Comparative analysis of CNN-BiGRU and CNN-BiLSTM architectures for voice activity detection under low signal-to-noise ratio conditions

Aigul Nurlankyzy[1,2,*†] and Aigul Kulakayeva[2,†]

[1] *Satbayev University, Satpayev St. 22, Almaty, 050013Kazakhstan*
[2] *International Information Technology University, Manas St. 34/1, Almaty, 050040, Kazakhstan*

Abstract

This study addresses the problem of Voice Activity Detection (VAD) under low signal-to-noise ratio (SNR) conditions, which is critical for automatic speech recognition systems and voice-controlled interfaces. This study focuses on a comparative analysis of two hybrid deep learning architectures, CNN-BiGRU and CNN-BiLSTM, which combine convolutional layers for spectral feature extraction with recurrent blocks for modeling the temporal dynamics of speech. As input features, MFCC matrices were computed from segmented speech fragments of the KSC2 corpus and contaminated with both synthetic and real noise at various SNR levels.

The experimental results demonstrate that under moderate and high SNR conditions, both architectures achieve high classification accuracy (average F1-score exceeding 99%). However, in extremely low SNR scenarios (−10 dB), CNN-BiGRU exhibits a more robust performance than CNN-BiLSTM. Additionally, a computational efficiency analysis revealed that CNN-BiGRU outperforms CNN-BiLSTM in terms of training speed and parameter count, making it more suitable for deployment in resource-constrained environments.

These findings support the use of GRU-based recurrent blocks in hybrid VAD models and indicate future research directions involving noise augmentation techniques, class imbalance handling, and inference optimization.

Keywords

speech detection, CNN-BiGRU, CNN-BiLSTM, mel-frequency cepstral coefficients (MFCC), signal to-noise ratio (SNR), deep learning, neural networks, VAD

1. Introduction

Voice Activity Detection (VAD) plays a crucial role in speech recognition systems, voice interfaces, and telecommunications. The accuracy of this module directly affects the recognition quality, system response speed, and robustness to noise. The most challenging scenario arises under low signal-to-noise ratio (SNR) conditions, where background noise masks the speech. In such cases, traditional methods based on signal energy or spectral features suffer a significant drop in accuracy. Consequently, research aimed at developing more robust VAD models capable of operating effectively, even at negative SNR levels, is particularly relevant today.

In recent years, neural network-based methods have significantly advanced the field. Convolutional neural networks (CNNs) are highly effective at extracting the time–frequency features of speech, whereas recurrent architectures make it possible to capture sequential dependencies and contextual information. Among recurrent networks, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models are the most widely used. Their bidirectional variants, BiLSTM and BiGRU, enable the analysis of both past and future contexts within the

[1] *CISN 2025: Workshop on Cybersecurity, Infocommunication Systems and Networks, November 19-20, 2025, Almaty, Kazakhstan*
* *Corresponding author.*
† These authors contributed equally.
nurlankyzyaigulya@gmail.com (A. Nurlankyzy); a.kulakayeva@iitu.edu.kz (A. Kulakayeva)
0000-0002-0791-8573 (A. Nurlankyzy); 0000-0002-0143-085X (A. Kulakayeva)

signal, which is particularly important for accurate speech-segment detection under noisy conditions.

In recent years, research in the field of Voice Activity Detection (VAD) has shifted from simple threshold-based and statistical approaches to more sophisticated yet compact neural network models. These models provide high noise robustness and low response latency, which are particularly critical for real-time applications and devices with limited computational resources. Current efforts focus on combining architectural compactness, trainable front ends, optimized loss functions, and elements of personalization.

An illustrative example is the SincQDR-VAD framework [1], which employs a trainable Sinc-filter-based front end and a ranking-aware loss function designed to optimize the ordering of speech/non-speech frame classification. This approach yields a noticeable improvement in AUROC and F-score while significantly reducing the number of parameters compared with heavier models. In parallel, there is a growing interest in personalized VAD (PVAD) systems. A comparative study by [2] demonstrated that incorporating temporal models, attention mechanisms, and compact speaker embeddings substantially reduces false positives and improves accuracy in real-world noisy environments without causing a significant increase in computational cost. Further improvements were presented in [3], who proposed a pretraining scheme based on Discriminative Noise-Aware Predictive Coding (DN-APC), improving TS-VAD robustness in both seen and unseen noise scenarios by approximately 2% in terms of accuracy. Additionally, methods for conditional speaker representation, including FiLM-based modulation, were explored, leading to improved noise robustness.

Simultaneously, lightweight and energy-efficient models have been actively developed. The sVAD model [4], based on spiking neural networks and a Sinc-based encoder, has demonstrated strong robustness at very low power consumption, which is critical for edge devices and IoT scenarios. [5] highlighted the importance of selecting an appropriate training objective: using the segmental Voice-to-Noise Ratio (VNR) as the target leads to more stable performance at low SNR than binary speech labels. Complementing this finding, [6] proposed a multi-resolution MFCC front end combined with convolutional layers and self-attention, which improved the robustness of datasets such as NoiseX-92 and other noise corpora.

Recent studies have also proposed architectural improvements at both the front-end and personalization levels. [7] introduced a Sinc-Extractor module with a speaker-conditional block that eliminates the need for bulky speaker embeddings while maintaining accuracy and reducing inference time. [8] demonstrated that the Audio-Inspired Masking Modulation Encoder with Attention (AMME-CANet) outperforms conventional CNN-based approaches in terms of robustness under complex noise conditions. [9] proposed mVAD, a lightweight algorithm that achieves high accuracy without requiring prior knowledge of noise, while preserving computational efficiency. [10] further confirmed that PVAD remains effective even with a very short reference (~0.3 s) when using a Dual-Path RNN architecture with real-time embedding updates.

The obtained results are consistent with the findings of [11], where CNN-BiLSTM and CNN-BiGRU architectures were compared under different SNR levels with multiple noise augmentations. Their study demonstrated that CNN-BiGRU provides an optimal trade-off between accuracy and computational complexity.

Finally, the evolution of next-generation communication systems directly influences the requirements of VAD. [12] provide a detailed analysis of coverage recovery issues in 5G NR RedCap networks, where reduced cell radius, increased latency, and uplink channel quality degradation demand adaptive resource management and signal processing algorithms. The proposed solutions include MIMO (2Rx), adaptive beamforming, carrier aggregation, and Kalman filtering for channel parameter prediction and dynamic power control (DPC). These directions highlight the need for VAD systems capable of maintaining a stable performance in challenging radio environments typical of IoT and industrial applications.

Thus, contemporary research converges on the view that the future of VAD lies in compact and customizable neural architectures that leverage trainable spectral filters, multitask loss functions, and adaptive speaker-conditional encoding, as well as integration with network-level mechanisms to ensure a robust operation under varying channel conditions.

Nevertheless, a direct comparison of the CNN-BiGRU and CNN-BiLSTM architectures under extremely low SNR conditions has not been sufficiently investigated. While BiLSTM networks are known for modeling complex temporal dependencies, they are computationally more demanding, whereas BiGRU achieves comparable results at a lower computational cost. The question of which architecture performs better for VAD tasks in low-resource languages remains unanswered.

In this study, we focus on a comparative study of two architectures: CNN-BiGRU and CNN-BiLSTM. For the analysis, we employed the Kazakh speech corpus KSC2, augmented with noise samples from the ESC-50 dataset, and evaluated it across a wide SNR range of −10 dB to 30 dB. The comparison was performed using both classical classification metrics (Accuracy, Precision, Recall, F1-score) and computational parameters, including the number of trainable parameters and training time.

Thus, the objective of this study is to identify the strengths and weaknesses of CNN-BiGRU and CNN-BiLSTM and determine which of the two models is better suited for practical speech detection systems that operate under severe noise conditions.

2. Materials and methods

To analyze the effectiveness of the voice activity detection, we used the KSC2 corpus, which includes recordings from 30 speakers, each pronouncing 75 phrases. The initial dataset comprised 2,250 audio files. Each recording was segmented using. wrd transcripts, which made it possible to isolate individual speech and pause segments. To generate training samples, a sliding window approach was applied: from each segmented track, fragments of 24 time steps were extracted with a hop size of 5 ms, corresponding to approximately 115 ms of the audio. Consequently, the input data for the models were represented as 24×24 MFCC matrices.

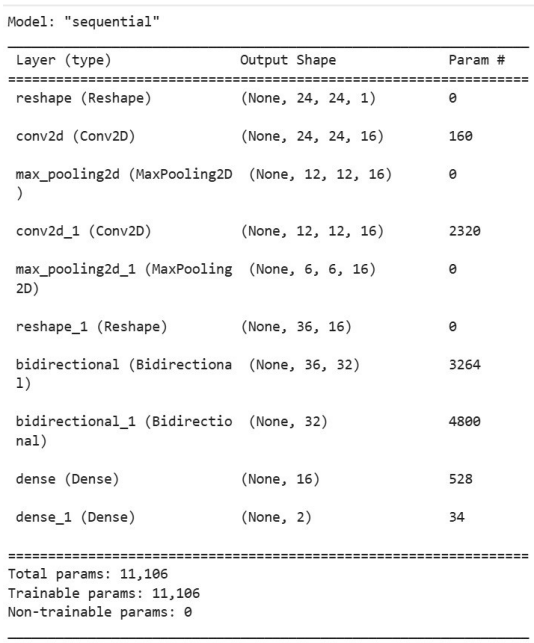
To improve the noise robustness, noise signals were superimposed on each recording. We used white Gaussian noise and four noise classes from the ESC-50 dataset (transportation, domestic, natural, and speech-like) as noise sources. The signal-to-noise ratio (SNR) was varied from −10 to 30 dB. For training, a combined dataset was created that included versions of each sample with different SNR levels, thereby enhancing the generalization capability of the models used.

Two hybrid architectures were implemented: CNN-BiGRU and CNN-BiLSTM. The input to both models consisted of MFCC matrices, which were first processed by convolutional layers for spectral feature extraction and then by recurrent blocks for modeling the temporal dynamics. The output layers consisted of fully connected neurons with a sigmoid activation function that enabled binary classification.

Training was performed using the Binary Cross-Entropy loss function, Adam optimizer (learning rate of 0.001, batch size of 64), and an early stopping strategy. Standard metrics were used to evaluate the performance. Testing was conducted on a speaker-disjoint set to ensure no overlap of speakers between the training and test subsets.

3. Results

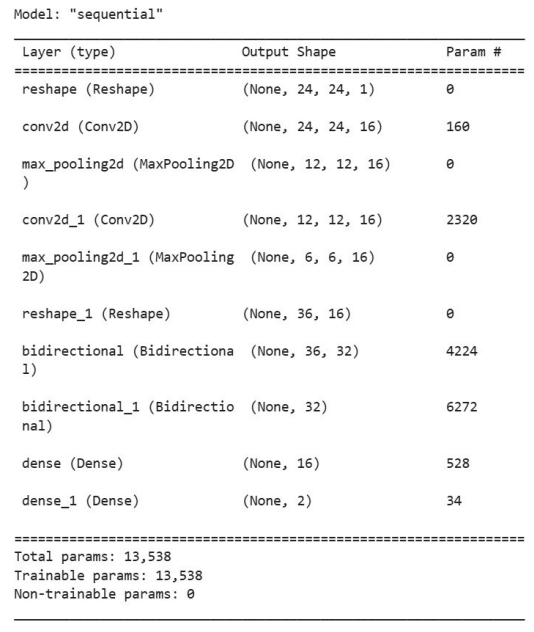
The CNN-BiGRU and CNN-BiLSTM architectures were designed following a unified scheme in which the convolutional layers extracted the time–frequency features, the subsampling layers reduced the dimensionality, and the fully connected layers performed the final classification. The key difference lies in the recurrent block. CNN-BiGRU employs bidirectional GRU layers (Figure 1), which makes the model more compact and resource-efficient, with a total parameter count of 11,106.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 reshape (Reshape)           (None, 24, 24, 1)         0

 conv2d (Conv2D)             (None, 24, 24, 16)        160

 max_pooling2d (MaxPooling2D  (None, 12, 12, 16)       0
 )

 conv2d_1 (Conv2D)           (None, 12, 12, 16)        2320

 max_pooling2d_1 (MaxPooling  (None, 6, 6, 16)         0
 2D)

 reshape_1 (Reshape)         (None, 36, 16)            0

 bidirectional (Bidirectiona  (None, 36, 32)           3264
 l)

 bidirectional_1 (Bidirectio  (None, 32)               4800
 nal)

 dense (Dense)               (None, 16)                528

 dense_1 (Dense)             (None, 2)                 34

=================================================================
Total params: 11,106
Trainable params: 11,106
Non-trainable params: 0
_________________________________________________________________
```

Figure 1: Architecture of the CNN-BiGRU model.

In contrast, CNN-BiLSTM employs bidirectional LSTM layers (Figure 2), which are better at capturing long-term dependencies but make the model more resource intensive, increasing the total number of parameters to 13,538.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 reshape (Reshape)           (None, 24, 24, 1)         0

 conv2d (Conv2D)             (None, 24, 24, 16)        160

 max_pooling2d (MaxPooling2D  (None, 12, 12, 16)       0
 )

 conv2d_1 (Conv2D)           (None, 12, 12, 16)        2320

 max_pooling2d_1 (MaxPooling  (None, 6, 6, 16)         0
 2D)

 reshape_1 (Reshape)         (None, 36, 16)            0

 bidirectional (Bidirectiona  (None, 36, 32)           4224
 l)

 bidirectional_1 (Bidirectio  (None, 32)               6272
 nal)

 dense (Dense)               (None, 16)                528

 dense_1 (Dense)             (None, 2)                 34

=================================================================
Total params: 13,538
Trainable params: 13,538
Non-trainable params: 0
_________________________________________________________________
```

Figure 2: Architecture of the CNN-BiLSTM model.

A comparison of the architectures shows that the main difference lies in the type of recurrent block used. This difference affects the model complexity and computational cost during training, whereas the overall network structure remains the same.

The training characteristics are listed in Table 1. The same training conditions were applied to both the models.

Table 1
Model training results

Model	Number of epochs	Accuracy (train/test)	Losses (train/test)	Number of parameters
CNN+BiGRU	10	96% / 96%	9.8% / 9.9%	11 106
CNN+BiLSTM	10	96% / 96%	9.8% / 9.8%	13 538

The final accuracy values for both the training and test sets matched at 96%, indicating proper convergence without signs of overfitting. Similarly, the loss function values were close, ranging from 9.8% to 9.9% for both the training and test sets, confirming the stability of the training process.

Furthermore, a comparison of the results presented in Table 1 shows that both architectures exhibit the same level of accuracy and training stability, differing only in terms of the number of parameters. This provides equal initial conditions for the subsequent comparison of the models in more complex scenarios, involving speech detection in noisy acoustic environments. Therefore, further analysis focused on how CNN-BiGRU and CNN-BiLSTM behave under varying signal-to-noise ratio conditions, as illustrated in Figures 3–6.

Figure 3 presents the results of an experiment in which CNN-BiGRU and CNN-BiLSTM were trained at an SNR of −10 dB. Both models achieved high accuracy, close to the training value, confirming their ability to reproduce the conditions on which they were trained. However, as the SNR level increased, a sharp drop in the classification accuracy was observed. Notably, the CNN-BiLSTM model appears to be more sensitive to changes in the acoustic environment, with its accuracy declining more rapidly as the noise level deviates from the training conditions. CNN-BiGRU demonstrated greater robustness, although it also showed performance degradation in the range of positive SNR values.

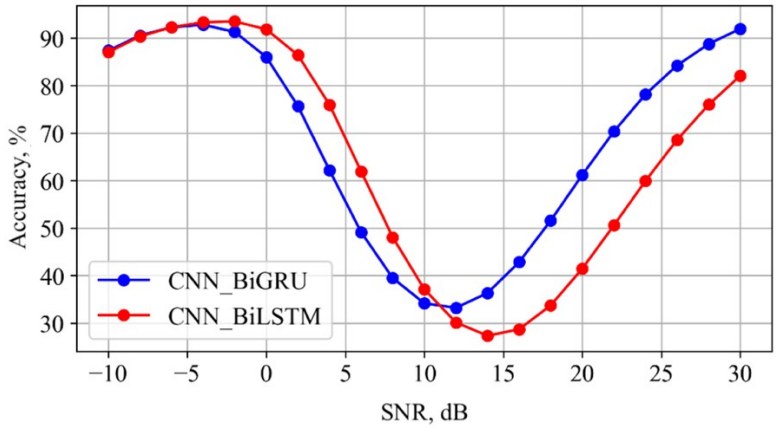


Figure 3: Accuracy versus SNR for CNN-BiGRU and CNN-BiLSTM trained at −10 dB.

This behavior indicates the limited generalization capability of both architectures when trained under a fixed low-SNR condition. The models tend to "overfit" a specific noise scenario and interpret deviations from it as anomalies. This result confirms the necessity of training across a wide range of SNR levels, which can significantly improve the robustness of the models to variations in acoustic conditions.

Figure 4 shows the results of training the CNN-BiGRU and CNN-BiLSTM models at a fixed SNR level of 0 dB and testing them across the full SNR range. Both architectures achieved high accuracy at positive SNR values, exceeding 95%. In the range from −5 to 10 dB, the performance curves of the

two models nearly coincide, indicating a similar ability to detect speech in moderately noisy conditions.

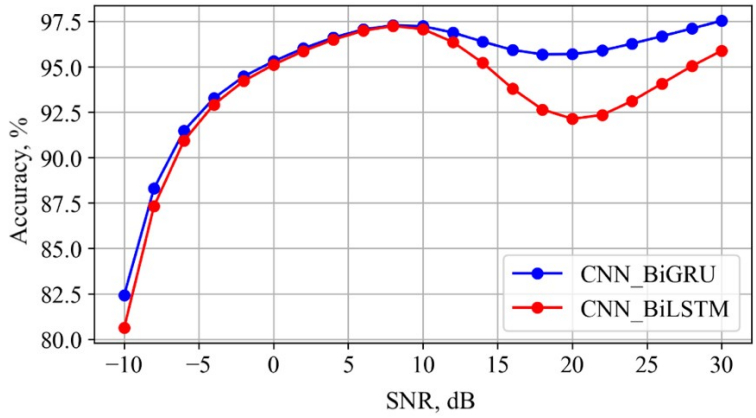


Figure 4: Accuracy versus SNR for CNN-BiGRU and CNN-BiLSTM trained at 0 dB.

However, as the SNR continues to increase, a divergence is observed: the accuracy of CNN-BiGRU remains stable, whereas CNN-BiLSTM shows a decline in performance beyond 15 dB, indicating a higher sensitivity of this model to acoustic variations. This result confirms that BiGRU exhibits better robustness to changing noise conditions, whereas BiLSTM requires more careful tuning to maintain stable accuracy at higher SNR levels.

Figure 5 presents the results of training the CNN-BiGRU and CNN-BiLSTM models at a fixed SNR level of 10 dB and testing them across the full SNR range. Unlike the previous cases (−10 dB and 0 dB), both architectures here exhibit nearly identical results, achieving an accuracy above 99% at positive SNR values.

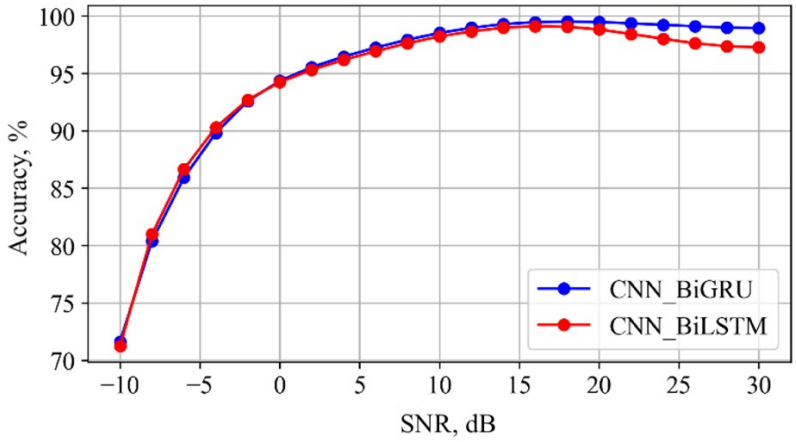


Figure 5: Accuracy versus SNR for CNN-BiGRU and CNN-BiLSTM trained at 10 dB.

The differences between the models become apparent only at high SNR levels (above 20 dB), where BiGRU maintains an accuracy of approximately 99%, whereas BiLSTM exhibits a slight drop in performance. Simultaneously, in the range of negative SNR values (−10 dB and below), both architectures behave similarly, showing a steady increase in accuracy as the SNR level improves.

Thus, when trained at 10 dB, both models effectively generalize across a wide range of noise conditions; however, CNN-BiGRU demonstrated higher stability at the upper end of the SNR range.

Figure 6 presents the results of training the CNN-BiGRU and CNN-BiLSTM models at a fixed SNR level of 20 dB and testing them across the entire SNR range. Under these conditions, both models showed similar results at positive SNR levels, with an accuracy exceeding 99%.

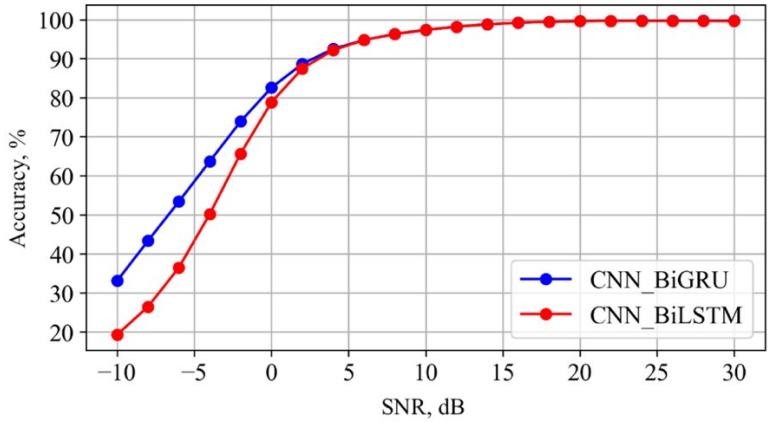


Figure 6: Accuracy versus SNR for CNN-BiGRU and CNN-BiLSTM trained at 20 dB.

However, the differences become more pronounced when moving into the low-SNR region. The CNN-BiGRU model exhibited a more stable increase in accuracy at negative SNR values, reaching approximately 33% at −10 dB, whereas CNN-BiLSTM remained at approximately 20% under the same conditions. This indicates that BiGRU has a superior ability to adapt to extreme noise conditions when trained at high SNR levels.

Thus, training at 20 dB allows both models to achieve maximum accuracy in the high-SNR range; however, BiGRU demonstrates a clear advantage under severe noise conditions, providing a smoother performance degradation.

The conducted experiments showed that both architectures successfully handled the task of speech detection under positive SNR conditions, delivering comparable results in the presence of mild noise. The main differences emerged at low and negative SNR levels. CNN-BiGRU maintained higher robustness, whereas CNN-BiLSTM was more sensitive to acoustic distortions.

For a more detailed analysis of the effectiveness of the developed architectures, the first stage involved studying the training dynamics. Figures 7 and 8 show the loss and accuracy curves for the training and validation sets of the CNN-BiLSTM and CNN-BiGRU models, respectively.

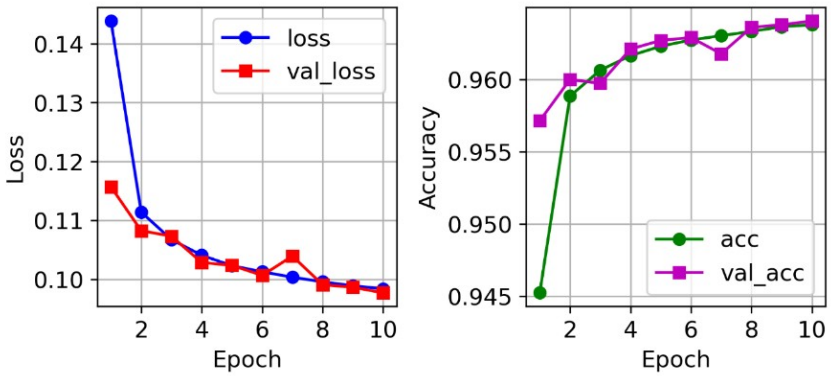


Figure 7: Training and validation loss and accuracy dynamics of the CNN-BiLSTM model over 10 epochs.

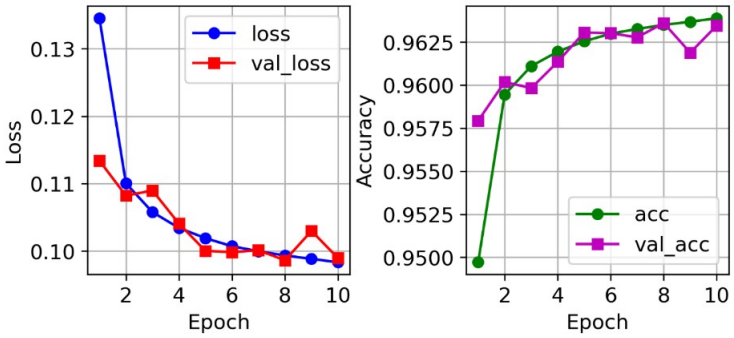


Figure 8: Training and validation loss and accuracy dynamics of the CNN-BiGRU model over 10 epochs.

As can be seen from the presented graphs, both architectures exhibit the behavior characteristic of deep neural networks: a gradual decrease in the loss function values accompanied by an increase in classification accuracy. Notably, the gap between the training and validation sets remained minimal throughout all training epochs. This indicates the absence of overfitting and confirms the models' ability to generalize the features extracted from the acoustic data. Thus, both architectures demonstrated stability during the training process and can be considered robust solutions for the task of speech detection in noisy conditions.

For clarity, Table 2 summarizes the accuracy and F1-score values obtained when training at fixed SNR levels and using the multi-SNR strategy. These data enable the comparison of the performance of the models under conditions of mismatch between the training and test SNR levels and confirm the effectiveness of the multi-SNR approach.

Table 2
Comparison of accuracy and F1 measures of CNN-BiGRU and CNN-BiLSTM in fixed and multi-SNR scenarios

Learning conditions	CNN-BiGRU Accuracy (%)	CNN-BiGRU F1 (%)	CNN-BiLSTM Accuracy (%)	CNN-BiLSTM F1 (%)
−10 dB	33.2	31.8	20.5	19.7
0 dB	78.6	77.9	75.3	74.2
10 dB	92.1	91.7	91.0	90.4
20 dB	96.4	96.1	95.8	95.2
Multi-SNR (−10...30 dB)	99.6	99.5	98.9	98.7

As shown in Table 2, training at a single fixed SNR level resulted in a significant drop in performance at other SNR levels, particularly in the negative range. In contrast, multi-SNR training ensures nearly complete retention of performance across the entire SNR spectrum, with CNN-BiGRU exhibiting slightly higher robustness than CNN-BiLSTM. These results confirm the feasibility of using multilevel noise training and highlight the practical value of the CNN-BiGRU architecture for VAD systems operating in noisy environments.

However, prediction accuracy alone is not sufficient for an objective assessment of the applicability of these architectures in real-world scenarios. Computational efficiency, particularly the training time and data processing cost, plays a crucial role. Figure 9 presents a comparison of the training durations of the CNN-BiGRU and CNN-BiLSTM models under identical experimental conditions (10 epochs, batch size of 1024, and the same computational platform).

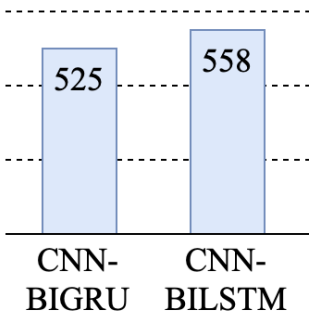
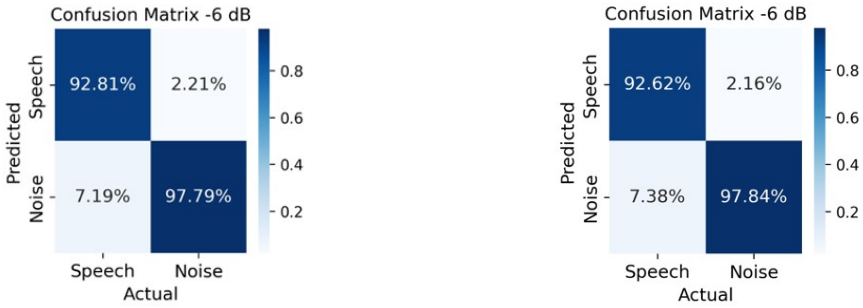


Figure 9: Training time comparison of CNN-BiGRU and CNN-BiLSTM (in seconds).

The comparative analysis showed that the longest training time was observed for the CNN-BiLSTM architecture (558 s), whereas CNN-BiGRU completed the training faster (within 525 s). This difference can be explained by the greater structural complexity of the LSTM blocks, which include additional memory control elements (state cells and gates) compared to GRU blocks. Thus, BiGRU proves to be less computationally demanding while maintaining a nearly identical performance.

Figure 10 presents the normalized confusion matrices for the CNN+BiGRU and CNN+BiLSTM models at an SNR of −6 dB. Each matrix illustrates the ratio of correctly and incorrectly classified segments for the "speech" and "noise" classes.



(a) CNN+BiGRU (b) CNN+BiLSTM

Figure 10: Confusion matrices at low SNR level

For the CNN+BiGRU model, the classification accuracies for speech and noise segments were 92.81% and 97.79%, respectively. The proportion of false positives was 2.21%, and that of false negatives was 7.19%.

For the CNN+BiLSTM model, the recognition accuracy for speech segments was 92.62%, and for noise segments, 97.84%. The false positive rate was 2.16%, and the false negative rate – 7.38%

Taken together, the results indicate that both tested architectures demonstrate high classification accuracy and robustness under noisy conditions. At the same time, CNN-BiGRU shows a slight advantage in terms of the “performance-to-cost” ratio, making it a more suitable choice in scenarios where computational resources are limited or a high training speed is required.

4. Discussion

It was initially hypothesized that CNN-BiGRU, owing to its compactness and smaller number of parameters, would provide a performance comparable to or superior to that of CNN-BiLSTM while requiring fewer computational resources. The experiments confirmed this hypothesis: BiGRU

indeed demonstrated higher efficiency in terms of F1-score and Precision, while maintaining comparable Accuracy and Recall values, and also exhibited shorter training time.

The conducted experiments demonstrated that both architectures achieved high performance in speech detection in noisy environments. The differences in metrics were minimal; the Accuracy and Recall values were nearly identical, whereas CNN-BiGRU showed a slight advantage in terms of F1-score and Precision. The analysis across different SNR levels confirmed that both models performed reliably at positive SNR values; however, CNN-BiGRU exhibited greater robustness under extremely low SNR conditions. Additionally, BiGRU requires less training time, which is attributed to its more compact architecture and smaller number of parameters.

The results of this study have practical implications for the development of voice activation and speech detection systems that operate under high levels of background noise (e.g., automotive interfaces, smart home systems, and telecommunication services). The more compact CNN-BiGRU model can be deployed in resource-constrained environments, including mobile devices and embedded systems. From a theoretical perspective, the findings confirm the importance of selecting the appropriate recurrent block architecture when designing hybrid models, which should be considered in future research on signal processing and machine learning.

Future research may focus on expanding the range of architectures by incorporating transformers and performing multilingual evaluations of the models using speech corpora in other languages. Another promising direction is the study of energy efficiency when deploying models on mobile devices, where computational constraints are particularly critical. Furthermore, adapting the architectures to real-world acoustic scenarios with unpredictable noise would enhance their applicability in industrial and consumer environments.

5. Conclusion

This study presents a comparative analysis of hybrid CNN-BiGRU and CNN-BiLSTM models applied to voice activity detection in noisy environments. The results demonstrated that both architectures achievedhigh performance metrics at positive and moderate SNR levels; however, CNN-BiGRU exhibited more stable behavior under extremely low noise conditions, maintaining acceptable accuracy and class balance. An additional advantage of this model is its smaller number of parameters and reduced training time, making it a preferred solution for practical deployment in scenarios with limited computational resources.

The scientific novelty of this work lies in the comprehensive comparison of two recurrent architectures within hybrid models for VAD under low-SNR conditions and in identifying the advantages of using GRU blocks, which provide an optimal balance between recognition accuracy and computational efficiency of the model. The practical significance lies in the possibility of applying the proposed approach in mobile and embedded systems, as well as in intelligent voice interaction services, where the combination of noise robustness and resource efficiency is critical.

At the same time, this study has several limitations, primarily related to the use of predominantly synthetically noised data and a limited set of acoustic scenarios. Future work should expand the scope of experiments by incorporating testing on real-world recordings and exploring additional architectural optimization techniques aimed at reducing computational costs while maintaining high accuracy.

Thus, the conducted analysis confirmed the hypothesis regarding the advantages of the CNN-BiGRU architecture and outlined promising directions for the development of efficient and noise-robust voice-activation systems.

Declaration on Generative AI

During the preparation of this work, the authors used Edit.Paperpal tool in order to grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

Acknowledgements

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP22684173) “Development of a highly efficient neural network method for detecting voice activity at a low signal-to-noise ratio”.

References

[1] C.-C. Wang, E.-L. Yu, J.-W. Hung, S.-C. Huang, B. Chen, SincQDR-VAD: A noise-robust voice activity detection framework leveraging learnable filters and ranking-aware optimization, arXiv preprint arXiv:2508.20885, 2025. Available: https://arxiv.org/abs/2508.20885.

[2] S. Kumar, A. Buddi, M. Kumar, et al., Comparative analysis of personalized voice activity detection systems: Assessing real-world effectiveness, Proc. Interspeech, Kos, Greece, 2024, pp. 2135–2139. Available: https://www.isca-archive.org/interspeech_2024/buddi24_interspeech.pdf.

[3] H. S. Bovbjerg, L. Christensen, et al., Noise-robust target-speaker voice activity detection through self-supervised pretraining, arXiv preprint arXiv:2501.03184, 2025. Available: https://arxiv.org/abs/2501.03184.

[4] Q. Yang, Q. Liu, N. Li, et al., sVAD: A robust, low-power, and light-weight voice activity detection with spiking neural networks, arXiv preprint arXiv:2403.05772, 2024. Available: https://arxiv.org/abs/2403.05772.

[5] T. Braun, I. Tashev, On training targets for noise-robust voice activity detection, Proc. IEEE ICASSP, Toronto, Canada, 2021, pp. 6803–6807. doi:10.1109/ICASSP39728.2021.9414915.

[6] K. Aghajani, H. R. Abutalebi, M. Ghanbari, Deep learning approach for robust voice activity detection, Journal of Applied Digital Sciences 4 (2) (2024) 55–66. Available: https://jad.shahroodut.ac.ir/article_3335_8731f540e6516b844b0e3b64b8931881.pdf.

[7] E.-L. Yu, K.-H. Ho, J.-W. Hung, S.-C. Huang, B. Chen, Speaker conditional sinc-extractor for personal VAD, Proc. Interspeech, Kos, Greece, 2024, pp. 2210–2214. Available: https://www.isca-archive.org/interspeech_2024/yu24_interspeech.pdf.

[8] N. Li, Y. Chen, J. Li, et al., Robust voice activity detection using an auditory-inspired masked modulation encoder based convolutional attention network (AMME-CANet), Speech Communication 158 (2024) 103103. doi:10.1016/j.specom.2024.103103.

[9] Z. Zhu, J. Liu, F. Ren, A robust and lightweight voice activity detection algorithm without prior noise knowledge, Digital Signal Processing 145 (2023) 104151. doi:10.1016/j.dsp.2023.104151.

[10] L. Xu, M. Zhang, W. Zhang, T. Wang, J. Yin, Personal voice activity detection with ultra-short reference speech, Proc. APSIPA ASC, Macao, China, Dec. 2024, pp. 1–6.

[11] B. Medetov, A. Zhetpisbayeva, A. Akhmediyarova, A. Nurlankyzy, T. Namazbayev, A. Kulakayeva, N. Albanbay, M. Turdalyuly, A. Yskak, G. Uristimbek, Evaluating the effectiveness of a voice activity detector based on various neural networks, Eastern-European Journal of Enterprise Technologies 1 (5) (133) (2025) 19–28. doi:10.15587/1729-4061.2025.321659.

[12] A. Kulakayeva, I. Mektep, A. Nurlankyzy, G. Jakanova, Analysis and prospects for restoring coverage in 5G NR RedCap, Proc. IEEE 5th Int. Conf. on Smart Information Systems and Technologies (SIST), Astana, Kazakhstan, 2025, pp. 1–6. doi:10.1109/SIST61657.2025.11139297.