Avaliação de modelos para detecção de ataques de *replay* usando diferentes bases de dados

Giovana Y. Nakashima[1], Higor D. C. Santos[1], Jone W. M. Soares[1], Mário Uliani Neto[1], Fernando O. Runstein[1], Ricardo P. V. Violato[1], Marcus Lima[2]

[1]CPQD - Centro de Pesquisa e Desenvolvimento, Campinas, SP, Brasil

[2]Pontifícia Universidade Católica de Campinas, SP, Brasil

{giovana.nakashima, higorcea, jonewisney, marcuslima3}@gmail.com
{uliani, runstein, rviolato}@cpqd.com.br

Abstract. *A replay attack is a speech forgery used in an attempt to authenticate a speaker. Deep neural networks have been proposed as methods for detecting fraudulent audio. In view of the use of these models in real applications, in addition to good learning performance it is expected that the models show good results with databases other than the one used for training. In this work two approaches were evaluated with three public databases, with results that indicate low generalization capacity of the models.*

Resumo. *Ataque de replay é uma falsificação de fala utilizada na tentativa de autenticação de locutor. Redes neurais profundas têm sido propostas como métodos para detecção de áudios fraudulentos. Tendo em vista a utilização desses modelos em aplicações reais, além de bom desempenho na aprendizagem, espera-se que o modelo obtido apresente bons resultados com bases de dados distintas da utilizada no treinamento. Neste trabalho, duas abordagens foram avaliadas com três bases de dados públicas, com resultados que indicam baixa capacidade de generalização dos modelos.*

1. Introdução

Sistemas de biometria de voz estão sendo amplamente utilizados nos mais diversos setores, como indústria automotiva, financeiro, saúde e educação [Khan et al. 2023]. Nesses casos, a autenticação do usuário é realizada por sistemas de verificação automática de locutor (*Automated Speaker Verification* – ASV), suscetíveis a ataques de falsificação (*spoofing*).

Ataque de *replay* consiste na apresentação a um ASV da reprodução de um áudio previamente gravado, com o objetivo de validar a fala do locutor como genuína. Esse tipo de falsificação ocorre de forma passiva e é difícil de ser detectado, uma vez que o sinal reproduzido apresenta semelhanças físicas (frequências, espectros, formas de ondas) ao original [Khan et al. 2023]. Para aumentar a confiabilidade, os sistemas ASV são combinados com sistemas que identificam a fala falsificada, também chamados de *antispoofing* [Alzantot et al. 2019].

A série de competições *Automatic Speaker Verification Spoofing and Countermeasures Challenge* (ASVspoof) promove, desde 2015, o desenvolvimento de métodos para

detecção de falsificação. A cada edição, uma base de dados é disponibilizada para ser utilizada no treinamento e na validação de contramedidas aos ataques [Alzantot et al. 2019, Lee et al. 2022, Nautsch et al. 2021].

De forma geral, os métodos iniciam com a extração de atributos (*features*), predominantemente baseados na análise espectral do áudio, como espectrogramas, cepstrogramas, coeficientes em escala mel e variações da análise de Fourier [Khan et al. 2023]. O processo de aprendizado ocorre utilizando-se esses atributos como entradas para um classificador, tipicamente uma rede neural. Frequentemente tem sido utilizadas redes convolucionais [Chettri et al. 2018, Korshunov et al. 2018] (*Convolutional Neural Network* - CNN) e suas variações, como a rede convolucional leve (*Light Convolutional Neural Network* - LCNN) [Lavrentyeva et al. 2017, Lavrentyeva et al. 2019] e a rede convolucional residual (*Residual Neural Network* - ResNet) [Alzantot et al. 2019, Zhang et al. 2021].

Usualmente, para utilização em uma aplicação real, é esperado que o modelo apresente uma boa capacidade de generalização [Korshunov and Marcel 2016], isto é, que seu desempenho não seja muito diferente quando comparados os resultados obtidos em dados distintos dos usados no treinamento.

O objetivo deste trabalho é estudar o desempenho de abordagens propostas à detecção de ataque de *replay* entre bases de dados distintas da utilizada no aprendizado das redes, de modo a contribuir para a discussão sobre generalização dos modelos.

2. Metodologia

Neste trabalho, duas abordagens de classificação foram avaliadas com três bases de dados públicas. O desempenho dos métodos foi mensurado pelo EER (*Equal Error Rate*), ponto de operação em que a taxa de falsa aceitação (*False Acceptance Rate* - FAR) e a taxa de falsa rejeição (*False Rejection Rate* - FRR) são iguais [Jain et al. 2008].

2.1. Bases de Dados

O estudo foi realizado com três bases de dados públicas: ASVspoof 2019[1], ASVspoof 2021[2] e REMASC[3]. As três bases foram gravadas em inglês e, portanto, a influência da variação de idioma não pode ser explorada nesse caso. O treinamento de todos os modelos neste trabalho utilizou o conjunto de treinamento da base de dados ASVspoof 2019.

As bases ASVspoof 2019 e ASVspoof 2021 são compostas por arquivos de áudio do tipo *flac*, com um canal e taxa de amostragem de 16kHz. Ambas possuem outros tipos de ataque além do ataque de *replay*, mas, para os experimentos deste trabalho, foram utilizados apenas os conjuntos relativos ao ataque denominado de acesso físico (*Physical Access* - PA), pois são os dados com o ataque de *replay*. Esses conjuntos são formados por 218.430 e 943.110 amostras, respectivamente.

A base de dados REMASC foi concebida visando sistemas controlados por voz (*Voice Controlled Systems* - VCS), em que a coleta do áudio ocorre a uma distância maior do locutor. Seu conjunto abrange 54.712 amostras, armazenadas em arquivos do tipo *wav*, multicanais, amostrados a 16kHz e 44kHz [Gong et al. 2019]. Os áudios foram

[1]https://datashare.ed.ac.uk/handle/10283/3336

[2]https://www.asvspoof.org/index2021.html

[3]github.com/ndmobilecomplab/replay-attack

padronizados em monocanais a 16kHz, aplicando a média dos canais e redução da taxa de amostragem (*downsampling*) quando necessário.

As três bases de dados disponibilizam arquivos de protocolo, indicando quais dados devem ser usados para treinamento e teste, bem como a identificação do locutor e do áudio e sua classificação original (genuíno ou falso). Além disso, fornecem metadados como tamanho do ambiente e características dos dispositivos utilizados para coleta.

2.2. Modelos

Neste trabalho, foram avaliadas dois tipos de arquitetura de redes neurais, ResNet e LCNN. A ResNet avaliada usa como atributos de entrada a magnitude do espectro em escala logarítmica (*Log-magnitude STFT - Short-Time Fourier Transform*). Foi utilizado um modelo pré-treinado disponibilizado publicamente [4].

Quanto à rede LCNN, foi utilizada uma implementação disponibilizada publicamente [5] e usada em um estudo que avaliou diversos atributos como entrada para a rede [Lee et al. 2022, Lee 2024]. Neste caso, como não há modelo pré-treinado disponível, o treinamento foi executado utilizando os seguintes atributos: análise discreta arbitrária de Fourier (arbitrary discrete Fourier analysis - ADFA), cepstrogramas (CEPS e CEPS1724), constant Q analysis (CQA), transformada discreta de cosseno (discrete cosine transform - DCT), análise discreta de Fourier em escala Mel (Mel-scale discrete Fourier analysis - MDFA) e espectrogramas (Spec e Spec1724). Os espectrogramas e cepstrogramas foram extraídos pela transformada rápida de Fourier (*Fast Fourier Transform* - FFT) utilizando uma janela de Blackman com comprimento 1024 (Spec e Ceps) e 1724 (Spec1724 e Ceps1724).

Ainda, como referência para comparação, foi usada a abordagem LFCC-LCNN disponibilizada como *baseline* para o desafio ASVspoof 2021 [Liu et al. 2023], que inclui um modelo pré-treinado.

3. Resultados e Discussão

Os treinamentos da abordagem RD-LCNN foram realizados ao longo de cem épocas e o foi escolhido o modelo obtido a partir da época com menor valor EER no conjunto de desenvolvimento da base ASVspoof 2019.

A Tabela 1 apresenta os resultados de todos os modelos avaliados nas três bases de dados, ASVspoof 2019, ASVspoof 2021 e REMASC. As colunas (a), mostram como referência os resultados relatados na literatura para os subconjuntos de desenvolvimento (Dev) e de avaliação (Eval) da base ASVspoof 2019 [Nautsch et al. 2021] e, nas demais colunas, os resultados obtidos nos experimentos deste trabalho.

O modelo pré-treinado da *baseline* LFCC-LCNN apresentou bom desempenho com o subconjunto *eval* da base de dados ASVspoof 2019, com EER = 2,43%, fato esperado, uma vez que o treinamento ocorreu com o subconjunto *train* dessa mesma base de dados. É importante observar que a mesma superou as *baselines* propostas em 2019, CQCC-GMM e LFCC-GMM, que apresentaram EER = 11,04% e EER = 13,54%, respectivamente [Nautsch et al. 2021]. Já com a base ASVspoof 2021, o desempenho (EER

[4]https://github.com/nesl/asvspoof2019

[5]https://github.com/shihkuanglee/RD-LCNN/tree/main

Tabela 1. Resultados do ERR (%) obtidos (ASVspoof 2019 (b), ASVspoof 2021 e REMASC) e reportados pela literatura (ASVspoof 2019 (a))

	ASVspoof 2019 (a)		ASVspoof 2019 (b)			ASVspoof 2021	REMASC
Modelo	**Dev**	**Eval**	**Train**	**Dev**	**Eval**	**Eval**	**-**
Baseline LFCC-LCNN	42,16	*	**	42,16	2,43	45,67	***
Spec - ResNet	3,85	3,81	**	3,85	7,07	43,21	48,49
ADFA - RD-LCNN	0,22	0,85	0,71	0,83	1,67	40,85	60,87
Ceps - RD-LCNN	**0,13**	**0,37**	**0,20**	**0,17**	**0,41**	37,13	49,21
Ceps1724 - RD-LCNN	0,28	0,71	0,23	0,28	0,80	49,77	51,66
CQA - RD-LCNN	0,35	0,74	0,75	0,76	0,97	46,72	49,78
DCT - RD-LCNN	1,44	2,90	1,58	0,69	11,66	38,15	56,85
MDFA - RD-LCNN	0,17	0,61	0,53	0,59	1,39	43,02	59,58
Spec - RD-LCNN	0,56	1,72	0,67	1,39	1,86	43,51	52,01
Spec1724 - RD-LCNN	0,20	0,92	0,58	0,79	1,65	35,65	50,54

* Resultado não disponível na literatura.

** Emprego de modelo pré-treinado; valor EER no conjunto train não reportado na literatura.

*** Processamento não realizado.

= 45,67%) foi similar ao relatado pela literatura (EER = 44,77%) [Liu et al. 2023] e, portanto, muito pior.

A abordagem ResNet para a base de dados ASVspoof 2019 do subconjunto *eval* apresentou resultado (EER = 7,07%) próximo ao da literatura (EER = 3,81%) e, embora 86% maior, ainda foi melhor que os das *baselines* do desafio de 2019: EER = 11,04% (B01 - CQCC-GMM) e EER = 13,54% (B02 - LFCC-GMM) [Nautsch et al. 2021]. Para as bases ASVspoof 2021 e REMASC observa-se um baixo desempenho, com valores EER acima de 40%.

Os modelos treinados da abordagem RD-LCNN apresentaram resultados para o subconjunto *dev* do ASVspoof 2019 próximos aos relatados na literatura. Para o subconjunto *eval* da base ASVspoof 2019, o atributo "DCT" expressou a maior diferença. A inferência nas bases de dados ASVspoof 2021 e REMASC também resultaram em valores EER muito piores, maiores que 35%.

4. Conclusão

Os resultados obtidos para a base de dados ASVspoof 2019 demonstram bom desempenho da *baseline* e das abordagens experimentadas, semelhantes aos relatados pela literatura. Observa-se que a baseline obteve ERR = 2,43% no ASVspoof 2019 - *eval*, e, embora não se tenha encontrado valor na literatura para efeito de comparação, esse resultado foi melhor que os de ambas as *baselines* do desafio de 2019.

Os altos valores EER obtidos com as bases de dados ASVspoof 2021 e REMASC ratificam a situação exposta por [Korshunov and Marcel 2016], apontando para uma baixa capacidade de generalização de todas as abordagens processadas.

Agradecimentos

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei no 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado PDI 03, DOU 01245.023862/2022-14.

Referências

Alzantot, M., Wang, Z., and Srivastava, M. B. (2019). Deep residual neural networks for audio spoofing detection. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-September:1078–1082.

Chettri, B., Mishra, S., Sturm, B. L., and Benetos, E. (2018). A study on convolutional neural network based end-to-end replay anti-spoofing. *arXiv*.

Gong, Y., Yang, J., Huber, J., MacKnight, M., and Poellabauer, C. (2019). Remasc: Realistic replay attack corpus for voice controlled systems. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2019-September, pages 2355–2359. International Speech Communication Association.

Jain, A. K., Flynn, P., and Ross, A. A. (2008). *Handbook of Biometrics*. Springer.

Khan, A., Malik, K. M., Ryan, J., and Saravanan, M. (2023). Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artificial Intelligence Review*, 56:513–566. 01.

Korshunov, P., Gonçalves, A. R., Violato, R. P. V., Simões, F. O., and Marcel, S. (2018). On the use of convolutional neural networks for speech presentation attack detection. In IEEE, editor, *2018 IEEE 4th international conference on identity, security, and behavior analysis (ISBA)*, pages 1–8.

Korshunov, P. and Marcel, S. (2016). Cross-database evaluation of audio-based spoofing detection systems. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 08-12-September-2016, pages 1705–1709. International Speech and Communication Association.

Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., and Shchemelinin, V. (2017). Audio replay attack detection with deep learning frameworks. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017-August, pages 82–86. International Speech Communication Association.

Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A., and Kozlov, A. (2019). Stc antispoofing systems for the asvspoof2019 challenge. *arXiv*.

Lee, S.-K. (2024). Arbitrary discrete fourier analysis and its application in replayed speech detection. *arXiv*.

Lee, S.-K., Tsao, Y., and Wang, H.-M. (2022). Detecting replay attacks using single-channel audio: The temporal autocorrelation of speech. In *Proceedings of 2022 APSIPA Annual Summit and Conference*. 2022 APSIPA Annual Summit and Conference.

Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., and Lee, K. A. (2023). Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 31:2507–2522.

Nautsch, A., Wang, X., Evans, N., Kinnunen, T., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M., Yamagishi, J., and Lee, K. A. (2021). Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *arXiv*.

Zhang, Z., Yi, X., and Zhao, X. (2021). Fake speech detection using residual network with transformer encoder. In *IH and MMSec 2021 - Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, pages 13–22. Association for Computing Machinery, Inc.