Text-Dependent Speech Biometrics - Evaluation of pre-trained ECAPA-TDNN and Wav2vec models with the BioCPqD and RedDots databases

Alcino Vilela R. Jr[1], Julia C. Colombo[1], Murilo M. Bergamaschi[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

{alcino.vilela, jccolombo24, murilomberg, marcuslima3}@gmail.com
{uliani, runstein, rviolato}@cpqd.com.br

Abstract. *This work addresses the challenge of text-dependent voice biometrics, evaluating different databases and classification models. We used the pre-trained ECAPA-TDNN and Wav2vec models on the BioCPqD and RedDots databases. Results showed very low error rates for both databases. It is also possible to observe that the performance of the Wav2vec model was significantly lower than that of ECAPA-TDNN.*

Resumo. *Este trabalho aborda o desafio da biometria de voz dependente de texto, avaliando diferentes bases de dados e modelos de classificação. Utilizamos modelos pré-treinados das arquiteturas ECAPA-TDNN e Wav2vec e aplicamo-os nas bases de dados BioCPqD e RedDots. Os resultados mostram que as taxas de erro são bastante baixas para ambas bases de dados. Também é possível observar que o desempenho do modelo Wav2vec foi muito inferior ao do ECAPA-TDNN.*

1. Introdução

Em um sistema de autenticação biométrica por voz, um usuário fornece inicialmente amostras de fala para que seja gerada uma representação única sua, chamada de *voiceprint*. Posteriormente, quando precisar se autenticar, o usuário fornece uma nova amostra de fala que é comparada com o *voiceprint* cadastrado para determinar se pertence ao mesmo indivíduo ou não [Chowdhury et al. 2018]. O que difere a biometria de voz dependente de texto, é que neste caso o sistema requer que uma frase específica, geralmente denominada *passphrase*, seja pronunciada pelo usuário tanto no momento de cadastramento quanto no de autenticação.

A biometria de voz tem uma ampla variedade de aplicações, das quais podemos destacar a autenticação de usuários em ligações telefônicas, em serviços digitais ou mesmo para controle de acesso físico [Jahangir et al. 2021]. Especificamente, a biometria de voz dependente de texto apresenta ainda as vantagens de oferecer mais segurança, uma vez que a própria *passphrase* pode ser vista como uma senha ou segredo, e permitir realizar o processo de autenticação com frases relativamente curtas, sem comprometer a acurácia do sistema.

Os último anos foram marcados por ganhos expressivos de acurácia das técnicas de reconhecimento automático de locutor, principalmente devido ao uso de redes neurais profundas, seja adotando uma abordagem mais tradicional, com uma etapa prévia de extração de atributos, como MFCC (*Mel-Frequency Cepstral Coefficients*), seja explorando diretamente as amostras do sinal [Jakubec et al. 2024], em uma abordagem *end-to-end*. Dentre as arquiteturas de rede de maior relevância, destacam-se a ECAPA-TDNN (*Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network*) [Desplanques et al. 2020] e o Wav2vec [Chen et al. 2021].

O treinamento desses sistemas se beneficiou, ou mesmo só foi possível, graças ao surgimento de grandes bases de dados e eles foram desenvolvidos para o caso mais genérico da biometria de voz independente de texto. No entanto, pouca atenção tem sido dada ao caso da biometria de voz dependente de texto, também conhecida por TD-SV (*Text-Dependent Speaker Verification*) [Tu et al. 2022]. Este trabalho buscou, portanto, avaliar o desempenho dessas redes no cenário dependente de texto.

2. Metodologia

Para avaliar o desempenho de técnicas modernas de reconhecimento de locutor, que não foram desenvolvidas especificamente para o caso da biometria dependente de texto, este trabalho adotou a abordagem ilustrada na Figura 1.

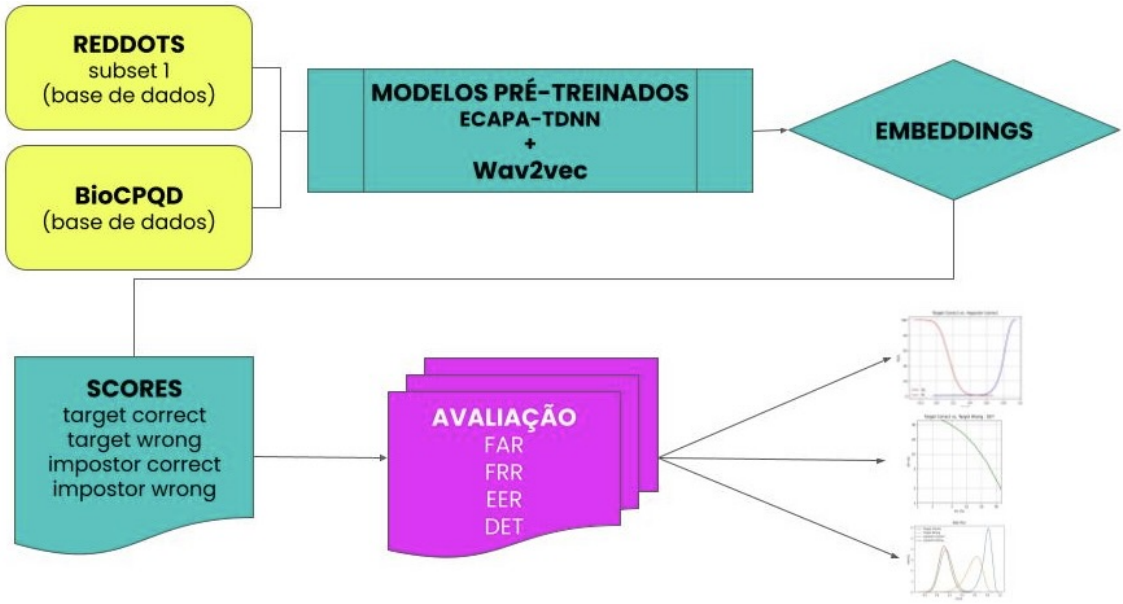

Figura 1. Fluxograma da abordagem utilizada

Para cada áudio das bases usadas (Seção 2.1), foi gerado um *embedding*. *Embeddings* são vetores de tamanho fixo que funcionam como representações numéricas que encapsulam as características do locutor de cada áudio da base em um formato que pode ser facilmente comparado e analisado.

Quando os locutores de duas amostras de áudio precisam ser comparados, seus *embeddings* são calculados e a similaridade entre esses vetores é medida, recebendo o nome de *score*. Se os *embeddings* forem similares, isso indica que os locutores das amostras de áudio também são semelhantes. Essa informação pode ser usada para aceitar ou

rejeitar a hipótese de que ambas as gravações contêm o mesmo locutor e, conforme o escopo deste trabalho, falando a mesma frase (dependente de texto).

Para isso, como medida de similaridade entre *embeddings*, foi utilizada a distância cosseno. Ou seja, o *score* de uma verificação biométrica é definido como a similaridade cosseno entre o *voiceprint* sendo testado e o *embedding* do áudio usado para autenticação.

Os protocolos de cada base, descritos na Seção 2.1, definem quais áudios podem ser usados para se criar o *voiceprint* de um locutor. No projeto foi decidido que seria calculada a média dos embeddings de cada áudio disponível para cadastramento e que esse vetor médio seria o *voiceprint* do locutor.

2.1. Bases de Dados

Para se avaliar um sistema TD-SV, é necessário uma base de dados com características peculiares: o mesmo locutor deve falar a mesma frase mais de uma vez em cada sessão de gravação e em diferentes sessões, diversos locutores devem falar essa mesma frase e, idealmente, várias dessas frases devem ser gravadas por todos os locutores.

Há poucas bases de dados disponíveis com essas características e, de acordo com nosso melhor conhecimento, podemos citar apenas as bases RedDots [Lee et al. 2015], RSR2015 [Larcher et al. 2012], Hi-Mia [Qin et al. 2019] e BioCPqD [Violato et al. 2013]. Dessas, não foi possível acessar RSR2015 e a base Hi-Mia não conta com um protocolo experimental definido. Por isso, neste trabalho adotamos apenas as bases RedDots e BioCPqD.

A base de dados RedDots foi gravada em inglês, com falantes nativos e não-nativos, e está subdivida em 4 partes, das quais somente a parte I é dedicada ao cenário dependente de texto. Este subconjunto da base é composto por 10 frases comuns a todos os 62 locutores, sendo 49 do sexo masculino e 13 do sexo feminino, oriundos de 21 países.

A base de dados BioCPqD conta com 222 locutores, sendo 124 do sexo masculino e 98 do sexo feminino. Todas as gravações são em português do Brasil. Cada locutor gravou até 5 sessões e, em todas elas, foram gravadas 3 repetições de uma mesma frase comum a todos os locutores. Esta frase foi então empregada nos testes em que a frase correta é esperada e outras 4 frases foram usadas como contra exemplo, ou seja, para os testes em que uma frase diferente da cadastrada é usada. Portanto, trata-se de uma condição bastante similar à da base RedDots.

Sendo assim, o protocolo define que os arquivos com a frase de interesse da primeira sessão de gravação de cada locutor são usados para seu cadastramento, isto é, para gerar seu *voiceprint*, e os testes são produzidos de acordo com a seguinte lógica:

- **target correto**: comparação do *voiceprint* do locutor alvo com os áudios contendo a frase de interesse das demais sessões do mesmo locutor;
- **target errado**: comparação do *voiceprint* do locutor alvo com os outros áudios das demais sessões do mesmo locutor;
- **impostor correto**: comparação do *voiceprint* do locutor alvo com os áudios contendo a frase de interesse de todas as sessões dos outros locutores;
- **impostor errado**: comparação do *voiceprint* do locutor alvo com os outros áudios de todas as sessões dos outros locutores.

2.2. Modelos

As técnicas atuais de verificação de locutores dependem de uma rede neural para extrair representações de locutores, chamadas de *embedding*. Uma das primeiras arquiteturas a ser bem-sucedida ficou conhecida como x-vector [Snyder et al. 2018] e é uma rede do tipo TDNN, que aplica agrupamento de estatísticas para projetar elocuções de comprimento variável em vetores de tamanho fixo (os *embeddings*) contendo as características dos locutores. Neste projeto empregamos uma evolução dessa arquitetura, que se tornou referência na área, chamada de ECAPA-TDNN [Desplanques et al. 2020].

O projeto adotou modelos pré-treinados da rede ECAPA-TDNN, cuja implementação está disponível no *toolkit* SpeechBrain [Ravanelli et al. 2024]. Foram usados tanto o modelo treinado com a base de dados VoxCeleb [Nagrani et al. 2019] (majoritariamente em inglês) quanto o modelo treinado com a CN-Celeb [Li et al. 2022], em chinês [1], permitindo assim avaliar o impacto do idioma nos resultados.

O Wav2vec 2.0 é um modelo do tipo *end-to-end*, pré-treinado de forma autosupervisionada no domínio do tempo (áudio bruto) para construir representações discretas da fala. Por isso, é uma arquitetura que dispensa a necessidade de uma grande quantidade de áudios rotulados para seu treinamento, tornando-se uma alternativa interessante a ser considerada. Neste projeto, foi usada uma implementação disponível publicamente [2], cujo modelo foi treinado com diversas bases de dados, contendo muitas línguas [Chen et al. 2021].

2.3. Método de avaliação

Para avaliar o desempenho do sistema, adotamos como métrica o EER (*Equal Error Rate*), que é definido como o ponto de operação de um sistema de reconhecimento biométrico em que a taxa de falsa aceitação (FAR - *False Acceptance Rate*) e a taxa de falsa rejeição (FRR - *False Rejection Rate*) são iguais [Jain et al. 2008].

O FAR é uma medida estatística que determina a probabilidade de um sistema de segurança biométrica permitir o acesso não autorizado de um usuário, medindo a porcentagem de entradas inválidas que são aceitas incorretamente. Às vezes é denotado como *False Match Rate* (FMR). O FRR representa a probabilidade do sistema não conseguir identificar corretamente uma pessoa, ou seja, a porcentagem de entradas válidas que são rejeitadas incorretamente. Às vezes é denominado como *False Non-Match Rate* (FNMR).

É importante lembrar que o EER é apenas um ponto de operação possível e pode não ser a melhor escolha para determinada aplicação. Pode-se escolher, por exemplo, um ponto com FRR baixo, o que torna o sistema fácil de usar, incentivando os usuários a adotar e continuar utilizando o sistema. Um ponto com FAR baixo, por outro lado, privilegia a segurança do sistema, mas pode levar dificuldades de utilização, resultando em frustração e possível abandono do sistema.

Além do EER, foi usada a curva DET (*Detection Error Tradeoff*) para comparar os sistemas avaliados. Ela consiste em um gráfico de taxas de erro para sistemas de classificação binária, plotando a FRR versus a FAR. Inclusive o próprio EER pode ser facilmente observado nessa curva.

[1] `https://huggingface.co/LanceaKing/spkrec-ecapa-cnceleb`
[2] `https://huggingface.co/microsoft/wavlm-base-plus-sv`

3. Resultados

Recapitulando, os modelos foram avaliados a partir dos seguintes quatro tipos de teste:

- **target correto (TC)**: locutor alvo falando a frase senha correta;
- **target errado (TE)**: locutor alvo falando a frase senha errada;
- **impostor correto (IC)**: locutor qualquer diferente do alvo falando a frase senha correta;
- **impostor errado (IE)**: locutor qualquer diferente do alvo falando a frase senha errada.

Os resultados são, então, obtidos, de acordo com três combinações desses testes. Observe que sempre a referência de interesse é o locutor alvo falando a frase senha correta (TC):

- **TC vs. IC**: simula a situação em que impostores, isto é, um locutor qualquer, tenta se passar por um usuário legítimo, conhecendo sua frase de acesso.
- **TC vs. IE**: simula a situação em que impostores, isto é, um locutor qualquer, tenta se passar por um usuário legítimo, sem conhecer sua frase de acesso.
- **TC vs. TE**: simula a situação em que o usuário legítimo tenta se autenticar usando uma frase de acesso diferente da cadastrada. O sistema poderia, por exemplo, interpretar isso como um alerta para uma situação de risco.

A Tabela 1 exibe as taxas de erro (EER) obtidas nesses três cenários com os três modelos e as duas bases de dados avaliadas.

Tabela 1. Tabela com os resultados obtidos, em termos de EER [%], com os três modelos aplicados às duas bases de dados testadas.

Base de Dados	Comparação	EER [%]		
		ECAPA-TDNN		WAV2VEC
		VoxCeleb	CN-Celeb	Várias
RedDots	TC vs. IC	1,8	5,8	13,1
	TC vs. IE	1,3	4,4	11,9
	TC vs. TE	27,3	29,6	34,0
BioCPqD	TC vs. IC	1,3	5,9	19,2
	TC vs. IE	0,9	4,5	17,6
	TC vs. TE	31,5	33,1	38,8

Analisando a base de dados RedDots com o modelo ECAPA-TDNN treinado com a base VoxCeleb, observa-se que, no cenário onde o locutor alvo foi comparado com um impostor falando a frase correta, o modelo apresentou um EER de 1,8%. Quando comparado com um impostor falando a frase errada, o EER foi ligeiramente menor atingindo 1,3%, indicando que o uso da frase específica teve um impacto positivo na acurácia do sistema.

Ao comparar o locutor alvo falando a frase correta com ele mesmo falando a frase errada, o EER aumentou significativamente para 27,3%, o que de certa forma era esperado, uma vez que trata-se da mesma pessoa.

O mesmo procedimento foi realizado utilizando o modelo ECAPA-TDNN treinado com a base CN-Celeb e a mesma tendência foi observada nos resultados. Ao compararmos os resultados do modelo treinado com a VoxCeleb (inglês) com a CN-Celeb (chinês), na base de dados RedDots (inglês), observa-se um resultado pior no segundo caso. Esses resultados demonstram que usar no treinamento do modelo áudios no mesmo idioma que será usado na sua inferência mostrou-se vantajoso.

Analisando agora os resultados obtidos na base BioCPqD com os modelos da ECAPA-TDNN em inglês e chinês, podemos dizer que as conclusões são as mesmas, com uma pequena melhora quando comparamos TC vs. IC com TC vs. IE, e um resultado muito pior quando se compara TC. vs. TE.

Além disso, é interessante notar que, mesmo se tratando de uma base em português, o resultado do modelo treinado com dados em inglês (VoxCeleb) foi muito bom e melhor do que o modelo treinado com dados em chinês, com uma diferença similar à observada com a base RedDots. Uma hipótese para esse resultado é que as línguas português e inglês são mais parecidas entre si, ambas não-tonais, do que o chinês, que é uma língua tonal.

Por fim, os resultados usando o modelo Wav2vec apresentam as mesmas tendências, mas são sistematicamente piores, indicando que, apesar de suas vantagens, essa arquitetura ainda precisa evoluir para competir com outras abordagens, ao menos neste cenário.

O gráfico da Figura 2 apresenta as curvas DET dos seis resultados (3 modelos aplicados a 2 bases de dados) no cenário que compara o target correto com o impostor errado (TC vs. IE). Esse gráfico é particularmente útil para se comparar o desempenho de diferentes sistemas em diversas condições, não apenas o EER, e, em aplicações práticas, ajuda a definir um ponto de operação mais adequado.

4. Considerações Finais

Esse trabalhou buscou avaliar sistemas de reconhecimento biométrico de locutor em um cenário de biometria dependente de texto. Para isso forma escolhidas duas arquiteturas diferentes: ECAPA-TDNN, baseada em uma abordagem mais tradicional de *machine learning*, em que há uma etapa de extração de atributos seguida de etapa de classificação em si, e Wav2vec, baseada em uma abordagem fim-a-fim, paradigma este que vem sendo explorado em diversas áreas da inteligência artificial.

Foram encontradas poucas bases de dados com as características necessárias para se avaliar um sistemas de reconhecimento biométrico de locutor dependente de texto. Delas, duas foram usadas neste trabalho: a RedDots e a BioCPqD. A primeira, em inglês, apesar de ter menos locutores, contém a diversidade de 10 frases de teste. Já a segunda, em português, conta com mais locutores, mas apenas 1 frase de teste para ser avaliada.

Além da diferença de idioma nas bases de teste, no caso da arquitetura ECAPA-TDNN foram testados dois modelos distintos, um treinado com a base VoxCeleb, em inglês, e outro treinado com a CN-Celeb, em chinês, o que permitiu avaliar a influência

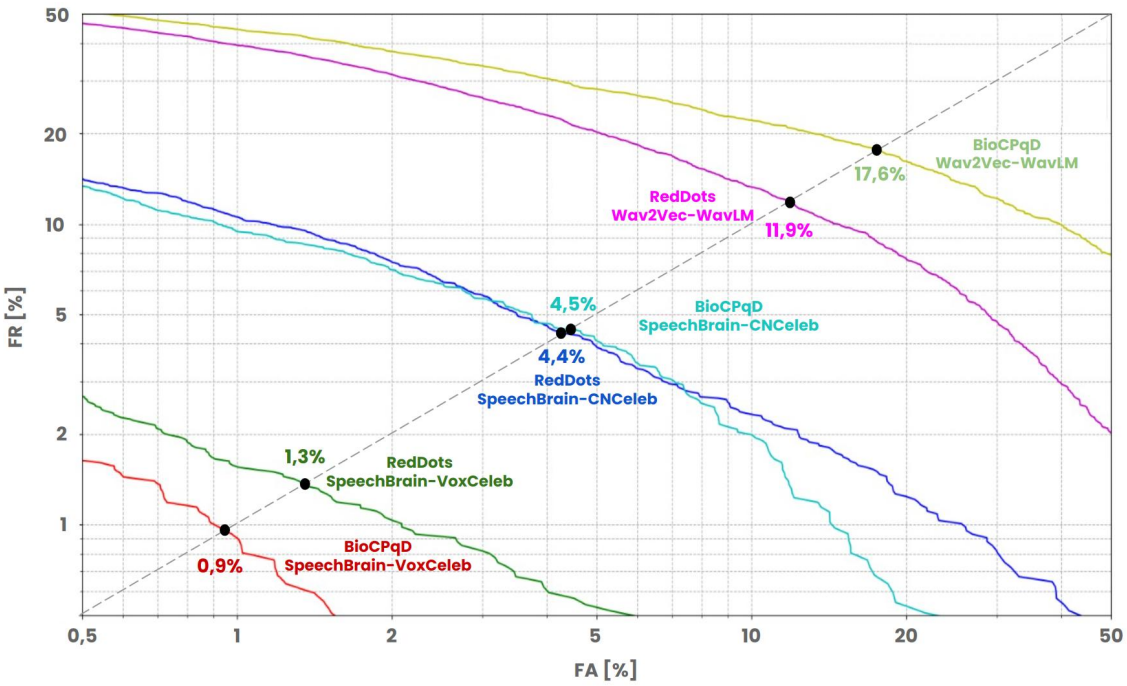

Figura 2. Curva DET dos 6 resultados de EER no cenário target correto vs. impostor errado (TC vs. IE).

do idioma nos resultados. O modelo Wav2vec foi treinado com áudios falados em diversos idiomas.

Os resultados mostram que, para este cenário, a abordagem clássica ECAPA-TDNN ainda apresenta um desempenho superior a abordagem fim-a-fim Wav2vec. Isso foi observado para as duas bases de dados de teste e considerando também os dois modelos da ECAPA-TDNN.

Em relação ao impacto do idioma, pode-se perceber que uma rede neural treinada com dados no mesmo idioma que os dados de teste apresenta um resultado consideravelmente melhor do que uma rede treinada em um idioma diferente. Também foi possível observar que, ainda que o idioma de treinamento seja diferente do idioma de teste, caso ambos sejam não-tonais o resultado é melhor do que quando o de treinamento é tonal e o de teste é não-tonal. Essas conclusões, no entanto, precisam de análises em uma diversidade maior de línguas para serem melhor validadas.

Analisando a questão da dependência de texto no momento da inferência, os resultados das comparações de target correto *versus* impostor correto (TC vs. IC) e target correto *versus* impostor errado (TC vs. IE) mostram que há pouca diferença de um locutor falar ou não a frase correta, o que pode facilitar o uso dessa tecnologia em casos práticos.

Entretanto, fica claro que os sistemas testados não são capazes de diferenciar quando o locutor correto fala ou não a frase correta (target correto *versus* target errado - TC vs. TE). Os modelos testados, de fato, não foram desenvolvidos com esse objetivo.

Em trabalhos futuros, poderia-se buscar formas de se fazer um *fine tuning* desses modelos, focando em diminuir o erro nesse caso. Outra opção seria explorar outros modelos, que levem em conta o conteúdo falado. Em um uso prático, uma outra camada de

segurança poderia ser obtida com o uso do reconhecimento de fala, para fazer a conversão de fala em texto, visando assim, ter a confirmação da frase falada pelo locutor. Por fim, uma evolução natural do trabalho seria conseguir explorar as outras duas bases de dados já identificadas, Hi-Mia e RSR2015.

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

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., and Wei, F. (2021). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv*.

Chowdhury, F. A. R. R., Wang, Q., Moreno, I. L., and Wan, L. (2018). *The Attention-based Models for Text-Dependent Speaker Verification*. Cornell University, arxiv:1710.10470 edition.

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). *The ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification*. Proc. Interspeech 2020.

Jahangir, R., Teh, Y. W., Nwekec, H. F., Mujtabad, G., Al-Garadi, M. A., and Ali, I. (2021). *The Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges*. Revista Elsevier 171.

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

Jakubec, M., Jarina, R., Lieskovska, E., and Kasak, P. (2024). Deep speaker embeddings for speaker verification: Review and experimental comparison. *Engineering Applications of Artificial Intelligence*, 127:107232.

Larcher, A., Lee, K. A., Ma, B., and Li, H. (2012). The rsr2015: Database for text-dependent speaker verification using multiple pass-phrases. In *Annual Conference of the International Speech Communication association (Interspeech 2012)*, Portland, United States.

Lee, K. A., Larcher, A., Guangsen, W., Patrick, K., Brummer, N., van Leeuwen, D., Aronowitz, H., Kockmann, M., Vaquero, C., Ma, B., Li, H., Stafylakis, T., Alam, J., Swart, A., and Perez, J. (2015). *The The RedDots Data Collection for Speaker Recognition*. Interspeech, sep 2015, dresden, germany edition.

Li, L., Liu, R., Kang, J., Fan, Y., Cui, H., Cai, Y., Vipperla, R., Zheng, T. F., and Wang, D. (2022). Cn-celeb: multi-genre speaker recognition. *Speech Communication*.

Nagrani, A., Chung, J. S., Xie, W., and Zisserman, A. (2019). Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*.

Qin, X., Bu, H., and Li, M. (2019). Hi-mia : A far-field text-dependent speaker verification database and the baselines.

Ravanelli, M., Parcollet, T., Moumen, A., de Langen, S., Subakan, C., Plantinga, P., Wang, Y., Mousavi, P., Libera, L. D., Ploujnikov, A., Paissan, F., Borra, D., Zaiem, S., Zhao, Z., Zhang, S., Karakasidis, G., Yeh, S.-L., Champion, P., Rouhe, A., Braun, R., Mai, F., Zuluaga-Gomez, J., Mousavi, S. M., Nautsch, A., Liu, X., Sagar, S., Duret, J., Mdhaffar, S., Laperriere, G., Rouvier, M., Mori, R. D., and Esteve, Y. (2024). Open-source conversational ai with speechbrain 1.0.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Tu, Y., Lin, W., and Mak, M.-W. (2022). A survey on text-dependent and text-independent speaker verification. *IEEE Access*, 10:99038–99049.

Violato, R. P. V., Neto, M. U., and Simões, F. O. (2013). *The BioCPqD: uma base de dados biométricos com amostras de face e voz de indivíduos brasileiros*. Cad. CPqD Tecnologia, Campinas, v. 9, n. 2, p. 7-18, jul./dez. 2013 edition.