Odyssey 2026
23-26 June 2026, Lisbon, Portugal

Assessing the Energy and Carbon Emissions of Neural Speaker Verification Model in Training and Inference

Hugo Leguillier[1,2], *Driss Matrouf*[1], *Guillaume Lechien*[2], *Mickael Rouvier*[1]

[1]Avignon University, LIA, UPR 4128, France
[2]Aday, France

firstname.lastname@univ-avignon.fr, glechien@aday.fr

Abstract

Deep-learning speaker verification (SV) increasingly relies on deep neural network backbones, whose environmental impact remains largely undocumented. In this paper, we conduct an evaluation of ResNet architectures trained on VoxCeleb2, varying depth, channel width, and stage distribution, and measure energy consumption and carbon footprint using node-level sensors. Results show a clear point of diminishing returns: deeper or wider models bring only marginal accuracy gains while energy consumption grows steeply. In contrast, mid-sized networks such as ResNet-50 and stage-concentrated variants achieve favorable trade-offs between performance and environmental impact. These findings provide actionable guidelines for designing energy-efficient SV systems.

Index Terms: speaker verification, energy consumption, carbon emissions

1. Introduction

Speaker verification (SV) aims to verify a speaker's identity based solely on their voice, with applications ranging from forensic analysis [1] to multimedia [2].

Modern speaker verification systems rely on Deep Neural Networks (DNNs) to extract fixed-dimensional speaker embeddings. These embeddings, extracted from hidden layers, generalize well to unseen speakers and have achieved strong performance in major evaluations (NIST SRE [3], VoxCeleb [4]). Over time, neural architectures such as TDNN, ResNet, ECAPA2, and ReDimNet have substantially increased in depth (from 9 to 293 layers), leading to significant performance gains. The intuition is that deeper networks can capture increasingly discriminative information, producing embeddings that are both robust and highly separable.

However, these performance gains do not come without trade-offs. As models grow deeper and larger, the increase in the number of parameters and arithmetic operations leads to substantially higher memory requirements, longer runtimes, and greater energy consumption, both during training and inference. This trend raises several challenges, including hardware constraints for embedded deployment, financial costs related to computing infrastructures, and the carbon footprint associated with electricity use [5]. These concerns are further amplified by the extensive GPU resources required to train deep networks and by repeated training runs for hyperparameter tuning or transfer learning. Therefore, it is essential to quantify the trade-offs among model complexity, performance, energy consumption, and carbon footprint in order to guide the design of more sustainable speaker verification systems.

In this context, evaluating speaker verification systems solely in terms of performance metrics such as EER and minDCF is no longer sufficient. It is also necessary to account for their energy consumption and carbon footprint in order to better understand the trade-offs between accuracy, model complexity, and environmental impact.

In this paper, we systematically evaluate ResNet-based SV models by varying three key structural dimensions: *depth* (from compact to very deep), *width* (scaling channel counts), and *stage distribution* (redistributing residual blocks across stages). Models are trained on VoxCeleb2 and tested both in-domain (VoxCeleb1-O, E, and H) and out-of-domain (CommonBench, CN-Celeb) [6].

The main contributions of this paper are as follows:

- **Comparison of ResNet variants:** We conduct a comparison of ResNet variants by varying three key structural dimensions—depth, width, and residual block distribution—in order to precisely characterize the trade-offs between speaker verification performance and environmental impact.
- **Evaluation across in-domain and out-of-domain datasets:** We evaluate these models on both in-domain and out-of-domain test sets using standard SV metrics (EER, minDCF), and relate these performance results to direct measurements of energy consumption and carbon footprint.
- **Recommendations for energy-efficient SV design:** Based on these analyses, we provide concrete recommendations for the design of energy-efficient SV systems by identifying the configurations that offer the best trade-off between performance, energy consumption, and carbon footprint.

The paper is organized as follows: Section 2 reviews the related work. Section 3 details the methodology for quantifying energy consumption and carbon emissions, specifying the chosen indicators (e.g., kWh, kg CO_2eq) and the measurement tools employed. Section 4 introduces the ResNet architectures. Section 5 presents the experimental setup and results, including in-domain and out-of-domain performance, training energy consumption, and an analysis of the trade-offs between architectural choices, verification accuracy, and environmental impact. Finally, Section 6 concludes the paper.

2. Related work

Recent work on sustainable machine learning has emphasized that model evaluation should not rely on accuracy alone, but should also account for energy consumption and carbon footprint. Early studies highlighted the financial cost and environmental impact of training large neural models and advocated for more transparent reporting of energy-related metrics [7, 8]. Subsequent work further showed that the carbon footprint of deep learning depends not only on model size, but also on hardware efficiency, datacenter infrastructure, geographic location,

10.21437/Odyssey.2026-34

and time-varying carbon intensity of electricity [5, 9].

In speech processing, explicit analyses of energy consumption and carbon footprint remain comparatively limited. Parcollet and Ravanelli [10] quantified the energy consumption and carbon footprint of training end-to-end ASR systems, showing that small performance gains may come at a disproportionately high environmental impact. Dinarelli et al. [11] studied a similar trade-off in end-to-end spoken language understanding, relating model performance to training time and electricity consumption. More recently, LeBenchmark 2.0 [12] highlighted the environmental impact of large-scale self-supervised speech pre-training, showing that increasing the amount of pre-training data can substantially raise energy consumption and carbon footprint. Kulkarni et al. [13] extended this perspective to large ASR systems at inference time by analyzing the energy consumption and carbon footprint of Whisper and MMS across multiple GPU platforms.

However, to the best of our knowledge, there is currently no published study specifically dedicated to analyzing the trade-off between speaker verification performance, energy consumption, and carbon footprint. Our work addresses this gap by focusing on ResNet-based SV systems and by studying how depth, width, and stage distribution affect the balance between accuracy and environmental impact.

3. Environmental Impact Indicators and Measurement Methodology

3.1. Environmental indicators

In this study, two indicators of environmental impact are considered: energy consumption and carbon footprint.

Energy consumption: measured in kilowatt-hours (kWh), this indicator reflects the total electricity consumption of a process. In practice, it mainly comes from GPU and CPU usage [14]. While hardware power ratings provide rough estimates, actual consumption depends on workload, utilization, memory access, and software efficiency [15]. Direct kWh measurement therefore offers a more accurate assessment during training.

Carbon footprint: expressed in kilograms (or grams) of CO_2 equivalent (CO_2e), this metric represents the total greenhouse-gas impact associated with the electricity consumed in computational tasks. It depends on the carbon intensity of local energy sources, each of which has a specific emission factor derived from life-cycle analyses [16]. In practice, the carbon footprint is estimated by converting energy consumption (kWh) into CO_2eq using regional conversion factors published by environmental agencies.

These two indicators provide a coherent framework for assessing the environmental impact of model training and inference.

3.2. Measurement tools

Energy consumption and carbon footprint are tracked using the Compute Energy and Emissions Monitoring Stack (CEEMS) [17], an open-source, platform-independent tool specifically developed to monitor energy usage and emissions of individual workloads on high-performance computing and cloud infrastructures.

CEEMS gathers fine-grained energy consumption data directly from node-level hardware sensors, specifically through two primary measurement mechanisms:

- **Intel's Running Average Power Limit (RAPL)**: this mechanism reports granular power metrics specifically for CPU and DRAM components at microsecond intervals, providing insights into computational resource consumption.
- **Baseboard Management Controllers via the IPMI-DCMI**: this interface provides node-level power measurements at lower frequency, covering additional system components beyond the CPU and DRAM.

Regarding carbon footprint estimation, CEEMS converts measured electricity consumption (in kWh) into CO_2eq using real-time and static emission factors. Real-time emission data are retrieved from trusted sources like RTE eCO_2 [18] (the French electricity transmission system operator) and Electricity Maps [19], providing accurate and dynamic grid-emission intensities. Additionally, static emission factors from authoritative sources such as Our World in Data (OWID) are used as reference benchmarks.

4. Speaker Embedding Extraction with ResNet

Many speaker verification (SV) systems are currently based on the ResNet architecture or its variants (e.g., Res2Net, ResNeXt, and ECAPA-style backbones). ResNet is particularly well suited to SV because it combines strong representational capacity with a flexible design that can be adapted to different computational budgets. Its architecture is organized into four stages, each composed of several residual blocks. Within each block, convolutional layers are followed by batch normalization and nonlinear activation functions, while a shortcut connection adds the input directly to the transformed output. This residual learning mechanism facilitates the optimization of deep networks by preserving gradient flow during backpropagation and enabling the training of substantially deeper models.

A major advantage of ResNet lies in its configurability. The architecture can be scaled in depth by varying the number of residual blocks, as illustrated in configurations ranging from 18 to 419 layers, and in width by adjusting the number of feature maps through width multipliers. In speaker verification, the convolutional backbone first produces frame-level representations from acoustic features such as filterbanks. These representations are then aggregated over time by a pooling layer to form a fixed-dimensional utterance-level representation. This aggregation step is essential because speech utterances are variable in duration, whereas the downstream scoring pipeline requires a compact and fixed-size embedding. The architecture is typically completed with a fully connected layer that maps the pooled representation into a discriminative speaker embedding space.

ResNet is trained on a speaker identification task, where the objective is to classify each speech segment into one of n known speaker classes in the training set. Once training is completed, the final classification layer is discarded, and a hidden layer at the utterance level is used to extract the speaker representation, commonly referred to as the speaker embedding. These embeddings are designed to capture speaker-specific characteristics while being robust to nuisance factors such as channel variability, background noise, or phonetic content. In practice, speaker embeddings can then be compared using similarity measures or back-end classifiers in order to perform verification.

System	Params (M)	Avg. EER	Avg. minDCF	Environmental		In-domain						Out-of-domain			
				Energy (kWh)	CO_2 (kg-eq)	Vox1-O		Vox1-E		Vox1-H		CommonBench		CN-Celeb	
						EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF
Depth scaling															
ResNet-419-D	133	**3.35**	**0.218**	895.67	13.734	**0.68**	**0.052**	**0.80**	**0.083**	1.42	**0.135**	3.48	0.330	**10.37**	**0.492**
ResNet-200-D	61	3.44	0.224	222.53	3.091	0.72	0.053	0.82	0.083	**1.42**	0.136	**3.46**	0.329	10.81	0.520
ResNet-101-D	38	3.54	0.229	135.27	4.263	0.71	0.067	0.82	0.087	1.45	0.136	3.50	**0.326**	11.24	0.528
ResNet-50-D	17	3.75	0.245	51.69	0.982	0.88	0.078	0.92	0.099	1.62	0.157	3.81	0.345	11.52	0.548
ResNet-34-D	16	3.90	0.254	46.2	0.661	0.99	0.082	1.00	0.103	1.75	0.169	3.94	0.359	11.83	0.555
ResNet-18-D	9	4.11	0.273	63.05	0.790	1.19	0.103	1.19	0.124	2.04	0.194	4.18	0.373	11.94	0.571
Stage-distribution scaling															
ResNet-50-D$_{[8,2,2,2]}$	11	3.80	0.248	76.67	1.127	0.97	0.086	0.98	0.101	1.67	0.160	3.81	0.346	11.59	0.548
ResNet-50-D$_{[2,8,2,2]}$	11	3.65	0.244	52.06	0.769	0.83	0.088	0.93	0.096	1.60	0.155	3.74	0.347	11.16	0.538
ResNet-50-D$_{[2,2,8,2]}$	17	3.68	0.242	49.33	0.728	0.84	0.072	0.90	0.094	1.57	0.153	3.71	0.347	11.39	0.544
ResNet-50-D$_{[2,2,2,8]}$	17	3.91	0.260	47.0	0.688	0.93	0.085	1.03	0.111	1.81	0.172	4.02	0.370	11.74	0.563
ResNet-50-D$_{[2,6,6,2]}$	16	3.65	0.233	53.82	0.796	0.76	0.054	0.88	0.092	1.52	0.146	3.66	0.343	11.41	0.533
Width scaling															
ResNet-50-W4	233	3.77	0.246	438.46	6.98	0.82	0.088	0.87	0.089	1.59	0.149	3.92	0.358	11.63	0.545
ResNet-50-W2	61	3.68	0.236	120.76	1.66	0.83	0.062	0.85	0.089	1.54	0.147	3.84	0.353	11.34	0.528
ResNet-50-W0.5	5	3.89	0.257	40.05	0.62	1.01	0.084	1.06	0.108	1.80	0.174	3.87	0.349	11.70	0.571
ResNet-50-W0.25	2	4.58	0.299	**34.35**	**0.47**	1.49	0.143	1.53	0.155	2.50	0.226	4.45	0.386	12.94	0.587

Table 1: *Speaker verification performance (EER % / minDCF) and environmental impact indicators (energy consumption in kWh and carbon footprint in kg CO_2eq) of different ResNet architectures across in-domain and out-of-domain benchmarks.*

5. Experiments and results analysis

5.1. Experimental setup

For these experiments, we use the Kiwano toolkit [20][1] and train ResNet-based speaker embedding extractors on VoxCeleb2 [21]. Training uses mini-batches of 512 and 3.5 s random crops. Standard data augmentation follows [22] with MUSAN [23], RIRs [24] and SpecAugment [25]. Inputs are 80-dimensional filterbanks, and embeddings are 256-dimensional. Models are trained for 42 epochs with AAM loss, SGD (momentum 0.9, weight decay 2×10^{-4}). Squeeze-and-Excitation is enabled on the first two stages [26]; the four stages use {128,128,256,256} feature maps.

We trained different ResNet models by adjusting depth, width, and stage distribution.

Depth Scaling: These models, referred to as *ResNet-X-D*, increase the number of layers (depth) in the network. The depth is denoted by X, which can take values such as 18, 34, 50, 101, 200, or 419. All variants share the same ResBlock design; their only difference lies in how many blocks are assigned to each stage, which directly determines the overall depth.

Width Scaling: These models, referred to as *ResNet-50-Wx*, adjust the number of feature maps in each layer by a constant factor x. For instance, the baseline *ResNet-50-D* uses feature maps $\{128, 128, 256, 256\}$ across its four stages. In contrast, *ResNet-50-W2* doubles these values to $\{256, 256, 512, 512\}$, while *ResNet-50-W0.5* reduces them to $\{64, 64, 128, 128\}$.

Stage Distribution Scaling: These models, referred to as *ResNet-50-D$_{[a,b,c,d]}$*, keep the total depth fixed (e.g., 50 layers) but redistribute residual blocks across stages. For example, in an early-concentrated variant *ResNet-50-D$_{[8,2,2,2]}$*, most blocks are shifted to the first stage, while later stages contain fewer residual blocks. Because the stages do not have the same width, this redistribution also changes the parameter count and the computational profile of the network.

We evaluate in-domain on VoxCeleb1-O/E/H (cleaned) [27] and out-of-domain on CommonBench [28] and CN-Celeb [29]. Metrics are EER and minDCF with $P_{\text{tar}}=10^{-2}$, $C_{\text{miss}}=C_{\text{fa}}=1$.

5.2. Hardware and computational cost

All experiments were conducted on the Jean Zay supercomputer. For consistency, all experiments were performed on the same GPU: an NVIDIA Tesla V100 32GB. Energy usage and runtime are recorded via the CEEMS infrastructure introduced in Section 3.2, which collects node-level measurements and derives per-job power usage and CO_2-eq emissions.

All experiments were conducted in France, where electricity had an average carbon intensity of approximately 50 gCO_2/kWh, significantly lower than in countries heavily reliant on fossil fuels, such as the United States (345 gCO_2/kWh) or China (536 gCO_2/kWh).

5.3. Performance of the models

Table 1 summarizes the results obtained with a family of ResNet architectures that vary along three structural dimensions: depth, channel width, and the distribution of residual blocks across stages. The depth ranges from 18 layers (9M parameters) to 419 layers (133M parameters). All models are evaluated on both in-domain and out-of-domain test sets.

For each system, the table reports the number of parameters, environmental-impact indicators (energy consumption and carbon footprint), and speaker verification performance in terms of EER and minDCF. Results are detailed for the in-domain datasets VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H, as well as for the out-of-domain datasets CommonBench and CN-Celeb. The average EER and average minDCF correspond to the mean values computed across all evaluation corpora.

[1] https://github.com/kiwano-toolkit/kiwano

Depth vs. Accuracy: increasing network depth (i.e., the number of hidden layers) generally improves performance, but with rapidly diminishing returns. From *ResNet-18-D* to *ResNet-419-D*, the average EER decreases from 4.11% to 3.35%, while the average minDCF drops from 0.273 to 0.218. However, most of the gains are observed between shallow models and mid-sized models: moving from *ResNet-18-D* to *ResNet-101-D/200-D* yields a significant improvement, whereas the transition from *ResNet-200-D* to *ResNet-419-D* results in only a marginal performance gain (from 3.44% to 3.35% in average EER). A similar trend is observed on out-of-domain datasets: all models exhibit degraded EER and minDCF on CommonBench and CN-Celeb, but the gap between *ResNet-200-D* and *ResNet-419-D* remains negligible, even though the latter requires more than four times the training energy cost. These results indicate that beyond a certain depth, adding more layers mainly increases energy consumption and carbon footprint without providing commensurate performance gains.

Width vs. Accuracy: beyond depth, we also investigate the impact of network width, measured by the number of channels in the convolutional layers. Increasing the number of channels tends to improve the average EER and average minDCF on in-domain datasets, but these gains do not consistently generalize to out-of-domain conditions. The *ResNet-50-W2* model achieves better performance than *ResNet-50-D* on VoxCeleb1-O/E/H and CN-Celeb, but performs worse on CommonBench. Similar trends were observed for deeper models such as *ResNet-101-D*. In contrast, extreme width expansion is less effective: *ResNet-50-W4* ($\sim$233M parameters) achieves an average EER of 3.77%, whereas *ResNet-50-D* achieves an average EER of 3.75%.

Narrower variants, such as *ResNet-50-W0.5* and *ResNet-50-W0.25*, yield worse EER and minDCF than the *ResNet-50-D* baseline. However, *ResNet-50-W0.5* (average EER of 3.89%) achieves performance comparable to that of *ResNet-34-D* (average EER of 3.90%), while using approximately three times fewer parameters and significantly less energy. This result highlights that a moderate reduction in network width can provide a favorable trade-off between performance and environmental impact.

Stage distribution vs. Accuracy: redistributing residual blocks across stages also affects performance. For *ResNet-50-D* (average EER of 3.75%), concentrating blocks in stages 2 and 3 improves both EER and minDCF: the *ResNet-50-D*$_{[2,8,2,2]}$ and *ResNet-50-D*$_{[2,6,6,2]}$ variants both reduce the average EER to 3.65%, while the *ResNet-50-D*$_{[2,2,8,2]}$ variant achieves 3.68%. In contrast, shifting residual blocks toward the first or the last stage degrades performance: *ResNet-50-D*$_{[8,2,2,2]}$ reaches an average EER of 3.80%, and *ResNet-50-D*$_{[2,2,2,8]}$ reaches 3.91%.

Overall, these results show that concentrating residual blocks in stages 2 and 3 of ResNet is more effective than allocating them to stages 1 or 4. Models following this strategy can reduce the number of parameters while matching or even surpassing the performance of the baseline model. This highlights the predominant role of stages 2 and 3 in the performance of speaker verification systems.

5.4. Training-time, energy and carbon footprint

Table 1 reports the energy consumption and carbon emissions for each ResNet variant.

Energy consumption and carbon footprint: Table 1 shows that energy use does not scale linearly with model size. For example, *ResNet-34-D* (16M parameters) consumes 46.2 kWh, while *ResNet-200-D* (61M) rises to 222.5 kWh. Beyond this point, the cost grows disproportionately: *ResNet-419-D* (133M) reaches 896 kWh, $4\times$ *ResNet-200-D*, for marginal EER and minDCF gains, emitting 13.7 kg CO_2-eq even under France's low-carbon grid (50 gCO_2/kWh).

Focusing on *ResNet-50-D* (17M, 52 kWh, 0.98 kg CO_2), width scaling shows the extremes of this trade-off. Doubling width (*ResNet-50-W2*, 61M) more than doubles energy demand (120.7 kWh, 1.66 kg CO_2), while extreme widening (*W4*, 233M) reaches 438 kWh and 6.98 kg CO_2, rivaling *ResNet-419-D*. Conversely, narrower variants (*W0.5*, 5M; *W0.25*, 2M) cut consumption to 40 and 34 kWh with emissions under 0.7 kg, but at the cost of much higher EER.

Stage-distribution models of *ResNet-50-D* show a different profile: redistributing blocks can lower cost, although the parameter count may also change depending on how blocks are allocated across stages. For instance, *ResNet-50-D*$_{[2,2,8,2]}$ and *ResNet-50-D*$_{[2,2,2,8]}$ consume only 49–47 kWh, with emissions around 0.7 kg CO_2, while mid-balanced variants (*[2,8,2,2]* and *[2,6,6,2]*) stay close to baseline at 52–53 kWh (0.77–0.80 kg). Thus, reallocating blocks to intermediate or later stages can slightly reduce energy consumption and carbon footprint, although the parameter count is not strictly constant across these variants. Overall, width scaling strongly amplifies environmental impact as parameters grow, whereas block concentration offers small but consistent efficiency gains at constant size.

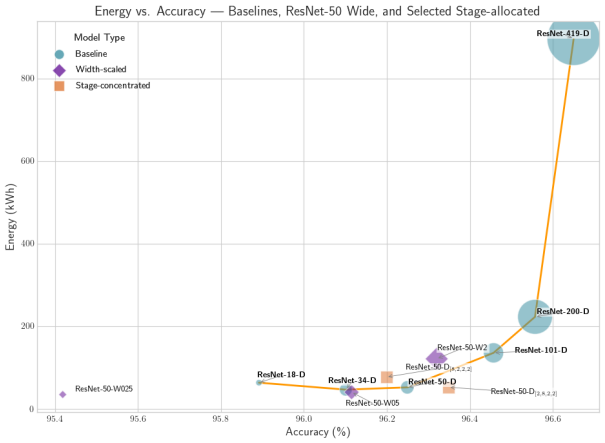

Figure 1: *Performance–energy trade-offs for ResNet-X-D architectures. The "elbow" is located around* ResNet-101-D/ResNet-200-D*, beyond which additional depth brings steep energy consumption for modest performance gains.*

Performance–energy trade-offs and sweet spot: Figure 1 plots a subset of representative models to highlight energy consumption vs. accuracy, revealing a clear "knee" where extra capacity no longer yields proportional gains. *ResNet-34-D* and *ResNet-50-D* emerge as optimal sweet spots, delivering strong accuracy with energy consumption up to an order of magnitude lower than very deep models (*ResNet-419-D*, 896 kWh). Redistributing residual blocks toward the central stages improves efficiency: the stage-concentrated variant *ResNet-50-D*$_{[2,8,2,2]}$ achieves a better accuracy–energy balance than the baseline, although its parameter count is lower. Narrow models (e.g.,

ResNet-50-W0.25, 2M) minimize energy consumption but suffer from degraded EER (4.58%).

Practical recommendations: Based on the results presented above, several practical recommendations can be drawn for the design of energy-aware speaker verification systems. First, balanced mid-sized architectures such as *ResNet-34-D* and *ResNet-50-D* should be preferred for general-purpose deployments, as they provide the most favorable compromise between verification performance and environmental impact. Second, narrower variants such as *ResNet-50-W0.5* are attractive for resource-constrained scenarios, since they substantially reduce energy consumption and carbon footprint while maintaining competitive accuracy. Third, very deep models beyond *ResNet-101-D*/*ResNet-200-D*, as well as extremely wide variants such as *ResNet-50-W4*, should be avoided unless their marginal performance gains clearly justify their much higher environmental impact. Finally, reallocating residual blocks toward intermediate stages appears to be a more effective design strategy than concentrating capacity in the earliest or latest stages, as it can improve the overall performance–efficiency trade-off.

5.5. Inference footprint

Table 2 shows the energy and emissions required to extract speaker embedding from the full CommonBench dataset (comprising over 9.7M pairs).

System	EER	minDCF	Energy (kWh)	CO_2 (kg-eq)	Duration (HH:MM)	Precision
ResNet-419	-	-	-	-	-	-
ResNet-200	3.456	0.329	8.51	0.1224	02:12:48	FP32
ResNet-101	3.502	0.325	5.03	0.0650	01:15:22	FP32
ResNet-50	3.814	0.345	3.11	0.0405	00:51:57	FP32
ResNet-34	3.935	0.359	2.67	0.0321	00:46:17	FP32
ResNet-18	4.181	0.372	**1.94**	**0.0239**	**00:36:33**	FP32
ResNet-419	3.475	0.330	8.58	0.1116	03:26:48	FP16
ResNet-200	3.462	0.327	5.45	0.0685	01:53:11	FP16
ResNet-101	3.502	0.325	3.13	0.0344	01:09:28	FP16
ResNet-50	3.813	0.345	2.30	0.0253	00:53:37	FP16
ResNet-34	3.935	0.359	1.94	0.0234	00:44:18	FP16
ResNet-18	4.181	0.372	**1.53**	**0.0184**	**00:37:48**	FP16

Table 2: *Inference-time performance, energy consumption, carbon emissions, and runtime of ResNet speaker verification models on CommonBench under FP32 and FP16 precision.*

We report results in both standard full precision (FP32) and mixed precision (FP16) modes. We observe that EER and minDCF are practically the same for FP16 and FP32. Energy usage scales with model size, with *ResNet-18* consuming as little as 1.94 kWh in FP32 and just 1.53 kWh in FP16. We note that *ResNet-419* was executed using only mixed precision (FP16), due to GPU memory limitations.

These results confirm that although inference costs are significantly lower than training, they remain non-negligible at scale. For large multilingual corpora like CommonBench, the difference between architectures is significant: *ResNet-419* emits almost 6 times more CO_2 than *ResNet-18* when using the same precision (FP16).

Switching to FP16 yields consistent gains across all models, with average energy savings of 25–35% and reduced inference times while maintaining the same performance. This highlights mixed precision as an effective low-effort optimization for energy-aware deployment of speaker verification systems. For real-world deployments in resource-constrained or carbon-sensitive environments, such differences need careful consideration too when selecting a backbone architecture.

6. Conclusions

This paper analyzed the trade-offs between accuracy, energy consumption, and carbon footprint in ResNet-based speaker verification by varying depth, width, and stage distribution. Results show that although deeper models improve EER and minDCF, these gains diminish sharply beyond *ResNet-101-D*/*ResNet-200-D*, while energy consumption and carbon footprint increase disproportionately.

Width scaling showed that moderate expansion (W2) can provide small in-domain gains, whereas extreme widening (W4) leads to a disproportionately high environmental impact. Conversely, narrower variants such as *ResNet-50-W0.5* and *ResNet-50-W0.25* substantially reduce energy consumption, at the expense of absolute accuracy.

Stage-distribution models further showed that reallocating residual blocks toward intermediate stages (stages 2 and 3) can improve the performance–efficiency trade-off, whereas concentrating capacity in the earliest or latest stages degrades performance.

Overall, our results suggest the following practical recommendations: balanced mid-sized models such as *ResNet-34-D* and *ResNet-50-D* should be preferred for general-purpose SV systems; narrower variants are attractive for resource-constrained deployments; and very deep or extremely wide architectures should be avoided unless their marginal performance gains clearly justify their additional environmental impact.

7. Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocations AD011013257R4 and AD011016050R1 made by GENCI.

8. References

[1] Campbell, Joseph P. and Shen, Wade and Campbell, William M. and Schwartz, Reva and Bonastre, Jean-Francois and Matrouf, Driss, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.

[2] A. Gresse, M. Rouvier, R. Dufour, V. Labatut, and J.-F. Bonastre, "Acoustic pairing of original and dubbed voices in the context of video game localization," in *Interspeech*, 2017.

[3] J. Alam, S. Barahona, D. Boboš, L. Burget, S. Cumani, M. Dahmane, J. Han, M. Hlavacek, M. Kodovsky, F. Landini *et al.*, "ABC SYSTEM DESCRIPTION FOR NIST SRE 2024."

[4] R. Makarov, N. Torgashov, A. Alenin, I. Yakovlev, and A. Okhotnikov, "Id r&d system description to voxceleb speaker recognition challenge 2022," *ID R&D Inc.: New York, NY, USA*, 2022.

[5] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon Emissions and Large Neural Network Training," 2021. [Online]. Available: https://arxiv.org/abs/2104.10350

[6] C. Rusti, A. Leschanowsky, C. Quinlan, M. Pnacek, L. Gorce, and W. Hutiri, "Benchmark dataset dynamics, bias and privacy challenges in voice biometrics research," 2023. [Online]. Available: https://arxiv.org/abs/2304.03858

[7] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," 2019. [Online]. Available: https://arxiv.org/abs/1906.02243

[8] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and Predicting the Carbon Footprint of Training

Deep Learning Models," 2020. [Online]. Available: https://arxiv.org/abs/2007.03051

[9] J. Dodge, T. Prewitt, R. T. D. Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, and W. Buchanan, "Measuring the Carbon Intensity of AI in Cloud Instances," 2022. [Online]. Available: https://arxiv.org/abs/2206.05229

[10] T. Parcollet and M. Ravanelli, "The Energy and Carbon Footprint of Training End-to-End Speech Recognizers," in *Interspeech 2021*, 2021, pp. 4583–4587.

[11] M. Dinarelli, M. Naguib, and F. Portet, "Toward Low-Cost End-to-End Spoken Language Understanding," 2022. [Online]. Available: https://arxiv.org/abs/2207.00352

[12] T. Parcollet, H. Nguyen, S. Evain, M. Z. Boito, A. Pupier, S. Mdhaffar, H. Le, S. Alisamir, N. Tomashenko, M. Dinarelli, S. Zhang, A. Allauzen, M. Coavoux, Y. Esteve, M. Rouvier, J. Goulian, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, "LeBenchmark 2.0: a Standardized, Replicable and Enhanced Framework for Self-supervised Representations of French Speech," 2024. [Online]. Available: https://arxiv.org/abs/2309.05472

[13] A. Kulkarni, A. Kulkarni, M. Couceiro, and I. Trancoso, "Unveiling Biases while Embracing Sustainability: Assessing the Dual Challenges of Automatic Speech Recognition Systems," in *Interspeech 2024*. ISCA, Sep. 2024, p. 4628–4632. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2024-2494

[14] R. Nana, C. Tadonki, P. Dokladal, and Y. Mesri, "Energy concerns with hpc systems and applications," 2023. [Online]. Available: https://arxiv.org/abs/2309.08615

[15] P. Delestrac, J. Miquel, D. Bhattacharjee, D. Moolchandani, F. Catthoor, L. Torres, and D. Novo, "Analyzing GPU Energy Consumption in Data Movement and Storage," in *ASAP 2024 - IEEE 35th International Conference on Application-specific Systems, Architectures and Processors*. Hong Kong, Hong Kong SAR China: IEEE, Jul. 2024, pp. 143–151. [Online]. Available: https://hal.umontpellier.fr/hal-04604802

[16] J. Dodge, T. Prewitt, R. T. D. Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, and W. Buchanan, "Measuring the carbon intensity of ai in cloud instances," 2022. [Online]. Available: https://arxiv.org/abs/2206.05229

[17] M. Paipuri, "Ceems: A resource manager agnostic energy and emissions monitoring stack," in *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, Nov. 2024, p. 1862–1866. [Online]. Available: http://dx.doi.org/10.1109/SCW63240.2024.00233

[18] Réseau de Transport d'Électricité (RTE), "écomix: Real-time electricity data in france," https://www.rte-france.com/en/eco2mix, 2025, accessed May 17, 2025.

[19] Electricity Maps, "Electricity maps: Real-time carbon intensity data," https://www.electricitymaps.com/, 2025, accessed May 17, 2025.

[20] M. Rouvier and P.-M. Bousquet, "Kiwano: A Cutting-Edge Open-Source Toolkit for Speaker Verification," in *Odyssey 2026*. ISCA, 2026.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018, arXiv:1806.05622.

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[23] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 2015. [Online]. Available: https://arxiv.org/abs/1510.08484

[24] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech*, 2019.

[26] M. Rouvier and P.-M. Bousquet, "Studying squeeze-and-excitation used in cnn for speaker verification," 2021. [Online]. Available: https://arxiv.org/abs/2109.05977

[27] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Interspeech*, 2017.

[28] J. Hintz and I. Siegert, "Commonbench: A larger scale speaker verification benchmark," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 17–20.

[29] Y. Fan, L. Chen, S. Kang, and et al., "Cn-celeb: A challenging chinese speaker recognition dataset," in *Proc. Interspeech*, 2019.