# Exploring the improvements of voice spoofing technology and its implications in the future

Ridwan Shahidul

Student Number 190056113

Project Supervisor: Syed Rafee

Big Data Science MSc

## Abstract

The voice anti-spoofing scene is currently gaining traction with the rise of voice deepfakes. However, there currently needs to be even more work done in this region due to the actions of malicious players. This study focuses on quantifying emerging technologies' efficacy against legacy anti-spoofing models.

Through a neural network trained on the 2021 ASVspoof deepfake dataset, the research assesses the model's performance on different datasets from various years, shedding light on legacy model adaptability. Results reveal diminishing performance over time, emphasising the need for advanced anti-spoofing technologies to counter evolving voice manipulation techniques and preserve trust and security in voice-based systems.

## Introduction

Voice spoofing or voice changing has been around for more than a decade, with countless people falling victim to its malicious nature. The three relevant types of voice spoofing techniques that researchers have identified are, replay, voice conversion, and speech synthesis. (Antispoofing, 2022)

Replay attacks consist of replaying audio snippets of a victim/ someone relating to the victim. This would be used to bypass voice authentication protocols that may protect personal information.

Voice conversion consists of an attacker reciting speech with the victim's voice features masked over it. This is a particularly dangerous technology as it allows for the attacker to effectively convey emotion and more 'human-like' speech by incorporating natural pauses and other disfluencies. (Corley & Stewart, 2008)

Speech synthesis consists of collecting a targets voice features and reconstructing speech by using a text-to-speech generator. This will be the focus of this paper; the models will be trained on this type of data and will be deployed on similar datasets.

There is a growing concern in the online community where there is uncertainty on whether it will be possible to discern human voices from spoofed voices. This is very much a problem that will affect the real world. There is still somewhat a level of 'uncanniness' with the spoofed voices that are available today. An uncanny valley some call.

This is especially an issue due to the increase in usage of Automatic Speaker Verification (ASV) systems. It is seen that the use of ASV systems is expected to quadruple from 2014 to 2025 (Khan et al., 2023). This makes a large and justified argument for the development of anti-spoofing technologies.

There are current technologies available and being produced to combat and identify these computer-generated voices. However, as technology improves, it will only become harder and harder for humans to discern between real and fake.

If it becomes impossible to rely on human-based classification, then it is only right to fight back with computers. Technology has been improving for both sides and will continue to do so. This is what one can call, the digital 'cat and mouse chase'.

To address the relationship between attacking and anti-spoofing technology, the paper will look to quantify how well emerging technologies are able to slip past legacy anti-spoofing models. This will be done by constructing a neural network based on the 2021 ASVspoof deepfake dataset, and it will then be deployed on various other datasets, both past and future, in relation to the training dataset.

This means the network will essentially take educated guesses at whether the audio sample is bonafide or spoofed. The accuracy of the model will then be calculated for each dataset. This will allow for better insight into the performance of legacy models in relation to newer datasets with updated technology.

## Related Work

There has been great work done in this field due to the imminent danger such technology poses on the greater public. Voice spoofs have already proven to be dangerous and will only become more efficient. "Mark Gorrie, the Asia Pacific Managing Director at cyber security software company Gen Digital, says **AI voice generators are going to keep getting better at tricking both people and security systems**." (Williams, 2023)

When this technology is put in the hands of bad actors and becomes more freely available, there will be an increase in the number of voice spoofing victims. Not only is this technology capable of economic damage, but it is also capable of social and political damage. (Klein, 2023)

Misinformation could be easily spread by puppeteering deepfakes of renowned political beacons such as presidents and ministers of countries. There is much at risk when deepfakes can be used as political weapons.

Currently, a large portion of the work done in this field revolves around making machine learning/ deep learning models based on certain publicly available datasets. (ASVSPOOF, 2021)

These types of datasets are fundamental to the research and development of anti-spoofing technology. These datasets are utilised by small-scale researchers all the way to large security companies for the development of this crucial technology. (ID R&D, 2023)

For example, ID R&D, a company that develops voice biometric detection systems among other things, developed the detection software with the lowest error rate out of all competitors, a value of 0.22%. Having a dataset like this in the first-place drives innovation from multiple sources, a clear and easy win for this technology.

Some researchers combine different types of neural networks to achieve better results. An example can be found in the paper (Chettri et al., 2019). This is where they combine different deep learning models into one, experimenting with different mixtures to attain different results. The report mentions that because the different models within the ensemble model utilise similar audio features from the voice sample, these features overlap and correspond to each other and therefore result in a more robust understanding of the audio sample by the ensemble model.

When the models are tested on Logical Access (LA) attacks, it was seen that all ensemble models were within a 0.2% difference in Equal Error Rate (EER). The best of them achieving an EER of 0.0%. When looking at the standalone deep learning models, their EERs can range from as little as 0.16% to as high as 13.58%. This clearly shows that the ensemble models are far superior.

However, this paper is from 2019, and between then and now (2023), there has been massive changes in the generative AI scene. Usually, these generative AI models would need large amounts of data to construct near-genuine voice models, however, a current breakthrough by Microsoft shows that they are able to construct voice models with only three seconds of voice speech. (Validsoft, 2023)

Breakthroughs such as these are concerning and have negative implications for what is to come. There will need to be aggressive research and development to combat these technologies in the future. Furthermore, these spoofing detecting technologies should be made widely apparent and available for the use of the public to mitigate the possible damages of deepfakes.

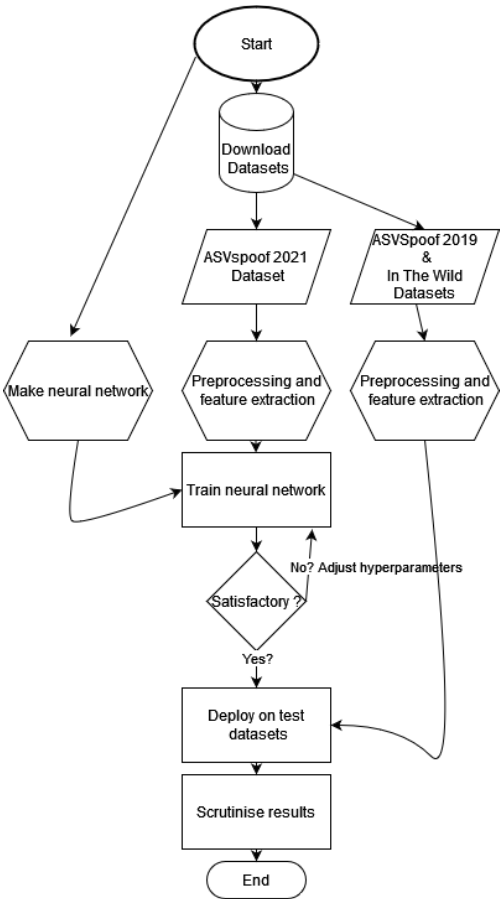
Similar sentiments can be derived from the paper produced by (Kassis & Hengartner, 2023). That advancements in voice authentication bypassing are currently at breakneck speeds. To be able to bypass voice authentication protocols with a 99% success rate is certainly a cause for concern.

They achieve this number by targeting weak points that are shared by most anti-spoofing technology. The model is therefore able to operate in real-time and works regardless of the anti-spoofing technique.

To reiterate, given the recent attention in this field, it is alarming that there has been this much progress made within this space. This again reinforces just how fast the technology is advancing and creates a bigger argument for development of spoofing countermeasures.

**Methodology**

The pipeline for making and testing the neural network is as follows in the flowchart below:



The pre-processing for audio data is extremely important. The processing pipeline that was used on the training dataset (2021 ASVspoof) is as follows:

- Balance data by deleting audio files so that there are equal numbers of bonafide and spoofed files.
    - This was done to prevent the model from becoming biased towards a certain class, in the case of the 2021 dataset, there were far more spoofed audio files compared to bonafide audio files.
- Remove silent portions of the audio snippets. The method to remove the silences was adapted from the code found in the paper (Chettri et al., 2019).
    - This was performed so that the model does not associate silences with any class, this makes the model focus on the actual audio features of the speech, as opposed to any other pattern.
- Extract Mel-spectrograms from the audio.
    - Mel-spectrograms were specifically chosen because of their focus on the lower frequencies. Humans are better suited at distinguishing lower frequencies, and so the Mel scale aims to mimic this by having a linear scale for lower frequencies and a logarithmic scale for higher frequencies, effectively giving more importance to the lower frequencies. (Doshi, 2021)
- Add zero-padding to all spectrograms so they all have the same dimensions.
    - This was done so that all the samples have the same dimensions. This is important as the model will not accept data points that have varying sizes.
- Split data into training and testing splits and reserve validation splits.

For the other datasets, the splitting of data was not performed. This is because the model only needs to be tested on the datasets, the training will already have been done on the ASVspoof 2021 dataset. The model is saved after training and ran on the other datasets and is asked to predict whether the audio samples are either bonafide or spoofed.

The architecture used was that of a Feed Forward Neural Network (FFNN) model, as seen in table 1. It was chosen for its simple architecture and ability to learn features quickly from data. The focus of this report is to identify patterns in the robustness of datasets throughout the years. There is little importance in maximising the accuracy of the trained model and it is not the objective of this report.

*Table 1 - Neural Network structure*

| Layer | Parameter |
|---|---|
| Dense | 256 hidden units |
| BatchNorm | - |
| Dropout | 0.6 |
| Dense | 128 hidden units, ReLU |
| BatchNorm | - |
| Dropout | 0.6 |
| Dense | 64 hidden units |
| BatchNorm | - |
| Dropout | 0.6 |
| Dense | 1 hidden unit, sigmoid |

The ReLU activation function was used because of its ability to resist the effects of the vanishing gradient problem (Aggarwal, 2023) that more traditional activation functions such as the sigmoid function are susceptible to. (Ide & Kurita, 2017)

Furthermore, a high dropout value of 0.6 was used due to the model's tendency to overfit on the dataset. This value was chosen during the hyperparameter tuning phase where the best dropout value of 0.6 was determined by looking at the best accuracy output.

In the end, the model was compiled using the Adam optimiser due to its ability to reliably converge faster. This can be seen in the paper (Jais et al., 2019). The use of this optimiser is widespread and was required in this use case because of the sheer amount of data that was processed.

The structure found in table 1 takes inspiration from the architecture found in, (Białobrzeski et al., 2019), and it is adapted using the Keras library as opposed to the TensorFlow library. This was done because of the simplicity of the Keras library, note that Keras is built on top of TensorFlow, however, it is simpler to use in the same way that Python is a high-level language.

Specifically, the Light-CNN architecture was used, the use of Dropout and Dense layers were specifically adapted from the source to better fit the use-case. Since the dataset is not that large, the dropout layer helps the model generalise better. The Dense layer allows the model to extract features more precisely from the data.

Results

*Table 2 - Trained Neural Network Performance*

| Accuracy (%) | Loss | Validation Loss | Validation Accuracy (%) |
|---|---|---|---|
| 96.26 | 0.7204 | 0.7636 | 96.08 |

*Table 3 - Performance of trained neural network on other datasets*

| Dataset | Accuracy (%) |
|---|---|
| ASVspoof 2019 | 89.65 |
| In The Wild (2022) | 37.15 |

**Discussion**

From training the initial ASVspoof 2019 dataset, the results in Table 2 show that the accuracy (96.26%) and validation accuracy (96.08%) are both high with little discrepancy. This indicates that there is little overfitting with the model and therefore suggests that the dropout layers are functioning as intended. Dropout in this case randomly deactivates some neurons; this ensures the model learns general patterns as opposed to dataset specific patterns.

When the trained model is deployed on the other datasets, there are wildly varying results.

It is important to conduct an analysis on this subject because it allows for the determination of its generalisability. Since there is a discrepancy of 52.5%, it shows that the model is either not able to generalise well or does not consider that the newer spoofing technologies mask the spoofing artifacts better.

On the older model it performs well with an accuracy of 89.65% as seen in table 3. This is to be expected as the dataset is built with older technology and therefore is not as robust as the newer ones. Even if the accuracy is relatively lower than the one found in the initial model, there is little importance in this discrepancy.

This is because, rather than looking at decimals, there is more importance placed in the identification of patterns. An interesting pattern arises when comparing the results from the 2019 dataset to the 2022 dataset. The accuracy of the model plummets to a mere 37.15%. Now when the outdated model is tested using a novel dataset, the model struggles to keep up with the new technology. When the model is run on datasets that were developed three years apart, a difference in accuracy of 52.5% is found.

This is an interesting and expected pattern because of how fast things are progressing in the space. Given the results it is evident that the research objectives have been met, this is because there is a prominent pattern found that reinforces the need for more advanced anti-spoofing techniques. It shows that the used dataset is currently not enough to construct reliable models. Therefore, there needs to be more emphasis on adapting models so that they can counteract the masking the new technologies are able to perform.

## Future Work

To improve on this work, there should have been more of an emphasis on having a wider variety of models built on the dataset. This was not possible due to time constraints, as training individual models not only requires training time, but there are also large amounts of time dedicated towards pre-processing the data and therefore was not feasible for this report.

There also could have been further emphasis into making a model that is trained on multiple datasets and is then compared to models trained on single datasets. This would allow for a better representation of what technologies are currently available as it will have training on all data.

Furthermore, a CNN model could have been used as opposed to the FFNN model found in the report. The data processing will be slightly different as the CNN will be fed PNGs as opposed to NumPy arrays. However, it would still be a valued approach as it is currently used widely in the field. This of course will be more resource intensive and therefore more expensive.

In the future, it would also be beneficial to produce an authentication system that incorporates multiple layers of authentication, such layers can be things such as location, time, and other contextual factors. This is of course a higher level of authentication that would be more suited for real life situations, as with a dataset there is only so much you can do.

## Conclusion

The results indicate that as time passes, it becomes harder and harder to rely on older models. This is because these older models are trained on outdated datasets. The performance of the new voice spoofing technology indicates that whatever feature/s the old model was scrutinising for artifacts becomes corrected in the newer models, as seen by the drop in performance as the years go by.

However, there is not enough data gathered during the research to come to this solid conclusion. Only having tested the model on datasets from two separate years makes the conclusion overzealous. Since this is somewhat a recent technology, this is to be expected as the availability of strong training data is scarce.

A larger spread of datasets would be useful to use as a better a comparison. There are similar sentiments found in the paper (Khan et al., 2023), where the researchers also struggle with finding similar datasets to test on. There is clear indication that there needs to be further work/

input into this field to allow for the advancement of anti-spoofing technologies.

## References

Antispoofing (2022) *Voice anti-spoofing: Origin and methods, Antispoofing Wiki*. Available at: https://antispoofing.org/voice-antispoofing-origin-types-and-preventive-techniques/ (Accessed: 01 August 2023).

Corley, M. and Stewart, O.W. (2008) ‘Hesitation disfluencies in spontaneous speech: The meaning of um’, *Language and Linguistics Compass*, 2(4), pp. 589–602. doi:10.1111/j.1749-818x.2008.00068.x.

Khan, A. *et al.* (2023) *Voice spoofing attacks and countermeasures: A systematic review, analysis, and experimental evaluation* [Preprint]. doi:10.21203/rs.3.rs-2557691/v1.

Williams, T. (2023) *Scammers are using artificial intelligence to trick people. here’s how to best protect yourself, ABC News*. Available at: https://www.abc.net.au/news/2023-04-12/artificial-intelligence-ai-scams-voice-cloning-phishing-chatgpt/102064086 (Accessed: 02 August 2023).

*ASVSPOOF* (no date) *ASVspoof*. Available at: https://www.asvspoof.org/ (Accessed: 11 June 2023).

Kassis, A. and Hengartner, U. (2023) ‘Breaking security-critical voice authentication’, *2023 IEEE Symposium on Security and Privacy (SP)* [Preprint]. doi:10.1109/sp46215.2023.10179374.

Ide, H. and Kurita, T. (2017) ‘Improvement of learning for CNN with Relu activation by sparse regularization’, *2017 International Joint Conference on Neural Networks (IJCNN)* [Preprint]. doi:10.1109/ijcnn.2017.7966185.

Jais, I.K., Ismail, A.R. and Nisa, S.Q. (2019) ‘Adam optimization algorithm  for wide and deep neural network’, *Knowledge Engineering and Data Science*, 2(1), p. 41. doi:10.17977/um018v2i12019p41-46.

Białobrzeski, R. *et al.* (2019) ‘Robust Bayesian and light neural networks for voice spoofing detection’, *Interspeech 2019* [Preprint]. doi:10.21437/interspeech.2019-2676.

Klein, C. (2023) *‘this will be dangerous in elections’: Political Media’s Next Big Challenge is navigating AI Deepfakes, Vanity Fair*. Available at: https://www.vanityfair.com/news/2023/03/ai-2024-deepfake (Accessed: 17 June 2023).

*Voice anti-spoofing* (2023) *ID R&D*. Available at: https://www.idrnd.ai/voice-anti-spoofing/ (Accessed: 17 April 2023).

*The rise of audio deepfakes: Implications and challenges* (2023) *ValidSoft*. Available at: https://www.validsoft.com/articles/the-rise-of-audio-deepfakes-implications-and-challenges/ (Accessed: 17 April 2023).

Doshi, K. (2021) *Audio deep learning made simple (part 2): Why Mel Spectrograms perform better*, *Medium*. Available at: https://towardsdatascience.com/audio-deep-learning-made-simple-part-2-why-mel-spectrograms-perform-better-aad889a93505 (Accessed: 18 March 2023).

Aggarwal, C.C. (2023) *Neural Networks and deep learning a textbook*. Cham: Springer International Publishing AG.

Chettri, B. et al. (2019) ‘*Ensemble models for spoofing detection in automatic speaker verification*’, Interspeech 2019 [Preprint]. doi:10.21437/interspeech.2019-2505.