FakeBrAccent: uma Base de Dados de Deepfakes de Áudios em Português com Diferentes Sotaques Brasileiros

Erick M. B. Santos[1], **Katarina Veljovic**[1], **Karin S. Komati**[2]

[1]Coordenação do Técnico em Internet das Coisas

[2]Programa de Pós-graduação em Computação Aplicada (PPComp)
Instituto Federal do Espírito Santo – Campus Serra, ES

{erickmiguelbsantos,katarinaveljovic123}@gmail.com

kkomati@ifes.edu.br

Abstract. *The article presents the FakeBrAccent dataset, aimed at detecting audio deepfakes in Brazilian Portuguese. Created from the BrAccent corpus, the dataset includes original samples and synthetic versions generated with the Speechify tool (zero-shot TTS and voice cloning). It covers five Brazilian accents — Southern, Northeastern, Fluminense, Carioca, and Baiano — and is available in two versions: FakeBrAccent-B, balanced (746 audio samples), and FakeBrAccent-D, unbalanced (1,545 audio samples).*

Resumo. *O artigo apresenta a base de dados FakeBrAccent, voltada para a detecção de deepfakes de áudio em português do Brasil. Criada a partir do corpus BrAccent, a base inclui amostras originais e versões sintéticas geradas com a ferramenta Speechify (zero-shot TTS e clonagem de voz). Contempla cinco sotaques brasileiros — sulista, nordestino, fluminense, carioca e baiano — e está disponível em duas versões: FakeBrAccent-B, balanceada (746 áudios), e FakeBrAccent-D, desbalanceada (1.545 áudios).*

1. Introdução

Tecnologias de aprendizado de máquina permitem criar *deepfakes*: mídia sintética realista em formatos de imagem, vídeo e áudio. *Deepfakes* de áudio são arquivos em que a voz é modificada ou inteiramente sintetizada por IA para simular outra identidade [Khanjani et al. 2023]. A produção dessas falsificações sonoras geralmente envolve os sistemas de conversão de texto para fala (*Text-to-Speech — TTS*), que transformam texto em áudio, e os sistemas de clonagem de voz (*Voice Cloning — VC*), que modificam características da voz de um falante para que soe como a de outro.

Deepfakes de áudio têm implicações diretas em três dimensões centrais: a segurança da informação, pela dificuldade crescente em distinguir conteúdos autênticos de fabricados; a privacidade individual, em razão do potencial uso indevido de identidades vocais; e a desinformação, facilitada pela circulação de conteúdos falsos com alta verossimilhança. Nesse cenário, o desenvolvimento de métodos para a detecção desses artefatos constitui um tema relevante nas investigações em processamento de linguagem e sinais [Seow et al. 2022].

Embora existam conjuntos de dados voltados à detecção de *deepfakes* de voz em idiomas como inglês, francês, japonês, mandarim, árabe e alemão, os recursos disponíveis

para a língua portuguesa ainda são limitados [Cuccovillo et al. 2022]. Até o momento, o único conjunto de dados publicamente acessível é o H-Voice [Ballesteros et al. 2020], que contém histogramas extraídos de gravações, sem fornecer os áudios originais. A ausência dessas formas de onda originais restringe a aplicação de técnicas de análise. No caso do português brasileiro, há ainda a questão do sotaque. A ausência de um conjunto de dados que contemple essas variações linguísticas dificulta a avaliação de modelos de detecção de *deepfakes* voltados aos sotaques de falas em português do Brasil.

Com base nessa lacuna, este trabalho propõe o conjunto de dados FakeBrAccent com cinco diferentes sotaques do português brasileiro: sulista, nordestino, fluminense, carioca e baiano, composto por duas versões: uma base balanceada, denominada FakeBrAccent-B, e uma base desbalanceada, denominada FakeBrAccent-D. As amostras originais foram obtidas a partir do corpus BrAccent [Batista et al. 2018], que também serve como referência para a geração das versões sintéticas, por meio de sistemas de TTS e clonagem de voz, preservando os sotaques presentes nas gravações originais. A estrutura do artigo é organizada da seguinte forma: a Seção 2 descreve o conjunto de dados FakeBrAccent; e a Seção 3 encerra o texto com as conclusões e indica possíveis desdobramentos para investigações futuras.

2. Materiais e métodos

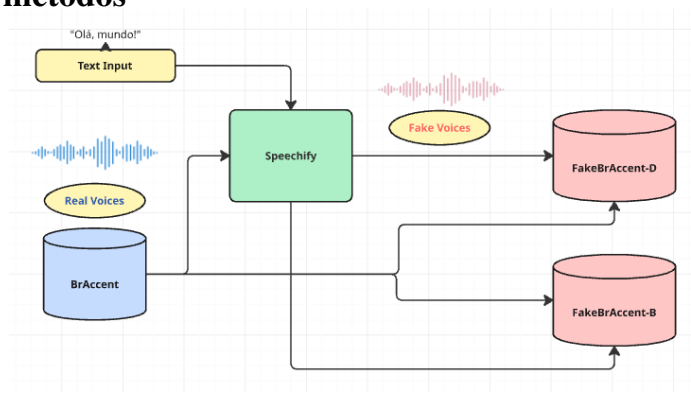

Figura 1. Fluxo de tarefas da geração da FakeBrAccent-B e FakeBrAccent-D

O fluxo deste estudo inicia com as amostras de áudio reais obtidas do conjunto de dados BrAccent, enquanto vozes sintéticas são geradas a partir de uma "Entrada de Texto" (por exemplo, "Olá, mundo!") inserida no Speechify[1]. Na geração dos áudios, utilizaram-se duas categorias de frases: (i) manuais, elaboradas intencionalmente para garantir controle de conteúdo e aderência ao tema; e (ii) automáticas, produzidas por algoritmo de geração aleatória, com extensão entre 100 e 150 caracteres por áudio. O Speechify é um sistema TTS de clonagem de voz *zero-shot*, ou seja, um sistema que aceita como entrada um texto e alguns segundos de amostra da voz do falante-alvo (proveniente do BrAccent) para produzir ondas sonoras semelhantes à voz desse falante [Azizah 2024]. Esse processo permite simular sotaques específicos durante a geração da fala sintética, preservando as características regionais originais. Tanto as vozes reais quanto as falsas geradas são então integradas ao conjunto de dados FakeBrAccent-B, a versão balanceada, e FakeBrAccent-D, a versão desbalanceada. O fluxo é ilustrado no diagrama da Figura 1.

[1] `https://speechify.com`

2.1. FakeBrAccent

BrAccent é um repositório público de gravações de falantes nativos do português brasileiro. A base inclui sotaques como sulista, nortista, nordestino, mineiro, fluminense, carioca e baiano, totalizando 1.743 amostras. Posteriormente, o trabalho de [Lopes et al. 2021] selecionou um subconjunto de 1.648 amostras, de acordo com a sua qualidade. Os sotaques mineiro e nortista não foram utilizados por contarem, cada um, com uma quantidade de pessoas falantes em quantidade insuficiente para o treinamento no Speechify.

Os áudios da BrAccent foram usados como base para gerar contrapartes sintéticas. As transcrições das falas originais foram sintetizadas com o TTS Speechify, formando pares real–sintético. O conjunto resultante, Fake BrAccent-B, abrange os sotaques Sulista, Nordestino, Fluminense, Carioca e Baiano, totalizando 746 áudios distribuídos de maneira balanceada entre versões reais e sintéticas (Tabela 1). Sotaques com amostras insuficientes na fonte original foram excluídos, além disso, a seleção considerou a qualidade dos áudios.

Tabela 1. Fake BrAccent-B

	Reais Fem.	Reais Masc.	Fakes Fem.	Fakes Masc.
Sulista	40	40	40	40
Nordestino	40	18	40	18
Fluminense	40	40	40	40
Carioca	40	35	40	35
Baiano	40	40	40	40
	200	173	200	173

Tabela 2. Fake BrAccent-D

	Reais Fem.	Reais Masc.	Fakes Fem.	Fakes Masc.
Sulista	285	330	40	40
Nordestino	161	18	40	18
Fluminense	63	51	40	40
Carioca	47	35	40	35
Baiano	102	80	40	40
	658	514	200	173

Adicionalmente, foi desenvolvida a base Fake BrAccent-D, nas mesmas estruturas da base anterior e com o objetivo de simular um cenário mais próximo de aplicações reais, onde o número de amostras pode variar entre categorias. Essa base não é balanceada entre áudios reais e sintéticos, totalizando 1.545 áudios, distribuídos de forma desigual entre os diferentes sotaques e com mais áudios reais, conforme Tabela 2.

Todos os arquivos — tanto reais quanto sintéticos — foram submetidos a um processo de uniformização, que incluiu ajustes de volume, conversão para formato único e eliminação de pausas prolongadas. Essa etapa garantiu homogeneidade na base utilizada durante os experimentos. A adoção do Speechify visou garantir uma base representativa dos padrões observados em aplicações reais. Ao final, ambas as bases de dados se encontram disponíveis na plataforma Kaggle[2][3]

[2]https://www.kaggle.com/datasets/erickmiguelsantos/fake-braccent

[3]https://www.kaggle.com/datasets/katarinaveljovic/fake-braccent-d

3. Conclusão

A criação do FakeBrAccent fornece uma ferramenta para o desenvolvimento e teste de algoritmos de detecção de *deepfakes* de voz especificamente adaptados às nuances do português brasileiro e suas variações de sotaque. Podendo ser utilizada na área de segurança para aprimorar sistemas de detecção de *deepfakes*, protegendo contra fraudes por voz; no combate à desinformação, se mostra fundamental para identificar conteúdos manipulados; e enquanto na pesquisa em linguística computacional, serve como recurso para estudos aprofundados sobre a variação linguística do português falado no Brasil, contribuindo para a criação de modelos de fala mais sofisticados e realistas.

Para futuras investigações, pretende-se expandir o FakeBrAccent para incluir uma gama mais ampla de sotaques brasileiros. Além disso, a base poderia ser enriquecida com a inclusão de áudios de mais de um sistema de conversão de texto para fala (*TTS*) e clonagem de voz, além do Speechify, para diversificar a origem dos *deepfakes* e testar a robustez dos modelos de detecção contra diferentes técnicas de síntese. E comprar diferentes modelos de IA para a classificação dos áudios em real e falso.

Agradecimentos

A professora Komati agradece ao CNPq pela bolsa DT-2 (nº 302726/2023-3) e pelo projeto nº 407742/2022-0; e agradece à FAPES pelo projeto nº 1023/2022 P:2022-8TZV6.

Referências

Azizah, K. (2024). Zero-shot voice cloning text-to-speech for dysphonia disorder speakers. *IEEE Access*, 12:63528–63547.

Ballesteros, D. M., Rodriguez, Y., and Renza, D. (2020). A dataset of histograms of original and fake voice recordings (H-Voice). *Data in brief*, 29:105331.

Batista, N. A. R. et al. (2018). Detecção automática de sotaques regionais brasileiros: A importância da validação cross-datasets. In *Anais do XXXVI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, pages 939–944, Campina Grande, PB. Sociedade Brasileira de Telecomunicações.

Cuccovillo, L., Papastergiopoulos, C., Vafeiadis, A., Yaroshchuk, A., Aichroth, P., Votis, K., and Tzovaras, D. (2022). Open challenges in synthetic speech detection. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

Khanjani, Z., Watson, G., and Janeja, V. P. (2023). Audio deepfakes: A survey. *Frontiers in Big Data*, 5:1001063.

Lopes, T., Andrade, J., and Komati, K. (2021). Comparação de serviços em nuvem para transcrição de fala na língua portuguesa em áudios com sotaques regionais brasileiros. In *Anais da IX Escola Regional de Informática de Goiás*, pages 96–109, Porto Alegre, RS, Brasil. SBC.

Seow, J. W., Lim, M. K., Phan, R. C., and Liu, J. K. (2022). A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing (Amsterdam)*, 513:351–371.