Evaluation of speaker identification performance with the X-Vector and ECAPA models and different speech corpora

José G. A. de Almeida[1], Marcelo P. Duarte[1], Thiago Kosciuk[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

{jgaalmeida, marcelopd20, thiago12.kosciuk, marcuslima3}@gmail.com
{uliani, runstein, rviolato}@cpqd.com.br

Abstract. *The performance of speaker identification systems under conditions other than those used in training the model helps predict the behavior of the system in real conditions. This article validates the performance of two speaker identification models, ECAPA-TDNN and X-Vector, with three different corpora: RedDots, VCTK and CN-Celeb. These corpora have different recording conditions, audio types and languages, being ideal for stressing the models. Using the TOP 1, TOP 3, TOP 5 and EER metrics, the results show that the ECAPA-TDNN outperforms the X-Vector in all conditions tested, but both models are impacted by the language of the data base, the type of speech and the speech variability.*

Resumo. *A performance de sistemas de identificação de locutor em condições diferentes daquelas do treinamento do modelo é de interesse para se prever o comportamento do sistema em condições reais. Neste artigo avaliou-se o desempenho dos modelos de identificação de locutores ECAPA-TDNN e X-Vector com três corpora distintos: RedDots, VCTK e CN-Celeb. Estes corpora tem condições de gravação, tipo de fala e idiomas diferentes, sendo ideais para estresse dos modelos. Usando as métricas TOP 1, TOP 3 e TOP 5 e EER, os resultados mostram que o ECAPA-TDNN superou o X-Vector em todas as condições testadas, mas ambos os modelos foram impactados pela língua da base de dados, pelo tipo de elocução e pela variabilidade de fala.*

1. Introdução

Desde meados da década de 2010, os sistemas de reconhecimento do locutor passaram a se basear cada vez mais em redes neurais profundas (DNNs - *Deep Neural Networks*). A partir de trabalhos pioneiros, como [Snyder et al. 2016], pode-se observar um aumento significativo no desempenho dos sistemas baseados em DNN quando comparado às técnicas anteriores, que exploravam outras técnicas de *machine learning* e, em alguns casos, dependiam mais de métodos de processamento de fala ditos clássicos.

A área também se beneficiou de evoluções propostas para outros campos do conhecimento, em especial Processamento de Linguagem Natural (NLP - *Natural Language Processing*), Reconhecimento de Fala (ASR - *Automatic Speech Regcognition*) e Visão Computacional (CV - *Computer Vision*). Como exemplos destas técnicas podemos citar *Residual Networks - ResNets*, *Squeeze-and-Excite - SE* e as *Transformer Networks*.

Tais redes profundas só podem ser treinadas com um grande volume de dados e, portanto, muito desse desenvolvimento foi acompanhado (ou precedido) do surgimento de bases de dados consideravelmente maiores, entre as quais destaca-se a VoxCeleb [Nagrani et al. 2019].

Este trabalho buscou avaliar o desempenho de dois modelos que são referências na área, X-Vector [Snyder et al. 2018b] e ECAPA-TDNN (*Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network*) [Desplanques et al. 2020], testando-os em 3 conjuntos de dados diferentes, a base RedDots [Lee et al. 2015], a base CN-Celeb [Li et al. 2022] e a VCTK [Veaux et al. 2019]. Especificamente, além da métrica tradicional, o EER (*Equal Error Rate*), também foi medido como esses sistemas de comportariam em um cenário de identificação de locutor, empregando a métrica TOP-N, com $N = 1$, $N = 3$ e $N = 5$.

2. Metodologia

A Figura 1 ilustra os passos adotados neste trabalho. Os experimentos foram realizados utilizando o *toolkit* SpeechBrain [Ravanelli et al. 2021], uma biblioteca livre e comunitária. Foram utilizados modelos disponibilizados no portal *HuggingFace*[1], já pré-treinados com o corpus VoxCeleb [Nagrani et al. 2019].

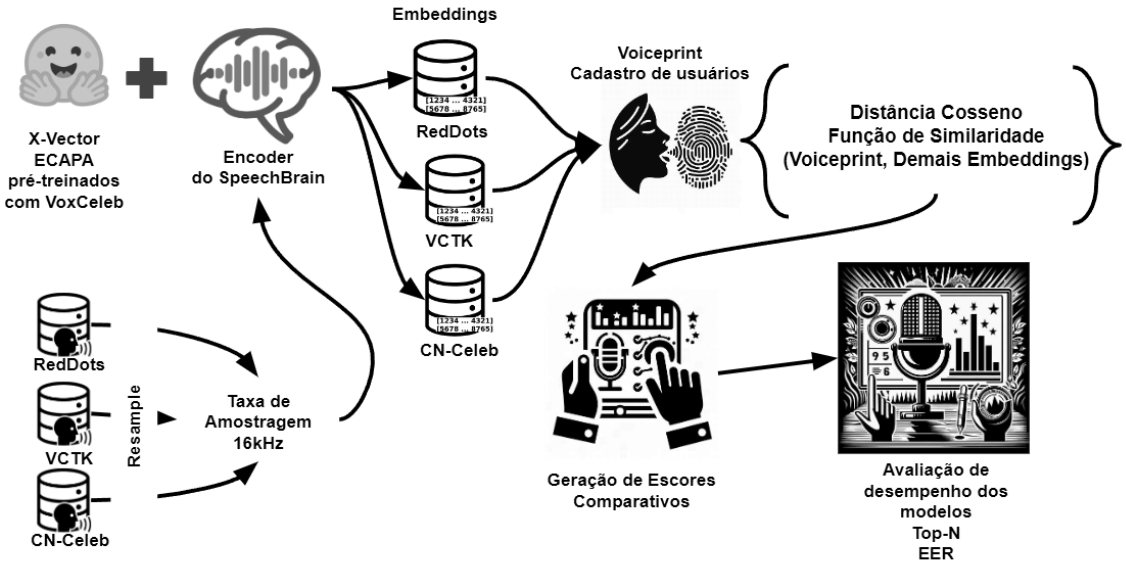

Figura 1. Pipeline utilizado para desenvolver o trabalho.

A seguir, são detalhados os modelos e as bases de dados utilizadas e o protocolo experimental adotado.

2.1. Técnicas de Reconhecimento de Locutor

A primeira DNN a ser considerada estado-da-arte em reconhecimento de locutor ficou conhecida como a arquitetura **X-Vector** [Snyder et al. 2018a, Snyder et al. 2018b]. Nesta

[1] `https://huggingface.co/speechbrain/spkrec-xvect-voxceleb` e `https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb`

técnica, a DNN é treinada para discriminar locutores em uma base de treinamento. Para isso ela é alimentada por parâmetros (tipicamente MFCC - *Mel Frequency Cepstral Coefficients*) extraídos de quadros do sinal de fala. Como o sinal tem duração variável, em certa camada da rede, a saída é acumulada (*pooling layer*) gerando uma única resposta para o sinal inteiro.

Uma vez a rede treinada, ela passa a ser usada como um extrator gerando uma representação do locutor em um sinal de fala, genericamente conhecido como *embedding*, mas neste caso chamado de X-Vector. Para isso, as camadas finais da rede treinada são ignoradas e é usado como *embedding* a saída de alguma camada após a *pooling layer*.

Por fim, o X-Vector obtido de dados de cadastramento de um locutor é comparado com o X-Vector obtido de dados de teste, geralmente através de PLDA (*Probabilistic Linear Discriminant Analysis*), gerando um *score* que indica a similaridade entre os X-Vectors e, portanto, entre o locutor presente nos dados de treinamento e o locutor presente nos dados de teste.

Especificamente, esta rede usa 23 coeficientes MFCCs extraídos de janelas de 25ms, com um passo de 10ms, que são submetidos uma rede neural de 9 camadas, sendo uma delas a camada de agregação (*pooling*), 5 camadas densas e 3 do tipo TDNN (*Time Delay Neural Network*). Os *embeddings* são extraídos na camada 7 da DNN. Originalmente, os *embeddings* ainda eram processados por técnicas de compensação de variabilidade e de redução de dimensionalidade, além de usar o já mencionado PLDA para obter um *score*. Atualmente é aplicado diretamente a similaridade cosseno para comparar os *embeddings*.

Após o X-Vector, a rede que passou a ser considerada o estado-da-arte foi a **ECAPA-TDNN**. Ela é usada até hoje como referência de comparação para os novos sistemas propostos. Esta rede segue os mesmos princípios da arquitetura do X-Vector, em que as camadas TDNN usada são constituídas por camadas convolucionais de apenas uma dimensão (CONV-1D), tal como no X-Vector, mas são seguidas de blocos compressão-excitação SE (*Squeeze-and-Excitation*), também de apenas uma dimensão combinados com blocos Res2Net (*ResNet* aprimorada) [Gao et al. 2019]. Estes dois blocos combinados (*SE + Res2Net*) foram chamados de *SE-Res2Block*. Além disso, a camada de *pooling* também foi aprimorada, bem como a função de perda usada no treinamento. Tudo isso permitiu à rede:

- Implementar mecanismos de atenção no canal (*Channel Attention*) através dos blocos *SE*. Este mecanismo permite à rede neural "focar", isto é, aprender mais informação e menos ruído, melhorando seu aprendizado das características relevantes.
- Melhorar a qualidade do treinamento pelo uso de Res2Net, e adicionalmente diminuir o número de parâmetros a serem aprendidos pela rede.

2.2. Corpora Utilizados

Neste trabalho, os dois modelos usados, X-Vector e ECAPA-TDNN, foram treinados com o corpus VoxCeleb [Nagrani et al. 2019], nas suas versões 1 e 2. Já os corpora utilizados para avaliação do desempenho dos modelos, como mencionado anteriormente, foram o RedDots [Lee et al. 2015], VCTK [Veaux et al. 2019] e CN-Celeb [Li et al. 2022]. A Ta-

bela 1 resume as principais características de todos estes corpora, cuja descrição é apresentada a seguir.

Tabela 1. Bancos de Dados Utilizados. * (Femi/Masc) %. ** Taxa de Amostragem.

Conjunto de Dados	Idioma	Origem	Falantes*	TA** (kHz)	Extensão	Acesso
VoxCeleb 1 e 2	Inglês	Americanos, ingleses, alemães	7000 (29/61)	16	wav, acc	Gratuito
RedDots	Inglês	16 países	62 (21/79)	16	linear pcm, 16 bit	Gratuito
VCTK	Inglês	Ingleses, escoceses, canadenses	110 (57/43)	48	wav	Gratuito
CN-Celeb	Chinês	Chineses	997	16	wav	Gratuito

2.2.1. VoxCeleb

VoxCeleb 1 foi construído a partir de vídeos públicos, principalmente do YouTube, incluindo falas de celebridades de vários contextos como entrevistas e *talk shows*. O conjunto de dados contém mais de 1.000 falantes únicos e aproximadamente 1 milhão de segmentos de áudio, totalizando mais de 2.000 horas de áudio. **VoxCeleb 2** é semelhante ao 1, também composto por vídeos públicos do YouTube, abrangendo uma ampla variedade de sotaques, idiomas e estilos de fala. Este conjunto é maior, com mais de 6.000 falantes distintos e mais de 1 milhão de segmentos de áudio, totalizando cerca de 3.500 horas de áudio. Os áudios têm duração entre 4 e 20 segundos, com 61% de vozes masculinas e 29% de vozes femininas de diversas nacionalidades.

Os dados passaram por revisão e correção para lidar com erros de alinhamento e transcrição, além de limpeza para remover dados ruidosos ou irrelevantes. Ambos os corpora têm padronização em termos de formato e taxa de amostragem, com áudios da versão 1 em ”wav”e da versão 2 em ”aac”, sempre amostrados a 16 kHz.

2.2.2. RedDots

Quando comparado aos outros corpus usados neste artigo, o RedDots é uma base de dados formada por um número de participantes relativamente baixo. A base contém 62 falantes ao todo (49 homens e 13 mulheres). Os áudios foram coletados pela internet com dispositivos móveis, são em inglês e de curta duração (3s) em média, usando falantes nativos e não nativos de 16 países diferentes (nenhuma amostra feminina é de país com inglês nativo).

Estes áudios foram coletados em mais de uma sessão de gravação (almejava-se 52 sessões), cada sessão podendo conter diversos áudios. O corpus é dividido em 4 partes.

- A parte I é composta por 10 sentenças comuns a todos os falantes e repetida a cada sessão. A fonte foi a **TIMIT**.
- A parte II é de 10 sentenças todas únicas para cada falante. As fontes foram **Gigawords**, **News** 2008, e **News** 2009.
- A parte III é composta por duas sentenças de escolha livre de cada falante, com frases providas pelo falante.
- A parte IV é composta por sentenças escolhidas e únicas para cada sessão. A fonte foi a **Wikipedia**.

O corpus contém os áudios em formato ”pcm”, que, para os fins deste trabalho, foram convertidos para ”wav”, com taxa de amostragem de 16 kHz. Esta base de dados, embora pequena, tem algumas variabilidades interessantes incorporadas:

- Variabilidade da voz de um locutor no tempo ao longo das 52 sessões;
- Ruídos e outros artefatos introduzidos pela coleta das fala por dispositivos móveis em condições reais.

2.2.3. VCTK

O corpus VCTK (versão 0.92) foi criado originalmente para apoiar um pacote de software voltado para clonagem de voz (CTRS VCTK). Este corpus tem 110 falantes da língua inglesa (63 mulheres e 47 homens) com diversos sotaques (inglês, americano, escocês, canadense e outros). Cada falante leu por volta 400 sentenças iguais dentre:

- Notícias de Jornal (Herald Glasgow), selecionadas por um algoritmo especial que aumenta a cobertura fonética e contextual [Veaux et al. 2013].
- ”The Rainbow Passage”. Uma conhecida passagem de teste fonético, disponível em [Fairbanks 1960].
- Um texto único projetado para coletar os sotaques dos falantes [Weinberger 2015].

Todas as falas foram gravadas com o mesmo *setup*, em estúdio, usando dois microfones, com amostras quantizadas com 24 bits e coletadas a 96kHz. Ao final, os áudios foram convertidos para 16 bits e 48kHz.

2.2.4. CN-Celeb

O CN-Celeb (versão 1) é uma base de dados em mandarim, semelhante à VoxCeleb, com áudios provenientes de vídeos de mídias sociais chinesas, como blogs, propagandas, entrevistas, filmes etc. de celebridades chinesas. Tem a característica de possuir mais de 11 gêneros de gravação, como canto, entrevista, discurso e outros, sendo a maioria de amostras de entrevistas, assim como o VoxCeleb. A taxa de amostragem é de 16 kHz e o formato é wav.

2.3. Protocolo Experimental

Conforme já mencionado, os modelos testados neste trabalho foram treinados usando a base de dados VoxCeleb. As bases de avaliação são, então, usadas para simular uma situação de autenticação biométrica, ou seja, parte dos áudios disponíveis deve ser usada para criar cadastros dos locutores, também chamados de *voiceprints*, e outra parte para simular tentativas de autenticação, isto é, para gerar *scores*. São essas informações que definem o chamado protocolo experimental.

As bases RedDots e CN-Celeb já fornecem arquivos indicando essa divisão dos áudios entre cadastramento e teste, sendo que os áudios utilizados para cadastro não são utilizados durante a verificação. No caso da RedDots, há três áudios disponíveis para a criação do *voiceprint* de cada locutor. Para cada áudio, foi gerado seu *embedding* e o *voiceprint* foi obtido pela média desses três *embeddings*.

Para a base VCTK, foi definido que os três primeiros áudios de cada locutor seriam usados para a criação do *voiceprint* (simulando o cenário de uso mais típico, em que um usuário primeiro faz seu cadastro no sistema, para só depois utilizá-lo para se autenticar), que foi calculado da mesma forma que no caso da base RedDots, ou seja, pela média dos respectivos *embeddings*. Todos os outros arquivos da base foram, então, empregados nos testes.

Uma vez gerados os *embeddings* de todos os áudios e computados os *voiceprints*, pode-se calcular a similaridade entre *voiceprints* e *embeddings* de teste, seguindo o protocolo definido. Nos protocolos há comparações tanto do *voiceprint* do locutor com falas dele mesmo (*target*), quanto com de outro falante (*impostor* ou *nontarget*), gerando um *score*. A métrica de similaridade utilizada para o cálculo do *score* foi a similaridade cosseno, definida na equação seguinte:

$$Sc(A, B) = cos(\theta) = \frac{A.B}{\|A\|.\|B\|}$$

A partir dos scores, o desempenho geral do modelo em análise foi avaliado utilizando as seguintes medidas:

- **EER** - *Equal Error Rate* - o ponto de operação onde a probabilidade de falsa rejeição (FRR - *false rejection rate*) de um falante é igual probabilidade de falsa aceitação (FAR - *false acceptance rate*) de um impostor.
- **TOP-N** - que mede a quantidade de vezes que o falante certo é reconhecido entre os N com maiores *scores*.

Para estas métricas, modelos mais precisos devem ter TOP-N maiores e EER menores.

3. Avaliação de Resultados

Os resultados apresentados na Tabela 2 e Figuras 2 e 3, fornecem uma visão detalhada do desempenho dos modelos ECAPA-TDNN e X-Vector em diferentes corpora e condições de gravação. A discussão a seguir analisa os principais achados para cada um dos corpora.

Para as três bases analisadas, RedDots, VCTK e CN-Celeb, e todos os seus subconjuntos, o modelo ECAPA-TDNN apresentou um resultado melhor do que o X-Vector em todas as métricas utilizadas.

Por exemplo, na Parte I da RedDots, o ECAPA-TDNN alcançou 97.35% em TOP 1, comparado a 87.16% do X-Vector, e um EER de 5.92% versus 18.59%. Para a VCTK, o TOP 1 foi de 98.30% para o ECAPA-TDNN contra 78.61% do X-Vector e um EER de 1.01% contra 5.94%. No caso da CN-Celeb, em "*speech*", o ECAPA-TDNN obteve 88.78% em TOP 1 e um EER de 7.07%, enquanto o X-Vector obteve 79,18% e 13,68%, respectivamente. Em "*singing*", o ECAPA-TDNN alcançou 15.46% em TOP 1, comparado a 11.16% do X-Vector.

Essas análises ressaltam a superioridade do modelo ECAPA-TDNN em relação ao X-Vector, especialmente em contextos mais desafiadores e diversos. O ECAPA-TDNN parece ser mais eficiente na identificação e discriminação de locutores, o que pode ser atribuído à sua arquitetura mais avançada e à capacidade de capturar características mais complexas da voz. O gráfico da Figura 3 resume os resultados acima.

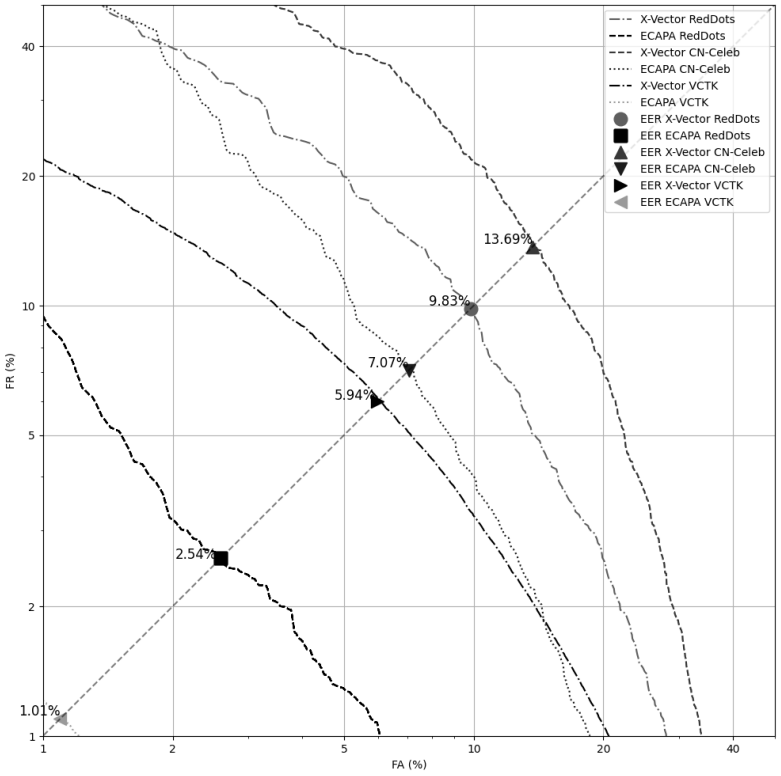

Figura 2. Curva DET para os melhores resultados de ECAPA e X-Vector

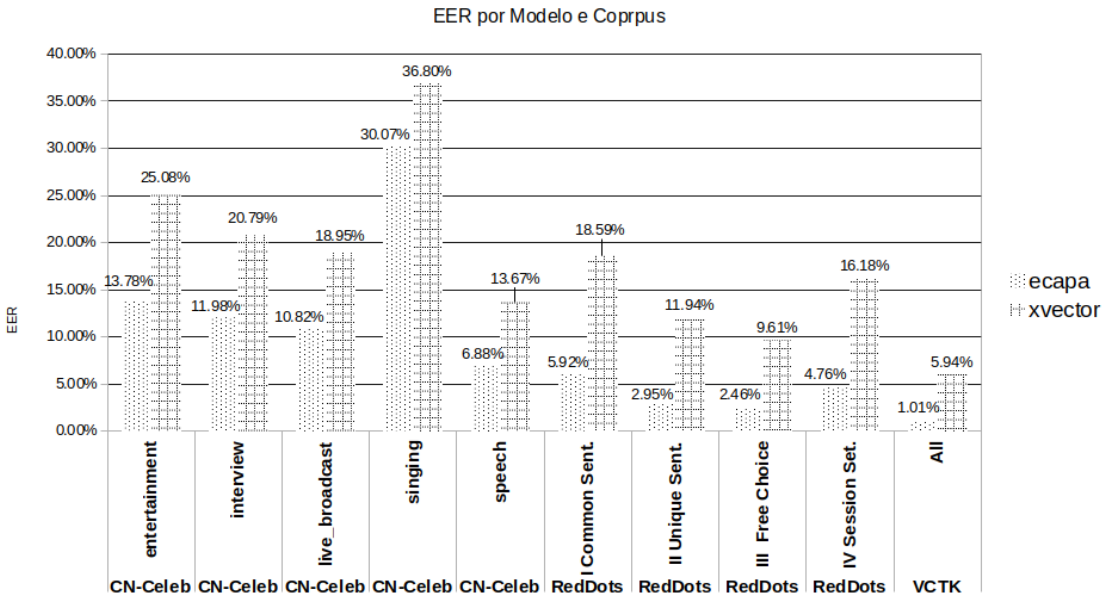

Figura 3. EER por Corpora e por Modelo

Tabela 2. Top de 1, 3 e 5 em porcentagem % de acerto e EER [%] para ECAPA e X-Vector sobre os corpora CN-Celeb, VCTK e RedDots.

	ECAPA				X-Vector			
PARTE	**TOP 1**	**TOP 3**	**TOP 5**	**EER**	**TOP 1**	**TOP 3**	**TOP 5**	**EER**
CN-Celeb								
entertain.	68.78	79.51	83.58	13.84	46.67	57.34	61.59	25.25
live_broad.	78.27	86.17	88.74	10.88	62.8	71.31	75.19	19.06
singing	15.46	24.46	28.51	30.21	11.16	18.66	21.81	36.92
interview	77.08	85.17	87.89	12.01	56.37	65.77	69.79	20.90
speech	**88.78**	**93.76**	**96.12**	**7.07**	79.18	86.1	89.41	13.68
all	69.2	77.73	80.88	15.50	52.1	61.54	65.58	23.30
VCTK								
all	**98.30**	**99.59**	**99.78**	**1.01**	78.61	90.07	93.44	5.94
RedDots								
I	97.35	98.46	99.01	5.92	87.16	92.06	93.46	18.59
II	98.99	99.56	99.73	2.95	91.15	93.32	94.18	11.94
III	**99.07**	**99.87**	**99.87**	**2.46**	97.07	98.93	99.33	9.61
IV	98.82	99.4	99.63	4.75	90.16	93.54	94.77	16.18

4. Conclusão

Os dados gerados mostram o potencial dos modelos baseados em DNNs e também apontam algumas das suas limitações. O modelo ECAPA é consideravelmente melhor que X-Vector, superando-o todos os cenários e métricas, com diferenças mais acentuadas nos corpora RedDots e VCTK.

Ambos os modelos sofreram substancial degradação de desempenho quando testados com a base de dados em chinês. Além disso, a degradação foi tanto maior quanto mais longe o modo da fala do corpus de treinamento. Digno de nota é o baixo desempenho no subconjunto *singing* do corpus CN-Celeb.

Os melhores resultados de todos os experimentos foram para a base VCTK, gravada em estúdio com alta taxa de amostragem e equipamento profissional, mostrando que a qualidade do áudio é importante.

Uma evolução natural deste trabalho seria explorar outros modelos, de arquiteturas diferentes ou até essas mesmas arquiteturas, mas treinadas com outro conjunto de dados. Também pode-se usar outras bases de teste, para avaliar diferentes condições de uso dos modelos.

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

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification.

In Meng, H., Xu, B., and Zheng, T. F., editors, *Interspeech 2020*, pages 3830–3834. ISCA. **arXiv:2005.07143**.

Fairbanks, G. (1960). Voice and articulation drillbook 2nd edn. New York: Harper & Row. pages 124-139.**ISBN-10:0060419903**.

Gao, S., Cheng, M.-M., Zhao, K., Zhang, X., Yang, M.-H., and Torr, P. H. S. (2019). Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*. **arXiv:1904.01169**.

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. In *Interspeech*, pages 2996–3000.

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*.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. **arXiv:2106.04624**.

Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., and Khudanpur, S. (2018a). Spoken Language Recognition using X-vectors. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, pages 105–111.

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

Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., and Khudanpur, S. (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 165–170.

Veaux, C., Yamagishi, J., and King, S. (2013). The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*, United States. Institute of Electrical and Electronics Engineers (IEEE). **DOI 10.1109/ICSDA.2013.6709856**.

Veaux, C., Yamagishi, J., and MacDonald, K. (2019). Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). **DOI 10.7488/ds/2645**.

Weinberger, S. (2015). Speech accent archive. https://accent.gmu.edu/. George Mason University.