IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

REAL-TIME VOICE CLONING USING DEEP LEARNING: A CASE STUDY

Hruthik B Gowda
Student of Computer Science and Engineering. Vidya Vikas Institute of Engineering and Technology, Mysore

Karun Datta Ramakumar
Student of Computer Science and Engineering. Vidya Vikas Institute of Engineering and Technology, Mysore

Sheethal.V
Student of Computer Science and Engineering. Vidya Vikas Institute of Engineering and Technology, Mysore

Sushma M
Student of Computer Science and Engineering. Vidya Vikas Institute of Engineering and Technology, Mysore

Dr. Madhusudhan G K
Computer Science and Engineering. Vidya Vikas Institute of Engineering and Technology, Mysore

- **ABSTRACT:**

Training a deep neural network for voice cloning typically involves using hours of professionally recorded audio from one speaker as input data. To change the cloned voice requires collecting an entirely new dataset and retraining the model - needless to say this incurs considerable expense. Recent research has developed an innovative three stage pipeline that solves these issues by allowing unseen voices to be cloned with just seconds of reference speech - all without necessitating template retraining! Furthermore, these studies have yielded highly natural sounding results which truly demonstrate its effectiveness. We intend on replicating their technique and making their methodology available open source for public use - our modified version will include adaptation with fresh vocoder models aimed towards boosting speed so that we can develop our platform into an efficient real time deep learning system capable of instantaneously performing voice cloning. Our work enables us to build upon Googles 2018 paper - we're only the second group to implement their methods publicly so far! Our system can accurately digitize and replicate any recorded speech utterance in just 5 seconds allowing for all extracted voices from this process to also perform text to speech. Our strategy involves reproducing all three stages of the model through a combination of our own implementations and open-source options. Our primary focus is on executing effective deep learning models while creating appropriate information pre-processing pipeline. Rather than focusing solely on the technical aspects of training these models lets evaluate both their benefits and drawbacks. A crucial factor is ensuring that this system can operate efficiently in real time - capturing a voice and producing speech faster than it takes to actually speak. Impressively this framework has the capability to duplicate voices not encountered during training as well as generate speech from previously unseen text.

1. INTRODUCTION:

In recent years, deep learning models have revolutionized the field of text-to-speech synthesis, allowing for more natural-sounding speech generation than traditional concatenative methods. Researchers have been focused on improving the effectiveness of these deep models, with a particular emphasis on making the speech sound more natural and training the models in an end-to-end fashion. Thanks to advancements in GPU technology, these models can now run inference hundreds of times faster than real-time on a mobile CPU, making them more practical for real-world applications. Several studies have shown that deep learning models can produce speech that is near-human in quality, with subjective metrics proving to be a better measure of speech naturalness than objective ones. While some argue that the limit of human nature has already been reached, there is still room for improvement in terms of naturalness, accuracy, and efficiency. Overall, deep learning models are transforming the field of text-to-speech synthesis, opening up new possibilities for natural language processing and human-machine communication. While a single-speaker TTS model's complete learning is theoretically a form of voice cloning, the purpose is rather to create a fixed model that can integrate new voices with little information to clone new speakers in text-to-speech synthesis, a commonly used approach is to use a pre-trained TTS template that can generalize the voice characteristics, and then condition it using a speaker encoder model. The speaker encoder model takes a reference speech as input and derives a low-dimensional embedding that represents the unique characteristics of the speaker. This embedding is then used to condition the TTS template to produce synthetic speech that sounds like the reference speaker. This method allows for voice cloning with minimal data requirements and can produce high-quality synthetic speech. Additionally, recent research has focused on using zero-shot learning methods to further reduce the data requirement for voice cloning and to improve the flexibility of the system. These developments have significant implications for applications such as personalized voice assistants and audio book narration. To derive low-dimensional embedding, a speaker encoder model leverages reference speech as input. Compared to generating separate TTS models for each individual, this approach proves more data-efficient while also being significantly faster and more affordable computationally. What's truly noteworthy is that the quantity of reference speech needed varies widely among methods, extending anywhere from several seconds to half an hour per speaker. The extent of similarity between the respective voices produced ultimately depends on this crucial aspect.

2. LITERATURE REVIEW

In [1]. The ability to replicate someone else's natural speech patterns using only a few audio samples has become increasingly popular among users seeking personalized speech interfaces aka voice cloning. This research delves into two techniques for achieving such functionality: speaker adaptation and speaker encoding. Speaker adaptation involves refining existing multi speaker generative models by tweaking them with minimal amounts of data from unheard of speakers; meanwhile the separate model created via speaker encoding produces newly embedded voices when combined with these same generative models. The study focuses specifically on voice cloning within sequence-to-sequence neural speech synthesis systems. Our contributions enhance this field significantly across three main areas: Firstly, we demonstrate how effective incorporating speaker adaptation can be into the process by fine tuning these preexisting models with data from unfamiliar speakers. Secondly, we propose a novel and innovative approach using speaker encoding that provides similar quality voices in subjective evaluations while taking significantly less time and computational resources than traditional methods. Lastly, our proposal is to utilize automated evaluation techniques that leverage neural speaker classification and speaker verification to assess the quality of voice cloning. Furthermore, we showcase the ability to achieve gender and accent transformation through embedding manipulations in our demonstration of voice morphing.

In [2]. The Research describes Tacotron 2, a neural network architecture for speech synthesis directly from text. The system is composed of a recurrent sequence-to-sequence feature prediction network that maps character embeddings to Mel-scale spectrograms, followed by a modified Wave Net model acting as a vocoder to synthesize time domain waveforms from those spectrograms. Our model achieves a mean opinion score (MOS) of 4.53 comparable to a MOS of 4.58 for professionally recorded speech. To validate our design choices, we present studies of key components of our system and evaluate the impact of using Mel spectrograms as the conditioning input to Wave Net instead of linguistic, duration, and features. We further show that using this compact acoustic intermediate representation allows for a significant reduction in the size of the Wave Net architecture. Index Terms- Tacotron 2, Wave Net, text-to-speech.

In [3]. The research We introduce a novel neural network-based system for synthesizing high quality speech in multiple speakers' voices. Our approach combines three distinct components: an independent speaker encoder network, a Tacotron 2 inspired sequence to sequence TTS synthesis network and a Wave Net powered neural vocoder. To create such high quality multispeaker TTS output our system employs discriminative training of the speaker encoder on noisy audio data sourced from thousands of unique personnel without transcripts. From mere seconds of reference audio per target speaker our encoder produces fixed-dimensional embeddings which are fed as conditions into the Tacotron 2-style synthesis models generation process. Finally, our Wave Net based vocoder takes the resulting Mel spectrograms and generates time domain waveform samples accordingly. Results show that even previously unseen target speakers can have their characteristic voices well emulated by our synthesized voice products via subjective listening tests and machine-driven evaluations based on speaker verification algorithms. With a university education under my belt, I recognize the significance of proper language usage and style when it comes to academic discourse. As such my aim is to convey the same message as before while rephrasing this passage four additional times utilizing various sentence structures. The study shows that the model can convincingly create speech from imaginary individuals who don't match those in the training set. This indicates that the model has acquired an authentic understanding of how speakers vary.

In [4]. The research a detailed explanation on the tasks and data used in the challenge, followed by a summary of submitted systems and evaluation results is presented. Four audio/text data sets and two track is provided as data. All audio data is mono, 44.1KHz sampling rate, 16 bits, equipped with transcripts. The language is Mandarin. Multi-speaker training set (MST): This part of data consists of two subsets, including the AIShell-3 [23] data set, called MSTAIShell in the challenge. The data set contains about 85 hours of Mandarin speech data from 218 common speakers, which is recorded through a high-fidelity microphone in an ordinary room with some inevitable reverberation and background noise. Target speaker validation set (TSV): For each track, there are two validation target speakers with different speaking styles. For track 1, each speaker has 100 speech samples, and for track 2, each speaker has 5 speech samples. Target speaker test set (TST): For each track, three target speakers with different speaking styles (different from those in TSV) are released for testing and ranking. Again, for track 1, each speaker has 100 speech samples, and for track 2, each speaker has 5 speech samples. Test text set (TT): This text set includes the sentences (with Pinyin annotations) provided to the participants for speech synthesis for the test speakers in TST. The sentences can be divided into three categories, namely, style, common, and intelligibility. The sentences in the style set are in-domain sentences for style imitation test. Track 1 (Few-shot track): The organizers provide two and three target speakers for voice cloning validation (TSV) and evaluation (TST) respectively. Track 2 (One-shot track): For track 2, requirements are the same as track 1, except that only 5 recordings are provided for each target speaker. The approaches used 1. Acoustic model- In the AR acoustic model category, the input phoneme sequence is first encoded by the encoder. Then the decoder generates the target spectral features in an autoregressive manner. In the submissions, Tacotron [1, 24] is the most popular one, where an encoder attention-decoder based architecture is adopted for autoregressive generation. 2. Vocoder- The vocoders used in the submitted systems can be divided into autoregressive and non-autoregressive as well. Specifically, 5 and 10 teams chose the AR and non-AR neural vocoders respectively 3. Speaker and style modelling- Robust speaker and style representations are crucial to model and generate the target voice with desired speaker identity and style.

The sentences in the style set are in-domain sentences for style imitation test. Track 1 (Few-shot track): The organizers provide two and three target speakers for voice cloning validation (TSV) and evaluation (TST) respectively. Track 2 (One-shot track): For track 2, requirements are the same as track 1, except that only 5 recordings are provided for each target speaker. The acoustic model approach involves encoding the input phoneme sequence with an encoder and then generating the target spectral features using a decoder in an autoregressive manner. The most popular system among submissions is Tacotron, which utilizes an encoder attention-decoder based architecture for autoregressive generation. In the submitted systems, Tacotron [1, 24] reigns as the most favored method. This model implements an encoder attention-decoder based architecture for autoregressive generation. The vocoders that were incorporated can be categorized into two groups: autoregressive and non-autoregressive. Among the submissions, five teams employed AR neural vocoders while ten opted for non-AR vocoders. Lastly, accurate representations of speaker and style are key to accurately create and model the target voice with specific speaker identity and style preferences.

3. METHODOLOGY

i)

Deep Learning: Deep learning is a powerful subset of machine learning that enables computers to learn from vast amounts of data and make predictions or decisions based on that learning. It involves the use of artificial neural networks, which are modelled after the structure of the human brain. Deep learning has transformed many areas of technology, including computer vision, natural language processing, and speech recognition. In particular, it has enabled breakthroughs in autonomous driving, allowing cars to detect and respond to their surroundings. It is also a key ii) technology behind voice control in a variety of consumer devices, making it possible for users to interact with their devices in a more natural and intuitive way. With its ability to learn from large datasets and make increasingly accurate predictions, deep learning has the potential to revolutionize many aspects of our lives.

iii)

Design: For effective rendering of any given text through pre-set vocal styles on our voice cloning platform, simultaneous inclusion of two crucial inputs is necessary – the first being complete textual data whilst the second one being sample audio corresponding to the elected vocal style. For streamlined performance by our technology, we require comprehensive insight into both volume patterns along with inherent linguistic features present in our input data set. With prominence on even usage among less technically adept end-users, implementation of a user-friendly interface becomes a high priority. Our system is primarily segregated into two key components - neural network training process & flask application integration. The neural network must be equipped to clone new vocal styles with remarkable accuracy while delivering natural-sounding speech outputs; whereas the flask application will be responsible for generating an interactive experience for users where they can supply their textual and voice data to produce desired outcomes.

Voice Cloning

1. The main aim of this initiative is to establish a system that can produce speech in any given speaker's voice from supplied text input. The procedure involves two principal stages; voice cloning and text to speech (TTS) synthesis. Choosing the most appropriate approach is imperative to achieve high levels of naturalness and comprehensibility in the ultimate output, which are key evaluation factors for TTS systems. Two primary methods exist for performing TTS conversion:

i) Concatenative approach:

- -Makes use of superior quality audio samples
- -Limited by data availability and lack of variation
- -Mixes fragments from various audio recordings to construct new synthesized speech. -The outcome comprises crisp and unambiguous speech but devoid of emotional intonation, which may not sound phonetically accurate. -In general, comprehensible but could have an artificial ringtone.

There are models that can be used here,

Wave net: Wave Net has emerged as a compelling option for individuals looking to generate raw audio waveforms using advanced machine learning techniques. With its ability to produce high quality speech that sounds more human like than other TTS models available today. Nevertheless, some challenges arise when attempting implementation in practical use cases due to the large datasets required for training purposes and significant computation cost involved.

Deep Voice: The Deep Voice neural TTS model, created by Baidu's Silicon Valley AI Lab (SVAIL), operates on a sequence-to-sequence learning framework with attention mechanisms. While it requires less data than Wave Net, it falls short in producing speech that is natural and easy to understand.

SV2TTS: Say hello to an advanced real-time voice cloning system that's changing the game! One of its most outstanding features is its capacity for mastering how to replicate any new speaker's voice without needing previous training samples via the ingenious application of zero-shot learning techniques. This state-of-the-art setup includes three distinct deep learning models that can be trained independently using varying data sources - giving rise to reduced reliance on high-quality multispeaker data. Thanks to this approach, the resulting synthesized speech is characterized by its exceptional quality and near-

instantaneous response times.

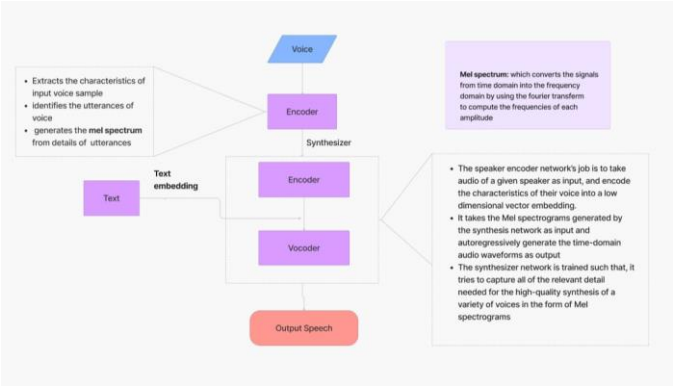

Fig:3.1. Flowchart

4. CONCLUSION

Real-time voice cloning with deep learning algorithms employed for real-time voice cloning technology has immense potential to revolutionize our interaction with computer systems. By mimicking a speaker's tone through text-to-speech function offers users an infinitely natural opportunity that allows for better engagement rates within said audio-based platforms. However; facing difficulties such as obtaining sufficient large quantities of exemplary training data while also attempting synthesized speech that remains both faithful to vocal delivery standards but still comprehensible aren't without its setbacks; still major strides have been made with new advancements like SV2TTS that have demonstrated significant potential in this field. With further refinements and research to follow, we can anticipate a proliferation of usage cases for real-time voice cloning technology across industries such as virtual assistants, gaming, and personalized voice interfaces.

REFERENCES

[1] Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples, 2018.

[2] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly,Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wave net on Mel spectrogram predictions. *CoRR*, abs/1712.05884, 2017. URL http://arxiv.org/abs/1712.05884

[3] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. CoRR, abs/1806.04558, 2018. URL http://arxiv.org/abs/1806.04558.

[4] Qicong Xie, Xiaohai Tian, Guanghou Liu, Kun Song, Lei Xie, Zhiyong Wu, Hai Li, Song Shi, Haizhou Li, Fen Hong, Hui Bu, Xin Xu https://arxiv.org/abs/2104.01818

[5] S. Shirali-Shahreza and G. Penn. Mos naturalness and the quest for human-like speech. In 2018 IEEE Spoken Language Technology Workshop (SLT), pages 346–352, Dec 2018. doi: 10.1109/SLT.2018.8639599..

[6] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, 33(2):443–445, April 1985. ISSN 0096-3518. doi: 10.1109/TASSP. 1985.1164550.

[7] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. End-to-end text- dependent speaker verification. CoRR, abs/1509.08062, 2015. URL http://arxiv. org/abs/1509.08062.

[8] S. Imai. Cepstral analysis synthesis on the Mel frequency scale. In ICASSP ’83. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 8, pages 93–96, April 1983. doi: 10.1109/ICASSP.1983.1172250.

[9] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis, 2018.

[10] Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. 02 2018.