Enhancing Deepfake Audio Detection for Vulnerable Populations

Sofya Diktas
Science and Society
Duke University
Durham, USA
Sofya.Diktas@duke.edu

Christian Fronk
Electrical and Computer Engineering
Duke University
Durham, USA
Christian.Fronk@duke.edu

Matthew LaRosa
Electrical and Computer Engineering
Duke University
Durham, USA
Matthew.LaRosa@duke.edu

Ashir Raza
Electrical and Computer Engineering
Duke University
Durham, USA
Ashir.Raza@duke.edu

***Abstract*—Deepfake detection models have increased in popularity as access to multi-modal generative AI applications have become increasingly available. While there has been a lot of emphasis on image and video detection given their prevalence on social media, audio deepfake detection research has been less emphasized. Although audio deepfake detection models do exist, the volume, quality, and diversity of data to train with is far less substantial compared to its other media counterparts. As a result, it is likely that many existing audio deepfake detection models do not adequately account for irregular speech patterns in vulnerable populations, such as children's voices and speech impediments. In this paper, we investigate the current state of deepfake audio detection by designing and training a detection model of our own using widely available real and fake adult voices. We test this baseline model against rarer voice types, confirming that our detection model does not perform well with stutters and children's voices. We then finetune the model first using only the children's voice dataset, which generalizes the model and improves performance against the stuttering dataset as well. Finally, we finetune again using both children's voices and stuttering, showing that intentional inclusion of low-occurrence audio datasets improves the overall accuracy of detection across all speech types. The full reproducible results can be found at: https://github.com/mrlarosa/AdversarialAIFP**

I. Introduction

With the public release of multimodal AI models, it has never been easier to generate synthetic media with no technical experience. It is now possible for anyone with access to the internet to convert text into images, speech, and videos in a matter of seconds at a relatively low cost. As generative AI technology advances, these deepfakes, or completely AI generated media artifacts, are becoming increasingly realistic and often impossible to distinguish from real content.

While there are many positive use-cases for generating deepfakes, this lowered barrier to entry, coupled with its increased realism, is leading to spikes in their generation for malicious purposes. According to the US Department of Homeland Security, deepfakes present an evolving threat to ”national security, law enforcement, financial, and societal domains.”[2] At a large scale, their use can cause mass disinformation campaigns, further impacting political domains and potentially leading to societal unrest. On the individual level, deepfakes are being weaponized by scammers to target and manipulate people.

Literature, research and commentary around deepfakes has primarily focused on images and videos. However, there is a significant risk to misuse and adversarial attacks posed by audio deepfakes that is less emphasized. In light of these threats, extremely accurate deepfake audio detection models will be important to combat misuse. However, compared to images and video, the quantity and quality of audio datasets available to train these models on is limited. Given this, current audio deepfake detection models are likely vulnerable to adversarial attacks from low occurrence audio, such as children's voices or stutters, due to the lack of available data for training.

In this project, we explore this vulnerability by building a binary classification model for detecting deepfake audio. We initially train this baseline model using widely available real and fake adult voices, which represent commonly occurring audio. Our model uses the Wav2Vec2 embedding model to vectorize the audio samples at 16kHz. It then passes them through a four layer, fully connected multi-layer perceptron (MLP) neural network to classify the deepfakes. After some hyper-parameter tuning, we test this baseline model on both children's and stuttering voices. The results confirm our suspicion that detection is poor on these out-of-distribution datasets.

To ensure fairness in our detection model across these vulnerable populations, we enhance our model's robustness through fine-tuning. We freeze the first two layers and fine-tune initially using only the children's data. Our theory is that training using even just one out-of-distribution dataset will generalize the model and improve performance for other irregular datasets that it hasn't been trained on. We then fine-tune using a combination of both the children's and stuttering datasets to improve overall performance of the model. With each tuning, the accuracy of the model improves for all voice types, ensuring the inclusion of marginalized communities in deepfake audio detection.

II. Related Work

Machine learning advancement has brought significant breakthroughs in voice synthesis and automatic speech recognition (ASR). However, these advancements also present new security challenges, such as the generation of realistic deepfake voices designed to fool humans and even neural network classifiers. These adversarial generations pose significant direct risks to society, such as enabling fraud by malicious parties. Furthermore, the ability to generate convincing deepfaked speech poses risks of enhancing the spread of misinformation, and raises ethical concerns regarding online content creation.

Adversarial attacks in deep learning settings were first introduced in the computer vision domain, with [5] citing the existence of adversarial examples that can be crafted by adding small, imperceptible, worst-case perturbations to a given image via gradient decent methods that would cause the given model to misclassify with high confidence. In the audio domain, adversarial examples have been shown to be able to be crafted based on these iterative, gradient-based attacks to make any given arbitrary waveform transcribe as any target phrase that the attackers may choose [4].

To address issues like this in audio settings, approaches like DeepSonar [10] have emerged. DeepSonar monitors layer-wise neuron behaviors of speech recognition models, and leverages these patterns to discern differences between real and synthesized voices. By focusing on the internal dynamics of deep neural networks, this method aims for a robust and interpretable approach to detecting fake voices, achieving high accuracy (98.1 percent) and low error rates (2 percent).

Similarly, [8] seeks to improve existing deepfake speech detection techniques. The authors work to improve upon frame-by-frame detectors by reducing inference time through the mixing of audio frames together. The authors find that this audio folding technique is effective on fully fake and partially fake datasets, reducing processing time down to 25 percent of the original inference time.

While existing methods like these provide detection capabilities for machine-generated voices, these methods make the assumption that the given input examples will be relatively "normal." That is, that these voices are not drawn from underrepresented groups such as children or those with a stutter. Our work seeks to investigate the possibility of training and fine-tuning a model that is capable of robust deepfake detection even in the presence of real and generated voices with these characteristics. For future work, we envision the creation of targeted attacks on this model, designed to stress test it to aid with future improvements.

III. Methodology

A. Dataset

In order to train and analyze the performance of our model, we draw from several existing datasets. First, we draw adult fake and real speech data from a Kaggle dataset titled "The Fake-or-Real Dataset." [1] This dataset aggregates speech data from multiple sources, including: Deep Voice 3 and Google Wavenet TTS for the deepfaked data, and the Arctic dataset, the LJSpeech dataset, and the VoxForge dataset for the real speech data. This dataset contains 195,000 utterances from real humans and computer generated speech, and an even split between real and fake data. We use this data for training the naive base model.

Obtaining the data for underrepresented speech sources, in our case children and those with a stutter, was much more challenging; especially so as we needed representative examples of this speech that was computer-generated. For the child speech data, we use the "Fastpitch child tts" dataset [7] to obtain synthetic child speech data, and the "Jibo Kids" dataset [9] for the real child speech data. The Fastpitch dataset contains 28,800 utterances from 40 speakers, synthetically converted to children's voices, and the Jibo Kids dataset contains 383 files from 110 children from pre-kindergarten through grade 1. We note that the samples in both of these datasets are less than ideal: the Fastpitch samples often contain very noticeable audio artifacts, and the Jibo Kids samples often contained periods of silence and instances of non-child voices. This is an unfortunate side effect of the rarity of child-centric datasets in general. In any case, we split this data evenly between real and fake for training and testing, and use this data to test the naive base models efficacy, and later fine-tune the model.

Lastly, we obtain synthetic stutter speech data from the LibriStutter dataset [2] and real stutter speech data from the UCLASS dataset [6]. The Libristutter dataset contains 100 hours of speech data across 10 synthetic speakers, and the UCLASS dataset contains 138 samples from 81 individuals who stutter, aged 5–47 years. We again split this real and fake data evenly for training and testing, and use this data to test the naive base models efficacy, and later fine-tune the model.

B. Model

Our model is composed of an audio feature encoder module $f : \mathcal{X} \rightarrow \mathcal{Z}$ which takes as input raw audio $\mathcal{X}$ and outputs latent speech representations $\mathbf{z} \in \mathbb{R}^{768}$. We then have a MLP binary classification module $g : \mathcal{Z} \rightarrow \{0, 1\}$. that takes in the latent representation of the audio and outputs a binary output that signifies whether or not the given input belongs to the class of real audio. Fig. 1 shows the overview of our data pipeline and model architecture.

To convert the audio into a vectorized, numerical feature vector that we can use for training, we utilize a pretrained

[1] https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset/data

[2] https://github.com/jordicapde/stutter-former

Wav2Vec2 module that takes in a .wav file and uses self-supervised learning to learn a good latent representation via a contrastive task, which has been fine-tuned on labeled data to be used specifically for downstream ASR tasks. [3]

In applying this to our binary classification task, we use a MLP classification module that takes in the flattened 768 x 1 vectors of each image and feeds the latent data through 4 fully connected layers, with four different sets of weights $\boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{W_3}, \boldsymbol{W_4}$ with relu activations after each layer except the last one over which we use a sigmoid to obtain a binary classification.

We used a learning rate of $\alpha = 0.0001$, hidden layer sizes of 64, 16, 8, 1 for $\boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{W_3}, \boldsymbol{W_4}$, and a batch size of 64. We then train for 100 epochs on the concatenated adult voices data from the "The Fake-or-Real Dataset" to get our base model.

C. Finetuning

For finetuning, we freeze the pretrained Wav2Vec2 module and the weights of the first two layers $\boldsymbol{W_1}, \boldsymbol{W_2}$ of the MLP Classification Module. We also lower the learning rate by a factor of 10 to 0.00001. This is to lessen the impact of catastrophic forgetting, wherein a model that is pre-trained on a first task and fine-tuned on a separate, equally difficult second task systematically forgets what it has learned on the first task.

We then introduce the concatenated children's voices data from the Fastpitch and Jibo Kids datasets for samples of synthetic and real children's voices, respectively. We also create a concatenated stuttering voices data from the Libristutter and UCLASS Kids datasets for samples of synthetic and real children's voices, respectively.

We then try two different variants of finetuning. In our first try, we first finetune on the children's voices data and then test on the stuttering data. Our hypothesis here is that by training on one out-of-distribution dataset, this training will generalize to good performance in another similarly out-of-distribution dataset. In our second try, we equally divide both the childrens voices and stuttering datasets into train and test sets. Then, we combine the two train sets as the training set for the finetuning and test on the combined test set. This is a more straightforward approach to this issue and will allow us to test whether or not training on one out-of-distribution dataset and testing on another will achieve similar results to training on both of them simultaneously. The finetuning process can be seen highlighted in light blue in Fig. 1.

IV. Results

The trained models support the hypothesis that out of distribution voices increase the generalization of the model, and thus accuracy across all data classes. Specifically, we examine the untrained model, a fine tuned model tuned with child speech data, and a fine tuned model tuned with a combination of the all the standard and underrepresented group data.

First, we examine the baseline model. This model is trained only on the standard voice data. Fig. 2 shows the training and validation loss across the 3 different voice groups. Notably, this model fails to generalize on the stuttering data. However, there is slightly better success with the children dataset. This is likely due to the poor quality of the synthetic children dataset. However, there is still significant room for improvement with both irregular speech groups.

Since the baseline model was able to perform better on the children's data, we next fine tuned the baseline model with the children's data. As shown in Fig. 3, this helped increase the performance on the stuttering dataset as well. Although it is empirically shown, it can also be explained theoretically as well. Specifically, one can consider both the stutter and children voices to be out of the standard distribution characterized by the standard voices dataset. By introducing the outliers, the model is forces to look at more complex features. Similarly, the out of distribution speech sets are likely closer to each other than either are to the standard speech. This provides the basis for the theoretical justification of the empirical results. This is evidence that making the dataset as diverse as possible not only helps the marginalized populations, but also helps the majority as well.

The last model was tuned with a combination of the children and stuttering datasets. Fig. 4 shows the result of the training. This combination shows significant gain for the irregular speech but small changes in error for the standard speech. This aligns with the previous results and showcases how including marginalized data results in stronger model performance.

V. Conclusion

This paper examines the how to detect AI generated audio. Specifically, we examine how out of distribution speech is impacted when models are only trained with standard data. This model is then tested on two irregular speech patterns: children and stuttering audio. By testing model performance on these irregular speech patterns, we showcase how AI audio detection fails to generalize. We note fine tuning on a single irregular dataset results in superior performance on other irregular data. This supports the collection and tuning on a wide variety of irregular speech patterns, as it increases performance of all marginalized group with insignificant effect on standard loss. Not only does this help the model performance, but results in a more equitable model.

VI. Ethics and Broader Impact

The real-world implications of deepfake audio are vast. Scammers are already using voice cloning AI to manipulate people, especially more vulnerable parts of the population, using impersonations. For example, there has been a rise in 'Grandparent Scams' in which bad actors use the voice of an elderly person's grandchild to threaten them for ransom. These scams can be incredibly convincing, even to the most cautious and aware person. According to the US Federal Trade Commission, Americans lost $2.7 billion from impostor scams in 2023 alone.[1] They also note that these numbers are likely

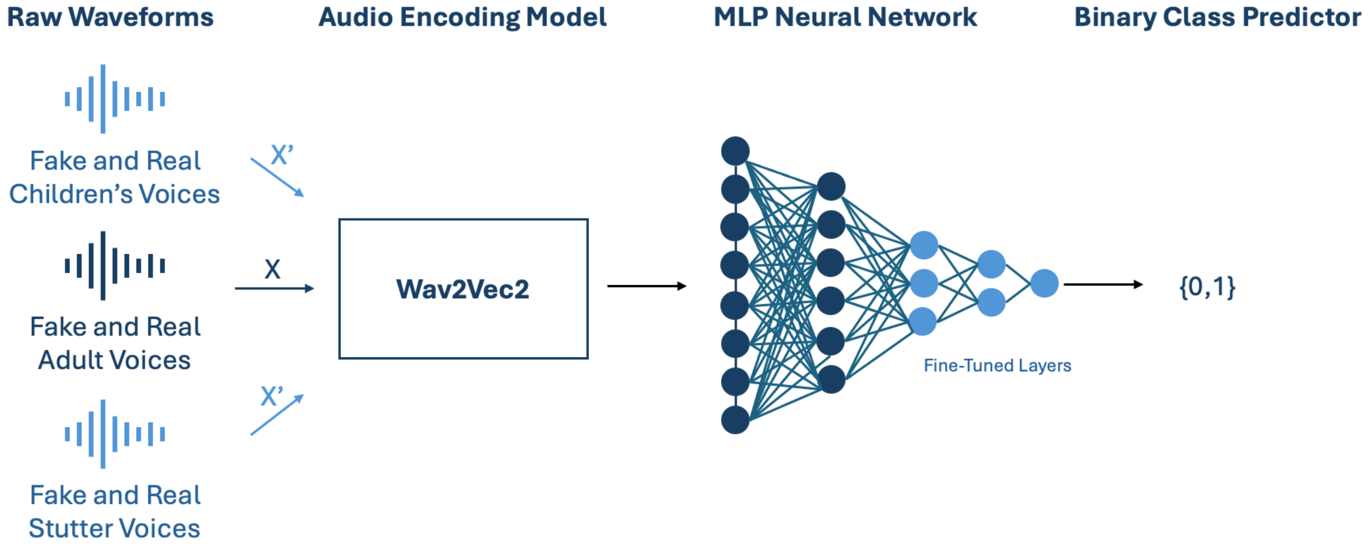

Fig. 1. Architecture for base model and finetuning.

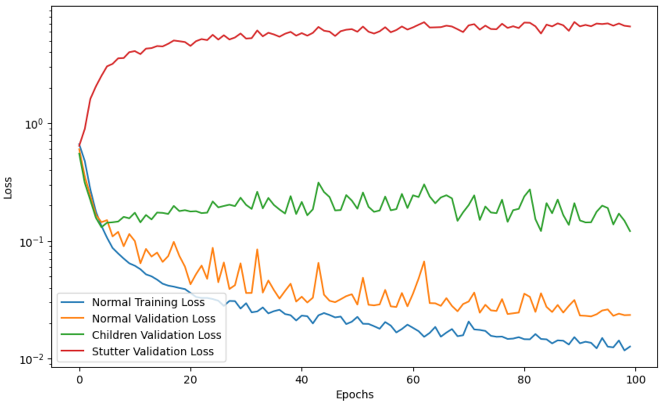

Fig. 2. Training and Validation losses when trained on standard voices.

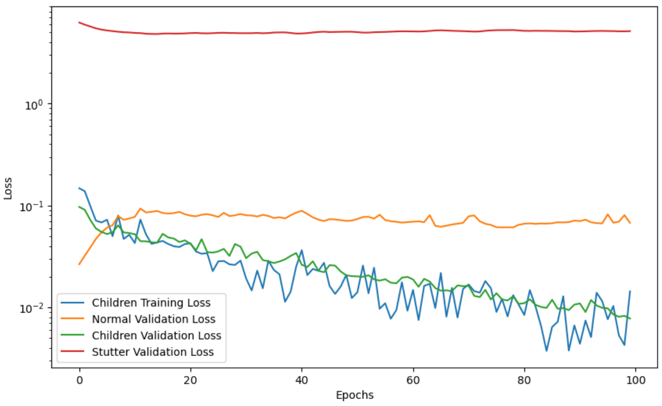

Fig. 3. Training and Validation losses when trained on children voices.

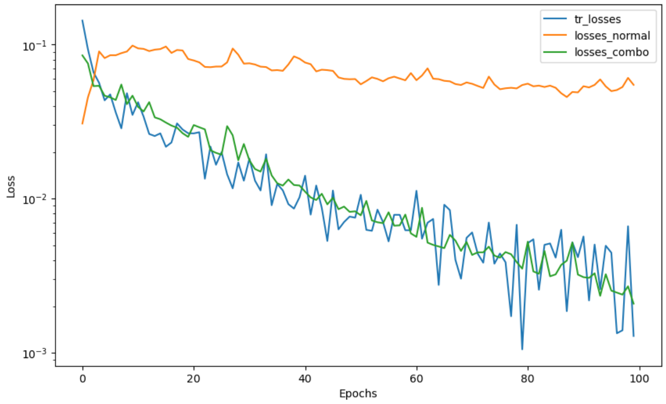

Fig. 4. Training and Validation losses when trained on a combination of children and stuttering voices.

an underestimate given many people do not end up reporting these scams out of shame, embarrassment, or fears that it will be too difficult to achieve accountability.

Beyond these scamming applications, there is also a huge vulnerability to adversarial attacks in voice assistants such as Alexa, Siri, or Google Nest. Malicious actors could use AI to generate inaudible frequencies of audio data to inject backdoor attacks or data poisoning attacks into models. This would enable them to manipulate these assistants to perform any number of unwanted actions. Considering many of these voice assistant devices are often connected to broader systems with access to people's other accounts and services, attacks like this could be weaponized to give adversaries an unprecedented level of access to people's personal lives. For example, one could imagine a backdoor attack that triggers Alexa to purchase a number of items on your amazon account, Siri to share your location with someone, or Google Nest to take control over all the smart devices in your home.

To combat these real threats from AI generated audio, robust deepfake audio detection models will undoubtedly need to be built into these systems. Even in a future where phone calls are monitored for AI generated audio as an enhanced means of warning people about scams, this will not be very helpful at preventing, for example, grandparent scams if the models are not equipped to detect fake children's voices. We wouldn't want a system that could easily avoid detection as long as the AI generated audio has a stutter to it.

Our work shows that ensuring this robustness can be quite difficult given the lack of quality, representative data to train with. Evidence suggests that less than 1% of the US population stutters and other speech impediments can be even rarer. [11] Although there are significantly more children in the world, privacy laws such as the European Union's General Data Protection Regulation (GDPR) have strict restrictions around the publishing and distribution of children's data on the internet. The lack of widely available access to these types of low-occurrence datasets imply that most models scraping the internet broadly for training data, will inherently be bad at detecting deepfakes of this form. To ensure that these vulnerable population are included in efforts to combat malicious attacks, future developers must make intentional efforts to include representative examples in training and skew their weight in the dataset to guarantee the model adjusts, despite their limited quantity.

References

[1] Anonymous. "As Nationwide Fraud Losses Top $10 Billion in 2023, FTC Steps Up Efforts to Protect the Public". In: *Federal Trade Commission*. 2024.

[2] Anonymous. "Increasing Threat of Deepfake Identities". In: *US Department of Homeland Security*. 2019.

[3] Alexei Baevski et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020. URL: https://arxiv.org/abs/2006.11477.

[4] Nicholas Carlini and David Wagner. "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text". In: *Proceedings of the 2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 532–547. DOI: 10.1109/SP.2018.00031. URL: https://arxiv.org/abs/1801.01944.

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *International Conference on Learning Representations (ICLR)*. 2015. URL: https://arxiv.org/abs/1412.6572.

[6] Peter Howell, Sue Davis, and Joanna Bartrip. "The University College London Archive of Stuttered Speech (UCLASS)". In: *Journal of Speech, Language, and Hearing Research* 52.2 (2009).

[7] Rishabh Jain and Peter Corcoran. *Improved Child Text-to-Speech Synthesis through Fastpitch-based Transfer Learning*. 2023.

[8] Davide Salvi, Paolo Bestagini, and Stefano Tubaro. "Synthetic Speech Detection through Audio Folding". In: *Proc. ACM International Workshop on Multimedia AI against Disinformation*. 2023.

[9] Natarajan Balaji Shankar et al. "The JIBO Kids Corpus: A speech dataset of child-robot interactions in a classroom environment". In: *JASA Express Letters* 4.11 (2024).

[10] Run Wang et al. "DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices". In: *Proc. ACM International Conference on Multimedia*. 2020.

[11] Ehud Yairi and Nicoline Ambrose. "Epidemiology of stuttering: 21st century advances". In: *Journal of Fluency Disorders* 38.2 (2013).