Detecção de tentativa de invasão por dados sintéticos em aplicações de biometria por voz

Wilson A. de Oliveira Neto[1,2], **Carlos Maurício S. Figueiredo**[1,2]

[1]Núcleo de Computação – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brasil

[2]Samsung Ocean Center (OCEAN)
Manaus – AM – Brasil

{wadon.snf, cfigueiredo}@uea.edu.br

Abstract. *Voice-based biometric systems are very common nowadays, especially with the popularization of voice command systems and digital assistants such as Google Assistant or Alexa. An important feature of these systems is to detect the user giving a command, as it controls the access to personal or sensitive information to your profile. Thus, as in face-based biometrics, audio biometrics can be attacked by synthetic data, where recordings can be presented as real data. This work presents a model based on deep neural networks capable of detecting this invasion technique. For the training, we used real data of recordings and synthetic data generated from the original recordings. We obtained satisfactory results, mainly due to the low rate of false acceptance and high rate of F1-Score, even in different environments and noises.*

Resumo. *Sistemas de biometrias baseados em voz são muito comuns hoje em dia, principalmente com a popularização de sistemas de comando por voz e assistentes digitais, como Google Assistant ou Alexa. Uma funcionalidade importante desses sistemas é detectar o usuário que está emitindo um comando, pois, assim controla-se o acesso à informações pessoais ou sensíveis ao seu perfil. Assim, como em biometria baseada em faces, biometria por áudio pode sofrer ataques por dados sintéticos, onde gravações podem se fazer passar por dados reais. Este trabalho apresenta um modelo baseado em redes neurais profundas capaz de detectar essa técnica de invasão. Para o treinamento utilizou-se dados reais de gravações e dados sintéticos gerados a partir das gravações originais. Obteve-se resultados condizentes com a literatura, principalmente pela baixa taxa de falsa aceitação e alta taxa de F1-Score, mesmo em diferentes ambientes e ruídos.*

1. Introdução

Sistemas Biométricos podem ser descritos como sistemas capazes de identificar unicamente pessoas baseando-se em características físicas, ou seja, utiliza-se de diversas propriedades do corpo humano, como, por exemplo, impressões digitais, geometria das mãos, fala, olhos ou reconhecimento facial [White 2014]. Com o poder de processamento dos *smartphones* e assistentes pessoais, como *Google Home* e *Alexa*, elevou-se a quantidade de sensores biométricos, permitindo várias formas de autenticação. Uma autenticação

bastante utilizada é a de voz, pois é uma interface de comunicação natural que não necessita de conhecimento aprofundado por parte do usuário [Portet et al. 2013].

Apesar do menor risco de vulnerabilidade, as técnicas de identificação biométricas não são imunes à falsificações [Abbas et al. 2018]. Existem ataques específicos, como *spoofing*, que possibilitam a obtenção de dados de usuários através de acesso não autorizado ou pelo sucesso na quebra de sistemas biométricos.

Os sistemas biométricos podem ser trespassados através de uso de réplicas artificiais. Clone de gelatina de impressões digitais, máscaras, fotos e gravações são alguns exemplos [Smith et al. 2018]. Cada réplica possui um grau de dificuldade maior, porém, todas as modalidades físicas utilizadas em biometria são vulneráveis [Rebera et al. 2014].

Para resolver esse problema foram criados sistemas de ponta a ponta com processamentos de sinal de entrada e saída de áudio [Sajjad et al. 2018]. A utilização de redes neurais convolucionais para detecção de invasão vem sendo cada vez mais utilizada [Lai et al. 2018] [Chetttri et al. 2018]. Assim como múltiplos fatores de autenticação [Ye et al. 2019].

Desse modo, apresentamos um modelo de redes neurais profundas que detecta invasões de sistemas biométricos baseados em som. O intuito é verificar se o áudio recebido é uma gravação ou não, utilizando as métricas acurácia, taxa de erro e F1-score, a partir da criação de áudios sintéticos da base de dados *Commom Voice*, bem como analisar quais ambientes possuem maiores chances de acerto.

O trabalho está organizado da seguinte forma: a seção 2 aponta os principais trabalhos relacionados com esse artigo, a seção 3 descreve o método proposto e as tecnologias utilizadas, a seção 4 apresenta o sistema e sua arquitetura, a seção 5 descreve os experimentos e resultados e, por fim, a seção 6 apresenta as conclusões e possibilidade de trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção, serão discutidos trabalhos relacionados que propõem alguma solução ou mecanismo para biometrias e detecção de presença baseadas em voz. Serão abordados, também, os seguintes temas: reconhecimento por voz, segurança digital e assistentes virtuais.

Existem diversos estudos a respeito de reconhecimento e biometria de pessoas baseados em voz, como proposto por Omid Ghahabi et al. [Ghahabi and Hernando 2014], que utiliza redes neurais profundas capazes de reconhecer usuários impostores baseado no limiar da frequência do som. Outro trabalho, proposto por Jin Yang et al. [Yang et al. 2015] realiza a biometria de usuários através de um dispositivo corporal capaz de detectar a pressão sanguínea da garganta.

Entretanto, a biometria de usuários possui diversos problemas. Na biometria por imagem o trabalho de Pravallika et al. [Pravallika and Prasad 2016] aborda a importância da segurança nos sistemas biométricos e como prevenir a utilização de dados sintéticos como fotos ou máscaras através da diferença de qualidade da imagem. Para Marcos Faundez-Zanuy et al. [Faundez-Zanuy et al. 2006], os problemas de áudio sintéticos podem ser resolvidos por meio de uma modificação no sinal chamada de marca d'água, porém permite que o áudio sintético seja utilizado no microfone que processa o sinal.

Como é visto no trabalho de Xinyu lei et a. [Lei et al. 2017] assistentes virtuais possuem problemas de segurança digital, um deles, é a não detecção física do usuário, permitindo a utilização de uma gravação de um usuário autorizado para realizar um comando. Contudo o sistema *VAuth* [Feng et al. 2017] se propõe a solucionar o problema de autenticação contínua utilizando um *wearable*, porém necessita que o *hardware* esteja em contato com a pele do usuário, a partir disto ele calcula a vibração e verifica se coincide com o sinal recebido no assistente de voz.

De forma geral, a maioria dos trabalhos faz uso de *hardware* para a detecção de presença do usuário. Similar aos trabalhos relacionados, utilizamos redes neurais convolucionais para aprender as características de um áudio gravado e de um real. Nossa proposta se distingue das anteriores pelo fato de realizar a identificação de presença baseada somente em *software*, permitindo que o modelo seja portado por qualquer dispositivo ou utilizado em qualquer assistente de voz, não sendo dependente de algum *hardware* específico ou modificado. A Tabela 1 apresenta de forma resumida a diferença entre os trabalhos correlatos.

Tabela 1. Resumo comparativo de trabalhos correlatos

Referência	**Hardware**	**Software**	**Tempo Real**	**Detecção de gravação**
[Ghahabi and Hernando 2014]		X	X	
[Yang et al. 2015]	X	X	X	X
[Faundez-Zanuy et al. 2006]	X	X	X	X
[Feng et al. 2017]	X	X	X	X
Nossa proposta		X	X	X

3. Método proposto

Nesta seção será apresentada a descrição do conjunto de dados, pré-processamento, ferramentas e arquitetura do algoritmo de *machine learning* utilizado nesse estudo. Foi empregada a abordagem de aprendizagem supervisionada para a classificação do som.

3.1. Base de dados

Common Voice é uma base de dados com mais de 14.000 horas de áudio validadas e rotuladas com suas respectivas frases. Os áudios foram gravados por pessoas reais que recitaram frases de licenças abertas ou cedidas [Mozilla 2019].

Começando em 2017, o projeto *Common Voice* é uma ação da *Mozilla* com objetivo de ensinar às máquinas como pessoas reais falam. Em 2018 foi lançada na plataforma *Kaggle* possuindo cerca de 500 horas de gravação. As frases são curtas e possuem aproximadamente 4 segundos de duração.

A base de dados que foi utilizada nesse trabalho tem cerca de 9 horas de áudios gravados, divididos em 8.069 arquivos, entretanto, essa base foi complementada sinteticamente, ou seja, o áudio original foi regravado a partir das seguintes fontes externas: celular, computador e caixas de som sem fio. Ao total possui-se 33.158 arquivos incluindo os áudios originais da base *Common Voice*.

Os rótulos desta base de dados foram divididos dessa maneira: 8.069 arquivos de áudio para a classe pessoa e 25.089 arquivos de áudio para a classe gravação.

3.2. Pré-processamento

Para que o processamento do sinal não perca frequências e espectros, utilizou-se a conversão para o formato WAV. O modelo atua utilizando quadros de $0,96$ segundos e 64 faixas de frequência na entrada para o log Mel espetrograma [Hori et al. 2018], extraídos a partir de um áudio com a taxa de $22Khz$, criado a partir da transformada de *Fourier* de tempo curto (STFT) [Zhao et al. 2015], com o tamanho de janela $0,025$ segundos e sobreposição de 40% [Hershey et al. 2017].

Os 3 primeiros segundos de cada áudio são lidos e pré-processados, na qual é gerada a saída de 3 dimensões: 64x258x1, sendo 64 a primeira dimensão, representando a faixa de frequência; 258 a segunda dimensão, sendo os 3 conjuntos concatenados de $0,96$ segundos retirando os $0,5$ segundos iniciais; e 1 é terceira dimensão, adicionada para que a rede convolucional possa lê-la como imagem.

3.3. Ferramentas

Foi utilizado a linguagem de programação *Python* na versão $3.6.5$ em conjunto com a biblioteca de abstração de *machine learning* Keras na versão $2.2.4$, aplicando o TensorFlow na versão 1.11 como *backend* de processamento.

Para a manipulação do áudio utilizou-se as bibliotecas *Pyaudio* $0.2.11$, responsável por capturar o som do microfone e $Librosa$ $0.6.2$ para realizar a etapa de pré-processamento.

Na manipulação e processamento dos dados foi utilizada a biblioteca *Numpy* $1.15.4$, *Pandas* $0.23.4$ e *Dask* $1.1.2$. O ambiente de programação foi o *Jupyter Notebook* $1.0.0$ para o treinamento do modelo e o *Visual Studio Code* $1.32.1$ para o pré-processamento e aquisição dos dados.

3.4. Arquitetura do modelo de aprendizagem

A Figura 1 é a representação gráfica da arquitetura da rede neural convolucional (CNN) e possui ao todo $41.002.850$ de parâmetros treináveis. Recebe o espectrograma de entrada com dimensão 64x258 *pixels* em escala de cinza e possui todos os *kernels* retangulares de dimensão 4x10, com 32 mapas de características em cada camada. Em cada camada de convolução é adicionada a propriedade *padding same*, para que não haja redução de dimensão da imagem.

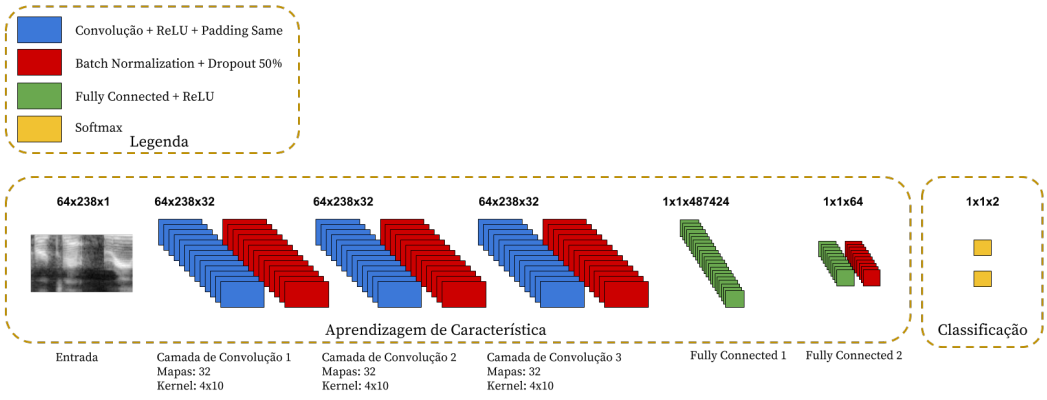


Figura 1. Arquitetura da rede neural convolucional.

Cada saída da camada de convolução está ligada à uma camada de *Batch Normalization*[Ioffe and Szegedy 2015] e uma de *Dropout* [Srivastava et al. 2014] com 50% de perda. Ao total foram 3 camadas de convolução e 2 camadas totalmente conectadas com a função de ativação *ReLU*. A camada de saída possui a função de ativação *Softmax* para representar a porcentagem de certeza em cada classe. O modelo foi compilado utilizando a função de entropia cruzada [Zhang and Sabuncu 2018] com a função otimizadora adam [Kingma and Ba 2014].

4. Sistema proposto

O programa é dividido em 4 partes, como mostra a Figura 2: Primeiramente, o modelo previamente treinado é carregado na memória do computador (I). Captura do áudio de duração fixa de 4 segundos, utilizando as seguintes configurações: tamanho de bloco 1024 *bytes*, taxa de $16Khz$ e canal mono (II). Criação de espectrograma do áudio gravado (III). Classificação do espetrograma (IV).

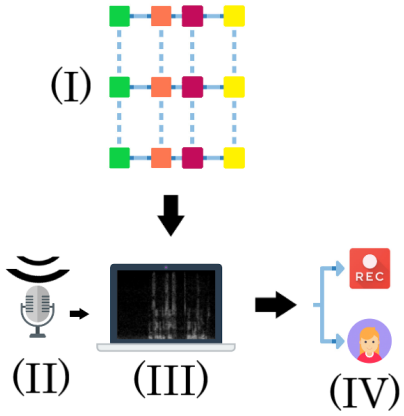


Figura 2. Arquitetura do sistema.

5. Experimentos e Resultados

A CNN foi executada em uma GPU GTX 1050 e os algoritmos de pré-processamento em um *notebook Samsung Odyssey* com CPU Intel i7 7700HQ 2.80GHz e 8GB de memória RAM.

Além disso, foram utilizados cinco auto-falantes, sendo eles duas caixas de som portáteis de modelo: JBL Go 2 e JBL *Charge* 3 , dois celulares de modelo: *Samsung Galaxy* $S7$ e *Samsung Galaxy* $S9$ e por fim, auto-falante nativo do *notebook* utilizado para processar os áudios.

5.1. Treino

O treinamento foi feito a partir da base de dados e arquitetura citados anteriormente, levou cerca de 3 minutos para treiná-lo, sendo necessário 30 épocas com a média de 55.75 segundos cada. Do total de 33.158 áudios processado, foram selecionados $16,248$ para treino, 6.962 para o teste e 9.948 para validação, ou seja, *holdout* de $50 - 20 - 30$.

Na Figura 3 são mostrados os dois valores próximos um do outro, no treinamento a acurácia final foi de $99,72\%$ e no teste $99,19\%$. Na Figura 4 é mostrado que o treino e teste estão próximos entre si e de 0, ou seja, a rede convergiu sem sobre ajuste.

Figura 3. Curva de acurácia

Figura 4. Curva de perda

Este módulo foi executado de maneira isolada, uma vez que, após o treino é possível serializar o modelo para ser utilizado posteriormente em outra situação, como em assistentes pessoais ou dispositivo embarcado.

5.2. Validação

A validação geral do modelo obteve $99,81\%$ para acurácia e $3,71\%$ para a taxa de erro. Esses valores são condizentes ao treino e teste e com isso é possível concluir a generalização do modelo. Para ficar mais claro, é necessário obter outras métricas como: precisão, revocação e F1-Score. Na Tabela 2 é possível ver os dados unitários das 3 métricas.

Tabela 2. Análise geral do modelo

	Precisão	**Revocação**	**F1-Score**
Pessoa	97%	99%	98%
Gravação	100%	99%	99%
Micro avg	99%	99%	99%
Macro avg	98%	99%	99%

As métricas para cálculo de sistemas de biometria são denotadas por: *False Accepted Rate* - FAR, *False Reject Rate* - FRR, *Equal Error Rate* - EER e Acurácia - ACC [Singh et al. 2018]. Na Tabela 3 é apresentada a média de cada métrica por meio da validação cruzada.

Tabela 3. Métricas biométricas

FAR	**FRR**	**EER**	**ACC**
0,24%	0,01%	0,27%	99,64%

Essas métricas são baseadas na matriz de confusão apresentada na Tabela 4. Na diagonal principal da matriz de confusão normalizada é possível observar os fundos das células mais escuros, o que demonstra alta densidade de valores. Este comportamento indica a classificação correta dos elementos de teste. Apesar do desbalanceamento da base

de dados, apresentaram-se resultados próximos de 1. Na Tabela 4 é possível, também, identificar o valor de verdadeiro positivo, falso positivo, verdadeiro negativo e falso negativo, que são respectivamente $0,9881$, $0,0106$, $0,9894$ e $0,0119$.

Tabela 4. Matriz de confusão

		Classe Predita	
		Pessoa	**Gravação**
Classe Real	**Real**	0,9881	0,0119
	Gravação	0,0106	0,9894

5.3. Procedimento Experimental

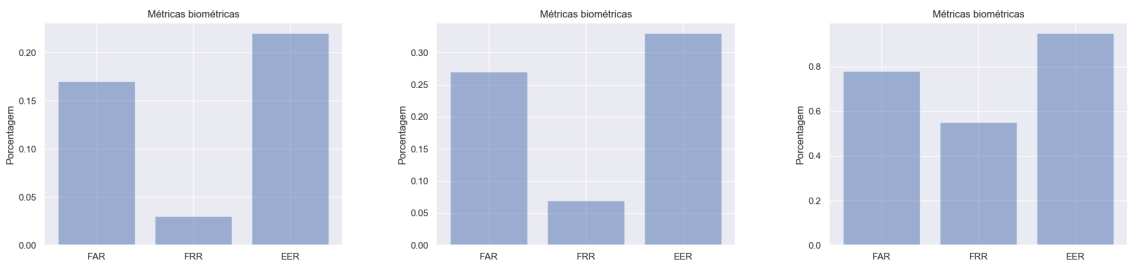
Os experimentos simularam uma situação real na qual um usuário fala em frente ao microfone e recebe um *feedback* da captura do áudio. Para tanto, o experimento foi realizado em diferentes ambientes e ruídos. O computador que processa e classifica o áudio é o mesmo utilizado para treinar o modelo anteriormente.

Para o experimento foram utilizadas 5 frases no idioma inglês, sendo repetidas duas vezes, uma para gravação e outra ao vivo. Nenhuma das frases abaixo foram utilizados no treino ou teste:

1. *Neural networks are really really cool, i do love them*;
2. *Go lang is probally the best programming lang that i know*;
3. *Kotlin is way better then java programming*;
4. *The new samsung smartphone are actually super cool*;
5. *Galaxy s10 came with brand new outfit*;

Esse conjunto de frases foi repetido por um total de 7 pessoas, sendo 2 deles por mulheres, houve variação na tonalidade e tempo de cada frase. Ao total o número de frases testadas foi 210. Cada conjunto de frases foi gravada e testada em um ambiente diferente. Os ambientes foram: sala fechada sem ruído, sala fechada com ruído ambiente de ar-condicionado e ambiente de trabalho aberto com cerca de 6 pessoas presentes conversando.

Cada ambiente obteve uma sequência de métricas normalizadas de 0 a 100, como pode ser observado nas figuras 5a, 5b e 5c.



(a) Sala sem ruído (b) Sala com ar-condicionado (c) Ambiente de trabalho

Figura 5. Resultados de experimentos reais

A Figura 5a possui os menores valores para as métricas FAR, FRR e EER comparado aos demais, pois o ruído ambiente é menor. Já na Figura 5b pode-se notar os valores de FAR, FRR e EER mais elevados quando comparados à figura anterior. Isso se dá ao fato de o ruído do ar-condicionado ter influenciado levemente na amostragem. A Figura 5c possui os maiores valores das métricas FAR, FRR e EER, evidenciado que os ruídos do ambiente influenciam diretamente na taxa de erro. Além disso, quanto maior o número de pessoas dentro do mesmo ambiente, tornando-o, consequentemente, mais barulhento, mais expressiva é a elevação nos valores.

6. Conclusão e trabalhos futuros

Este artigo apresentou um sistema capaz de validar a presença do usuário real em um ambiente baseado em voz. Através de uma base de dados pública gerou-se dados sintéticos capazes de representar a diferença entre áudios produzidos por um ser humano presente, de áudios produzidos por dispositivos sonoros, como caixas de som e celulares. As classes rotuladas nesse trabalho são pessoa e gravação.

Os experimentos mostraram que o modelo treinado obteve a acurácia média de $99,72\%$ nos treinos e a acurácia média de $98,14\%$ nos experimentos. Os valores são significativos pois no campo da biometria já se tornam resultados satisfatórios e possivelmente comercialmente aceitos. Em experimento com pessoas reais, com frases diferentes das usadas no treino, observou-se um desempenho similar aos demais experimentos, visto que os erros foram muito baixos.

Assim podemos concluir que a rede neural foi capaz de identificar características do áudio quando este é gerado por alto falantes. Em uma primeira análise, comparando áudios podemos observar que alto falantes limitam respostas em frequências e introduzem reverberação, sendo essas características detectadas pela rede neural.

Como trabalho futuro, pretende-se diminuir a taxa de falsos positivos, além de aumentar a base de dados para outros idiomas levando em consideração que diferença fonética dos idiomas influencia no reconhecimento de voz [Mary Zarate et al. 2015]. Além disso, pretende-se entender melhor quais as principais características são encontradas em alto falantes, aprendidas pela rede neural de forma a prever possíveis ataques de contorno [Wu et al. 2012].

Referências

Abbas, G., Humayoun, S. R., AlTarawneh, R., and Ebert, A. (2018). Simple shape-based touch behavioral biometrics authentication for smart mobiles. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, AVI '18, pages 50:1–50:3, New York, NY, USA. ACM.

Chetttri, B., Mishra, S., L. Sturm, B., and Benetos, E. (2018). Analysing the predictions of a cnn-based replay spoofing detection system. pages 92–97.

Faundez-Zanuy, M., Hagmüller, M., and Kubin, G. (2006). Speaker verification security improvement by means of speech watermarking. *Speech Communication*, 48(12):1608 – 1619. NOLISP 2005.

Feng, H., Fawaz, K., and Shin, K. G. (2017). Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, MobiCom '17, pages 343–355, New York, NY, USA. ACM.

Ghahabi, O. and Hernando, J. (2014). i-vector modeling with deep belief networks for multi-session speaker recognition.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. (2017). Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135.

Hori, C., Alamri, H., Wang, J., Winchern, G., Hori, T., Cherian, A., Marks, T., Cartillier, V., Gontijo Lopes, R., Das, A., Essa, I., Batra, D., and Parikh, D. (2018). End-to-end audio visual scene-aware dialog using multimodal attention-based video features.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Lai, C.-I., Abad, A., Richmond, K., Yamagishi, J., Dehak, N., and King, S. (2018). Attentive filtering networks for audio replay attack detection.

Lei, X., Tu, G., Liu, A. X., Li, C., and Xie, T. (2017). The insecurity of home digital voice assistants - amazon alexa as a case study. *CoRR*, abs/1712.03327.

Mary Zarate, J., Tian, X., Woods, K. J. P., and Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, 5:11475 EP –. Article.

Mozilla (2019). Common Voice by Mozilla — Common Voice. `https://mzl.la/voice`. [Online; accessed 2019-01-10].

Portet, F., Vacher, M., Golanski, C., Roux, C., and Meillon, B. (2013). Design and evaluation of a smart home voice interface for the elderly: Acceptability and objection aspects. *Personal and Ubiquitous Computing*, 17(1):127–144.

Pravallika, P. and Prasad, K. S. (2016). Svm classification for fake biometric detection using image quality assessment: Application to iris, face and palm print. In *2016 International Conference on Inventive Computation Technologies (ICICT)*, volume 1, pages 1–6.

Rebera, A. P., Bonfanti, M. E., and Venier, S. (2014). Societal and ethical implications of anti-spoofing technologies in biometrics. *Science and Engineering Ethics*, 20(1):155–169.

Sajjad, M., Khan, S., Hussain, T., Muhammad, K., Sangaiah, A. K., Castiglione, A., Esposito, C., and Baik, S. W. (2018). Cnn-based anti-spoofing two-tier multi-factor authentication system. *Pattern Recognition Letters*.

Singh, N., Agrawal, A., and Khan, P. R. (2018). Voice biometric: A technology for voice based authentication. *Advanced Science, Engineering and Medicine*, 10.

Smith, M., Mann, M., and Urbas, G. (2018). *Biometrics, Crime and Security*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

White, J. M. (2014). Chapter 12 - access control. In White, J. M., editor, *Security Risk Assessment*, pages 149 – 160. Butterworth-Heinemann, Boston.

Wu, Z., Siong, C. E., and Li, H. (2012). Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In *INTERSPEECH*.

Yang, J., Chen, J., Su, Y., Jing, Q., Li, Z., Yi, F., Wen, X., Wang, Z., and Wang, Z. L. (2015). Eardrum-inspired active sensors for self-powered cardiovascular system characterization and throat-attached anti-interference voice recognition. *Advanced Materials*, 27(8):1316–1326.

Ye, D., Zhang, T.-Y., and Guo, G. (2019). Stochastic coding detection scheme in cyber-physical systems against replay attack. *Information Sciences*, 481:432 – 444.

Zhang, Z. and Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 8778–8788. Curran Associates, Inc.

Zhao, Y., Zou, Z., Wu, L., and Li, Y. (2015). Frequency detection algorithm for frequency diversity signal based on stft. In *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, pages 790–793.