BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Methods for Realtime Voice Deepfakes Creation

Kambulat Alakaev*

Abstract

In addition to being an entertainment, voice deepfakes represent a significant threat to both voice biometrics systems and the average person. Numerous deep learning models for creating voice deepfakes are in the public domain and anyone can use them for fraudulent purposes. However, the fraudster faces a problem in the form of the time needed to generate a deepfake. Long delays when trying to fraud someone during a telephone conversation with a person can raise suspicions. We took several existing open-source tools for creating deepfakes and tested them on computers of different performance. The result revealed a model capable of synthesizing speech within a second. The ability to generate the voice of a specific person almost instantly can lead to increasing in the number of fraudsters using this technology.

*xalaka00@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Since the number of scammers is growing, and their actions aimed at obtaining certain benefits by illegal means are becoming more sophisticated, voice deepfakes are an attractive tool that can be used to deceive or gain access to people's personal data.

Numerous deep learning models for creating voice deepfakes are in the public domain and anyone can use them for fraudulent purposes. However, to conduct various kinds of deception using deepfakes, such as convincing a person to make a money transfer on behalf of his acquaintance or relative through a telephone conversation will require several properties from the tools that generate deepfakes:

- Voice must be generated in real or near-real time
- A deepfake must be of sufficient quality and sound relatively natural.

Both of these points are necessary to avoid arousing suspicion in the other person. It was decided to put ourselves in the fraudster's shoes and focus on finding a way to create real-time voice deepfakes and what the speech synthesis time might depend on.

We chose to analyze the dependence of the deepfake creation time of different tools on the computing power of the device where they were launched. And also if only by increasing the performance of these devices it is possible to achieve real-time speech synthesis.

Existing commercial solutions are not suitable for our purposes, as all computation takes place remotely and the voices that can be used to generate speech are often predetermined without the ability to add one's own.

Two popular open-source speech synthesis tools were chosen for our goals:

- Real-Time-Voice-Cloning(RTVC)[1]
- Coqui TTS[2]

The experiment was conducted using five text-to-speech(TTS) and one voice conversion(VC) models. A set of four computers of different performance was prepared to measure the deepfake generation time for each model. For the TTS models, the testing was separated with respect to the length of the target text to determine if this could also affect the model output time.

The results have showed that it is possible to achieve near-real-time deepfake generation, which could potentially pose a serious threat in the form of unauthorized access to people's personal data or fraudulent phone calls.

2. Tested models

The popular RTVC tool provides only one model with multiple speakers, the Coqui TTS tool contains a

wide set of pre-trained text-to-speech models and one voice conversion model. Selected models:

- RTVC model(Tacotron 2, Wavenet)[3][4]
- VITS (trained with VCTK dataset)[5]
- VITS (trained with LJ Speech dataset)[5]
- Glow-TTS[6]
- YourTTS[7]
- FreeVC[8]

All tests and time measurements were performed with the same input data on each device several times and the result of each test is the average time required to generate a deepfake by a given model on a given device.

3. Experiment results and optimization

Based on the results of the experiment, which can be seen in Figure 1, Glow-TTS was identified as a model capable of generating speech within approximately one second. However, the Coqui TTS console tool, which contains this model and an API to access it, has one drawback in terms of real-time deepfake generation: for each new speech synthesis, it is necessary to restart the application, enter the startup parameters and the target text, and wait for the model to load. This is quite a big waste of time.

It was decided to write our own program, which is an interface to the Coqui TTS program that allows continuous text input to generate deepfakes by a selected model from those available in Coqui TTS without having to run the tool again for each voice synthesis. The abstract principle of the program is shown in Figure 2.

4. Deepfake quality testing

We have evaluated the quality of the generated speech by the Glow-TTS model in two ways: first, to define how well deepfakes detection methods can determine if the speech is artificial. Second, to involve human evaluation - real people - to determine if they can detect any signs that the speech is not genuine. For the quality evaluation with the detection methods were chosen four pretrained models:

- Resemblyzer[9]
- RDINO[10]
- CAM++[11]
- Eres2Net[12]

Experiments with deepfake detection models were conducted in two types: text-dependent and text-independent tests. In the text-dependent tests, the original speech and deepfake were created with the same target text. In the text-independent tests, the target text in the generated deepfake was different from the text in the original speech. Similarity coefficients were calculated for each pair of "original - deepfake" in each experiment type. The similarity coefficient for each experiment type was taken separately for each model.

Average result for the text-dependent test through the all models is about 80-82% similarity. Text-independent tests resulted in 79% of similarity. An output of the Resemblyzer model is shown in Figure 3.

The second part of the experiment was an online survey used to assess people's ability to distinguish deepfakes created by Glow-TTS from real speech. Survey had two parts. In the first part, participants listened to ten mixed audio files, half of which contained real speech and the other half generated speech. Their task was to identify which recordings were real and which were synthetic. In the second part participants were given a 30-second recording of a real voice. Then, following the same format as in the first part, they listened to ten new audio files. Again, their task was to discriminate between the real and synthesized speech. There were 13 participants in the survey. As a result, in an overwhelming number of questions, participants were almost 100% likely to identify which recordings were real and which were synthetic.

5. Conclusions

The program created with the Coqui TTS tool under the hood is able to generate voice deepfakes in a time close to real time. The model used can fool recognition models, but it can't fool a real person. Further breakthroughs in voice deepfakes could pose a significant threat to the security of personal data or funds that could potentially be accessed using such software tools.

Acknowledgements

I would like to thank my supervisor Mgr. Kamil Malinka, Ph.D. for his guidance, suggestions, help while working on this project.

References

[1] Corentin Jemine. Real-time voice cloning. Master thesis, Université de Liège, 2019.

[2] Coqui.ai. Coqui tts, 2022.

[3] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng

Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. online, 2018.

[4] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. 9 2016.

[5] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. online, 2021.

[6] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. online, 2020.

[7] Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. online, 2023.

[8] Jingyi li, Weiping tu, and Li xiao. Freevc: Towards high-quality text-free one-shot voice conversion. online, 2022.

[9] Resemblyzer, 2019.

[10] Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, and Qian Chen. Pushing the limits of self-supervised speaker verification using regularized distillation framework. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[11] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. 2023.

[12] Hui Wang Luyao Cheng Qian Chen Jiajun Qi Yafeng Chen, Siqi Zheng. An enhanced res2net with local and global feature fusion for speaker verification. 2023.