for alignment calculated by:

$$e_{ij} = \text{Score}(s_i, h_j) \tag{4}$$

The alignment score e_{ij} can be determined using a feed-forward neural network, making sure that the model focuses on the most pertinent sections of the input sequence for producing each output. In Equation (5), the context vector c_i at the decoder for step i is a weighted combination of the outputs from the encoder:

$$c_i = \Sigma_{j=1}^{T} \alpha_{ij} h_j \tag{5}$$

3.2.5 Two-Layer Pre-Net

The context vector c_i is input into a 2-layer Pre-Net that includes two linear projection layers. This Pre-Net is designed to convert the context vector into a format that is appropriate for the LSTM layers in the decoder.

3.2.6 Two LSTM Layers

The decoder employs two LSTM layers to produce mel-spectrogram frames in an autoregressive fashion. At every step, the LSTM calculates the hidden state s_i using the prior mel-spectrogram frame y_{i-1} along with the context vector c_i as shown in the Equation (6):

$$s_i = \text{LSTM}(y_{i-1}, c_i) \tag{6}$$

3.2.7 Linear Mapping

The results from the LSTM layers are subsequently input into a linear projection layer to forecast the mel-spectrogram frames. The mel-spectrogram frame y_i at step i is defined as:

$$y_i = \text{FullyConnected}(s_i, c_i) \tag{7}$$

3.2.8 5 Conv Layer After-Net

The produced mel-spectrogram frames are further enhanced by a Post-Net that comprises five convolutional layers. This post-processing stage enhances the quality of the predicted mel-spectrograms by refining the spectral details.

3.2.9 HiFi-GAN

Ultimately, the mel-spectrograms are fed into HiFi-GAN, a neural vocoder, that transforms the mel-spectrograms into waveform samples to produce high-quality audio output.

3.2.10 Termination Token

A stop token serves to signify the conclusion of the sequence, enabling the model to conclude the creation of mel-spectrogram frames at the right moment. Every element in this framework is crucial for transforming the input text into a high-quality audio waveform, as the attention mechanism guarantees alignment while the LSTM layers produce the mel-spectrogram in an autoregressive fashion. The HiFi-GAN additionally generates the waveform, producing realistic audio from the anticipated mel-spectrogram frames.

3.3 Training Procedure

3.3.1 Tacotron2 Fine-Tuning: The model was pre-trained on the LJ Speech Dataset and fine-tuned on the custom dataset of the target speaker's voice. In this phase, we used RMS Loss to measure the variance between the predicted and target mel-spectrograms and guided the model to minimize the discrepancies. The model was fine-tuned over 19 epochs with a batch size of 32, employing a decaying learning rate to balance convergence speed and stability.

3.3.2 HiFi-GAN Training: In a nutshell, HiFi-GAN was pre-trained as a Universal model and fine-tuned using the novel mel-spectrograms obtained through Tacotron2. In training HiFi-GAN, Mean Square Error Loss and GAN Loss are used together to stimulate audio waveforms that would resemble natural speech, as long as the generator cannot be differentiated from real speech in audio quality and minimizes the errors.

3.4 Inference Process

3.4.1 Text-to-Speech Conversion: Supplied with a sequence of input text, Tacotron2 produces a matching mel-spectrogram. Following this, HiFi-GAN manipulates the spectrogram, transforming it into a believable audio waveform.

3.4.2 Post-Processing: We also made use of post-processing techniques : denoising to remove the remaining background noise introduced by synthesis and super-resolution to improve the higher frequency details of the speech, making for clearer and more natural-sounding speech.

3.5 Evaluation Metrics

3.5.1 Root Mean Square (RMS) Loss: RMS Loss was employed to gauge the difference between forecasted and real mel-spectrograms, steering the model towards concentrating on exact mel-spectrogram forecasts for precise speech synthesis.

$$RMSLoss = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (8)$$

In Equation (8):

- y_i represents actual mel-spectrogram value at the i^{th} point.
- $\hat{y}_i$ represents predicted mel-spectrogram value at the i^{th} point.
- N is the total number of data points.

3.5.2 Encoder-Decoder Alignment: We visualize the attention alignments of Tacotron2 to confirm that the model indeed mapped the input text sequences into the corresponding audio segments. Such evaluation assisted in the confirmation of clarity and naturalness of synthesized speech. Greater the intensity of colour, greater is the attention.

3.5.2.1 Significance of Encoder-Decoder Alignment

3.5.2.1.1 Precise Mapping: The alignment ensures that each text element corresponds accurately to the appropriate segment of the generated speech. This is especially important for maintaining the natural flow and intonation of the voice.

3.5.2.1.2 Quality Assurance: Visualizing the attention alignments helps diagnose issues in the model, such as whether it's focusing on the correct parts of the input text. Proper alignments indicate that the model is learning to associate text and speech components effectively.

3.5.2.1.3 Performance Evaluation: By observing alignment, researchers can assess whether the model is robust across different texts or whether it struggles with certain sentence structures.

3.5.3 Mean Opinion Score(MOS): MOS is a strong subjective quality assessment measurement in speech synthesis and telecommunication systems, widely applied for the perceived quality of synthesized or transmitted speech. The MOS is obtained from listening tests where the speech samples are rated according to a predefined scale by human listeners. In this context, it will be useful in evaluating several aspects of naturalness, clarity, and intelligibility in synthesized speech, which even an objective measure like RMS Loss cannot capture to their full potential.

$$\text{MOS} = \frac{1}{M}\sum_{j=1}^{M} S_j \tag{9}$$

In Equation (9):

- S_j represents score given by the j^{th} listener.
- M is the total number of listeners.

4. Results and Discussion

The proposed speech synthesis system-that combined Tacotron2 with HiFi-GAN-delivered successful implementation of natural and high-fidelity speech generation. Features such as efficient pre-processing, accurate transcription using Whisper, and realistic voice output through HiFi-GAN have enhanced overall performance and user experience significantly. The model shows significant improvements over its naturalness and expressiveness in speech synthesis with clear and intelligible results that are very close to human-like speech.

Our system attains a final **RMS validation loss** of 0.183548 and an **MOS** of 4.52, similar to 4.53 obtained by Shen. J. et al. [1], producing high-quality speech with natural intonation and clarity. The *Fig 2* below shows the Encoder vs Decoder alignment of the trained model.

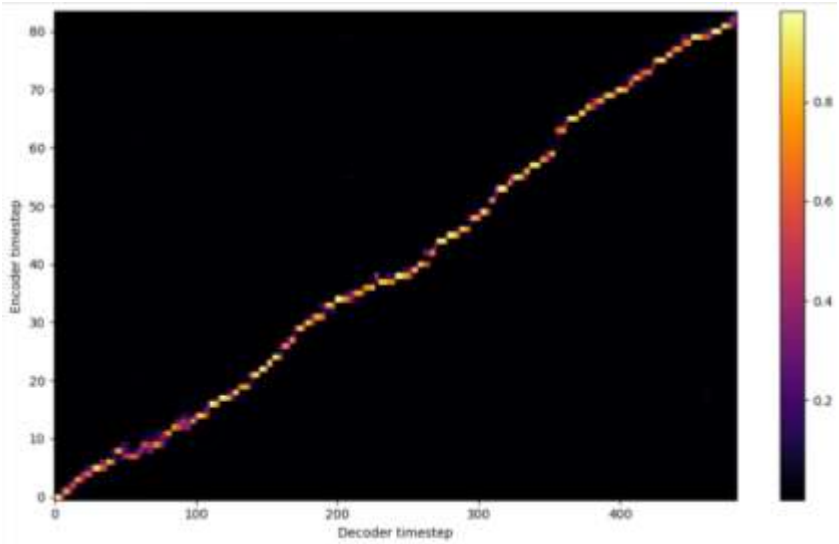

Fig 2 Encoder vs Decoder Alignment

5. Conclusion

Through extensive experimentation we are demonstrating that the proposed system could generate natural sounding expressive speech with high-quality features mirroring human's intonation and rhythm properties of mel-spectrogram close up. The performance has led to a final RMS validation loss value at 0.183548 and MOS of 4.52, which resulted in having clear and comprehensible synthesized speech. Our system further enhances audio quality by denoising and super-resolution techniques by removing background noise and higher frequency details. The listening tests proved the naturalness and clarity of the synthesized speech, making it suitable for various practical applications such as virtual assistants, audiobooks, and voice cloning.

To put it in a nutshell, the work presented in this dissertation provides a strong base from which future enhancements using deep learning for speech synthesis can be made. Some of the future studies might include increasing the range of emotions in synthesized speech, decreasing latency, and providing support for more languages and dialects. Moreover, the use of voice cloning to make bad things happen, and other ethical concerns, are going to be very important in the future.

References

[1] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2017, December 16). Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions. https://doi.org/10.48550/arXiv.1712.05884

[2] Arik, S. O., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural Voice Cloning with a Few Samples. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1802.06006

[3] Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., & Wu, Y. (2018). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1806.04558

[4] Qian, K., Zhang, Y., Chang, S., Yang, X., & Hasegawa-Johnson, M. (2019, May 14). AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. arXiv.org. https://doi.org/10.48550/arXiv.1905.05879

[5] Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R., Jia, Y., Rosenberg, A., & Ramabhadran, B. (2019, July 9). Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. arXiv.org. https://arxiv.org/abs/1907.04448

[6] Sutoyo, R., Chowanda, A., Kurniati, A., & Wongso, R. (2019). Designing an Emotionally Realistic Chatbot Framework to Enhance Its Believability with AIML and Information States. Procedia Computer Science, 157, 621–628. https://doi.org/10.1016/j.procs.2019.08.226

[7] Ruggiero, G., Zovato, E., Di Caro, L., & Pollet, V. (2021). Voice Cloning: a Multi-Speaker Text-to-Speech Synthesis Approach based on Transfer Learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2102.05630

[8] Zhang, M., Zhou, Y., Zhao, L., & Li, H. (2021). Transfer Learning From Speech Synthesis to Voice Conversion With Non-Parallel Training Data. IEEE/ACM Transactions on Audio Speech and Language Processing, 29, 1290–1302. https://doi.org/10.1109/taslp.2021.3066047

[9] Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2022). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence, 53(4), 3974–4026. https://doi.org/10.1007/s10489-022-03766-z

[10] Naik, V., Mendes, A., Kulkarni, S., Naik, S., & Verlekar, S. P. (2022). Voice Cloning in Real Time. International Journal for Research in Applied Science and Engineering Technology, 10(8), 1443–1446. https://doi.org/10.22214/ijraset.2022.44524

[11] Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F., Sarti, A., Stamm, M. C., & Tubaro, S. (2022). Deepfake Speech Detection Through Emotion Recognition: A Semantic Approach. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp43922.2022.9747186

[12] Yao, J., Lei, Y., Wang, Q., Guo, P., Ning, Z., Xie, L., Li, H., Liu, J., & Xie, D. (2022). Preserving background sound in noise-robust voice conversion via multi-task learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2211.03036

[13] Qiu, Z., Tang, J., Zhang, Y., Li, J., & Bai, X. (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

[14] Barrington, S., Barua, R., Koorma, G., & Farid, H. (2023, July 15). Single and Multi-Speaker Cloned Voice Detection: From perceptual to learned features. https://doi.org/10.48550/arXiv.2307.07683

[15] Champion, P. (2023, August 5). Anonymizing Speech: Evaluating and designing speaker anonymization techniques. https://doi.org/10.48550/arXiv.2308.04455

[16] Roman, R. S., Pierre, F., Alexandre, D., Teddy, F., Tuan, T., & Hady, E. (2024). Proactive detection of voice cloning with localized watermarking. https://doi.org/10.48550/arXiv.2401.17264