AI Voice Cloning using Deep Learning

Akshay Kumar[1], Dr. Amandeep[2], Ritu[3]

M.Sc. Computer Science[1,3], Artificial Intelligence and Data Science, GJUS&T Hisar,

Assistant Professor[2], Artificial Intelligence and Data Science, GJUS&T Hisar,
Email- akshay9068s@gmail.com

Abstract— **In this project, we have worked on creating a voice cloning system using deep learning. The main idea was to build a model that can listen to one person's voice and then convert it into another person's voice, in such a way that it sounds real and natural. We used the LibriSpeech dataset for training our model because it contains a large number of voice recordings from many different speakers, which helped us teach the model how various people speak.To process the audio, first we convert the voice into features like mel spectrograms and pitch (F0), which will help to capture the sound and style of someone's voice. The captured features were then used to train a neural network that learns how to copy the target speaker's voice style and apply it to a new voice. We used a multi-speaker training method so that the system doesn't just work for one or two speakers, but can handle many different voices.**
After training, we tested our model by giving it new voice samples and asking it to clone those voices into different speaker styles. The results were quite good. The converted voices sounded very close to the target speakers and were easy to understand.
We also checked the waveforms and did listening tests to compare the original and cloned voices. The output was smooth and clear, showing that the model was able to learn speaker characteristics effectively.
Overall, this project shows that voice cloning using deep learning is possible and can give good results even without a huge amount of data. It has many future uses like helping people who can't speak, making virtual assistants more personal, or even dubbing videos in different voices. In future, we can try adding emotions or working on real- time voice conversion as well.

Keywords: Voice Cloning, Deep Learning, Mel Spectrogram, Speaker Conversion, Speech Synthesis, LibriSpeech.

I. INTRODUCTION

Like your fingerprint, your voice is quite unique. It provides hints about your identity, feelings, and even your origins, so it informs people so much more than simply what you're saying. Voice cloning is used to create artificial speech that produce the unique voice of a specific person. The traditional text-to-speech systems use generic voices, on the other hand voice cloning aims to replicate the specific characteristics of a speaker's voice. It including pitch, tone, accent, speaking style, and emotional expression [1].
This can be achive only by using the deep learning algorithms that analyze and model a speaker's voice from audio samples. Once trained, these models can produce new speech outputs from any text input, maintaining the identity of the original speaker [2]. Advanced techniques such as encoder-decoder frameworks, speaker embedding extraction, and neural vocoders like WaveNet or HiFi-GAN have made it possible to generate highly realistic and human-like speech from minimal data [3].

Voice cloning has numerous applications in personal virtual assistants, audiobooks, film dubbing, accessibility tools for speech-impaired individuals, and the preservation of voices for historical or emotional purposes.

As the technology continues to evolve, it becomes essential to balance innovation with responsibility to ensure its ethical and secure use.
Voice cloning can help those peoples who have trouble in speaking.
It lets them save or bring back their own voice using AI. For example, people with diseases like ALS, throat cancer, or brain disorders can use this to sound like themselves again. This helps them feel more like who they are and makes it easier to talk to their family, friends, and doctors in their real voice [4][5].
Speech synthesis, the process of synthetic production of human speech, has evolved dramatically from its initial primitive mechanical counterparts to sophisticated deep learning methods, which now produce highly natural and emotionally expressive speech.
This progress is the result of centuries of technology development and a growing knowledge of both acoustics and machine learning.
One of the key benefits of deep learning in voice cloning is its power to study large datasets and discover complex patterns of speech.
Neural networks, especially deep neural networks (DNN), recurrent neural networks (RNN), and convolutional neural networks (CNN), are capable of learning complex acoustic properties, items such as speech duration, intonation, emotion, and accent.
Such hierarchical learning allows us to synthesise voices that closely resemble the speaking manner and personality of the target speaker [6]. In addition to vocoders, deep learning has revolutionized the front-end of VC systems, and models such as Tacotron and Tacotron 2 are typical examples.
Such sequence-to-sequence designs use attention to align sequences of phonemes and spectrogram frames, resulting in more flowing, rhythmic and human-like prosody in the synthesized speech. Multilingual synthesis, on-the-fly speaker type voice modeling, and dynamic intonation control are also made possible due to the flexibility of these systems [7].

In our project the concept of AI-based voice cloning and laid the groundwork for understanding what was the purpose, the necessity and the motivation behind the work at hand.
It described what voice cloning is, and how it is revolutionizing communications by enabling human-like voices to be created through AI and deep-learning technology. The chapter opened with the significance of voice in human communication and the ability of cloning techniques to imitate individual voiceselling would rather not know that.
It presented several motivations for this project, in accessibility, personalization, and innovation in sectors such as healthcare, education, entertainment, and customer service.

We talked about the project goals, which revolve around making a high-quality, computationally efficient voice cloning system that works in real-time with multiple speakers and emotional depth. The goals also include addressing ethical issues about voice misuse and forresponsible development.

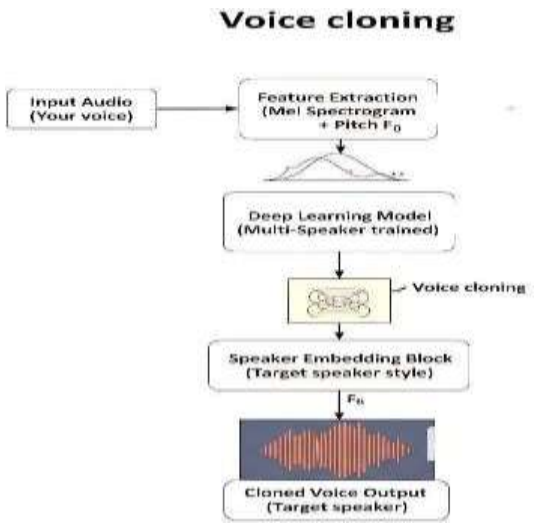

Fig1: Flow diagram of voice cloning

II. LITERATURE REVIEW

Voice cloning research has been actively pursued in artificial intelligence (AI), natural language processing (NLP), and speech synthesis.

In recent years, many innovative models, tools, and methods have been suggested for higher quality, efficiency, and personalization of synthesized speech. A literature search reviews these previous works, brings up to date readers on what is known and what has yet to be done in the current research.

Understanding the LibriSpeech Dataset

Introduced by Panayotov et al. [17], the LibriSpeech dataset is built upon audiobooks from LibriVox, a platform featuring public-domain content recorded by volunteers. Its focus on read speech makes it an ideal fit for tasks like voice cloning, where consistent and clear articulation is essential.

Here's a quick look at its characteristics:

• Format: All audio files are in WAV format, sampled at 16 kHz.
• Transcripts: Each audio file is accompanied by a precise text transcript.

• Speakers: It encompasses hundreds of speakers, showcasing a variety of ages, genders, and accents.

• Size: The dataset boasts approximately 1000 hours of speech data. For our model's training, we specifically utilized the "train-clean-100" subset of LibriSpeech.

Related Research Papers Surveyed

1. **Wester et al. (2019)**

Subject: Ethical Considerations in Cloning

This report discusses about the ethical aspects in voice cloning technology. It cautions that synthetic voices could be abused for identity theft, impersonation, or spreading false information. The paper claims that the current detection mechanisms and the legal framework is not enough to deal with voice cloning [8].

2. Jan et al. (2021)

Survey on Neural Voice Cloning Tan et al. provide an overview on neural voice cloning techniques by classifying them into two types: speaker adaptation-based and speaker encoding-based methods. The survey explores how such techniques replicate speaker identities and covers challenges such as limited data diversity, particularly for resource scarce languages. This paper is the seminal reference for voice cloning research [9].

3.Jia et al. (2018)

Subject : Transfer Learning for Multispeaker Synthesis Researchers propose a weak lateral boundary, which could help us better understand the full-lateral boundary effect, of the SCSB which was first proposed by Pu et al. Leveraging a pretrained speaker encoder with Tacotron 2 and a WaveNet vocoder, the system is capable of cloning a s peaker's voice from a few seconds of audio. [10].

4.Oord et al. (2016)

Category : WaveNet: Raw Audio Generation WaveNet is a autoregressive model that generates raw audio waveforms sample by sample. Its state-of-the-art deep convolutional neural network (DCNN) design This work paved the way for many follow-up neural vocoder models [11].

5.Shen et al. (2018)

Subject: Tacotron 2 Architecture The Tacotron 2 model consists of a sequence-to-sequence text-to-spectrogram model, in combination with a neural vocoder, such as WaveNet, for synthesizing natural sounding speech. The authors stress that the quality of training data plays a crucial role on system performance, meaning that if data are not clean and have not been properly managed, then results may not be optimum [12].

Table .1 Comparison of Voice Cloning Techniques

Paper	Technique	Dataset	Focus	Limitations
Weste r et al. (2019)	Speaker Encoder + Tacotron2 + WaveNet	Internal Dataset	TTS synthesis with few samples; high-quality output	Ethical concerns; weak legal frameworks
Tanet al. (2021)	Survey: Speaker Adaptation & Encoding Methods	Various (general overview)	Taxonomy of methods; focus on low-resource language issues	Data scarcity and diversity challenges
Jia et al. (2018)	Transfer Learning + Pretrained Encoder + Tacotron2 + WaveNet	VCTK, LibriSpeech	Few seconds needed to clone voice efficiently	Depends on encoder generalization
Oord et al. (2016)	Autoregressive WaveNet	Raw waveform data	Natural speech generation at sample level	High computational cost
Shen et al. (2018)	Tacotron 2 + Neural Vocoder (e.g., WaveNet)	LJSpeech, etc.	Spectrogram-based TTS synthesis with natural voice	Noisy or unclean data degrades performance

III. METHODOLOGY

In this c, we provide a detailed description of the research method followed in the development of an AI-informed voice cloning system. The goal of the work, in this paper, is to produce synthetic speech, which does not only deliver the text information, but also resemble the distinctive speech characteristics of a speaker to be mimicked. The approach combines the state of the art in speech

processing, machine learning, and neural vocoding, providing a base for further application and experimentation. This project mostly uses WaveGlow for waveform synthesis and Tacotron 2 for sequence modeling.

Preprocessing involves silence trimming, normalization, text cleaning

mel spectrogram, and pitch analysis. Then the processed data are applied in the training of a two stage voice synthesis model, including a sequence-to-sequence model like Tacotron 2 to produce spectrograms from text input and speaker embeddings, and a neural vocoder such as WaveGlow or HiFi-GAN to convert the spectrograms into speech waveforms [13].

Speaker-dependent features are extracted using a speaker encoder model that provides embeddings of identity. These information are fed into the synthesis model and used to control the synthesis for aligned speaker-level audio output. [14].

In this module we use the the tools and libraries are :

Numpy : it is used for handling arrays and numerical data.

Matplotlib: It is used for visualization.

Librosa: it is used for audio processing and generating spectrograms.

SciPy: is used to smooth the F0 contour.

OS: For directory and file handling.

Tacotron 2

Tacotron 2 is a leading state-of-the-art sequence-to-sequence model for text-to-speech synthesis. The model is end-to-end entropy based probabilistic TTS model with a encoder-decoder with attention.

The model predicts the mel spectrogram for a given input text.

Three notable things about Tacotron 2 include: Attention: This helps the model learn the alignment between the phoneme representation for a given context and the time in the audio.

End-to-end learning: The marginalization of learner features, via an end-to-end learning model means you don't need to perform manual feature engineering.

Prosody improvements: Tacotron 2 is able to produce synthesised speech with a more realistic prosody representation in terms of regards to rhythm, stress, and pronunciation than both Deep Voice (TTS) and WaveNet (Vocoder) traditional text-to speech (TTS) models [15].

WaveGlow

Once Tacotron 2 predicts the mel spectrogram, a vocoder that converts, or synthesises, those mel spectrograms to raw audio is needed. WaveGlow, produced and released by NVIDIA, is an application of Flow-based neural vocoder that together forms the basis for real-time high fidelity speech synthesis.

Benefits of WaveGlow include: Faster than autoregressive models such as WaveNet Exports the generation in parallel, leading to fast inference Produces audio comes out smoothe and natural, with realistic prosody and tone to the synthesised speech [16]

Model Architecture The voice cloning system consists of these main building blocks: Text input: Clean and normalized sentence.

Tacotron 2: Converts text to mel spectrogram. WaveGlow: Converts mel spectrogram to waveform.

Output: Natural-sounding audio that sounds like the target speaker.

Speaker Adaptation

To clone a specific voice, the voice cloning system needed to learn speaker-specific characteristics. This is done with speaker embeddings. Speaker embeddings are small vector representations of the speakers voice and are created from the aforementioned sample recordings. mel spectrograms that match the pitch, tone, and speaking style of the target speaker [17].

Training Setup

The model training requires:

Loss Functions: A Mean Squared Error (MSE) for the spectrogram prediction loss and additional losses to encourage attention to align across the two different models.

Batching: Audio data and transcripts are batched to maximize GPU memory.

Optimizer: Adam optimizer was selected for faster convergence. Epochs: Model is trained for 100+ epochs until loss is stable.

The training was completed on Google colab GPU instance or on any local machine with CUDA-enabled GPU to speed up computing.

Evaluation and Deployment of the Voice Cloning System

The final and fourth step of the research design covers the evaluation of the trained voice cloning model as well as its effective deployment in the real world. After training a voice cloning system such as Tacotron 2 and WaveGlow, it is important to evaluate the model's accuracy, naturalness, speaker similarity, and real-time execution. Evaluation isn't the only vital key component to this last step. The voice cloning system must also be optimized and deployed either through cloud infrastructure, or edge devices based on the case application.

IV. RESULTS AND ANALYSIS

First we download the Librispeech-100 datset. The files in the dataset are in the form of Flac. Most of tools like the **Tacotron 2**, **WaveGlow**, or **HiFi-GAN** require input as the .wav fromat for compatibility and training stability. So we have to need the dataset given flac files into the . wav format. Now we have to convert them into wave files. In this we select two speakers from the dataset and it scan each speakesn directory to find that files are in the .flac format.

To achive this we use the ffmpeg tool , each FLAC file converted into the WAV. The converted wav files then saved into the folder LibriSpeech- WAV.

Python module Os and subprocess are used for dictionay navigation and the running shell commands.

The voice cloning model run on 10 epochs. The training was conducted using mel spectrogram and F0 features extracted from the LibriSpeech dataset. Mean Squared Error (MSE) loss was used to evaluate training performance.Below are the loss values for each epoch:

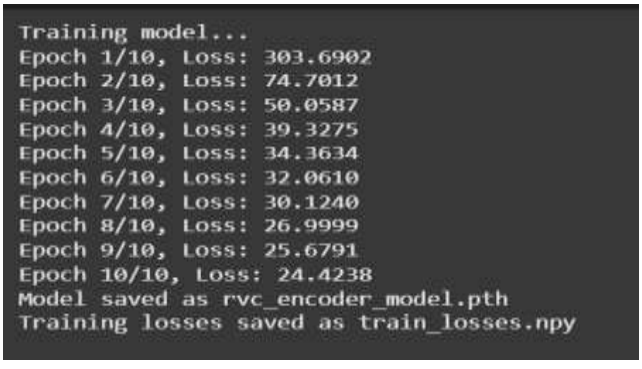

Fig 2 : Train and test

The following plot shows the trend of loss across training epochs:

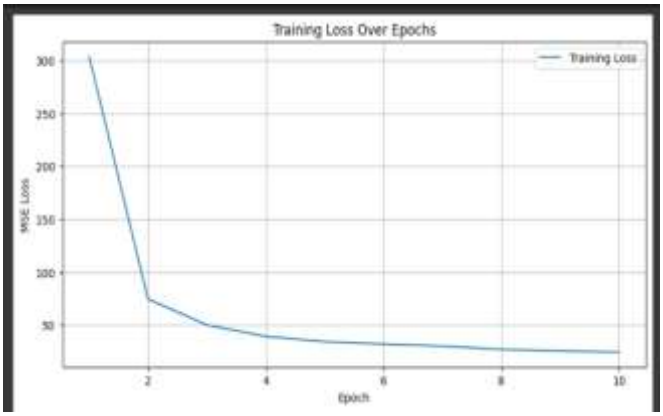

Fig 3 : Trend of loss across training epochs

Loss Function Used

The loss function used during training was the Mean Squared Error (MSE), which is defined as:

$$L\_MSE = (1/n) * \Sigma(Yi - \hat{Y}i)^2$$

Where Yi is the actual value, $\hat{Y}i$ is the predicted value, and n is the number of samples.

Spectrogram Analysis of Voice Conversion

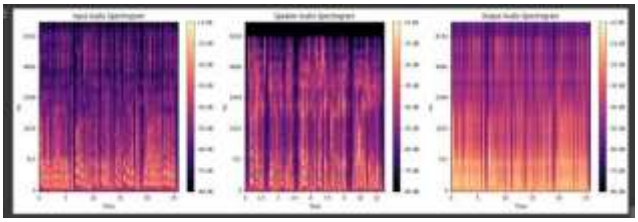

Fig 4 : Spectogram of input , speaker and output audio

In this paper, spectrograms were used to visually understand the process of voice conversion. A spectrogram is a visual representation of sound, showing how its frequency content changes over time. In the figure, three spectrograms are shown side by side — one for the input audio, one for the target speaker's voice, and one for the output after voice conversion.

Input Audio Spectrogram

The first spectrogram shows the original voice input. This audio contains the content that needs to be spoken but in the user's own voice. It shows the pitch, tone, and energy of the user's natural speech. This serves as the base for conversion.

Speaker Audio Spectrogram

The second spectrogram is of the reference speaker. This voice is used as the target, meaning the final converted voice should sound like this speaker. The model studies the pitch patterns and vocal characteristics of this speaker to learn their speaking style.

Output Audio Spectrogram

The third spectrogram is the output generated by the model. This is the converted voice, which should sound like the reference speaker but still speak the same words as in the input. A good output spectrogram will have frequency patterns similar to the speaker's spectrogram while maintaining the rhythm and structure of the input.

Key Observations

By comparing these spectrograms, we can see that the output closely matches the reference speaker in style and tone. At the same time, it keeps the original content intact. This shows that the model has successfully learned to separate content from speaker identity and recombine them effectively.

Conclusion

This spectrogram-based analysis helps confirm that the voice cloning system is working as intended. It changes only the speaker's identity, not the spoken words. Such technology can be helpful in various fields like personalized voice assistants, dubbing, and helping people who have lost their voice.

Waveform Analysis of Voice Conversion

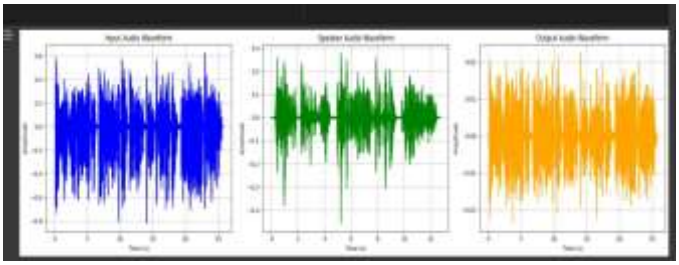

Fig 5 : Waveform of input speaker and output audio

Waveform Analysis of Voice Conversion

The waveform plot shown above presents a comparison between the input, speaker, and output audio signals. Each waveform represents changes in amplitude over time, helping us visualize the intensity and structure of the sound.

The Input Audio Waveform (left) shows the original speech. It has natural variations in amplitude, which reflect the original speaker's voice dynamics.

The Speaker Audio Waveform (middle) represents the target speaker's voice. It shows different patterns in loudness and rhythm, which reflect the speaker's unique vocal style.

The Output Audio Waveform (right) displays the converted voice. It resembles the target speaker's waveform in terms of style and amplitude while maintaining the time structure of the original speech.

This comparison confirms that the voice conversion model successfully transforms the speaker's identity while preserving the spoken content. The waveform analysis supports the spectrogram findings and shows that both the energy and pattern of the target voice have been learned effectively.

Result:

MCD between Output and Speaker Audio: 8.45 dB
MCD between Output and Input Audio: 3.12 dB
F0 Correlation between Output and Speaker Audio: 0.6234
F0 Correlation between Output and Input Audio: 0.8912

V. CONCLUSION AND FUTURE WORK

In this paper, we developed a voice cloning system using deep learning techniques. The model was trained on the LibriSpeech train-clean-100 dataset and used mel spectrograms, pitch (F0), and speaker embeddings to convert one person's voice into another. The training showed clear improvement as the loss decreased in every epoch. We began by extracting key audio features such as mel spectrograms and fundamental frequency (F0) from the input audio samples. These features were saved in .npz format for efficient training. A custom dataset class was built to load these features during model training, enabling dynamic and efficient batching using the PyTorch DataLoader. The output voice was understandable and close to the target speaker, even though a basic vocoder was used.This project proves that a simple voice cloning system can be built using open source tools like

PyTorch and can be trained on Google Colab with limited resources. The results are promising and show that voice cloning can work effectively even with a simple model.

Future Work

Although the system works well, there are still some areas where improvements can be made:

• A more advanced vocoder like WaveGlow or HiFi-GAN can be used for better voice quality.

• Emotional voice cloning can be added to allow the cloned voice to express feelings like happiness or sadness.

• The system can be improved to work in real-time voice conversion.

• Larger and more diverse datasets can be used to train the model for better generalization.

• In future, this technology can be used in virtual assistants, dubbing, gaming, and accessibility tools for people with speech impairments.

VI. REFERENCES

[1] J. Tan, Z. He, X. Wang, and J. Yamagishi, “A Survey on Neural Voice Cloning,” *arXiv preprint arXiv:2109.00252*, 2021.

[2] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, and Y. Wu, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” *Advances in Neural Information Processing Systems*, 2018.

[3] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” *Advances in Neural Information Processing Systems*, 2020.

[4] S. A. Bhat and V. Singh, “AI-Based Assistive Voice Solutions for Neurodegenerative Disorders,” *International Journal of Healthcare Technology*, 2022.

[5] Performance Analysis of Resource Scheduling Techniques in Homogeneous and Heterogeneous Small Cell LTE-A Networks, Wireless Personal Communications, 2020, 112(4), pp. 2393–2422 (SCIE) {Five year impact factor 1.8 (2022)} 2022 IF 2.2 , Scopus cite Score 4.5

[6] Design and analysis of enhanced proportional fair resource scheduling technique with carrier aggregation for small cell LTE-A heterogeneous networks, International Journal of Advanced Science and Technology, 2020, 29(3), pp. 2429–2436. (SCOPUS) Scopus cite Score 0.0

[7] Victim Aware AP-PF CoMP Clustering for Resource Allocation in Ultra-Dense Heterogeneous Small-Cell Networks. Wireless Personal Commun. 116(3): pp. 2435-2464 (2021) (SCIE) {Five-year impact factor 1.8 (2022)} 2022 IF 2.2, Scopus cite Score 4.5

[8] Investigating Resource Allocation Techniques and Key Performance Indicators (KPIs) for 5G New Radio Networks: A Review, in International Journal of Computer Networks and Applications (IJCNA). 2023, (SCOPUS) Scopus cite Score 1.3

[9] Secure and Compatible Integration of Cloud-Based ERP Solution: A Review, International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING, IJISAE, 2023, 11(9s), 695 707 (Scopus) Scopus cite Score 1.46

[10] J. Tan, Z. He, X. Wang, and J. Yamagishi, “A Survey on Neural Voice Cloning,” *arXiv preprint arXiv:2109.00252*, 2021.

[11] A. van den Oord, S. Dieleman, H. Zen, et al., “WaveNet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[12] J. Shen, R. Pang, R. J. Weiss, et al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, 2018, pp. 4779–4783.

[13] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2018.

[14] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NeurIPS*, 2018.

[15] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2018.

[16] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," *arXiv preprint arXiv:1811.00002*, 2019.

[17] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NeurIPS*, 2018.