BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024

Methods for Realtime Voice Deepfakes Creation

student: Kambulat Alakaev   supervisor: Mgr. Kamil Malinka, Ph.D.

MOTIVATION

Voice deepfakes are often used as a scouting tool, but also pose a risk to voice biometrics systems and individuals. Numerous deep learning models for creating voice deepfakes are in the public domain and anyone can use them for fraudulent purposes. But can such models be used in real time? Or do the open-source models have the ability to generate real-time speech?

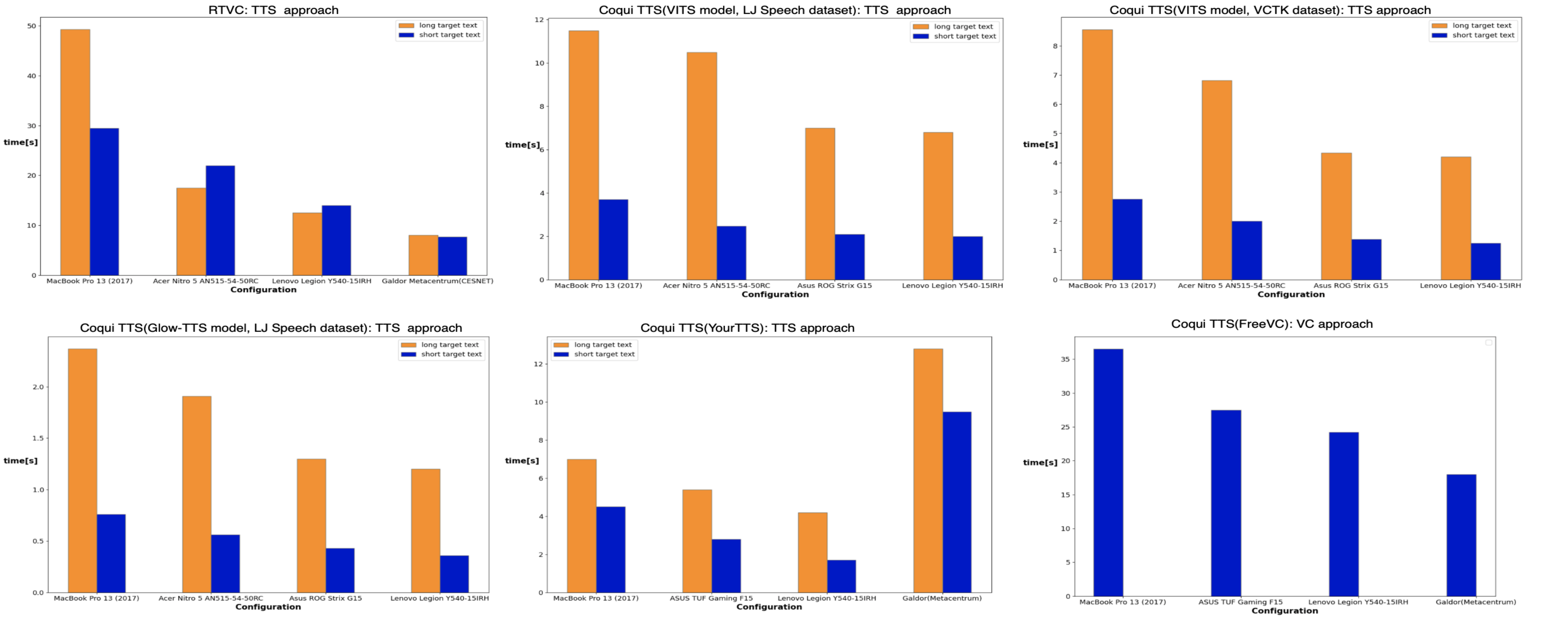
TIME VS COMPUTING POWER

From the position of a fraudster, open-source speech synthesis tools were chosen. It was necessary to determine whether there is a dependence between the time to create a deepfake using the selected tools on devices of various computing power. Used tools:

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

EXPERIMENT

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.


RTVC: TTS approach

Coqui TTS(VITS model, LJ Speech dataset): TTS approach

Coqui TTS(VITS model, VCTK dataset): TTS approach

Coqui TTS(Glow-TTS model, LJ Speech dataset): TTS approach

Coqui TTS(YourTTS): TTS approach

Coqui TTS(FreeVC): VC approach

Figure 1: The dependence of various models on the computing power of the devices on which they were run.

EXPERIMENT RESULTS

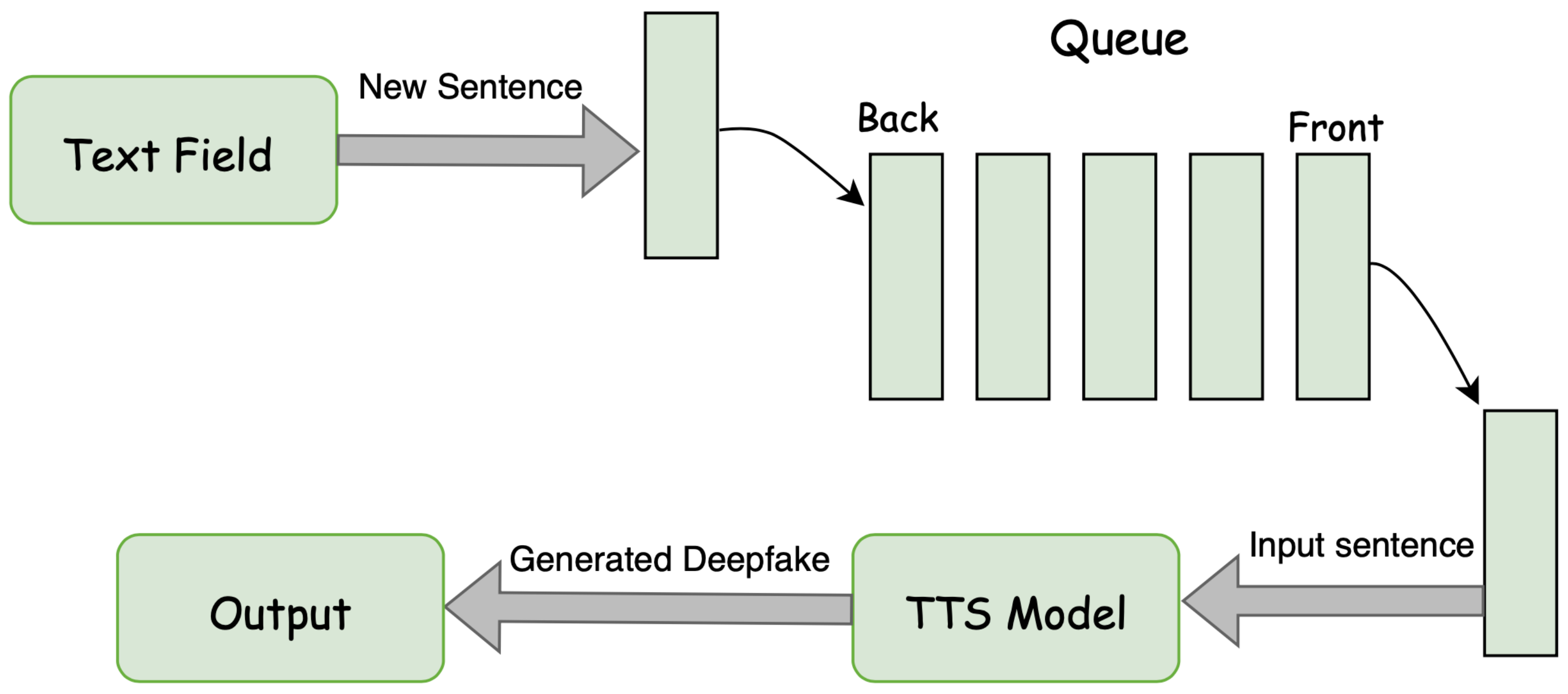
Figure 1 shows that the best model in terms of deepfake generation time is Glow-TTS, which is able to synthesize speech within a second, which is almost equal to real time. However, the Coqui TTS tool in which this model is implemented has one drawback in terms of real-time speech synthesis, which is that for each new synthesis, the console application(Coqui TTS) must be restarted and input data must be entered.

DESIGN A PROGRAM

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 generation. Glow-TTS [3] is used as the primary model. The abstract principle of the program is shown in Figure 2.

MODEL QUALITY TESTING

The Glow-TTS model was tested for its ability to fool voice dipfake detection models and to deceive humans. The following were selected as the detection models: Resemblyzer[4], RDINO[5], CAM++[6], Eres2Net[7].
The models had to estimate the similarity coefficient of a real voice and its deepfake. People were offered an online survey, where they had to determine which voice recordings were fake and which were real.


Text Field

New Sentence

Queue

Back

Front

Input sentence

TTS Model

Generated Deepfake

Output

Figure 2: The abstract principle of the designed program.

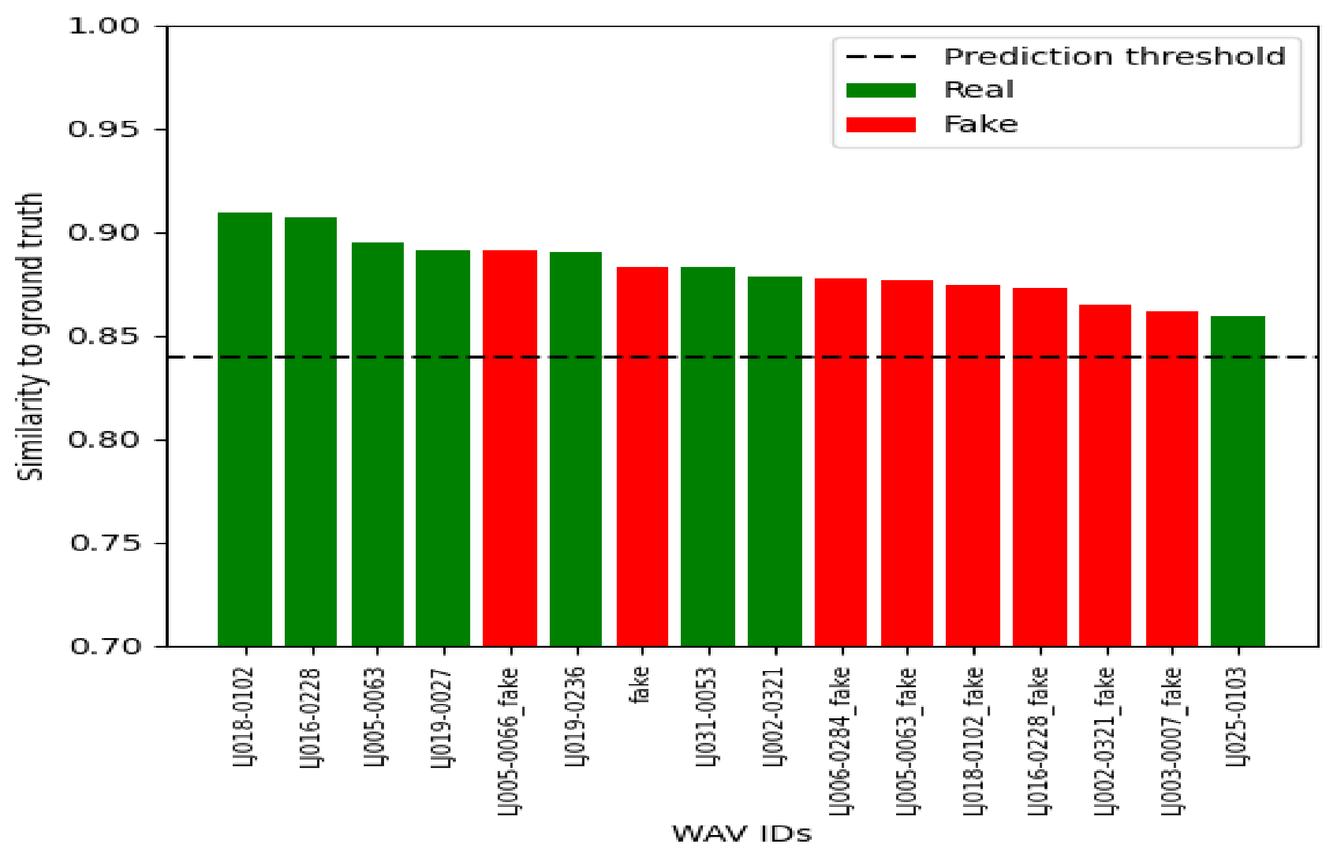


Figure 3: Resemblyzer output. Each column is an audio recording.

MODEL QUALITY TESTING RESULTS

The results for all models were about 80-82% similarity. The test takers in almost 100% of cases correctly determined which were fake and which were real speech.
An output of the Resemblyzer model is shown i n Figure 3.

CONCLUSION

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

REFERENCES

- [1] Jemine, C. Real-Time Voice Cloning [online]. Liège, Belgium, 2019. MASTER THESIS. Université de Liège. Available at: https://matheo.uliege.be/handle/2268.2/6801?locale=en
- [2] Coqui.ai, Coqui TTS[online]. Available at: https://github.com/coqui-ai/TTS
- [3] Kim, J., Kim, S., Kong, J. and Yoon, S. Glow-TTS: A Generative Flow forText-to-Speech via Monotonic Alignment Search[online]. Available at: https://arxiv.org/abs/2005.11129
- [4] Resemblyzer. 2019 [cit. 2024-02-20]. Available at: https://github.com/resemble-ai/Resemblyzer/tree/master
- [5] Chen, Y., Zheng, S., Wang, H., Cheng, L. and Chen, Q. Pushing the limits of self-supervised speaker verification using regularized distillation framework[online]. Available at: https://arxiv.org/abs/2211.04168
- [6] Wang, H., Zheng, S., Chen, Y., Cheng, L. and Chen, Q. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. [online]. Available at: https://arxiv.org/abs/2303.00332
- [7] Yafeng Chen, H. W. L. C. Q. C. J. Q. An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification. [online]. Available at: https://arxiv.org/abs/2305.12838