ISSN No:-2456-2165

Deep Learning-Based Analysis of a Real-Time Voice Cloning System

Ndjoli Isaaka Luc Emmanuel
School of Computer Science and Applications
Reva University
Bangalore, India

Kusha K. R
Prof., School of Computer Science and Applications
Reva University
Bangalore, India

Abstract:- Voice technology has emerged as a hotspot for deep learning research due to fast advancements in computer technology. The goal of human-computer interaction should be to give computers the ability to feel, see, hear, and speak. Voice is the most favorable approach for future interactions between humans and computers because it offers more benefits than any other method. One example of voice technology that is capable of imitating a particular person's voice is voice cloning. Real-time voice cloning with only a few samples is proposed as a solution to the issue of having to provide a large amount of samples and having to endure a long time in the past for voice cloning. This strategy deviates from the conventional model.

For independent training, different databases and models are used but for joint modeling, only three models are used. The vocoder makes use of a novel type of LPCNET that works well on certain samples and low-performance devices.

Keywords:- *Real-Time, Samples, Voice Cloning.*

I. INTRODUCTION

In recent years, deep learning models have proved to be as a dominant force in a variety of applied machine learning in recent times. One such domain is text-to-speech (TTS), which involves generating artificial speech from textual inputs. Achieving natural-sounding voice synthesis, with accurate pronunciation, lively intonation, and minimal background noise, necessitates training data of high quality. However, the amount of reference speech required to clone a voice varies significantly between methods, ranging from half an hour per speaker to just a few seconds.

To contribute to this field, we have reproduced the framework proposed by Jia et al. and made our implementation. We have also integrated a real-time model based on the work of Kalchbrenner et al. into the framework. In this paper, we provide a comprehensive overview of AI based TTS techniques, with a specific focus on the development and current state of TTS systems, emphasizing speech naturalness as the primary evaluation metric. We then present the findings of Jia et al. (2018) and discuss our own implementation. Finally, we conclude by presenting a toolbox that we have developed to interface with the framework.

A. Research Background

Our method for cloning voices in real time relies primarily on (Jia et al.,2018) (SV2TTS is referred to in this paper). A framework for zero-shot voice cloning that only demands five seconds of reference speech is developed. This paper is just one of many that Google has published in the Tacotron series. Interestingly, the SV2TTS paper is based on three major Google earlier works and does not introduce much new information. the loss of GE2E (Wang et al., 2017, van den Oord et al., 2017), and WaveNet2016. The entire architecture is a three-stage pipeline whose steps correspond to the order in which the models were listed earlier. Many of Google's current TTS tools and features, such as Google Assistant and Google Cloud services, employ these same models. As of May 2019, we do not know of any open-source reimplementation of these models that include the SV2TTS framework.

B. System Overview

This study breaks down the system into three parts: encoder, synthesizer, and vocoder as depicted in Figure 1.

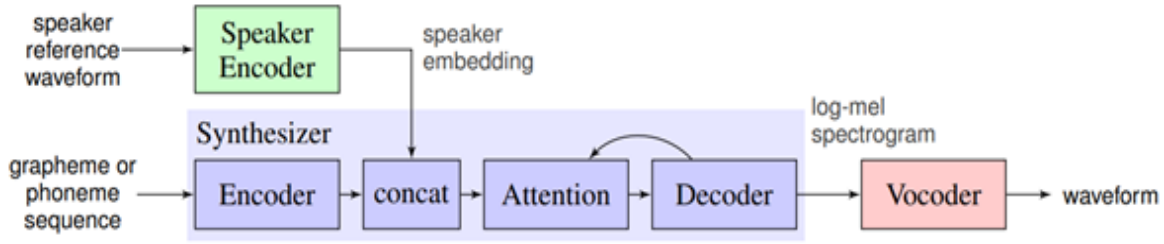

Fig. 1. Model overview. During inference, The SV2TTS framework employs a broad view of the Tacotron's design that has been altered to support voice conditioning, as indicated by the blue blocks in the figure cited from (Jia et al., 2018).

C. Network structure

➢ *Speaker Encoder*

High-quality audio input is not required for this study. So a vast corpus of several distinct speakers is used to train the encoder. Strong anti-noise capability and ease of capture of human acoustic characteristics are two advantages of this choice. Additionally, the encoder has undergone GE2E loss training, which is utilized for tasks related to speaker verification. The method for embedding output is specified in the task.

- *Model architecture*

The model is a three-layer LSTM with a projection layer of 256 units and 768 hidden nodes. Although there is no mention of a projection layer in any of the papers, our understanding is that it is merely a fully connected layer with 256 outputs per LSTM that is applied to each LSTM output repeatedly. For the purposes of simplicity, rapid prototyping, and a lower training load, we used 256-unit LSTM layers directly when we first implemented the speaker encoder. This last part is important because, despite using a larger dataset, the authors' model was trained using 50 million steps, which is technically difficult for us to replicate. We have not had the time to train the larger model because this smaller model performed exceptionally well. The model receives 40 log-mel spectrograms with a 10ms step and a 25ms window width. The output is a 256-element vector that represents the last layer's L2-normalized hidden state. Prior to normalization, our approach additionally includes a ReLU layer with the intention of making embeddings sparse and hence simpler to comprehend.

- *Generalized End-to-End loss*

A speaker verification task is used to train the speaker encoder. One common use of biometrics is speaker verification, in which a person's voice is used to verify their identity. A layout is made for an individual by getting their speaker inserted from a couple of expressions. The term for this procedure is enrolment. A user identifies himself with a brief utterance at runtime, and the system compares that utterance's embedding to the embeddings of the enrolled speakers. The user is identified when their similarity reaches a certain threshold. In order to improve the model, the GE2E loss simulates this process. During training, the model calculates the embeddings e_{ij} $(1 \le i \le N, 1 \le j \le M)$ of M utterances of fixed duration from N speakers. A speaker embedding c_i is derived for each speaker: $c_i = \frac{1}{M}\sum_{j=1}^{M} e_{ij}$. The similarity matrix $S_{ij,k}$ is the result of the two-by-two comparison of all embeddings e_{ij} against every speaker embedding c_k $(1 \le k \le N)$ in the batch. This measure is the scaled cosine similarity:

$$S_{ij,k} = w \cdot \cos(e_{ij}, c_k) + b = w \cdot e_{ij} \cdot ||c_k||2 + b(1)$$

where w and b are parameters that can be learned. Figure 2 depicts the full procedure. Since two L2-normed vectors' cosine similarity is essentially their dot product from a computing perspective, the rightmost side of equation 1. When an utterance matches the speaker, an ideal model should produce high similarity values $(i = k)$ and lower values of similarity elsewhere $(i \neq k)$.The loss in this direction is the sum of row-wise softmax losses.

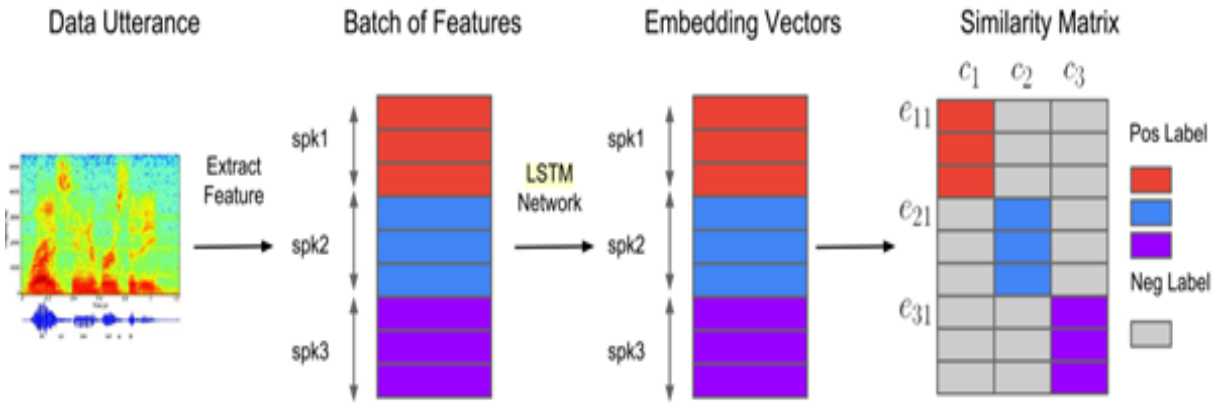

Fig. 2. The process of creating the similarity matrix during training. The data for this graph was taken from (Wan et al., 2017).

In a training batch, the utterances have a fixed duration of 1.6 seconds. These are samples of partial utterances taken from the dataset's longer complete utterances. Although the model architecture is capable of handling inputs of varying lengths, it is reasonable to anticipate that it will perform at its best with utterances that are the same length as those that are observed during training. As a result, an utterance is divided into 50 percent overlapped segments of 1.6 seconds at inference time, and the encoder forwards each segment separately.

The utterance embedding is created by averaging and normalizing the resulting outputs. This is delineated in Figure 3. Curiously, the SV2TTS authors recommend for training windows of 1.6 seconds but 800 milliseconds at inference time. Like in GE2E, we would like to keep 1.6 seconds for both.

ISSN No:-2456-2165

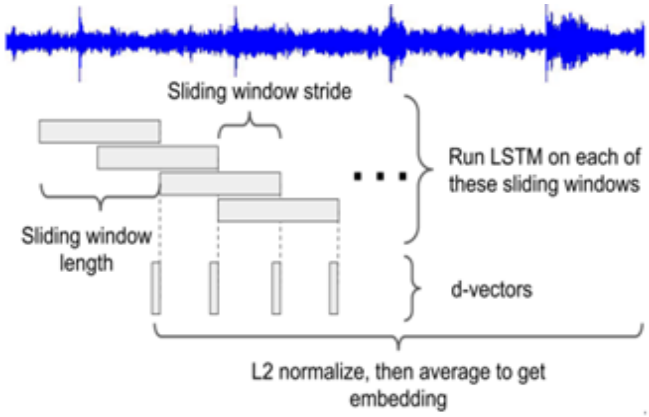

Fig. 3. Calculating the embedding of an entire utterance. The d-vectors are essentially the model's unnormalized outputs. This illustration is based on (Wan et al., 2017).

- *Experiments*

We use the webrtcvad python package for Voice Activity Detection (VAD) to avoid prevent predominantly quiet segments while sampling partial utterances from complete utterances. This produces a binary flag over the audio that indicates whether or not the segment is spoken. Short spikes in the detection are smoothed out by applying a moving average to this binary flag, which is then binarized once more. Last but not least, we dilate the flag with s + 1 kernel size, where s is the longest amount of silence that can be tolerated. The unvoiced parts are then cut out of the audio. We chose the value s = 0.2s because it kept the natural prosody of speech. Figure 4 demonstrates this procedure. Normalization is the final pre-processing step that is perform to the audio waveforms to compensate for the dataset's varying speakers' volume.

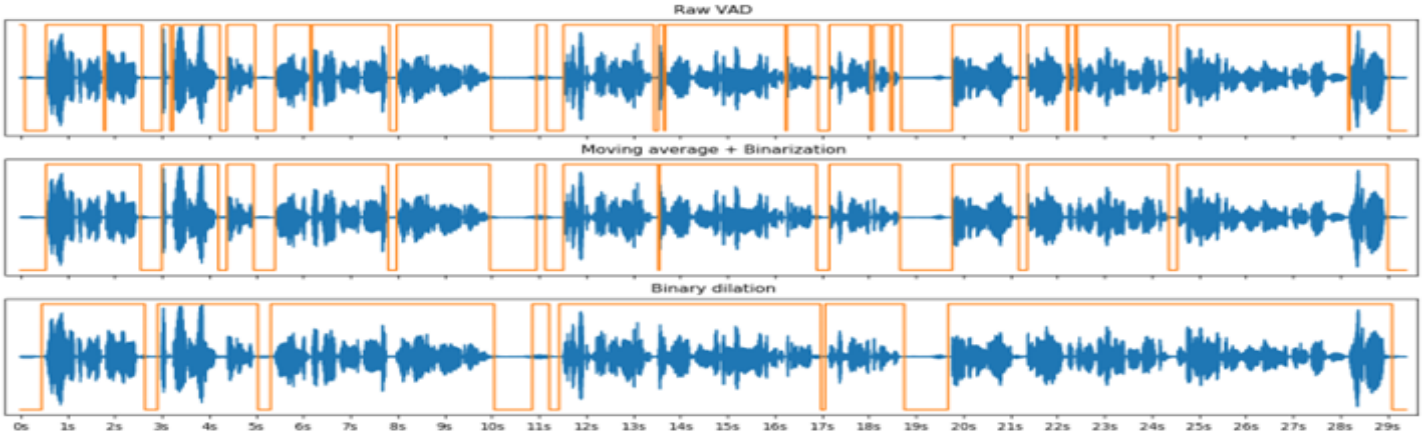

Fig. 4. The steps to silence removal using VAD, from top to bottom. The orange line is the two-voice banner, with the top value implying that the segment is voiced and the lower value implying that it is unvoiced.

➢ *Synthesizer*

The synthesizer utilizes an improved Tacotron 2. Tacotron 2 make s use of a modified network to replace Wavenet. To make a single encoder frame longer, each character in the text sequence is embedded as a vector before being convolved. In the meantime, input the appropriate phoneme sequence to speed up convergence and enhance pronunciation. Encoder output frames are produced bidirectional LSTM to transfer these encoder frames. In order to generate the decoder input frame, the encoder output frame is the focus of the attention mechanism. The model is autoregressive because every single decoder input frame is linked to the output of the preceding decoder frame.

This cascading vector traverses two LSTM layers before being projected onto a single MEL spectrum frame. with only one direction. Frame generation is halted when another projection prediction network of the same vector emits a value above a predetermined threshold. Before becoming a MEL spectrum, the entire sequence of frames goes through a residual network. Figure 5 depicts this architecture. The synthesizer's target MEL spectrogram has more acoustic properties than the speaker encoder. They are fed into an 80-dimensional MFCC in 12.5-millisecond steps from a 50-millisecond window. This module makes use of the thchs30 data set. Tsinghua University has created a 30-hour data set called Thchs30. The data set was picked because it has pinyin, which makes processing it easier. Each transcript must be processed in one step to obtain alignment before it can be processed into libspeech format, which takes a lot of time.

➢ *Vocoder*

To transform the time-domain waveforms produced by the synthesis network into synthesized Mel spectrograms, we employ the sample-by-sample autoregressive WaveNet as a vocoder. The architecture, which consists of 30 dilated convolution layers, is identical to that described in. The speaker encoder's output has no direct influence on the network. The synthesizer network's Mel spectrogram predicts all of the relevant information for high-quality voice synthesis, making it possible to build a multispeaker vocoder by training on data from multiple speakers.

The model is conditioned using any speech audio that has not been transcribed and does not need to match the text in order to be synthesized during inference. It is possible to train it on audio from speakers outside the training set

ISSN No:-2456-2165

because the characteristics of the speakers that are used in synthesis are inferred from the audio. In practice, we discover that zero-shot adaptation to novel speakers can be achieved by synthesizing new speech with the corresponding speaker characteristics using a single, brief audio clip. We assess how well this procedure applies to previously unseen speakers in Section 3. Figure 5, which depicts spectrograms created from a variety of 5 second speaker reference utterances, provides an illustration of the inference process. The spectrogram of the synthetic male speaker exhibits formants and a substantially lower fundamental frequency, as shown by the denser harmonic spacing at the low frequencies, which can be seen in the mid-frequency peaks that are present during vowel sounds like the "i" at 0.3 seconds. The top male F2 is in mel channel 35, whereas the middle speaker's F2 appears to be closer to channel 40. Sibilant sounds exhibit similar variations; for instance, in the male voice, the "s" at 0.4 seconds includes more energy in lower frequencies than in the female voice. The speaker embedding also partially captures the characteristic speaking rate, as indicated by the lower row's greater signal duration compared to the upper two rows. The reference utterance spectrograms are shown in the right column. can be the subject of similar observations.

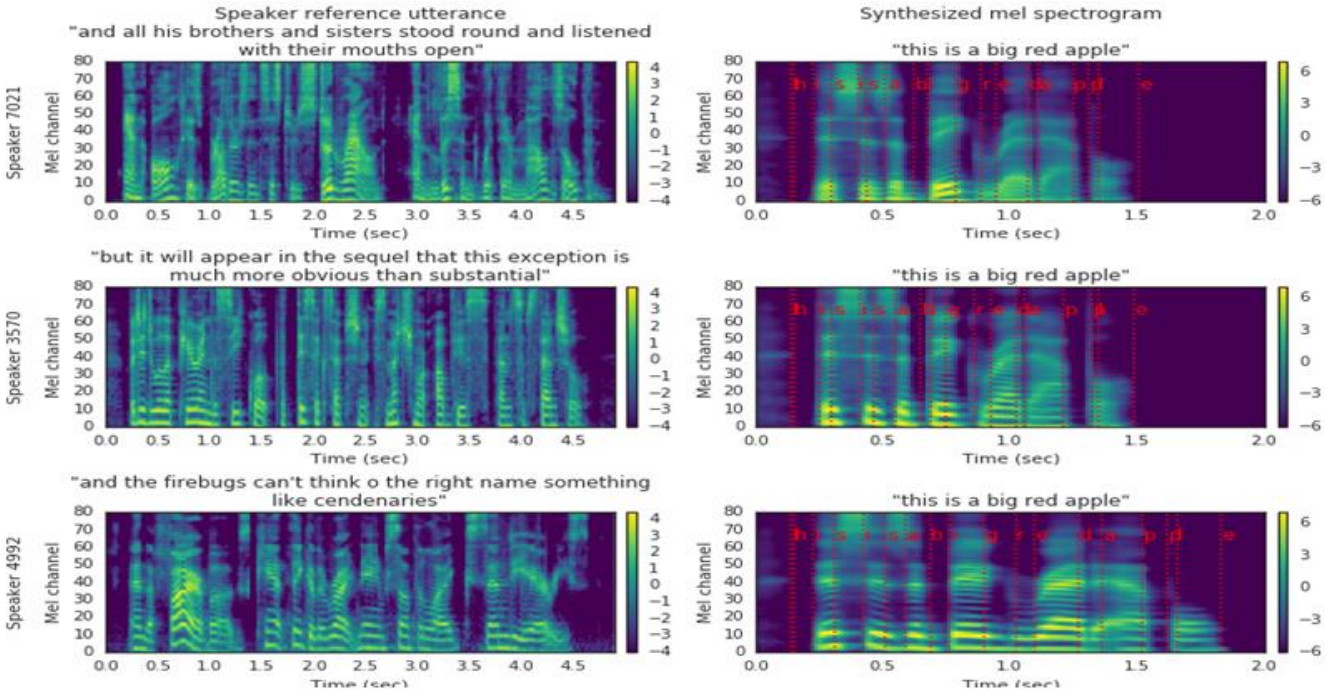

Fig. 5. As an example, use the suggested approach to synthesize a statement in several voices. Mel spectrograms for the speaker embedding reference utterances (left) and the synthesizer outputs (right) are displayed. The text alignment to the spectrogram is highlighted in red. Three speakers are kept outside the train sets: one male (top) and two female (focus and base).

II. EXPERIMENTAL RESULTS AND ANALYSIS

One important aspect of voice clone research is testing the method and evaluating how well it works. In order to boost the voice clone's performance, it is critical to devise a reliable and effective evaluation strategy. Nowadays, primarily objective and subjective methods are used to test and evaluate the performance of the voice clone method.

A. A thorough and impartial evaluation

In terms of MFCC and spectrum, the test-generated cloned speech was compared to the original voice:
As an example, the content of STCMD00044A is: "the man asked me if I would like to" for male

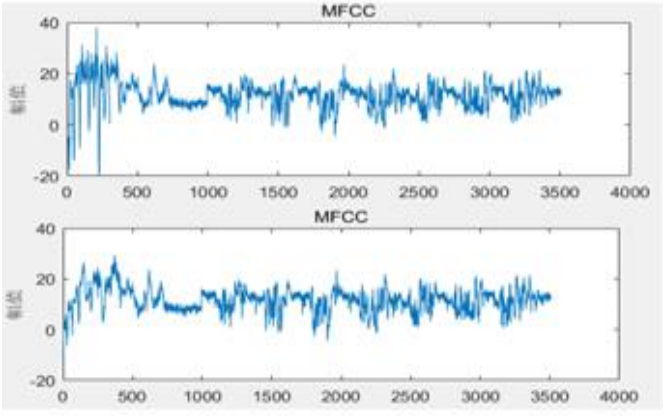

Fig. 6. Male proto-speech and cloned voice MFCC

ISSN No:-2456-2165

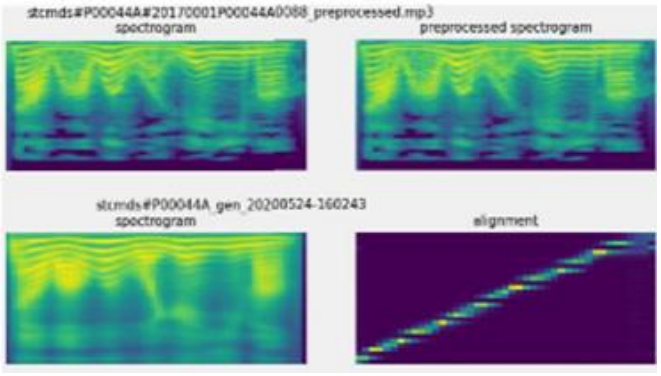

Fig. 7. Male speech spectrum alignment and comparison

As an example, the content of STCMD00052I is: "prepare the stock gap in advance" for women.

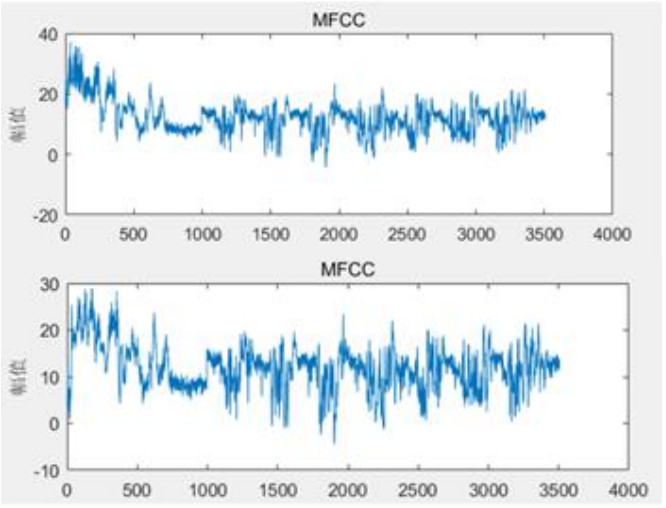

Fig. 8. Female proto-speech and cloned voice MFCC

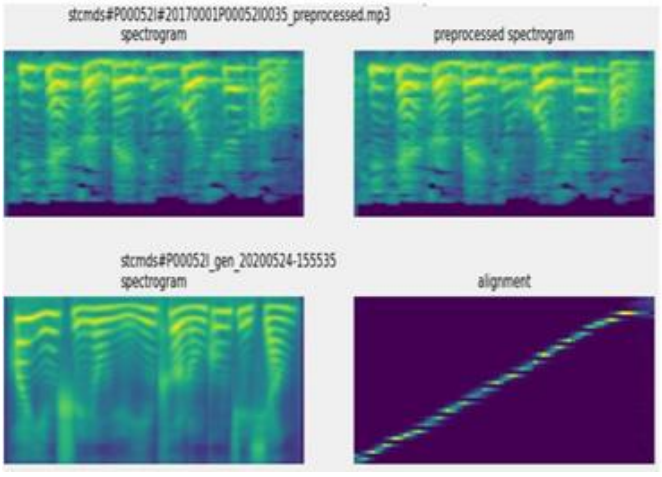

Fig. 9. Female speech spectrum alignment and comparison

The cloned speech and the original speech are nearly identical in the middle and back parts, as shown in the image above; however, the beginning part is false. In the future improvement, this aspect can be improved further. Because they were trained with fewer male voice data, female voice clones performed better than male ones. High spectral similarity and speech-to-text alignment.

B. Subjective evaluation and analysis

In the subjective evaluation, people's subjective feelings are used to test pronunciation. The cloning effect is typically evaluated using the subjective Mean opinion score (MOS) method, which takes into account both the quality of the speech and the similarity of the characteristics of the speakers. MOS check: the primary standard of the MOS test is to request the commentator to score the emotional sentiments from the test discourse as indicated by five grades, which can be utilized for the abstract assessment of the discourse quality and the similitude of the speaker's attributes. The MOS score is the sum of all test statements and reviewers' averages. In general, it can be broken down into five levels, with 1 point representing the least understandable and 5 points representing the most natural. We recruited ten Internet users to provide subjective evaluations.

TABLE 1 SCORES FROM THE FEMALE VOICE MOS TEST

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

TABLE 2 SCORES FROM THE MALE VOICE MOS TEST

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

The aforementioned data demonstrate that the impact of both the cloning of male and female voices still differ. This is due to a number of factors: 1. The female voice is typically sharper and has greater penetration, making it easier for computers to extract data from it. 2. Male voice cloning does not have the same impact as female voice cloning because there is insufficient male voice data in the study's training database.

III. CONCLUSION

In conclusion, we have developed a real-time voice cloning system that, at present, has not been made publicly accessible. While we are satisfied with the overall outcomes of our framework, we acknowledge the presence of some artificial prosody. Comparatively, approaches utilizing more reference speech time have demonstrated superior voice cloning capabilities. Moving forward, we intend to enhance our framework by incorporating recent advancements in the field that extend beyond the scope of this research.

Looking ahead, we predict the emergence of even more powerful voice cloning techniques in the near future. Our ongoing efforts will focus on refining our framework, fixing the limitations identified, and exploring potential avenues for improvement. By remaining up to date on the newest advances, we aim to contribute to the continuous progress and availability of advanced voice cloning technologies.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my guide, Prof Kusha K R, for his constant encouragement, insightful suggestions, and critical evaluation of my work. I am also

ISSN No:-2456-2165

grateful to my lecturer, Ass. Prof. Shreetha Bhat, for her valuable guidance, unwavering support, and constructive feedback.

Additionally, I would like to extend my gratitude to my mentor, Dr. Hemanth K. S, for his invaluable support, encouragement, and expert advice. His mentorship, insights, and guidance were crucial in shaping my research and inspiring me to push the boundaries of my knowledge.

Finally, I would be negligent if I did not mention my family, particularly my parents and siblings. Their confidence in me has maintained motivation and spirits strong throughout the process.

REFERENCES

[1]. Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Advances in neural information processing systems, 31.
[2]. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in Proc. Int. Conf. Acoust., Speech, Signal Process., 2000, pp. 1315–1318.
[3]. Aaron van den Oord, Sander Dieleman, Heiga Zen, et al. Wavenet: A Generative Model for Raw Audio[J/OL]. arXiv Preprint arXiv:1609.03499, 2016.
[4]. Master thesis: Automatic Multispeaker Voice Cloning. Corentin Jemine. 2019, pp. 10-29.
[5]. Sercan O Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. arXiv preprint arXiv:1802.06006, 2018.
[6]. Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In Interspeech, pages 1086–1090, 2018.
[7]. Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C Cobo, Andrew Trask, Ben Laurie, et al. Sample efficient adaptive text-to-speech. arXiv preprint arXiv:1809.10460, 2018.
[8]. Rama Doddipatla, Norbert Braunschweiler, and Ranniery Maia. Speaker adaptation in dnnbased speech synthesis using d-vectors. In Proc. Interspeech, pages 3404–3408, 2017.
[9]. Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612, 2017.
[10]. Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: an ASR corpus based on public domain audiobooks. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pages 5206–5210. IEEE, 2015.
[11]. Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018.
[12]. Artificial Intelligence at Google – Our Principles. https://ai.google/principles/, 2018.
[13]. Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit, 2017.
[14]. Jose Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2Wav: End-to-end speech synthesis. In Proc. International Conference on Learning Representations (ICLR), 2017.