Problem

* Deepfake Audio generation techniques are able to imitate person’s speech, generate believable voice clips using only second long samples[1]. These generated synthetic audio is harder to detect with human ears than using machine-based detection method[1]
* Speech Synthesis and voice cloning[18]
* In order to access personal data for stealing private audio data of individuals, attackers sometimes infiltrates the organization/company/account by having the user click on malicious link (email, social media direct message, text message)[2.
* Crimes: Fake news, empowering criminals, super-change political mayhem, spam calls and whitle-collar crime[4]

Technique:

* 1. Text to Speech Text to speech synthesis(TTS)

Produce voice based on text so sound like target identity

* 1. Voice conversion (VC)

Transform the speech produced by source speaker, make the voice seem like it is spoken by target speaker

I. Defend Method - Detection Method – Dataset

Dataset

Based on several papers. There are three major datasets that has been used for evaluation of model performance:

1. ASVSpoof [7,11]
2. AVSpoof dataset[7,11]
3. More recent dataset: Fake or Real (FoR) dataset[7]

* 198,000 utterances including both real and synthetic speech generated using the latest algorithms[7]

1. **Google recently made** a [vast dataset of its own synthetic speech](https://www.blog.google/outreach-initiatives/google-news-initiative/advancing-research-fake-audio-detection/) available to researchers who are working on deepfake detection. This trove of training data can help AI systems find and recognize the hallmarks of fake voices.[4]

I. Defend Method - Detection Method

As mentioned before the fakes are getting better as technology advances and it is hard for human ears to identify the nuances between human-generated speech and machine-generated or alternated speech. In order to defend against Deepfake Audio, users can run the audio content through a deepfake detector, let the detector determine whether the audio is ad Deepfake or not[1,3]

One article mentions to train computer for inaudible hints that the voice couldn’t have come from an actual person[4]

Result from[7]

The detection method can be divided into 2 categories

1. Feature vector extraction and then input features to train models
2. Image based classification

Diagram, schematic

Description automatically generated

Feature-based approach:

1. Convert audio file into a feature-based dataset consisting of various spectral features of the audio sample[7]
   1. Step1. Input the audio file using librosa library -> output a time series and the sampling rate : which represents the digital form of the audio file
   2. Step2. Using librosa library functions output 37 features of audio as a vector
   3. [librosa](https://librosa.org/doc/latest/core.html" \l "module-librosa" \o "librosa) is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.
   4. Done using sliding window technique
   5. Refer to the paper[7] for specific information of each 37 features

Diagram

Description automatically generated

1. Feed the 37 spectral features into the machine learning algorithms for the classification of audio as fake or real[7]
2. Machine Learning Algorithm: Support Vector Machine (SVM), Light Gradient Boosting Machines (LGBM), Extreme Gradient Boost (XGBoost), K-Nearest Neighbors (KNN) and Random Forest (RF)[7]
3. Hyperparameters were optimized using GridSearchCV

Evaluation Metrics

1. Equal Error Rate (EER) [13]

* A statistic used to show biometric performance, typically when operating in the verification task
* The EER is the location on a ROC or [DET](https://www.innovatrics.com/glossary/detection-error-tradeoff-det/) curve where the [false acceptance rate](https://www.innovatrics.com/glossary/false-accept-rate-far/) and [false rejection rate](https://www.innovatrics.com/glossary/false-reject-rate-frr/) are equal. In general, the lower the equal error rate value, the higher the accuracy of the [biometric system](https://www.innovatrics.com/innovatrics-abis/). [12]

1. Tandem detection cost function (t-DCF) [13] : a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification [14]
2. Validation accuracy : on data that used to validate the generalisation ability during the

training process

1. Test accuracy : on data that did not use for training
2. Precision = TP/TP+FP [8]

Based on Confusion matrix ( True/False) (Positive/Negative)[8]

Measurement of ability to classify positive samples in the model[8]

If the goal is to classify all as well as negative samples as positive[8]

1. Recall = TP/TP+FN[8]

Measurement of how many positive samples were correctly classified by model[8]

If goal is to detect only all positive samples[8]

1. F-score

Measure of model’s accuracy on a dataset[9]

Used to evaluate binary classification systems (positive/negative)[9]

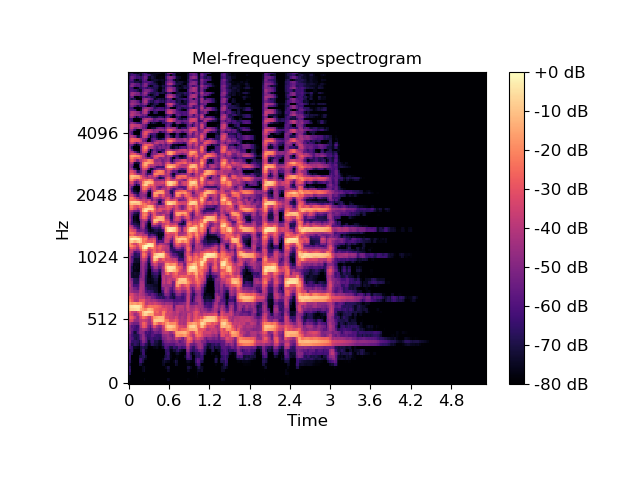
A perfect model has an F-score of 1[9]

Diagram

Description automatically generated with low confidence

Image-based approach

1. Convert audio samples into melspectrograms
2. Melspectrograms:
   1. Graphic representation of signal strength in terms of signal’s intensity
   2. Shows frequency variation of acoustic signals with respect to time



1. Input the melspectrograms into deep learning algorithms for classification
2. Deep Learning Architecture: Convolutional Neural Networks (CNN), Spatial Transformer Network (STN), and Temporal Convolutional Network (TCN)
3. TCN:
   1. is a sequential network which is better for capturing features in data which is also sequential in nature
   2. 1D Fully Convolutional Network (FCN) is used wherein the length of all hidden layers is kept equal to the length of the input layer along with zero padding which has a length of kernel size is equal to 1
4. STN:
   1. Allows spatial manipulation of the data
   2. Helps remove spatial invariance

Diagram

Description automatically generated

Based on the performance metrics presented in [8]. TCN achieves the highest test accuracy of 0.92. Results shows that Image-based classification models performs better than feature-based classification models[8]

Another Aspect:

Differences between human-generated and machine-generated speech[11]

1. Characteristics of human speech[11]
   1. Features extracted from voiced and unvoiced speech: Pitch, formants, jitter, shimmer, energy, loudness, zero-crossing rate, spectral entropy[11]
      1. Pitch : high/low of the sound
      2. Formant: frequency range, reflects sound special quality
      3. Jitter: slight unsteadiness in voice/sound
      4. Zero crossing rate: the rate at which a signal changes from positive to zero to negative or from negative to zero to positive
   2. Inhalation sounds are generally more natural and less affected by voluntary influences: breath sounds are bio-signatures that can be sued to identify speakers[11]
   3. Features that capture the fine-level inconsistencies and nuances of the speech production process could consistently exhibit differences between synthetic speech and genuine speech[11]

Table

Description automatically generated

Results show that spectral entropy of F0 sequence is a good indicator that captures statistical difference between synthetic speech and natural speech across datasets [11]

F0 sequence is trimmed to remove the zero values at the beginning and end of the sequence, then plot the spectral entropy distributions of ASVspoof data’s train/dev/eval set and find consistency patterns, also compute the distribution from the FoR dataset[11]

Spectral Entropy: spectral power distribution, a measure of signal irregularity

Chart

Description automatically generated

1. Models

Proposed global modulation feature model accuracy from 90% to 98%[11]

Used CNN as the baline model

Spectro-temporal feature: the 2D-DCT on log-Mel spectrogras

Best baseline of EER 4.03%[11]

Table

Description automatically generated

[8]

Proposed global modulation feature model accuracy from 90% to 98%[11]

Used CNN as the baline model

[1] Refers to Reference [10] R. Reimao and V. Tzerpos, "FoR: A Dataset for Synthetic Speech Detection," 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2019, pp. 1-10, doi: 10.1109/SPED.2019.8906599.

Conclusion:

Image-based classification approach is better

Image-based features and distributions serves as better features.

Highest Accuracy 92% - 98% in use of global modulation feature and CNN based approach

Pindrop

* Pindrop: an Atlanta company that sells voice authentication to big banks and insurance companies, is also developing defenses, worried that the next wave of attacks on its clients will involve deepfake audio.[4]
* One key to detecting fakes, according to the company: sounds that seem normal, but that people aren't physically capable of making.[4]
* An example from Pindrop CEO Vijay Balasubramaniyan: If you say "Hello, Paul," your mouth can only shift from the "o" to "Paul" at a certain speed. Spoken too fast, "the only way to say this is with a 7-foot-tall neck," Balasubramaniyan says.[4]
* Pindrop, the audio biometrics company, is developing synthetic voices in order to [train its own defenses to detect them.](https://www.axios.com/deepfake-audio-ai-impersonators-f736a8fc-162e-47f0-a582-e5eb8b8262ff.html)[5]

Labelling

* The videos/voice should be labeled is something is detected as being manipulated[3]
* hold those platforms who host and make deepfakes available to the public accountable and responsible for them[6]
* If a post has not had any type of trusted source or context provided, then correct labeling of the content should be clear to the viewer that the content source has been verified, is still being analyzed, or that the content has been significantly modified.[6]

Detection and Classification serves as the first step.

Then we can perform content labelling

1. Items that are classified as Deepfake should be labeled with Deepfake or other information. It should be clear to the viewer that the content has been significantly modified or not
2. Items that has no trusted source or has not been tested should also be labelled clear, that the content is not known whether trusty or not. So the viewers are being remined to be more careful with the information received
3. Item that passed the classification detection model can be labeled as trusty.

Raise user awareness (Company)

* Enforce strict verification procedure[1]
* Practice verification when recognize someone’s voice or face[1]
* Educate employees for the verification procedure[1]
* Train employees to properly answer, identify and react to suspicious calls[2]
* Engineers can develop and implement security training program[2,6]
* make sure your security training includes identifying modern tactics like deepfakes and mobile phishing – especially while people work remotely. Since we can’t walk down the hall to validate communication from a co-worker, encourage your employees to reach out over different channels.[6]
* sending a message through a collaboration system to verify that an unusual phone call was legitimate.[6]

Secure digital profile (Organization/Individual)

* Keep software up to date (security updates)[2]
* Routinely check privacy settings[2]
* Strong password[2]
* Account security: Multi-factor authentication on every account[1]
* Never navigate to a website from phone call or email link (check validity)[2]
* Check validity of callers (Solution: request the caller’s email address to send a email to check their identity, it can be similar to what websites has done, after entering correct password, they would send a security code though text, phone call or email)[2]
* Look up caller though employee system[2]

Individual awareness in data protection

* Limiting public presence on social media[1]
* Enabling privacy restrictions[1]
* Prevent scammers form easily stealing your voice[1]

How to evaluate the effectiveness of such defense

Enforce verification procedure and educate employees: [1]

Test employees by having them receive live calls from trained professionals who can emulate the tactics of real attackers[1]

What does it mean to audio/voice anonymization

Result from [16]

Speaker anonymization methods are a growing research area due to the common use of voice interfaces coupled with growing privacy requirements[16]

Voice cloning techniques (DeepFake Audio), Voice authentication systems

Need solution for anonymization of individual voices

Voice data

* Constitute personally identifiable information
* Contain user-sensitive information

Previous Method

1. Voice conversion to transform speaker’s voice to a new special speaker identity
   1. Require parallel training data
2. Ceptral frequency warping transformation, apply transformation function in the spectral domain, de-identifying the voice
3. Decomposing the audio into the identity components, x-vectors and non-identifying components and Replace x-vector with a pseudo x-vector

Metrics

1. EER
2. Log-likelyhood-ratio cost fnction
3. Discrimination loss component of this

Minimize the discriminating power of the classifier against a dataset

EER of 50% an Cmin of 1.00 are optimal

Word error rates are caculated using a TDNN-F acoustic model

Diagram

Description automatically generated

1. Input audio
2. Derives three components from the audio
   1. X-vector extracted by TDNN (Time delay neural network)
   2. The bottleneck features obtained by applying an automatic speech recognition acoustic model
   3. Pitch information (F0)
3. The pseudo x-vector represents the new identity of the speaker and is used for all utterances intended to be spoke by that identity
4. A speech synthesis module use F0 an BN features and the pseudo-x-vector to generate melspectrograms, outputs mel-filterbanks
5. NSF model processes these filterbanks, along with the F0 and the pseudo x-vector, generating the anonymized audio

Problem: in the generation process of the pseudo x-vector, averaging of several x-vectors reduce diversity of voices and lead to the reduction in entropy

x-vector generation

1. Learn the properties of x-vector space suing PCA – Principle Component Analysis on a large x-vector dataset
   1. PCA maps original data matrix onto an orthogonal space
   2. Let us extract a set of factors and then select a subset of those factors, reduce the dimensionality of the problem, those factors are the principle components (PCs) extracted through PCA)
2. Fit a generative model on the PCA-reduced space, in order to sample from it

Sample from the GMM(Gaussian Mixture Model) in the PCA reduced space and then apply the PCA inverse transform

* GMM: a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distribution with unknown parameters
* Inverse PCA:

Result

Table

Description automatically generated

(O-O) : baseline

(O-A): Original enrollment and anonymized trail

* Examines the difference between the original voices and an anonymized version of them
* EER decreasing up to 6.69%

(A-A): anonymized enrollment and anonymized trail

* Each of the enrollment and trail utterances are anonymized but to different identities

Improving the diversity of anonymize voices

C approaches a perfect score in many scenario, indicating strong anonymization

WER results 0.36% to 3.63%

Result from [15]

Voice Anonymization

* Modify the speaker identity vectors[15]
* Speaker identification algorithms can be used to determine the identity of a speaker through speech data[15]
* Convert the identity of one speaker to an anonymous speaker by removing voice biometrics of the speaker[15]
* Reduce the risk concerning the exploitation of voice-related personal information
* But at the same time, preserving the linguistic content of speech data[15]

Technique:

* User voice conversion techniques to convert the identity of one speaker to an anonymous speaker[15]

Voice Modulation[15]

* Alter various features of voice data (pitch, intonation) to create different speech styles. Attempts to remove the identity of the speaker from a given speech
* Example such as in TV news to ensure the anonymity of suspects and witnesses
* Implemented suing acoustic filters to alter the spectral characteristics of a given speech
  + Problem: the original speech can be easily recovered sing inverse filters
* Through synthesis: use fundamental frequencies(F0s) and aperiodicities(Aps) with spectrograms, the resultant speech can have different speaker characteristic

These two methods is not designed to retain the comprehensibility of the modulated speech

Need captions (TV news)

May not sound normal voice, cannot be use for training speech recognition sytems

Voice Conversion[15]

1. Convert the identity of the speaker of an input speech to that of the target speaker while retaining the linguistic content of the input speech
2. Simple form: requires parallel data for training (one-to-one speaker conversion)
   1. Problem: parallel data is expensive to collect, impractical
3. Variational autoencoders (VAEs) can be used to many-to-many voice conversion
   1. Using a single model and non-parallel training data
   2. VAEs can be combined with GANs to enhance the quality of converted speech, the decoder of VAE is shared with the generator of GAN
   3. Extended to include the cycle-consistency loss to improve voice quality for non-parallel training data
   4. CycleVAE-GAN

Diagram

Description automatically generated

1. z is the latent vector which corresponds to the linguistic information of the input speech
2. x is the input speech given by source speaker
3. x’
   1. is the reconstructed speech given I­X is the source speaker identity vector
   2. is the converted speech given I­X being the target speaker (here means the anonymous speaker) identity vector
4. x’’ is the converted back speech which should recover the original input speech x
5. dashed line represent the cyclic conversion path that produces x’’
6. When the speaker identity vector being replaced with an anonymous speaker identity vector, the speech can be anonymized

VAE

A picture containing text, clock

Description automatically generated

1. The model receives x as input
2. The encoder compresses it into the latent space
3. The decoder receives the information sampled from the latent space and produce x’, the decoder aims to make x’ as similar as possible to x

By minimizing the loss function of VAE (refer to [15] for the formula),

* the encoder is trained to extract latent vector z (contains the linguistic information of the input speech)
* the decoder is trained to reconstruct the input speech from the latent vector z and the source speaker identity I­X

The target speaker identity vector: a one-hot vector containing 1 for the target speaker and 0s for other speakers, which is fed into the decoding process of the VAE to convert the source speaker speech to the target speaker speech

Optimize the VAE module by minimizing the loss function of CycleAVE (refer to [15] for the formula) which is a combination of the VAE loss function and the cycle-consistency loss function

GAN part:

The GAN module is used to train the CycleVAE-GAN model

The decoder of the VAE is considered as the generator of the GAN

The discriminator of GAN helps the generator( VAE decoder\_ to produce a speech similar to that of the target speaker

Anonymous Speaker Identity Vector

Table

Description automatically generated

a) One-hot vector: used for converting to target speaker

minimize the voice biometric of a given speaker

Use uniform values

* 1. As in b) assign 0 to the source speaker and assign the value of 1/(n-1) to other speakers uniformly in the anonymous speaker identity vector
  2. As in c) assign -1 to the source speaker and assign the value of 2/(n-1) to other speakers uniformly in the anonymous speaker identity vector
  3. As in d) assign -1/(n-2) to the source speaker and 1/(n-2) to the other speaker
  4. All leads to summation of 1: the decoder of the VAE is trained to handle a speaker identity vector is a unit vector

Use nonuniform values

Table

Description automatically generated

1. As in a) use the cosine similarity of each pair of vectors
2. As in b) use the inverse of the cosine similarity values, setting the source speaker value to 0 and normalizing the values to obtain a sum of 1
3. Determination of k-farthest speakers and uniformly assigning the value of 1/k to them and 0s to others:
   1. C) one-farthest speaker is set to 1 and others are set to 0
   2. D) two-farthest speaker is set to 0.5 each and other are set to 0

Result

Data: VCC2016

Models

1. Gaussian Mixture Model (GMM)
2. Deep Neural Network (DNN)

Table

Description automatically generated

Reduced the identification accuracy of speech data to 0.1%-9.2%

While retain the speech recognition accuracy to 78.2%-81.3%[15]

Results from [17]

The previous two models targets speaker anonymization while preserving the linguistic content of the audio samples. In some scenarios we might want to conceal the linguistic content as well in order to protect user privacy.

The method mentioned in [17] achieves three objectives

1. Speaker de-identification

* Speaker anonymization

1. Content obfuscation

* Blur method

1. Scene preservation

* Used for scenarios where background need preservation , examples: monitor, analyze and mitigate urban noise pollution. Audios retrieved from acoustic sensors passively collect data in public places[17]

Model Steps

* 1. Extract voices from the mix of voice and background using deep neural network (U-Net), separates into 2 audio signals: the voice and the residual background
  2. Blur the separated voice to remove both speaker identifiable information and obscure linguistic content.
  3. Recombine the blurred voice with the background signals, resulting the anonymized resynthesis

Diagram

Description automatically generated

Evaluation metric

1. Content obfuscation: successfully obscure content, the efficacy of the process is a function of the quality of separation
2. Speaker anonymization: correct identification 29%
3. Scene preservation: MFCC-intersion method is the one that best preserve the scene

Separation is a crucial step in the overall system

Blur method:

1. Low pass filtering:
   1. Filter passes low frequencies and blocks high frequencies
   2. Most of the voice content is localized to the high frequencies
   3. A picture containing rectangle

      Description automatically generated
   4. Low-pass filter the separated voice at 250 Hz
2. MFCC inversion
   1. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum").
   2. Represent the spectral envelope of an audio signal
   3. Low order coefficients have been used in speech recognition systems because they broadly retain phonetic information while discarding much of the identifiable characteristics
   4. Compute MFCC coefficient and choose to retain only the first 5 coefficients for inversion, which is insufficient for speech recognition but still capture the general spectral envelope

Both blur method can be achieved using librosa

Results from [18]

[17] uses blur method to achieve both speaker anonymization and speech content desensitization, after the blur method, linguistic content is destroyed to a large extent.

[18] shows a method called speech sanitizer, which achieves three objectives

1. desensitize speech content

2. speaker anonymization

3. preserve the utility of the sanitized speech

Privacy Risks

1. Identity privacy breach: Attacker would be able to train voice models of this person based on speech samples, attackers may extract voiceprints and generate fake speeches that sound like us through speech synthesis and voice cloning
2. Speech content privacy breach: Attacker can analyze the speech content and learn more detailed information about the person

Diagram

Description automatically generated

Desensitize speech content[18]

1. Sensitive keyword determination
   1. Personalized keywords determination
      1. TF-IDF(term frequency-inverse document frequency)
         1. Idea: a person frequently using a certain word that is not popular among the others usually implies that the word is highly related to the person
         2. A term with higher TF-IDF is more likely to reveal the person’s private information
         3. Refer [18] for specific calculation formula
   2. Common keywords
      1. Collect reports from users by a designed efficient aggregation algorithm, which use a keyword list, makes it difficult for the provider to earn information with respect to specific users. Anonymous individual vote
      2. Estimate the frequency of each word
   3. Personalized keyword determination may overlook commonly used sensitive words that have low TF-IDF values, common key word determination might overlook words that are highly sensitive to only a small group of people. Thus these two methods are complementary and both should be put into use together.
2. using DTW-based keyword spotting to localize sensitive words
   1. DTW ( dynamic time warping)
      1. Scan input audio stream and compute the distance between audio signals
      2. Apply DTW algorithm to calculate the distance of two signals
      3. Use cosine distance of the audio’s STFT features as the distance metric
      4. If the distance between speech segment being scanned and the keyword sample is below a certain threshold, the segment is considered as an occurrence of the keyword
      5. Considering linking problem, the keyword detected by app will update the store keyword sample with this new audio sample
3. substitute with safe words
   1. randomized safe word substitution
   2. split a safe word set into different categories, after the keyword is detected, determine which category the keyword belongs to, and randomly picks a safe word from that category for the substitution

Anonymize user’s voiceprints – speaker anonymization[18]

Use voice conversion mechanism

Idea: utilize voice conversion to change the speaker’s pitch to hid the voiceprint

* in the bilinear function tunes the extent of distortion of the output voice
* would produce a deeper (low-pitched) output voice
* woud produce a sharper (high-pitched) output voice
* The output foice is not distorted at all if alpha = 0
* Select the best alphs that bring a considerable drop in the recognition accuracy
* Best proper range is [-0.1,-0.08] U[0.08,0.1]

Compound Warping functions

To prevent reversing attacks since warping functions are invertible (filinear function)

Use two independent parameters alpha, beta in combination, prohibits the attacker from conducting the reucing attack

A pair of alpha and beta

Compound two different warping functions

1. Alpha
2. Beta: deeper when beta<0, sharper when bta>0
3. H(w,a,b) = g(f(w,a),b)

Result:

1. Keyword detection accuracy up to 90%
2. Accuracy of speaker recognition on the output speech is under 0.2
3. Speech recognition accuracy maintained at 0.72 – 0.75

Conclusion

How to defend Deepfake Audio

Three aspect

1. Prevent the generation of good Deepfakes that does speech synthesis and voice cloning
   1. Speaker anonymization, prevent Deepfake algorithms from extracting speaker voice biometrics to clone voices
2. Be able to detect the Deepfakes
   1. Feature vector extraction
   2. Image based classification
3. Raise user awareness
   1. Security training program

Audio/Voice Anonymization

Two Privacy Risks

1. Identity privacy breach: Attacker would be able to train voice models of this person based on speech samples, attackers may extract voiceprints and generate fake speeches that sound like us through speech synthesis and voice cloning

Solution: Speaker Anonymization

1. Speech content privacy breach: Attacker can analyze the speech content and learn more detailed information about the person

Solution: Speech content desensitization

Relation between voice anonymization and deepfake audio defense

Overlapping: speaker anonymization

Defend Against DeepFake Audio Audio/Voice Anonymization

Speaker Anonymization

Speech Content Desensitization

DeepFake Audio Detection

Raise employee/Individual Awareness

Some Vocab

* Phishing: “practice of sending emails appearing to be from reputable sources with the goal of influencing or gaining personal information.” [n. 网络仿冒，网络钓鱼]
* Vising: [n. 网络电话诈骗]

Reference

[1] Seorg, “Deepfakes: How to defend yourself from attack,” Security Boulevard, 20-May-2020. [Online]. Available: https://securityboulevard.com/2020/05/deepfakes-how-to-defend-yourself-from-attack/.

[2] Social-Engineer, “Secure it – keep your digital profile safe from Vishers and phishers,” Social, 20-Nov-2020. [Online]. Available: https://www.social-engineer.com/secure-it-keep-your-digital-profile-safe-from-vishers-and-phishers/.

[3] J. Vincent, “Deepfake detection algorithms will never be enough,” The Verge, 27-Jun-2019. [Online]. Available: https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work.

[4] K. Waddell, “Researchers are figuring out how to detect audio deepfakes before it's too late,” Axios, 03-Apr-2019. [Online]. Available: https://www.axios.com/2019/04/03/deepfake-audio-ai-impersonators.

[5] K. Waddell, “This ai-generated Ellen DeGeneres Voice is the future of deepfakes,” Axios, 03-Apr-2019. [Online]. Available: https://www.axios.com/2019/04/03/deepfake-audio-ai-ellen-degeneres.

[6] C. Franklin, “Defending against deepfakes: From tells to crypto,” Dark Reading, 17-Sep-2020. [Online]. Available: https://www.darkreading.com/edge/defending-against-deepfakes-from-tells-to-crypto.

[7] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, “A Deep Learning Framework

for Audio Deepfake Detection,” Arabian journal for science and engineering (2011), vol.

47, no. 3, pp. 3447–3458, 2021, doi: 10.1007/s13369-021-06297-w.

[8] “Precision and recall in Machine Learning - Javatpoint.” [Online]. Available: https://www.javatpoint.com/precision-and-recall-in-machine-learning.

[9] “F-score,” DeepAI, 17-May-2019. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/f-score#:~:text=The%20F%2Dscore%2C%20also%20called,positive'%20or%20'negative'.

[10] R. Reimao and V. Tzerpos, "FoR: A Dataset for Synthetic Speech Detection," 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2019, pp. 1-10, doi: 10.1109/SPED.2019.8906599.

[11] Y. Gao, “Audio Deepfake Detection Based on Differences in Human and Machine

Generated Speech,” ProQuest Dissertations Publishing, 2022.

[12] “Equal error rate (EER) - biometric glossary,” Innovatrics, 05-May-2021. [Online]. Available: https://www.innovatrics.com/glossary/equal-error-rate-eer/.

[13] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, “Does Audio

Deepfake Detection Generalize?,” 2022.

[14] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, “T-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification,” The Speaker and Language Recognition Workshop (Odyssey 2018), 2018.

[15] I.-C. Yoo, K. Lee, S. Leem, H. Oh, B. Ko, and D. Yook, “Speaker Anonymization for

Personal Information Protection Using Voice Conversion Techniques,” IEEE access, vol.

8, pp. 198637–198645, 2020, doi: 10.1109/ACCESS.2020.3035416.

[16] H. Turner, G. Lovisotto, and I. Martinovic, “Generating identities with mixture models for

speaker anonymization,” Computer speech & language, vol. 72, p. 101318–, 2022, doi:

10.1016/j.csl.2021.101318.

[17] A. Cohen-Hadria, M. Cartwright, B. McFee, and J. P. Bello, “Voice Anonymization in

Urban Sound Recordings,” in 2019 IEEE 29th International Workshop on Machine

Learning for Signal Processing (MLSP), 2019, pp. 1–6. doi:

10.1109/MLSP.2019.8918913.

[18] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, “Speech Sanitizer: Speech Content

Desensitization and Voice Anonymization,” IEEE transactions on dependable and secure

computing, vol. 18, no. 6, pp. 2631–2642, 2021, doi: 10.1109/TDSC.2019.2960239.