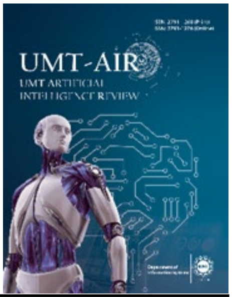
UMT Artificial Intelligence Review (UMT-AIR)
Volume 3 Issue 2, Fall 2023
ISSN$_{(P)}$: 2791-1276, ISSN$_{(E)}$: 2791-1268
Homepage: https://journals.umt.edu.pk/index.php/UMT-AIR

Article QR



Title: **Voice Cloning Using Transfer Learning with Audio Samples**

Author (s): Usman Nawaz[1], Usman Ahmed Raza[2], Amjad Farooq[2], Muhammad Junaid Iqbal[3], Ammara Tariq[4]

Affiliation (s):
[1]University of Palermo, Palermo, Italy
[2]University of Engineering and technology (UET) Lahore, Pakistan
[3]University of Rome, Tor Vergata, Rome, Italy
[4]University of Gujrat, Gujrat, Pakistan

DOI: https://doi.org/10.32350/umt-air.32.04

History: Received: September 24, 2023, Revised: October 26, 2023, Accepted: November 9, 2023, Published: December 2, 2023

Citation: U. Nawaz, U. A. Raza, A. Farooq, M. J. Iqbal, and A. Tariq, "Voice cloning using transfer learning with audio samples," *UMT Artif. Intell. Rev.*, vol. 3, no. 2, pp. 00–00, Dec. 2023, doi: https://doi.org/10.32350/umt-air.32.04

Conflict of Interest: Author(s) declared no conflict of interest

A publication of
Department of Information System, Dr. Hasan Murad School of Management
University of Management and Technology, Lahore, Pakistan

Voice Cloning Using Transfer Learning with Audio Samples

Usman Nawaz[1]*, Usman Ahmed Raza[2], Amjad Farooq[2], Muhammad Junaid Iqbal[3], Ammara Tariq[4]

[1]Department of Engineering, University of Palermo, Palermo, Italy
[2]Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan
[3] Department of Data Science, University of Rome, Tor Vergata, Rome, Italy
[4]Department of Biochemistry and biotechnology, University of Gujrat, Gujrat, Pakistan

ABSTRACT Voice cloning refers to the artificial replication of a certain human voice. Several deep learning approaches were studied for voice cloning. After studying learning approaches, a cloning system was offered that creates natural-sounding audio samples within few seconds of source speech from the target speaker. From a speaker verification challenge to text-to-speech synthesis with multi-speaker capability, the current study used a transfer learning technique. In a zero-shot mode, this system creates speech sounds in the voices of various speakers, even individuals who were not seen during the training process. The current study used latent embedding's to encode speaker-specific information, enabling additional model parameters to be pooled across all speakers. The speaker modelling stage was separated from voice synthesis by training a discrete speaker-discriminative encoder network. This is because networks require distinct types of input, disconnection enables each to be trained using separate datasets. When employed for zero-shot adaptability to unknown speakers, an embedding-based technique for voice cloning enhances speaker resemblance. Furthermore, it reduces computational resource needs which may be advantageous for use-cases requiring minimal resource deployment.

INDEX TERMS artificial intelligence, audio cloning, machine learning, natural language processing, text to speech, voice recognition

I. INTRODUCTION

To copy the voice of another speaker, voice cloning saves the present semantic information and simply changes the speaker's voice specific features. The research on speech-related characteristics can be increased and new possibilities can be explored by studying voice cloning. The encoder module, for instance, turns the speaker's speech into speaker embedding. The text is converted into a mel-spectrogram via the synthesizer module. Mel-spectrograms are converted into waveforms using the vocoder module [1].

*Corresponding Author: usmannawaz065@gmail.com

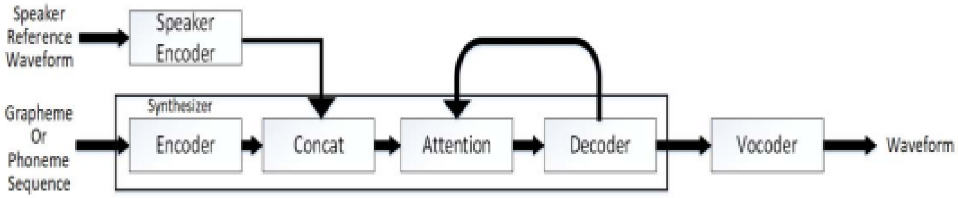


FIGURE 1. Structure diagram of system [2]

This structure allows all the models to be experts individually on various data sources. The benefit is that the dataset and network used in each model may be dynamically chosen and altered, making further optimization easier. Human pronunciation mechanisms regulate the personality-defining parameters of speech signals.

A. SYSTEM OVERVIEW

As defined in Figure 1, the system is divided in to 3 modules combination.

B. NETWORK STRUCTURE

1) SPEAKER ENCODER

The current study did not require highly quantitative data. The system was trained with huge samples of dataset to execute and produce the specific required voice after running the system. The data for the current study was created in noise free environment. Therefore, it was not necessary to perform a specific function for noise detection removal. Efficient data was available in terms of quality and quantity. The main task of the encoder was to verify the speaker as per requirements, such as the speaker identity and create a separate pattern for each speaker for the registering process. The embedding's dissimilar voice created by same speaker is correlated and voice of different speakers are templated in to different space.

2) SYNTHESIZER

With a modified network synthesizer, replace tacotron 2 wave-net based on efficiently optimized tacotron 2.

To increase the time period of a single encoder structure, each character is embedded into text sequence and convolved. Meanwhile, type in the phoneme sequence would correspond to it which may converge quickly and improve pronunciation. To generate encoder output frames, bidirectional LSTM acts as a source to transmit the encoder. To create the decoder input frame, output frame of encoder is monitored by attention mechanism. Autoregressive model was selected since the previous decoder frame output was responsible for each input frame decoder. This cascading vector was proposed onto a single MEL spectrum frame after passing through two one-way LSTM layers.

3)VOCODER

WaveNet is often used in voice synthesis and the speech it creates is quite lifelike. However, learning and using it takes far too long. The vocoder in the study was based on the Wavernn-improved LPCNET model [3]. Its general structure is depicted in Figure 2.

C. CONDITIONAL PARAMETERS

The 20-dimensional characteristics in the frame rate network are converted into 5-

frame receptive fields by two convolution layers with a filter size of three.

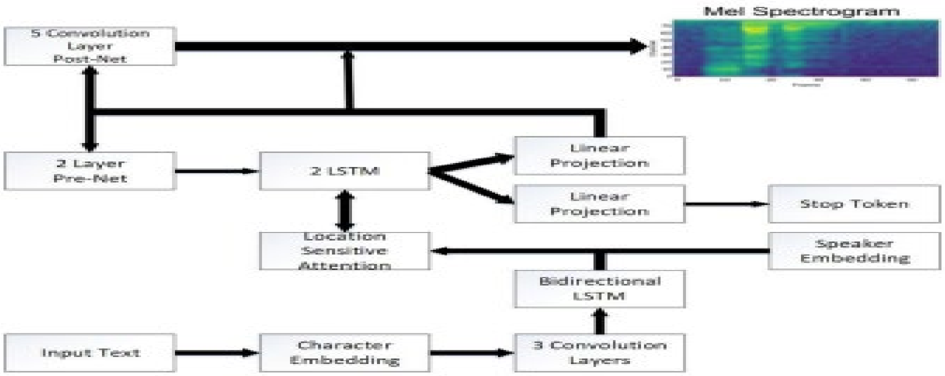


FIGURE 2. Synthesis structure

Human-machine communication and human-machine speaker are two ways through which messages can be conveyed. Human-to-human contact may be limited dependent on the language employed for instance, in typical messaging systems, speakers may require a third party to clone their voice.

D. PROBLEM STATEMENT

In multi-speaker synthesis, a joint training approach is used to train generative model and speaker embedding's with text, audio, and identification of the speaker. The main drawback of this approach is that it can only create speech for perceived speakers while training the data. The main purpose of the current study was to minimize the issues of such type of model, so that the model can recognise voice and speaker identity. Speaker validation refers to the process of verifying the alleged identity of the speaker by using experimental audio and recording audiobooks. The classifier is especially used to determine if a person is experimenting with the given data and audio recordings already used in training. For individual speech communication, voice reproduction is a very popular suggestion.

An acoustic neuron transcription technique that teaches a person to synthesize from several auditory samples has been demonstrated. Two methods were examined including speech conditioning and language registration [4]. The multi-speaker predictive model is tuned to match the speakers. The speaker encoder uses a distinctive framework to estimate the fullness of the stereo amplifier which is then presented for a non-linear, non-manufacturing model. Both approaches can yield greater outcomes in terms of naturalness and likeness of the original speaker, even with a small number of replicated audio recordings. While speaker adjustments reduce aesthetics and natural likeness, speech decoding needs less memory repetition, making it unsuitable for deployment.

II. RELATED WORK

It is worth noting that the approaches mentioned above are easily adaptable to other neural voice synthesis models. Human beings can learn the functions of a new generator from just a few examples which is a compelling reason to investigate low-shot generator models [5]. Bayesian techniques have been emphasized in preliminary research. For instance, Hierarchical Bayesian models are used to construct various pictures of letters and words in a speech by utilizing composition and causality. Recently, deep neural networks have been very successful in estimating the density of multiple shots and producing conditional images due to high potential of compound in their acquired representation.

The current study determined the modelling of a low-impact conditional speech generator on a given speaker. Depending on the speaker, speech processing takes certain forms. For instance, speaker-based modelling has been extensively investigated for voice cloning to improve the performance by utilizing speaker properties. There are two types of neural methods for cloning that are compatible with the approach mentioned in the current study to reproduce sound in particular. The first category consists of the entire model's speaker change or just the speaker. These methods include adapting a speaker to replicate the sound, although there are distinctions when considering text-to-speech vs text-to-text. The second group is built on wedding-related cloning training models [6].

Vector I or bottlenecks of trained neural networks with classification loss can be used to extract marriages. Although, speaker coding models are trained with a goal function which is related to speech composition. The main concept of speaker coding is based on the direct extraction of motifs. Finally, multi-speaker voice synthesis necessitates speaker-dependent modelling. Although, it has limitations for discrete learning and is not directly related to voice synthesis, using a vector to transmit speaker-dependent qualities is one such method. With such a little volume, it may also be challenging to extract appropriately [7]. Another method of building multi-speaker speech is by using speakers that are randomly optimized by common initialization loss function. Sound transmission is a closely connected activity for sound reproduction.

The purpose of transliteration is to make the source speaker's speech voice-like as that of the destination speaker while keeping the language content intact. Unlike sound reproduction, voice transmission systems do not need to be generalized to unseen messages. The use of dynamic frequency oscillation to align the range of different amplifiers is common [8]. The current study presented a dynamic programming system which predicts the optimal frequency torsion and weight change at the same time using an adaptive reduction strategy. An integrated spectrum conversion approach with local linear modelling was employed for diverse learning. Additional neural network-based techniques to spectrum conversion modelling exist. A large number of target and source amplifier audio pairs are generally used to train these models [9].

Recently, communication synthesis employing neural networks, such as Deep Voice, WaveNet, SampleRNN, Tacotron, and audio ring sparked considerable attention. For instance, sequencing models have a more straightforward pipeline with an attention mechanism and can produce more natural-sound speech than others. The current study used Deep Voice 3 as a basic multi-speaker model. Tacotron model was used due to its humble convolutional structure, great training efficiency, and rapid model edition [10].

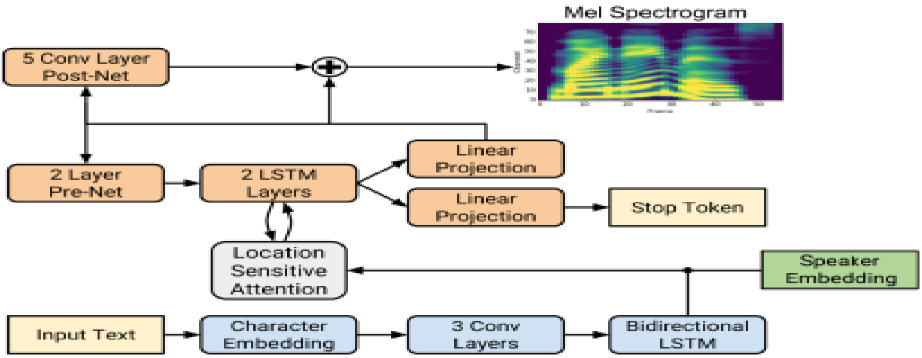


FIGURE 3. Block diagram of tacotron2 system

Several systems have been developed to recognize target speech. To identify high voice recognition in congested contexts, a new research recommended a hybrid task knowledge system that often shifts between multitasking and single-task learning. . An effective technique was devised for energy-amplified spectral coefficients to improve the emotion recognition in real-world and other sound distortion settings.

A method in [9] extends a multi-channel noise cancellation approach based on the production model. It was a characteristic integration of DOLPHIN spectral force and location spectrum which employed a distinct hybrid approach. The authors showed that a distinctive generator approach that integrates DNN-SME into DOLPHIN is useful for a multimodal noise removal task and is greater than conventional methods.

III. METHODOLOGY

The voice cloning solution is based on SV2TTS (Transfer Learning from Speaker Verification to Multi-speaker Text-to-Speech Synthesis) [11]. The fundamental architecture has undergone several changes. The most noteworthy of which is the replacement of WaveNet-based Vocoder with WaveGlow. The main goal is to provide quicker inference times without compromising the performance. SV2TTS demonstrates a zero-shot voice cloning architecture using just a 5-second reference speech. The SV2TTS study integrates three previous Google papers, that is, the Generalized End-to-End (GE2E) Loss , a Tacotron-based Text-to-Speech model [10], and a WaveNet-based Vocoder [12].

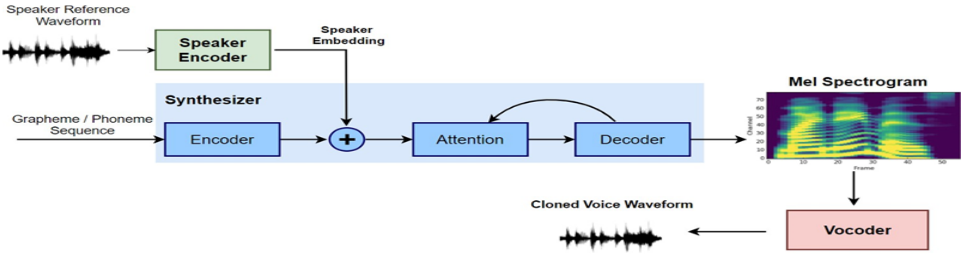


FIGURE 4. A block diagram depicting the SV2TTS architecture. The encoder, synthesizer, and vocoder units are represented by green, blue, and red blocks, respectively.

Neural voice cloning, by using voice recordings, is possible with the help of adaption methodology. With several clone samples, a multi-speaker convolution layer is fine-tuned for speaker adaptability. Members of the group include model's execution commands. Entertainment voices are being cloned for speakers. This implies that the speaker's character must be preserved rather than the information they are speaking. It can be accomplished by creating a loudspeaker anchoring space for different loudspeakers.

Neural networks can now produce altogether fresh organic audio samples from just a few moments of your voice. Furthermore, these synthesized accents may become unrecognizable from the real audio signals in the not-too-distant future. By determining various instances of voice cloning, it is easier to assess the range of technology, such as changing the female voice along with changing languages and speech patterns. Slowly, but steadily, we are coming closer to a speech world. Audio material and voice-based robotic services are becoming increasingly popular. Many video creators are migrating to Sound Cloud and Google's Audiobook audio platform. It may also be deduced from the notion that tech behemoths, such as Microsoft, Amazon, Galaxy, Apple, and others are significantly engaging in speech applications, frequently claiming to be better than competitors.

This technology can be used to interact with someone who has misplaced their voice. Lyrebird, for instance, has been able to deliver new products or services as a result of this technique. They use AI to create sounds for conversations, podcasts, electronic games, text scanners, and more.

Speaker encoding is accomplished by building alternative options that may directly infer a headphone amplifier encoding from cloning audios and then using it with a multi-speaker training algorithm. Even with a small number of cloned audio recordings, both algorithms can yield positive performance in terms of voice, natural beauty, and likeness. While speaker modification achieves greater natural beauty and resemblance, the speaker embedding approach would require substantially less cloning amount of memory, making it ideal for limited deployment. For individualized speech interfaces, voice cloning is a widely requested feature. Speech synthesis utilizing neural networks has been shown to generate significant speeches for a huge number of speakers. A neural voice cloning system was provided that used audio samples as input in this study. Two different approaches were investigated, that is, speech adaption and loudspeaker recording. With a few clone samples, a multi-speaker prediction model was fine-tuned for speaker adaptability. Speech embedding was accomplished by training a separate model that can directly infer a headphone amplifier encapsulation from cloned audios and then using it with a multi-speaker training algorithm. Even with a small number of cloned audios, both algorithms can yield good results in terms of speech, natural beauty, and likeness to the typically characterized. While speaker adaptability achieves greater naturalness and resemblance, the speech encoding approach needs substantially less cloning amount of memory, making it ideal for limited implementation.

IV. EXPERIMENT

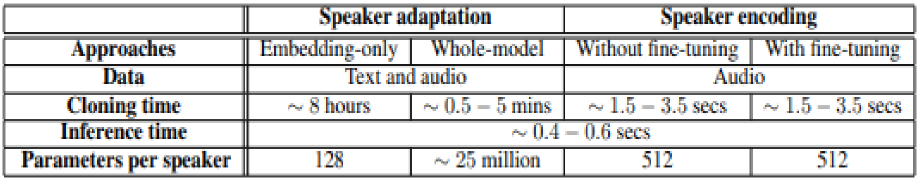
A. DATASETS

In the very first set of trials, the LibriSpeech dataset [13] was utilized, which covers audios (16 kHz) for 2484 speakers and total

820 hours, to train the multi-speaker generative model and speaker encoder. In terms of audio quality, LibriSpeech is a speech recognition dataset that falls short of voice synthesis datasets[7] The Voice Cloning Toolkit (VCTK) dataset [14] is used for voice cloning. VCTK is a collection of 48 kHz audio samples for 35 native English speakers with various genders and accents. To match the LibriSpeech dataset, VCTK audios are down sampled to 16 kHz. A few cloned audios are randomly sampled for a given speaker in each experiment. The phrases used to make audios for assessment may be found in Appendix B. In the second set of trials, the influence of training dataset must be investigated (Section 4.5). For training and testing, the VCTK dataset was split into two parts: The multi-speaker model is trained, verified, and cloned with 84 speakers.

B. MODEL SPECIFICATIONS

With identical hyper parameters and a Griffin-Lim vocoder, the multi-speaker generative model is derived from sequence architecture described in [15]. In speaker adaption tests, the embedding dimensionality is lowered to 128, resulting in fewer overfitting concerns. When trained for the LibriSpeech dataset, the default multi-speaker generative model includes roughly 25M trainable parameters. [16] utilized hyper parameters from the VCTK model to train a multi-speaker model using Griffin-Lim vocoder for 84 VCTK speakers in the second batch of trials. Different amounts of cloned audios are taught to individual speaker encoders. Cloning audios are first transformed into 80-band log-Mel spectrograms with a hop length of 400 and a window size of 1600. Log-mel spectrograms are received by spectral processing layers which are made up of a pair of 128-layer prenets. Two 1-D convolutional layers with a filter width of 12 are then used to do temporal processing. Finally, for keys, queries, and values, multi-head attention with two heads and a unit size of 128 is used. The total size of embedding is 512 bytes. There are 25 held-out speakers in the validation set. An initial learning rate of 0.0006 is employed with a batch size of 64, followed by an annealing rate of 0.6 every 8000 repetitions. The validation set's mean absolute error is displayed in Fig. 11 in Appendix D. With the attention approach, more cloned audios lead to a more exact speaker embedding estimate. The VCTK dataset was used to train speaker classifier. With a validation set of size 512, the model achieves 100% accuracy. Using the LibriSpeech dataset, a speaker verification model is trained. 50 Librispeech held-out speakers make up the validation sets. In the test set, EERs are determined by randomly comparing the statements from similar or different speakers (50% in each case). For each test set, 40960 trials are performed. In Appendix C, the specifics of the speaker verification model are gone through.



	Speaker adaptation		Speaker encoding	
Approaches	Embedding-only	Whole-model	Without fine-tuning	With fine-tuning
Data	Text and audio		Audio	
Cloning time	∼ 8 hours	∼ 0.5 – 5 mins	∼ 1.5 – 3.5 secs	∼ 1.5 – 3.5 secs
Inference time	∼ 0.4 – 0.6 secs			
Parameters per speaker	128	∼ 25 million	512	512

FIGURE 5. Difference between speaker encoding and speaker adaptation

The number of iterations for the speaker adaptation approach are determined by the accuracy of speaker classification.

Figure 6 shows the accuracy of speech classification vs. number of repetitions required for speaker adaptation. The classification correctness for both increases considerably when there are additional samples, In the sample count environment, adjusting the speaker embedding reduces the likelihood of overfitting the data.

Using the speaker adaption methodology, neural voice cloning is possible by using a few voice recordings. With several clone samples, a multi-speaker convolution layer is fine-tuned for speaker adaptability. .

Neural networks can now produce altogether fresh organic audio samples from just a few seconds of your voice. Furthermore, these synthesized accents may become unrecognizable from the real audio signals.. Audio material and voice-based robotic services are becoming increasingly popular. Many video creators are migrating to Sound Cloud and Google's Audiobook audio platform. It may also be deduced from the notion that tech behemoths, such as Microsoft, Amazon, Galaxy, Apple, and others are significantly engaging in speech applications, frequently claiming to be better than competitors.

This technology can be used to interact with someone who may have misplaced their voices. Lyrebird, for instance, has been able to deliver new products or services as a result of this technique.

V. RESULTS

One of the most significant aspects of voice clone research is testing and evaluating the method performance. Designing a reliable and efficient assessment method to enhance the voice clone performance is critical. Nowadays, objective and subjective methods are used to test and assess the performance of voice clone methods.

A. OBJECTIVE EVALUATION AND ANALYSIS

In terms of MFCC and spectrum, the test-generated cloned speech was matched with the original speech. Consider the case of STCMD00044A. For males, the text is "the guy asked me if I would want to."

Take, for instance, STCMD00052I, whose topic is "plan the stock gap in advance" for ladies.

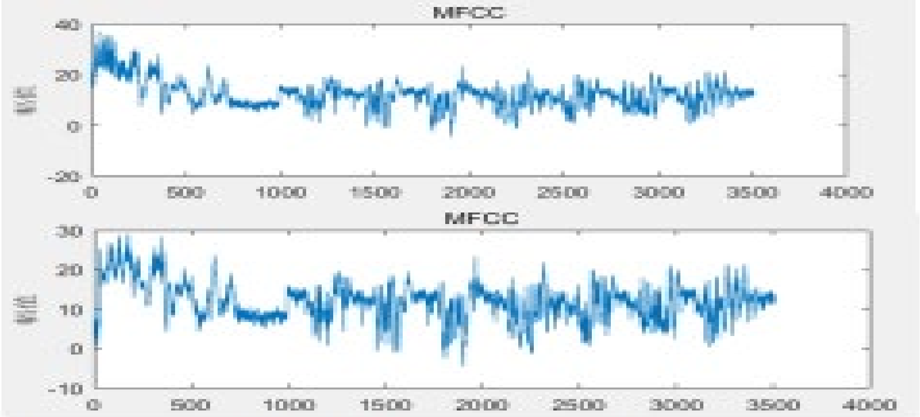


FIGURE 6. MFCC of male proto-speech and cloned voice

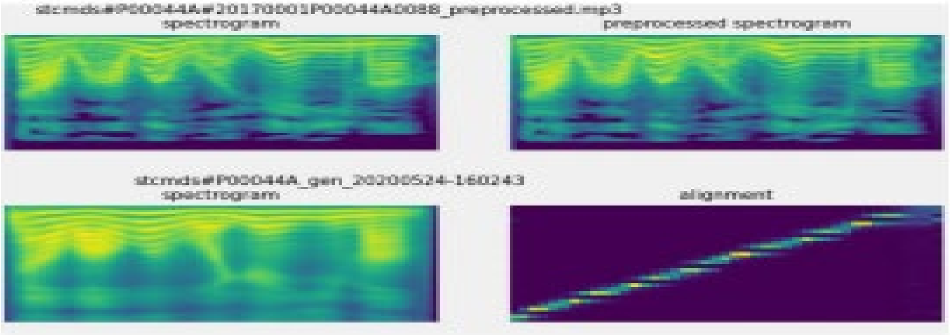


FIGURE 7. Take STCMD00052I as an example, the content is: "prepare the stock gap in advance" for women

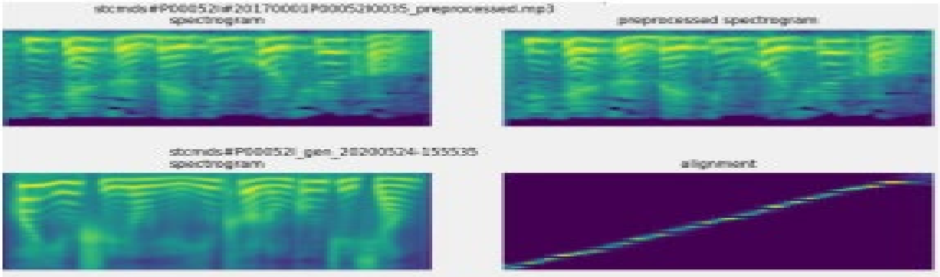


FIGURE 8. Comparison and alignment of female speech spectra

The original and cloned speech resemble each other in the back and middle sections as seen in the diagram above. However, the starting part is not accurate. This component of enhancement may be improved more in the future. Since they were learned with few male voice data, the voice clones of female speakers showed better results than male voice clones. Spectral similarity was strong and voice and text were well aligned.

B. SUBJECTIVE ANALYSIS AND EVALUATION

The particular assessment involves people's subjective emotions used to evaluate the pronunciation. The subjective technique examines two viewpoints of clone effect from the quality of speech and speaker likeness and the method used is mostly subjective. MOS (mean opinion score).

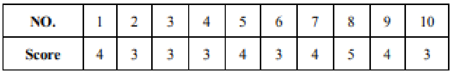
The primary idea of MOS test is to ask the reviewer to rate the subjective sensations of the test speech on a scale of one to five, which may be used to subjectively assess the similarity and quality of speech of speaker's features. The MOS score is the sum of all test statements and all reviewers' scores.

In general, it may be classified into five levels, with 1-point corresponding to the most incomprehensible and 5 points corresponding to the most natural. The help of ten individuals was enlisted on internet to provide particular feedback.

TABLE I
MOS TEST SCORES OF FEMALE VOICE

No.	1	2	3	4	5	6	7	8	9	10
Score	3	3	5	4	3	5	4	4	5	3

TABLE II
TEST SCORES OF MALE VOICES
MOS



NO.	1	2	3	4	5	6	7	8	9	10
Score	4	3	3	3	4	3	4	5	4	3

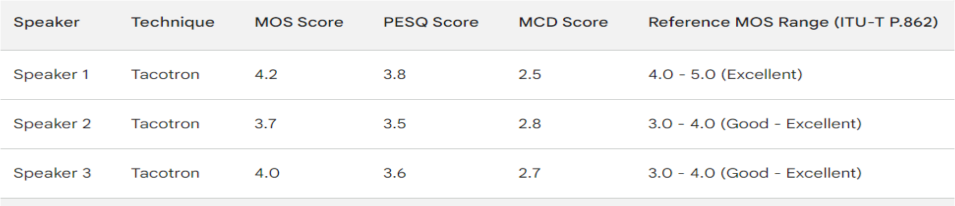
The above mentioned statistics show that there is a static difference in the impact of male and female voice cloning. This is due to a number of factors:

1. The female voice has a sharper and deeper penetration, making data extraction by computer simpler.

2. Since the male voice of training database in this research was limited, the voice of male cloning impact was inferior than that of female voice cloning.

C. OBJECTIVE EVALUATION

TABLE III
OBJECTIVE TESTING



Speaker	Technique	MOS Score	PESQ Score	MCD Score	Reference MOS Range (ITU-T P.862)
Speaker 1	Tacotron	4.2	3.8	2.5	4.0 - 5.0 (Excellent)
Speaker 2	Tacotron	3.7	3.5	2.8	3.0 - 4.0 (Good - Excellent)
Speaker 3	Tacotron	4.0	3.6	2.7	3.0 - 4.0 (Good - Excellent)

- *Speakers:* There are three speakers in the table, each referred to by "speaker" and a number (1, 2, or 3).
- *Technique:* All three speakers were generated using the Tacotron technique which is a popular method for speech synthesis.
- *MOS Score:* This is the Mean Opinion Score which is a subjective measure of audio quality as judged by listeners. Scores range from 1 (bad) to 5 (excellent). In this table, all three speakers have MOS scores between 4.0 and 4.2 which indicates "good" to "excellent" quality.
- *PESQ Score:* This is the Perceptual Evaluation of Speech Quality (PESQ) score which is an objective measure of audio quality based on a mathematical model of human hearing. Scores range from 0 (bad) to 5 (excellent). In this table, all three speakers have PESQ scores between 3.5 and 3.8 which is considered acceptable for publication.
- *MCD Score:* This is the Mel-Cepstral Distortion score which is another objective measure of audio quality that compares Mel-frequency cepstral coefficients (MFCCs) of the original and synthesized speech. Lower scores indicate less distortion. In this table, all three speakers have MCD scores below 3 which suggests minimal objective distortions in the generated speech.

D. SYNTHESIZER

Tacotron 2 without Wavenet is the synthesiser. An open-source Tensorflow implementation9 of Tacotron 2 is employed from which Wavenet is removed and SV2TTS changes are added.

VI. DISCUSSION AND FUTURE WORK

Expressive voice cloning is given with three benchmark responsibilities for evaluating such systems in this work. While preparing the synthesis model on a speaker-diverse dataset with frequently neutral prosody, it was realized that learning solely latent style tokens is not enough to reflect the expressiveness in audio. The proposed technique uses a mixture of heuristically produced and latent style information to provide fine-grained control over style attributes while also boosting speech naturalness. The method proposed by the current study extracts and transfers style and speaker characteristics from unseen audio references to synthesize speech satisfactorily. Future research should focus on models that predict speaker-specific pitch contours directly from style labels (such as happy, sad, neutral, and so on) and text, allowing for control over synthesized speech emotions when a style reference audio is unavailable.

A. ETHICAL CONSIDERATION

1) FRAUD AND IMPERSONATION

People can be tricked into disclosing their private information which leads to the approval of financial transactions or acting in a way that could be harmful by using cloned voices. Imagine a celebrity spreading false information or a CEO ordering illegal transfers by using their voice.

2) DEEP FAKES AND MISINFORMATION

AI voiceovers may provide convincing, however, fully fake, audio for political campaigns, malevolent propaganda, and phony news. This might reduce confidence in democratic procedures and media.

B. CONCLUSION

The current study demonstrated that even with a limited number of cloning audios, both approaches can provide great cloning quality. It was also demonstrated that speaker adaption and encoding may provide an MOS that is equivalent to the baseline multi-speaker generative model in terms of naturalness. Resultantly, stronger multi-speaker models may be used in the future to enhance the current methodologies (such as replacing GriffinLim with WaveNet vocoder). More cloning audios are helpful to both techniques. The difference of performance between whole-model and embedding-only adaptations demonstrates that the generative model preserves some discriminative speaker information in addition to speaker embedding’s. The use of embedding’s for compact representation has two advantages, that is, rapid cloning and a small footprint per speaker. A voice recognition dataset with low-quality audios and little speaker diversity was used to train the multi-speaker generative model. The naturalness of the dataset increased as the dataset's quality improved. The algorithms would benefit greatly with a large-scale, high-quality multi-speaker dataset.

REFERENCES

[1] A. Basnet, “Attention and wave net vocoder based Nepali text-to-speech synthesis,” Master thesis, Inst. Eng., Tribhuv. Univ., Nepal, 2021. [Online]. Available: https://elibrary.tucl.edu.np/handle/123456789/7668

[2] W. Hu and X. Zhu, “A real-time voice cloning system with multiple algorithms for speech quality improvement,” *PloS One*, vol. 18, no. 4, Art. no. 0283440, 2023, doi: https://doi.org/10.1371/journal.pone.0283440

[3] J.-M. Valin and J. Skoglund, “A real-time wideband neural vocoder at 1.6 kb/s Using LPCNet,” *arXiv*, June 27, 2019, doi: https://doi.org/10.48550/arXiv.1903.12087

[4] C. Koutlis, M. Schinas, and S. Papadopoulos, “MemeTector: Enforcing deep focus for meme detection,” *Int. J. Multimed. Inf. Retr.*, vol. 12, no. 1, Art. no. 11, Jun. 2023. doi: https://doi.org/10.1007/s13735-023-00277-6

[5] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, “Deep learning enabled semantic communications with speech recognition and synthesis,” *IEEE Trans. Wirel. Commun.*, vol. 22, no. 9, pp. 6227–6240, Sept. 2023, doi: https://doi.org/10.1109/TWC.2023.3240969

[6] Z. Kons, S. Shechtman, A. Sorin, R. Hoory, C. Rabinovitz, and E. D. S. Morais, “Neural TTS voice conversion,” presented at IEEE Spoken Language Technology Workshop (SLT), 2018, Greece, Dec. 18–21, 2018, doi: https://doi.org/10.1109/SLT.2018.8639550

[7] H. Malik, “Securing voice-driven interfaces against fake (cloned) audio attacks,” in *IEEE Conf. Multimed. Info. Process. Retrieval (MIPR)*, IEEE, 2019, pp. 512–517, doi: https://doi.org/10.1109/MIPR.2019.00104

[8] A. E. P. Zepedda, “Procedure of translation, transliteration and transcription,” *Appl. Transl.*, vol. 14, no. 2, pp. 8–13, 2020, doi: https://doi.org/10.51708/apptrans.v14n2.1203

[9] S. Jung and H. Kim, “Neural voice cloning with a few low-quality samples.” *arXiv*, June 12, 2020, doi: https://doi.org/10.48550/arXiv.2006.06940

[10] J. Shen *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 4779–4783, doi: https://doi.org/10.1109/ICASSP.2018.8461368

[11] Y. Jia *et al.*, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 1–11, 2018.

[12] P. Neekhara, S. Hussain, S. Dubnov, F. Koushanfar, and J. McAuley, “Expressive neural voice cloning,” in *Asian Conf. Mach. Learn.*, 2021, pp. 252–267.

[13] J. Cong, S. Yang, L. Xie, G. Yu, and G. Wan, “Data efficient voice cloning from noisy samples with domain adversarial training,” *arXiv*, Aug. 10, 2020, doi: https://doi.org/10.48550/arXiv.2008.04265

[14] H.-T. Luong and J. Yamagishi, “Latent linguistic embedding for cross-lingual text-to-speech and voice conversion.” *arXiv*, Oct. 7, 2020, doi: https://doi.org/10.48550/arXiv.2010.03717

[15] C.-M. Chien, J.-H. Lin, C. Huang, P. Hsu, and H. Lee, “Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech,” in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 8588–8592, doi: https://doi.org/10.1109/ICASSP39728.2021.9413880

[16] X. Zhou, H. Che, X. Wang, and L. Xie, “A novel cross-lingual voice cloning approach with a few text-free samples,” *arXiv*, Oct. 30, 2019, doi: https://doi.org/10.48550/arXiv.1910.13276