E-ISSN: 2962-6110
P-ISSN: 2964-7843
JSRET
jsret.knpub.com
Journal of Scientific Research, Education, and Technology

Conformer-Performer: An Efficient Architecture for Voice Activity Detection

Echa Apriliyanto, Anita Fira Waluyo
Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

ABSTRACT

Voice Activity Detection (VAD) is a crucial pre-processing step for speech technologies, yet standard Conformer architectures suffer from quadratic computational complexity. This study introduces the Conformer-Performer, a novel architecture utilizing the Fast Attention Via positive Orthogonal Random features (FAVOR+) mechanism to achieve linear complexity. The objective was to develop an efficient VAD model maintaining high accuracy for resource-constrained applications. The model was trained on the multilingual FLEURS dataset using a robust teacher-student framework. Uniquely, we utilized the restored FLEURS-R corpus to generate high-fidelity labels while training on noisy FLEURS inputs, effectively enforcing noise-invariant boundary detection. Experimental results demonstrate that the Conformer-Performer achieves an F1-score of 98.29%, statistically comparable to the standard Conformer's 98.41%, while achieving a 7.8-fold reduction in peak GPU memory usage and a 3.46-fold speedup in CPU inference time. Furthermore, this training strategy enabled the model to significantly outperform the SileroVAD teacher on the original test set. These findings confirm that the Conformer-Performer offers a compelling balance of accuracy and efficiency, suitable for real-time, on-device speech processing.

Keywords: *Voice Activity Detection, Conformer, Performer, Deep Learning, Linear Attention*

Corresponding author
Name: Echa Apriliyanto
Email: echa.apriliyanto.dev@gmail.com

INTRODUCTION

Voice Activity Detection (VAD) serves as the fundamental gatekeeper in modern speech processing pipelines, tasked with the critical binary classification problem of distinguishing speech segments from non-speech silence or background noise. The accuracy of this front-end module is paramount, as errors at this stage propagate downstream, degrading the performance of subsequent high-level tasks such as Automatic Speech Recognition (ASR), speaker diarization, and speech enhancement (Sharma et al., 2022). In real-world scenarios, audio streams are rarely clean, they are often contaminated by non-stationary noise, reverberation, and competing acoustic events. Therefore, an effective VAD system must be robust enough to filter out irrelevant acoustic information while preserving the integrity of the speech signal to reduce the computational load on downstream applications (Ball, 2023).

Historically, VAD methodologies relied on statistical signal processing techniques utilizing hand-crafted features paired with models like Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs) (Sharma et al., 2022). While computationally inexpensive, these methods lacked the discriminative power for complex, non-stationary noise environments (Ball, 2023; Braun & Tashev, 2021). The advent of Deep Learning introduced Convolutional Neural Networks (CNNs) to capture local spectro-temporal patterns and Recurrent Neural Networks (RNNs) for sequential modeling (Jia et al., 2021; Wilkinson & Niesler, 2021). However, RNNs suffer from sequential processing constraints that limit parallelization. To address this, the Conformer architecture emerged, hybridizing the locality of CNNs with the global modeling capabilities of Transformers to capture both local interactions and global dependencies (Gulati et al., 2020).

This architecture has proven highly effective for ASR and has been successfully adapted for VAD tasks. However, the deployment of Conformer-based VADs on resource-constrained edge devices faces a critical theoretical bottleneck, that is the standard Multi-Head Self-Attention (MHSA) mechanism scales quadratically with the sequence length. Specifically, calculating the attention matrix requires time complexity of $O(L^2 d)$ and space complexity of $O(L^2 + Ld)$, where L is the sequence length and d is the hidden dimension (Choromanski et al., 2022). This quadratic scaling creates a massive memory footprint when processing long audio sequences, prohibiting the deployment of these powerful models on devices like smartphones or embedded IoT systems where low latency is non-negotiable (Köpüklü & Taseska, 2022).

This research addresses the efficiency-accuracy trade-off by proposing the Conformer-Performer, a novel architecture that fundamentally rethinks the attention mechanism within the VAD context. We posit that the quadratic bottleneck of the standard Conformer is unnecessary for the VAD task. Instead of calculating the full attention matrix, we integrate the "Performer" attention mechanism, which utilizes the Fast Attention Via positive Orthogonal Random features (FAVOR+) kernel. This method provides an unbiased approximation of the standard softmax attention kernel but scales linearly with sequence length. By mathematically decoupling the query and key matrix multiplications using random feature maps, the Performer architecture eliminates the need to materialize the massive attention matrix. This reduces the time complexity to $O(Lrd)$ and space complexity to $O(Lr + Ld + rd)$, where r is the number of random features (Choromanski et al., 2022). This effectively linearizes the attention mechanism ($O(L)$), allowing the model to process indefinitely long sequences with a significantly reduced memory footprint.

To ensure that this architectural efficiency does not come at the cost of detection accuracy, we employ a rigorous training methodology designed for robustness. The model is trained on the massive multilingual FLEURS dataset, which provides n-way parallel speech data across 102 languages, ensuring that the VAD learns universal speech characteristics rather than over-fitting to a specific language phonology (Conneau et al., 2022). Furthermore, we adopt a "teacher-student" knowledge distillation framework to overcome the scarcity of labeled VAD data in diverse conditions (Dinkel et al., 2021). We utilize the enterprise-grade SileroVAD as the teacher model to generate high-quality frame-level

pseudo-labels on the restored FLEURS-R speech corpus (Ma et al., 2024; Silero Team, 2021). This clean supervisory signal guides the lightweight Conformer-Performer student model, a method empirically shown to improve generalization in noisy conditions by allowing the student to learn from the teacher's robust representations (Luckenbaugh et al., 2021). Finally, to enforce noise invariance, we implement a comprehensive data augmentation pipeline, utilizing spectrogram augmentation to mask time and frequency bands, forcing the model to rely on robust contextual cues rather than specific spectral artifacts (Park et al., 2019; Wang et al., 2022).

In summary, this research makes several key contributions to the field of efficient speech processing. First, we propose the Conformer-Performer VAD, replacing the quadratic attention mechanism with the linear-complexity FAVOR+ kernel to effectively solve the computational bottleneck of standard Conformers. Second, we demonstrate through rigorous evaluation that this architecture achieves a 7.8-fold reduction in peak GPU memory usage and a 3.46-fold speedup in CPU inference compared to a standard Conformer, making it viable for on-device deployment. Finally, we validate the model's robustness on the multilingual FLEURS dataset and the challenging, out-of-domain AVA-Speech dataset, confirming that efficiency gains do not compromise detection accuracy.

METHOD

1. Data Sources and Preparation

To ensure the proposed model generalizes effectively across diverse linguistic and acoustic environments, this study utilized a composite dataset approach. The primary training corpus was the FLEURS dataset, a massive multilingual parallel speech corpus covering 102 languages, which allows the model to learn universal phonetic representations rather than overfitting to specific language characteristics (Conneau et al., 2022). To facilitate our teacher-student training strategy, we also utilized FLEURS-R, a computationally restored version of the FLEURS dataset that provides higher signal fidelity for generating reliable supervisory labels (Ma et al., 2024). To simulate realistic "in-the-wild" conditions, the clean speech data was dynamically augmented with diverse noise sources during training. We incorporated background noise, music, and babble from the MUSAN corpus (Snyder et al., 2015), environmental soundscapes from the ESC-50 dataset (Piczak, 2015), and urban acoustic events from the UrbanSound8K dataset (Salamon et al., 2014). Additionally, reverberation effects were applied using impulse responses from the Room Impulse Response (RIR) dataset to model varying spatial acoustics (Ko et al., 2017).

2. Research Design

The overall research methodology follows a rigorous pipeline designed to balance efficiency and accuracy. The process begins with data aggregation, followed by a teacher-student labeling process where the robust teacher model supervises the efficient student model. The complete training and augmentation workflow is illustrated in Figure 1.

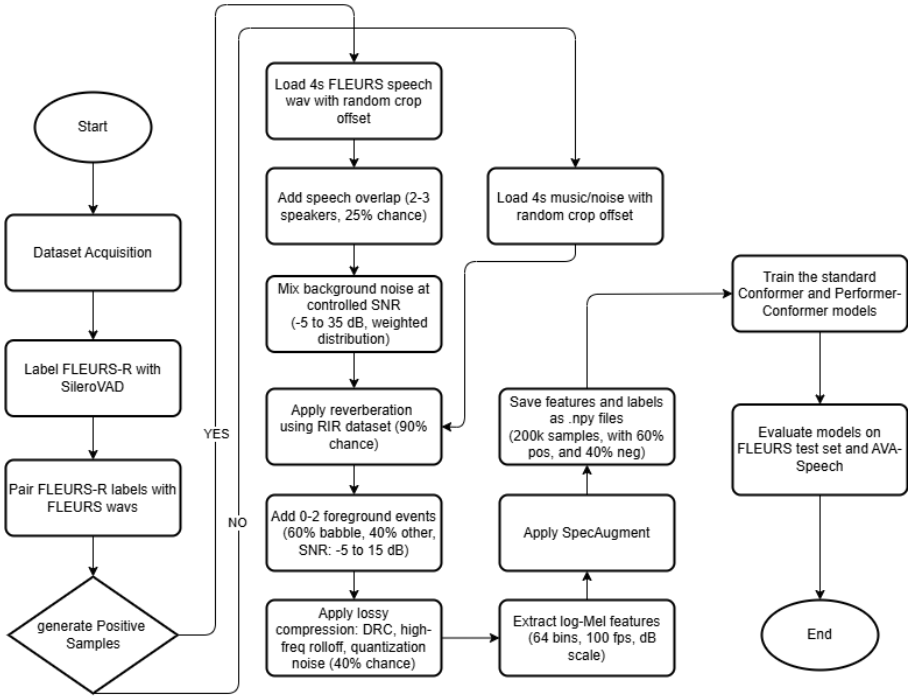

Figure 1. *Research Framework*

3. Feature Extraction

The audio preprocessing pipeline was standardized to ensure input consistency across all datasets. All raw audio inputs were first resampled to a 16 kHz monaural format. We extracted Log-Mel spectrograms using 64 Mel filterbanks spanning a frequency range of 20 Hz to 7600 Hz. These features were computed using a 30 ms window with a 10 ms hop length, resulting in a temporal resolution of 100 frames per second. To mitigate channel variability and loudness differences between recordings, we applied per-utterance Cepstral Mean and Variance Normalization (CMVN) prior to feeding the features into the neural network. A sample of the extracted feature representation is shown in Figure 2.

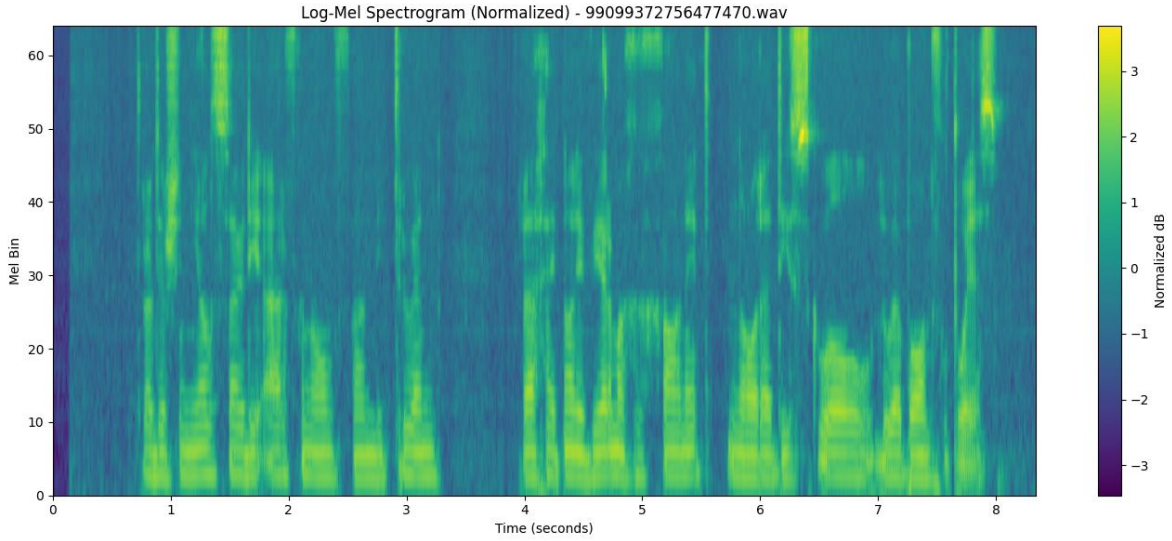

Figure 2. *A sample of Log-Mel spectrogram feature representationrethink*

4. Conformer-Performer Architecture

The core contribution of this study is the Conformer-Performer, a novel architecture designed to address the computational bottlenecks of standard Voice Activity Detection (VAD) models. While the standard Conformer backbone effectively captures local and global dependencies, its Multi-Head Self-Attention (MHSA) mechanism scales quadratically with sequence length due to the pairwise computation of the attention matrix (Gulati et al., 2020). To resolve this, we replaced the standard attention module with the Performer mechanism, which utilizes the Fast Attention Via positive Orthogonal Random features (FAVOR+) kernel (Choromanski et al., 2022).

The FAVOR+ mechanism fundamentally rethinks the attention operation by applying a kernel approximation method to the softmax function. Instead of computing the exact similarity between every query and key pair, which forces the materialization of the massive attention matrix, FAVOR+ maps the input vectors into a randomized orthogonal feature space. This mapping enables the use of the associative property of matrix multiplication to mathematically decouple the query and key interactions. Conceptually, the model first aggregates the keys and values into a compact global context vector and then queries this context linearly. This reduces the time and space complexity to a linear scale ($O(L)$) while maintaining a theoretically unbiased approximation of the standard attention (Choromanski et al., 2022). The detailed mathematical formulation of this linearization process is provided in the Appendix. The architectural integration of this module is depicted in Figure 3.

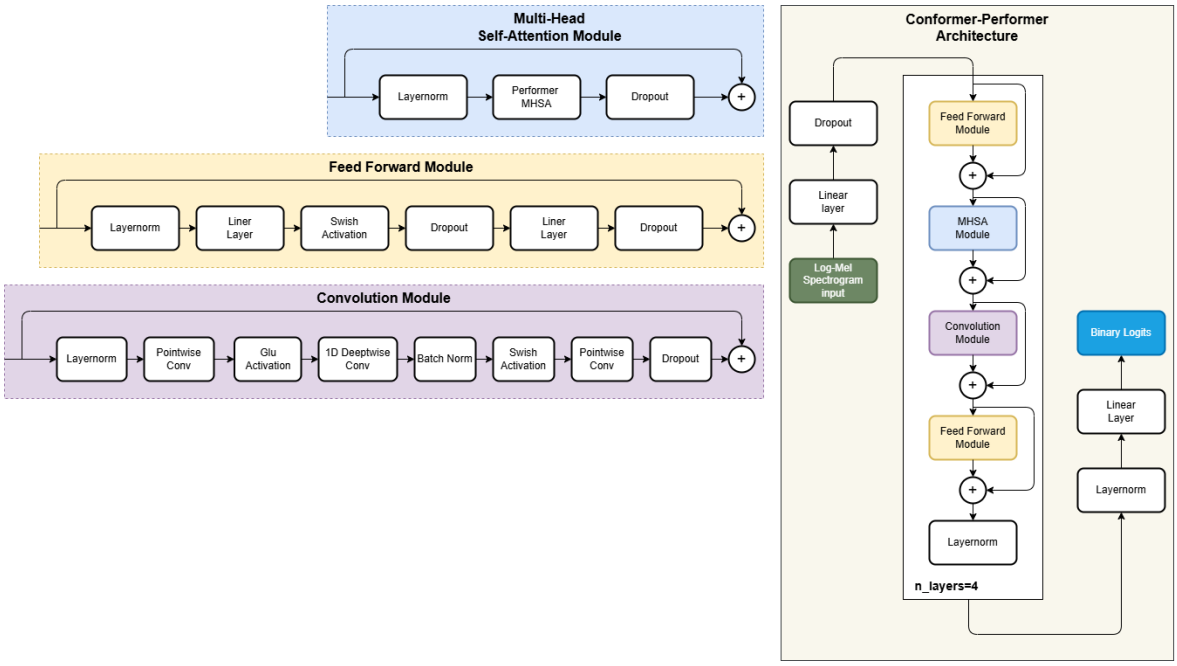

Figure 3. *Conformer-Performer architecture*

The final model configuration consisted of a model dimension of 64, 2 attention heads, and 4 stacked Conformer blocks, utilizing a kernel size of 31 for the depthwise convolution modules. The detailed hyperparameter settings are listed in Table 1.

Table 1. *Hyperparameter Configuration of the Conformer-Performer*

Parameter	Value	Description
Model Dimension	64	Base dimensionality (d_{model})
Attention Heads	2	Number of parallel attention heads
Conformer Blocks	4	Number of stacked layers
FFN Dimension	256	Feed-forward hidden dimension
Conv Kernel Size	31	Depthwise convolution size
Random Features	32	Random features for FAVOR+ (r)
Dropout	0.2	Regularization rate

5. Training Configuration

We employed a "teacher-student" knowledge distillation framework to overcome the scarcity of high-quality VAD labels for diverse acoustic conditions. The enterprise-grade SileroVAD model served as the teacher, generating frame-level pseudo-labels on the high-quality FLEURS-R dataset (Ma et al., 2024; Silero Team, 2021). These labels provided a clean supervisory signal to train the student Conformer-Performer model on the original, noisy FLEURS data. The model was trained to minimize the Binary Cross-Entropy with Logits loss using the AdamW optimizer with a base learning rate of $1e^{-4}$ and a cosine annealing

schedule with a 2000-step warmup. To further prevent overfitting, we applied spectrogram augmentation, which randomly masks blocks of time steps and frequency channels during training (Park et al., 2019).

6. Evaluation Metrics

The performance of the proposed architecture was evaluated using a comprehensive set of classification and efficiency metrics. Classification accuracy was assessed using the F1-score, Area Under the Receiver Operating Characteristic (AUROC), and Equal Error Rate (EER) on both the in-domain FLEURS test set and the out-of-domain AVA-Speech dataset (Chaudhuri et al., 2018). To quantify the architectural efficiency gains, we measured the Real-Time Factor (RTF) on CPU and the Peak GPU Memory usage during inference. These efficiency metrics are critical for determining the viability of the model for deployment on resource-constrained edge devices (Köpüklü & Taseska, 2022).

RESEARCH RESULT

1. VAD Performance on In-Domain Data (FLEURS)

The primary evaluation of the proposed Conformer-Performer architecture was conducted on the multilingual FLEURS test set to assess its efficacy on read speech, matching the training domain. The quantitative results, presented in Table 2, compare the model against the Standard Conformer baseline and the SileroVAD teacher model. The data indicates that the Conformer-Performer achieved an F1-score of 98.2%, which is statistically comparable to the 98.4% achieved by the Standard Conformer. Similarly, the Area Under the ROC Curve (AUROC) for the Conformer-Performer was 0.993, showing a negligible difference of 0.001 compared to the Standard Conformer (0.994). In terms of error rates, the Conformer-Performer recorded an Equal Error Rate (EER) of 2.9%, whereas the SileroVAD baseline recorded a significantly higher EER of 18.6%.

Table 2. *Model Performance Metrics on the FLEURS Test Set*

Model	Accuracy	F1	Precision	Recall	AUROC	EER
Conformer-Performer	0.975	0.982	0.983	0.982	0.993	0.029
Conformer	**0.976**	**0.984**	0.983	**0.984**	**0.994**	**0.027**
Silero	0.757	0.815	0.943	0.717	0.890	0.186

2. Qualitative Analysis of Noise Robustness

To investigate the significant performance gap between the proposed model (2.9% EER) and the SileroVAD baseline (18.6% EER), we visualized detection boundaries on a representative sample from the FLEURS test set. As detailed in the methodology, the Conformer-Performer was trained on noisy FLEURS inputs using high-fidelity labels from the restored FLEURS-R corpus. This mismatch forces the model to learn noise-invariant features and focus on the structural rhythm of speech rather than transient acoustic energy. Figure 4 compares the detection probability curves of the models against the ground truth (Green

blocks). The SileroVAD baseline (Purple curve) exhibits significant instability and is characterized by rapid oscillations and frequent threshold crossings during speech segments. This suggests the baseline reacts to short-term acoustic fluctuations and results in a lower F1-score of 0.807 for this sample.

In contrast, the Conformer-Performer (Orange curve) demonstrates remarkable temporal consistency. It produces smooth and confident plateaus near probability 1.0 during speech with clean drops during silence. By effectively filtering out the artifacts that confuse the baseline, it achieves the highest accuracy (0.978) and F1-score (0.983) on this sample. This slight performance edge over the Standard Conformer (Pink curve, F1 0.980) confirms that the linear attention mechanism successfully captures the intentional structure of human speech while ignoring environmental interference.

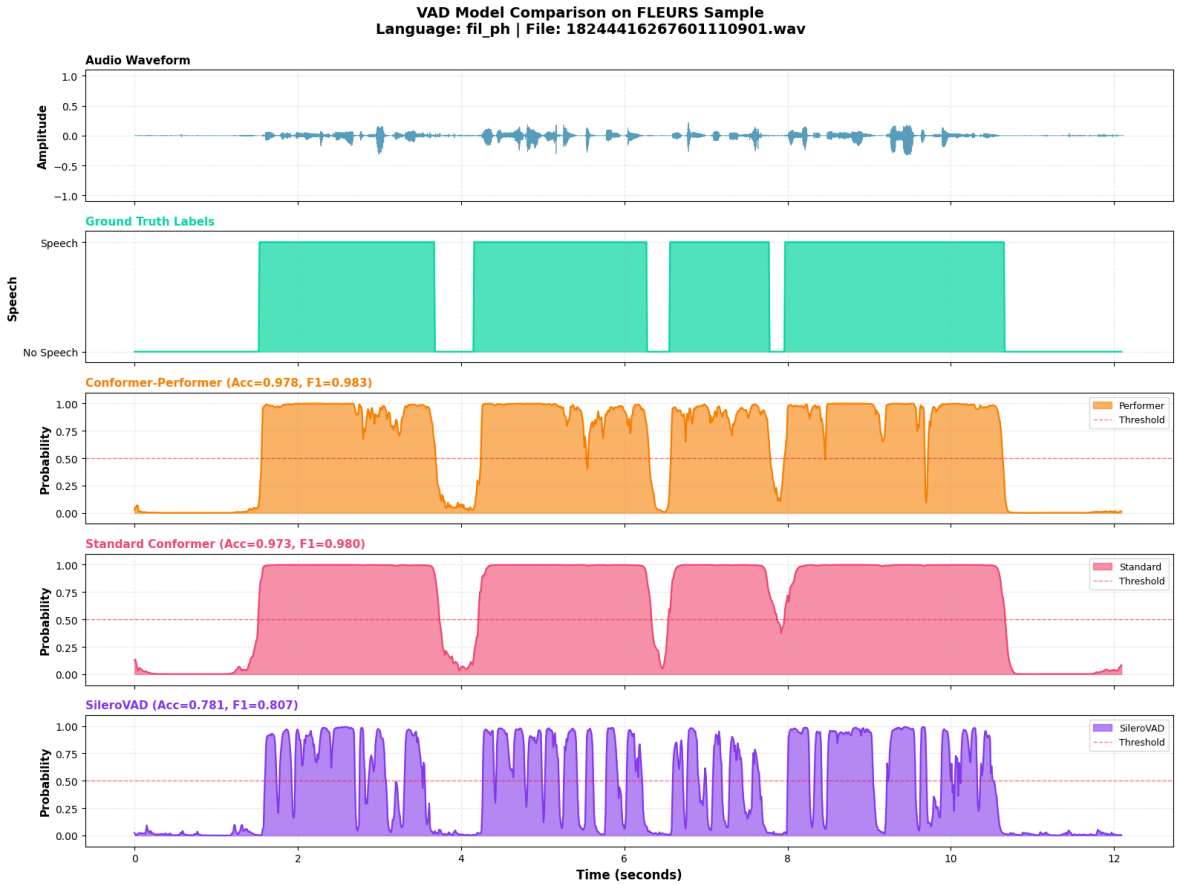

Figure 4. *Frame-Level Probability Comparison against Ground Truth.*

3. Computational Efficiency Analysis

To validate the architectural efficiency improvements, we measured the resource consumption of the models during inference. Table 3 details the Peak GPU Memory usage, Real-Time Factor (RTF), and average CPU inference time per 4-second audio chunk. The Conformer-Performer required 38 MB of Peak GPU Memory, representing a 7.8-fold

reduction compared to the 303 MB required by the Standard Conformer. Regarding latency, the Conformer-Performer achieved an average CPU inference time of 87 ms, which corresponds to a 3.46-fold speedup over the Standard Conformer (302 ms). The SileroVAD model remained the fastest on CPU (52 ms) and had a memory footprint of 136 MB, which is higher than the Conformer-Performer but lower than the Standard Conformer.

Table 3. *Computational Efficiency Benchmarks on the FLEURS Test Set*

Model	Peak GPU Memory (MB)	RTF	Avg. CPU Inference Time (ms)	Parameters
Conformer-Performer	**38**	0.007	87	461569
Conformer	303	0.024	302	**396801**
Silero	136	**0.004**	**52**	462594

4. Generalization to Out-of-Domain Data (AVA-Speech)

To assess zero-shot generalization, the models were evaluated on the AVA-Speech dataset, which consists of complex movie audio distinct from the training distribution (Chaudhuri et al., 2018). As shown in Table 4, the Conformer-Performer maintained robust performance with an F1-score of 0.796 and an AUROC of 0.872. The Standard Conformer achieved a slightly higher F1-score of 0.801 and an AUROC of 0.876. Both Conformer-based architectures significantly outperformed the SileroVAD baseline, which recorded an F1-score of 0.697 and an AUROC of 0.774 on this dataset.

Table 4. *Zero-Shot Generalization Performance on the AVA-Speech Dataset*

Model	F1	Precision	Recall	AUROC
Conformer-Performer	0.796	0.846	**0.753**	0.872
Conformer	**0.801**	**0.878**	0.736	**0.876**
Silero	0.697	0.755	0.647	0.774

5. Benchmarking Against State-of-the-Art

We further compared the proposed model against established benchmarks provided in the original AVA-Speech study (Chaudhuri et al., 2018). Table 5 presents the True Positive Rate (TPR) at a fixed False Positive Rate (FPR) of 0.315. The Conformer-Performer achieved a TPR of 0.838. This performance exceeds that of the lightweight "tiny 320" baseline (TPR 0.810) and the "RTC vad" baseline (TPR 0.722). The Standard Conformer achieved a TPR of 0.850. The "resnet 960" model, which was trained directly on the in-domain AVA-Speech data, achieved the highest TPR of 0.917.

Table 5. *Comparison of TPR (at FPR=0.315) Against Published Benchmarks on AVA-Speech Dataset*

Model	TPR (at FPR=0.315)
Conformer-Performer	0.838
Conformer	0.850
Silero (Silero Team, 2021)	0.721
resnet_960 (Chaudhuri et al., 2018)	**0.917**
tiny_320 (Chaudhuri et al., 2018)	0.810
RTC_vad (Chaudhuri et al., 2018)	0.722

DISCUSSION

1. Interpretation of Findings

The experimental results strongly validate the primary hypothesis of this study, demonstrating that the quadratic computational bottleneck of standard Transformer-based VADs can be effectively mitigated using linear attention mechanisms without compromising discriminative power. The Conformer-Performer maintained an F1-score of 98.2% on the FLEURS dataset, representing a negligible degradation of only 0.12% compared to the Standard Conformer. This finding is significant as it empirically demonstrates that the FAVOR+ kernel provides a sufficiently high-fidelity approximation of the softmax attention surface for speech tasks. By preserving the model's ability to capture global context through linear attention, we successfully decoupled the computational cost from the sequence length. Crucially, the 7.8-fold reduction in Peak GPU memory and the 3.46-fold speedup in CPU inference time confirm that the theoretical linear complexity of the Performer translates directly into tangible benefits for edge deployment. This aligns with recent efforts in the field to optimize Transformer architectures for mobile devices, where memory footprint is often the primary constraint preventing the deployment of high-capacity models (Köpüklü & Taseska, 2022).

2. Relationship to Previous Studies

When placed in the context of existing literature, the Conformer-Performer occupies a unique niche between ultra-lightweight statistical models and heavy Transformer architectures. Previous efficiency-focused architectures, such as MarbleNet, have utilized depthwise separable convolutions to achieve low parameter counts (Jia et al., 2021). However, these convolutional approaches often lack the global context modeling capabilities required for complex acoustic scenes. Our model retains the global modeling capacity of the Conformer while approaching the inference speeds of purely convolutional benchmarks. Similarly, while recent works utilizing Spiking Neural Networks (sVAD) have demonstrated extreme power efficiency, they typically suffer from lower absolute accuracy in high-noise environments compared to Transformer-based approaches (Yang et al., 2024). The Conformer-Performer offers a balanced alternative, providing server-grade accuracy suitable for high-quality downstream tasks while maintaining mobile-ready efficiency.

Furthermore, the proposed model significantly outperformed the SileroVAD baseline which served as the teacher. This substantial performance gap highlights the efficacy of the student network in generalizing beyond the teacher's capabilities when trained on a larger, more diverse dataset like FLEURS.

3. Generalization and Robustness

The robustness of the proposed model is further evidenced by its zero-shot performance on the AVA-Speech dataset. Despite being trained exclusively on read speech from the FLEURS corpus, the Conformer-Performer successfully generalized to the domain of movie audio, achieving a True Positive Rate of 0.838. This strong generalization can be attributed to the teacher-student training strategy employed. Distilling knowledge from a robust teacher allows the student to learn noise-invariant representations that are difficult to acquire from noisy labels alone (Luckenbaugh et al., 2021; Zhou et al., 2021). Additionally, the application of spectrogram augmentation played a critical role in preventing the model from overfitting to specific language phonologies or recording channels, a common pitfall in multilingual VAD systems (Park et al., 2019). The ability to maintain high performance across such a distinct domain shift confirms that the linear attention mechanism captures fundamental speech characteristics rather than superficial dataset artifacts.

4. Limitations and Future Work

Despite the clear advantages, this study identified a specific limitation regarding GPU latency. While the Conformer-Performer drastically reduced memory usage, its inference speed on GPU was slightly slower than the Standard Conformer. This counter-intuitive result is likely due to the computational overhead of generating and projecting the random features for the FAVOR+ kernel, which can be less optimized in current CUDA implementations compared to standard matrix multiplications (Choromanski et al., 2022). However, the massive gains in CPU speed make it superior for non-accelerated edge devices which are the primary target for this technology. Future research should focus on investigating quantization-aware training to further compress the model parameters for microcontroller deployment. Additionally, extending this linear-attention architecture to related downstream tasks such as speaker diarization, where long-sequence modeling is even more critical, represents a promising direction for future investigation.

CONCLUSION

This study successfully validates the Conformer-Performer as a robust and efficient architecture for Voice Activity Detection that effectively resolves the quadratic bottleneck of standard Transformer models. By integrating the linear complexity FAVOR+ attention kernel, the proposed architecture achieved detection performance statistically equivalent to the Standard Conformer on both the multilingual FLEURS dataset and the out-of-domain AVA-Speech benchmark. Crucially, the cross-domain teacher-student training strategy paired noisy FLEURS inputs with high fidelity labels from the restored FLEURS-R corpus. This approach enabled the model to learn noise-invariant representations and resulted in a 2.9%

Equal Error Rate compared to the 18.6% rate of the SileroVAD baseline. This accuracy was maintained while delivering a 7.8-fold reduction in peak GPU memory usage and a 3.46-fold speedup in CPU inference time. These findings confirm that linear attention mechanisms coupled with robust denoising supervision offer a compelling pathway for deploying high-capacity speech models on resource-constrained edge devices.

REFERENCES

Ball, J. (2023). *Voice Activity Detection (VAD) in Noisy Environments*. http://arxiv.org/abs/2312.05815

Braun, S., & Tashev, I. (2021). *On training targets for noise-robust voice activity detection*. http://arxiv.org/abs/2102.07445

Chaudhuri, S., Roth, J., Ellis, D. P. W., Gallagher, A., Kaver, L., Marvin, R., Pantofaru, C., Reale, N., Reid, L. G., Wilson, K., & Xi, Z. (2018). *AVA-Speech: A Densely Labeled Dataset of Speech Activity in Movies*. http://arxiv.org/abs/1808.00606

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., & Weller, A. (2022). *Rethinking Attention with Performers*. http://arxiv.org/abs/2009.14794

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., & Bapna, A. (2022). *FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech*. http://arxiv.org/abs/2205.12446

Dinkel, H., Wang, S., Xu, X., Wu, M., & Yu, K. (2021). *Voice activity detection in the wild: A data-driven approach using teacher-student training*. https://doi.org/10.1109/TASLP.2021.3073596

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). *Conformer: Convolution-augmented Transformer for Speech Recognition*. http://arxiv.org/abs/2005.08100

Jia, F., Majumdar, S., & Ginsburg, B. (2021). *MarbleNet: Deep 1D Time-Channel Separable Convolutional Neural Network for Voice Activity Detection*. http://arxiv.org/abs/2010.13886

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017). *A STUDY ON DATA AUGMENTATION OF REVERBERANT SPEECH FOR ROBUST SPEECH RECOGNITION*.

Köpüklü, O., & Taseska, M. (2022). ResectNet: An Efficient Architecture for Voice Activity Detection on Mobile Devices. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022-September*, 5363–5367. https://doi.org/10.21437/Interspeech.2022-820

Luckenbaugh, J., Abplanalp, S., Gonzalez, R., Fulford, D., Gard, D., & Busso, C. (2021). Voice activity detection with teacher-student domain emulation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *6*, 4521–4525. https://doi.org/10.21437/Interspeech.2021-1234

Ma, M., Koizumi, Y., Karita, S., Zen, H., Riesa, J., Ishikawa, H., & Bacchiani, M. (2024). *FLEURS-R: A Restored Multilingual Speech Corpus for Generation Tasks*. https://doi.org/https://doi.org/10.48550/arXiv.2408.06227

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2019-September*, 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Piczak, K. J. (2015). ESC: Dataset for environmental sound classification. *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference*, 1015–1018. https://doi.org/10.1145/2733373.2806390

Salamon, J., Jacoby, C., & Bello, J. P. (2014). A dataset and taxonomy for urban sound research. *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, 1041–1044. https://doi.org/10.1145/2647868.2655045

Sharma, M., Joshi, S., Chatterjee, T., & Hamid, R. (2022). A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows. In *Neurocomputing* (Vol. 494, pp. 116–131). Elsevier B.V. https://doi.org/10.1016/j.neucom.2022.04.084

Silero Team. (2021). *Silero Models: pre-trained enterprise-grade STT / TTS models and benchmarks*. GitHub. https://github.com/snakers4/silero-models

Snyder, D., Chen, G., & Povey, D. (2015). *MUSAN: A Music, Speech, and Noise Corpus*. http://arxiv.org/abs/1510.08484

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Wang, C., Song, Y., Liu, H., Liu, H., Liu, J., Li, B., & Yuan, X. (2022). Real-Time Vehicle Sound Detection System Based on Depthwise Separable Convolution Neural Network and Spectrogram Augmentation. *Remote Sensing*, *14*(19). https://doi.org/10.3390/rs14194848

Wilkinson, N., & Niesler, T. (2021). *A Hybrid CNN-BiLSTM Voice Activity Detector*. http://arxiv.org/abs/2103.03529

Yang, Q., Liu, Q., Li, N., Ge, M., Song, Z., & Li, H. (2024). *sVAD: A Robust, Low-Power, and Light-Weight Voice Activity Detection with Spiking Neural Networks*. http://arxiv.org/abs/2403.05772

Zhou, H., Du, J., Chen, H., Jing, Z., Xiong, S., & Lee, C. H. (2021). Audio-visual information fusion using cross-modal teacher-student learning for voice activity detection in realistic environments. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *6*, 4550–4554. https://doi.org/10.21437/Interspeech.2021-592

APPENDIX

Mathematical Formulation of FAVOR+

The standard Multi-Head Self-Attention (MHSA) computes the output Y using the softmax kernel, as defined by (Vaswani et al., 2017). To analyze its complexity, (Choromanski et al., 2022) define the operation in matrix form:

$$Attention(Q,K,V) = D^{-1}AV, \quad \text{where } A = \exp\left(\frac{QK^T}{\sqrt{d}}\right), \quad D = \text{diag}(A\mathbf{1_L}) \tag{1}$$

Here, A represents the attention weights and D is the normalization matrix (the row-sum of A). Computing A explicitly requires $O(L^2d)$ complexity. The FAVOR+ mechanism approximates the kernel A using random feature maps ϕ, such that $A(i,j) \approx \phi(q_i)^T\phi(k_j)$. By defining transformed matrices $Q' = \phi(Q)$ and $K' = \phi(K)$, the operation can be rewritten using the associative property of matrix multiplication:

$$\widehat{Attention}(Q,K,V) = \widehat{D^{-1}}\left(Q'(K'^TV)\right) \tag{2}$$

Crucially, the normalization term $\widehat{D}$ is also computed linearly without materializing the attention matrix:

$$\widehat{D} = \text{diag}\left(Q'(K'^T\mathbf{1_L})\right) \tag{3}$$

This decomposition reduces the complexity to linear time $O(Lrd)$, where r is the number of random features, while providing an unbiased estimation of the original softmax attention (Choromanski et al., 2022).