Text Vocal Reader Cloning

Mir Adil Raza[1] Dr Ganesh D[2]
[1,2]Department of MCA
[1,2]Jain University, Bangalore, Karnataka, India

Abstract— Voice cloning is the assignment of figuring out how to orchestrate the voice of a concealed speaker from a couple of tests. While current voice cloning techniques accomplish promising outcomes in Text-to-Speech (TTS) union for another voice, these methodologies come up short on the capacity to control the expressiveness of orchestrated sound. In this work, we propose a controllable voice cloning strategy that permits fine-grained authority over different style parts of the incorporated discourse for a concealed speaker. We accomplish this by expressly molding the discourse blend model on a speaker encoding, pitch form and inactive style tokens during preparing. Through both quantitative and subjective assessments, we show that our system can be utilized for different expressive voice cloning errands utilizing a couple of deciphered or untranscribed discourse tests for another speaker. These cloning errands incorporate style move from a reference discourse, combining discourse straightforwardly from text, and fine-grained style control by controlling the style molding factors during induction.

Keywords: WaveNet, TTS, Vocoder, Spectogram

I. INTRODUCTION

Ongoing examination efforts in voice cloning have zeroed in on combining an individual's voice from a couple of reference sound examples. While such a framework can create discourse from text for another speaker, it leaves out authority over different style parts of discourse. Unequivocal command over the style parts of cloned discourse is alluring for a few applications, for example, voice-overs in energized films, blending practical and expressive discourse for DeepFake recordings, interpreting discourse starting with one language then onto the next while protecting talking style and speaker character, promotion crusades with expressive discourse in numerous voices and dialects (and so forth) Expressive voice cloning frameworks can likewise help make customized discourse interfaces with voice aides in cell phones, vehicles, and home aides. Since discourse fills in as an essential correspondence interface between AI specialists and people, the capacity to talk expressively is a truly attractive quality for voice cloning frameworks. Moreover, such frameworks can possibly engage people who have lost their capacity to talk.

II. WORKING METHODOLOGY

The three main components Speaker Encoder, Mel Spectrogram Synthesizer and Vocoder are all trained separately

A. WaveNet

WaveNet is a deep neural network for generating raw audio. It was created by researchers at London-based artificial intelligence firm DeepMind

B. Speaker Encoder

Speaker encoding is based on training a separate model to directly infer a new speaker embedding, which will be applied to a multi-speaker generative model. In terms of naturalness of the speech and similarity to the original speaker, both approaches can achieve good performance, even with a few cloning audios

C. Mel Spectrogram Synthesizer

A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. When applied to an audio signal, spectrograms are sometimes called sonographs, voiceprints, or voicegrams

D. Vocoder

A vocoder combines a recording of a human voice with a synthesized waveform to produce a robot-like effect. The Audacity free, open-source audio editing program includes a vocoder plug-in that you can use to produce this effect. ... The vocoder then modulates the left-hand channel with the right-hand one

III. WORKING FLOW

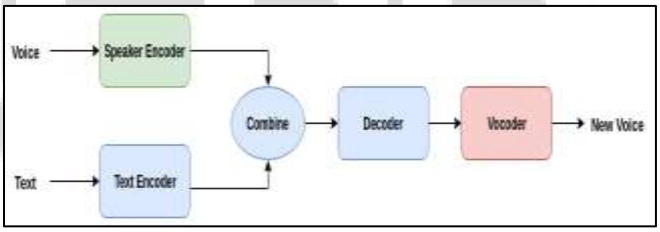

Fig. 1: work Flow

Fig. 2: View Page

IV. RESULT AND DISCUSSION

1) Given a small audio sample of the voice we wish to use, encode the voice waveform into a fixed dimensional vector representation
2) Given a piece of text, also encode it into a vector representation. Combine the two vectors of speech and text, and decode them into a Spectrogram
3) Use a Vocoder to transform the spectrogram into an audio waveform that we can listen to.

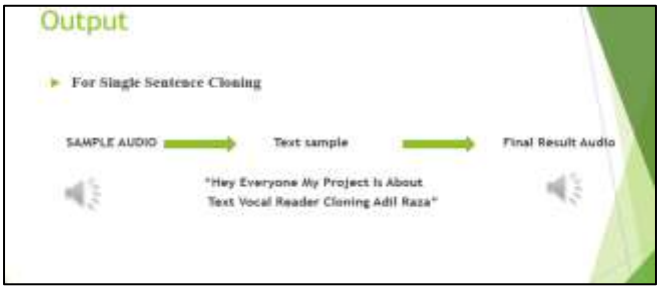

Fig. 4: final result

V. CONCLUSION

This paper has introduced WaveNet, a profound generative model of sound information that works straightforwardly at the waveform level. WaveNets are autoregressive and consolidate causal filters with expanded convolutions to permit their open fields to develop dramatically with profundity, which is critical to demonstrate the long-range transient conditions in sound signs. We have shown how WaveNets can be adapted on different contributions to a worldwide (for example speaker character) or nearby way (for example etymological highlights). At the point when applied to TTS, WaveNets delivered tests that beat the current best TTS frameworks in abstract effortlessness. At last, WaveNets showed exceptionally encouraging outcomes when applied to music sound displaying and discourse acknowledgment

ACKNOWLEDGMENT

I should convey my real tendency and obligation to Dr MN Nachappa and Asst. Prof: Dr Ganesh D and undertaking facilitators for their effective steerage and consistent inspirations all through my assessment work. Their ideal bearing, absolute co-action and second discernment have made my work gainful.

REFERENCES

[1] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. In NeurIPS. 2018.
[2] Mengnan Chen, Minchuan Chen, Shuang Liang, Jun Ma, Lei Chen, Shaojun Wang, and Jing Xiao. Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding. In INTERSPEECH, 2019.
[3] Wei Chu and Abeer Alwan. Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In ICASSP. IEEE, 2009.
[4] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In INTERSPEECH, 2018.
[5] E. Cooper, C. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In ICASSP, 2020.
[6] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. 2002
[7] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In NIPS, 2017.
[8] Daniel W. Griffin, Jae, S. Lim, and Senior Member. Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoustics, Speech and Sig. Proc, 1984.
[9] Y. Huang, L. He, W. Wei, W. Gale, J. Li, and Y. Gong. Using personalized speech synthesis and neural language generator for rapid speaker adaptation. In ICASSP, 2020.
[10] Keith Ito. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.