Voice Command Recognition for 3D Endless Games Using Hybrid Transformer LSTM

Oddy Virgantara Putra[1*], Adrik Fikhtiyaril Amro[2], Adimas Arya Alief Riarta[3], Alvin Arya Pangestu[4]

[1,2,3,4]Department of Informatics, Faculty of Science and Technology, Universitas Darussalam Gontor, Ponorogo, East Java, Indonesia
Email: [1*]oddy@unida.gontor.ac.id, [2]adrikfikhtiyaarilamro41@student.cs.unida.gontor.ac.id, [3]adimasaliefriarta@student.cs.unida.gontor.ac.id, [4]alvinaryapangestu@student.cs.unida.gontor.ac.id

(Naskah masuk: 10 Nov 2025, direvisi: 13 Dec 2025, diterima: 17 Dec 2025)

Abstract

Voice command recognition has become increasingly important for enabling natural human–computer interaction in gaming and embedded systems. However, achieving accurate and noise-robust recognition on small-scale datasets remains challenging due to the limited availability of data and computational resources. To address this, this paper presents a systematic study on architectural choices for small-scale speech command recognition. We compare three neural architectures, BiLSTM, Transformer, and a sequential hybrid, in a controlled framework using identical MFCC front-ends and standardized noise augmentation. Experiments include various configurations, varying sampling rates, and noise types, rigorously evaluated using repeated cross-validation to ensure reliability. The results show that the hybrid architecture achieves superior accuracy, clearly outperforming the standalone BiLSTM and standalone Transformer baseline architectures. The hybrid model exhibits lower variance across cross-validation folds and initialization processes, and demonstrates significantly higher training throughput compared to the BiLSTM model. The model demonstrates excellent robustness to acoustic noise; variants trained with pink and white noise augmentation show comparable accuracy, confirming robust feature learning under diverse data augmentation conditions. These findings support the hypothesis that BiLSTM's local temporal modeling complements Transformer's global self-attention, enabling more effective capture of multi-scale temporal patterns in short utterances. Real-time deployment feasibility was confirmed through integration with a 3D game engine, achieving an average total response time of 5.23 ms, demonstrating the model's suitability for low-latency interactive applications.

Keywords: Speech Recognition, BiLSTM, Transformers, Hybrid Model, MFCC.

I. INTRODUCTION

The 3D gaming industry is growing rapidly. One of the major advances in this field is the use of voice commands to control games. To validate this concept, we developed a prototype game. The prototype utilized 3D character assets capable of four actions: jump, slide, move right, and move left. The primary motivation for this work was to enable voice-based control for this application.

Of course, implementing a voice command system requires a fast response, and most importantly, it must remain robust in noisy acoustic environments. Acoustic disturbances can originate from various sources, such as cooling fans, background conversation, or the game's own audio output. If the system misclassifies a command or exhibits excessive latency, the game may become uncontrollable.

Previous research has explored various methods of voice command classification [1], [2]. Long Short-Term Memory (LSTM) has long been relied upon for modeling temporal sequential data such as audio [3]. Recently, the Transformer architecture, which is dominated by self-attention mechanisms, has revolutionized tasks in the audio processing domain. However, the performance of Transformers trained from scratch is not guaranteed to achieve superior performance when trained from scratch on limited-volume datasets [4]. This presents a research gap for a controlled performance comparison between these two architectures [4]. To address this, we propose a hybrid BiLSTM-Transformer architecture and evaluate it using key comparative parameters: classification accuracy, training throughput, and real-time response latency. To ensure the reliability of the findings, we used a rigorous validation methodology. Each model configuration was evaluated using Stratified 5-Fold Cross-Validation repeated three times with different random initialization seeds (a total of 15 training processes per configuration) [5]. The expected target of this research is to

identify an architecture that not only achieves superior accuracy (>95%) compared to standalone models but also meets the strict real-time requirement (<50 ms) for interactive gaming. This study empirically identifies hybrid architecture as the most accurate and robust solution for this small-scale, four-class classification task.

This article is organized as follows: Section II describes our main contributions. Section III details the research questions. Section IV reviews related work. Section V outlines the methodology. Section VI presents the results. Section VII discusses the findings.

II. CONTRIBUTIONS

This work advances voice command recognition research through three interconnected contributions that address critical gaps in small-scale, resource-constrained ASR systems:

A. A Systematic Benchmark of a BiLSTM-Transformer Hybrid for Voice Command Classification

We propose and evaluate a compact hybrid architecture based on the sequential integration of a BiLSTM, a projection layer, and a Transformer encoder [4], [6]. The BiLSTM module models local temporal dependencies in MFCC features using bidirectional context. A linear projection then maps the 256-dimensional BiLSTM outputs to a 128-dimensional embedding space, which serves as input to the Transformer encoder. The Transformer encoder applies self-attention to capture global acoustic patterns over the entire utterance. This design is motivated by the hypothesis that local pattern extraction (BiLSTM) and global context modeling (Transformer) provide complementary information for short-utterance classification [7]. The hypothesis is evaluated through a systematic comparison against standalone LSTM and Transformer baselines across 12 configurations (3 architectures × 2 sampling rates × 2 noise types), all using a fixed 30-coefficient MFCC front-end. Each configuration is assessed with stratified 5-fold cross-validation across 3 seeds (15 runs per configuration; 180 trained models total). To the best of our knowledge, this is one of the first empirical studies to train a BiLSTM-Transformer hybrid architecture from scratch for four-class voice command recognition using fewer than 10,000 training samples.

B. Systematic Multi-Parameter Evaluation Framework

Second, this study introduces a systematic evaluation framework that jointly analyzes model architecture and signal-processing parameters for small-scale voice command recognition. Three representative architectures are examined: LSTM, Transformer, and the proposed BiLSTM-Transformer hybrid, capturing the spectrum from parameter-efficient recurrent models to attention-based and hybrid designs. A factorial design is applied, varying two key factors: sampling rate (8 kHz vs. 16 kHz) and noise type (white vs. pink). This results in 12 distinct experimental configurations when applied across the three architectures. Each configuration is evaluated using stratified 5-fold cross-validation and three independent random seeds (36, 38, 42), yielding 180 trained models in total. This design enables a statistically reliable analysis of both main effects and interaction effects between architecture and front-end parameters. To the best of our knowledge, this constitutes the first comprehensive sensitivity study that links architectural choices and signal-processing configurations in a unified framework for small-scale voice command classification, providing evidence-based guidance for hyperparameter selection.

C. Empirical Insights Into Accuracy-Efficiency Trade-offs

Third, we provide empirical analysis of the trade-off between recognition performance and computational efficiency. The study compares: (i) an LSTM baseline at 8/16 kHz, (ii) a Transformer baseline at 8/16 kHz, and (iii) the proposed BiLSTM-Transformer hybrid at 8/16 kHz. In addition to accuracy-based metrics, we report parameter counts, training throughput (samples/sec), and per-epoch wall-clock time. These results offer practical guidance for deploying ASR models in low-resource scenarios, highlighting when a hybrid architecture is preferable to purely recurrent or purely attention-based models in low-data, low-resource scenarios [8].

III. RELATED WORKS

A. An Effectivity of MFCC and LSTM

Mel-frequency cepstral coefficients (MFCCs) have long been a standard feature for audio signal analysis, valued for their ability to capture perceptually relevant spectral information. While 13 coefficients are conventionally used as a standard, prior optimization studies (e.g., Yan et al., 2025) demonstrate that increasing the dimensionality to approximately 30 coefficients capture finer spectral details, yielding peak accuracy for biomedical and respiratory sound tasks [9]. These gains generalize across models: using an LSTM-based classifier to validate MFCC tuning also led to improved accuracies of 6–14% on multiple datasets[9]. MFCC features remain effective in broader domains (e.g., urban acoustic event classification with high accuracy) [10].

Likewise, recurrent neural networks such as Long Short-Term Memory (LSTM) and its bidirectional variants (BiLSTM) have become essential for modeling temporal dynamics in audio analysis. Bidirectional LSTM (BiLSTM) architectures can capture long-range context more effectively than unidirectional models. Advances illustrate how optimization of MFCC parameters and adoption of enhanced BiLSTM architectures improve effectiveness across health, multimedia, and navigation applications [3].

B. Transformers in Audio Processing

Transformers introduced by Vaswani et al. (2017), replaces recurrence with self-attention. Reviews, such as Zaman et al. (2025), document the rapid adoption of Transformers across audio tasks. Attention enables modeling

of remote dependencies and parallel processing advantages over LSTMs (Karmarkar et al., 2024).

C. Gap Research and Positioning

Transformer-based architectures now dominate large-scale benchmarks in natural language processing and speech recognition, consistently achieving state-of-the-art performance. However, Zaman et al. (2025) [4] highlight that these gains come with substantial computational costs and a strong reliance on large-scale training data. Loubser et al. (2024) [8], who investigated small-scale Transformers trained from scratch for ASR in under-resourced languages, further show that achieving competitive performance without extensive pre-training remains challenging. These findings raise a critical question for resource-constrained scenarios for compact classification tasks with limited data, might more efficient architectures such as LSTM-based models be more appropriate than Transformer-only solutions?

Noise robustness poses an additional challenge for practical deployment of voice command recognition systems. Pizzi et al. (2026) [11] demonstrate that noise-based data augmentation improves robustness not only under noisy conditions but also against adversarial perturbations in modern ASR systems, including Transformer-based architectures. Their results support the use of principled noise augmentation as a core component in designing robust ASR pipelines, especially for real-world environments.

Despite these advances, three key gaps remain insufficiently addressed:

- Gap 1. Lack of controlled LSTM-Transformer comparison on small-scale tasks (matched features, augmentation, rigorous validation).
- Gap 2. Limited study of hybrid architectures for accuracy efficiency trade-offs under fixed, well-motivated feature dimensionality.
- Gap 3. Limited multi-parameter analysis of noise type and sampling rate interactions (white vs. pink; 8 vs. 16 kHz) under low-resource voice command classification.

This study systematically addresses all three identified gaps. By implementing a standardized evaluation framework with fixed feature dimensionality, we provide the missing controlled comparison between LSTM and Transformer architectures (Gap 1). Furthermore, the proposed BiLSTM-Transformer hybrid is specifically designed and evaluated to optimize the accuracy-efficiency trade-off, directly targeting Gap 2. Finally, our factorial experimental design, which varies sampling rates and noise types, offers the improved multi-parameter analysis required to close Gap 3.

IV. METHODOLOGY

The overall workflow of the proposed system, from audio input to classification output, is illustrated in Figure 1 Proposed Hybrid BiLSTM-Transformer System Architecture. This pipeline ensures a standardized processing flow for all architectural comparisons. The system processes raw audio (1s @16kHz) through MFCC extraction (30 coefficients), followed by a BiLSTM for local temporal features, a linear projection, and a Transformer Encoder for global context, culminating in a classification head for the four commands.

A. MFCC Extraction Features

The Speech is resampled (8 kHz or 16 kHz) and framed with a 25 ms window, 10 ms hop, Hamming window. We compute log power spectra, apply a Mel filterbank, then DCT to obtain MFCCs (fixed at 30 coefficients). The choice of 30 is explicitly based on findings by Yan et al., which indicate that increasing coefficients beyond the standard 13 to approximately 30 significantly improves feature robustness and classification accuracy [9]. Unless stated, no pre-emphasis. SR-dependent framing in code: 16 kHz → n_fft=512, win_length=400, hop_length=160, 8 kHz → n_fft=256, win_length=200, hop_length=80. Frame-wise MFCCs are aggregated (mean pooling) for utterance-level classification [12].

B. Bidirectional Long Short-Term Memory

We use a BiLSTM to model temporal dependencies over 30-D MFCC sequences. Two bidirectional LSTM layers (hidden size 128 per direction, dropout 0.2) produce a global representation passed to a classification head (LayerNorm, Dropout, Linear). Orthogonal and Xavier uniform initializations stabilize training [13]. Experiments at 8 kHz and 16 kHz with fixed 30-D MFCCs and white/pink noise augmentation (SNR 5–20 dB).

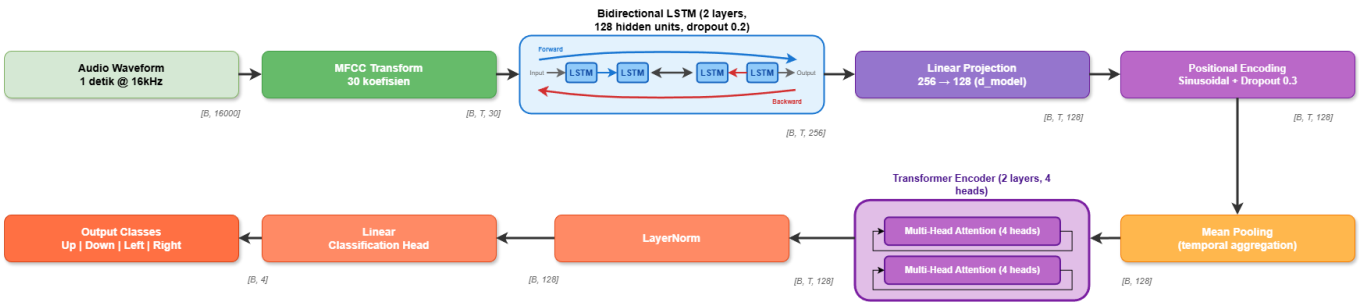

Figure 1. Proposed Hybrid BiLSTM-Transformer System

C. Stratified K-Fold Cross-Validations

In this study, we used *a* K-fold cross-validation scheme (K=5). Specifically, we implemented a Stratified K-Fold CV, where all data is partitioned into 5 folds without overlapping, while maintaining a proportional distribution between classes ('left', 'right', 'up', 'down') in each *fold*. The use of this stratified method is essential in classification tasks to ensure fair evaluation and reduce bias, especially if the distribution of data is unbalanced [5]. The model is trained on K-1 folds and validated on the remaining folds. The key to our methodology is that the entire 5-fold CV process is repeated three times using different random seeds (36, 38, 42). This multiple-run approach is crucial because it allows us to measure and quantify the performance of the model that can arise purely from the initialization of random weights [14].

D. Transformer

We employ an encoder-only Transformer for speech-command classification using sequences of MFCC frames as input (Fig 1). This choice follows recent overviews showing that Transformers capture long-range temporal dependencies in audio, work well on spectrogram-like inputs, and can be trained efficiently via self-attention and feed-forward stacks for detection and classification tasks [4]. Each utterance is converted to a sequence $\mathbf{X} \in \mathbb{R}^{T\times n_{\mathrm{MFCC}}}$of MFCC vectors (we use $n_{\mathrm{MFCC}} = N_{\mathrm{MFCC}}$in the code). Using MFCCs is consistent with compact, effective front ends for end-to-end ASR/audio models and has been used successfully in small-scale Transformer pipelines [4]. Frames are linearly projected to the model dimension d_{model} via self. Input proj. Because self-attention is order-agnostic, we add sinusoidal positional encoding (PositionalEncoding) before the encoder stack, providing the network with absolute and relative position information standard practice in Transformer audio pipelines. We use a stack of num_layers Transformer encoder blocks (multi-head self-attention plus position-wise FFN, GELU, residual connections, and LayerNorm; batch_first=True, norm_first=True). This follows practices summarized in recent surveys and lightweight ASR designs demonstrating that compact Transformers can be competitive while using far fewer parameters than large wav2vec-style models [4], [8]. The encoder outputs $\boldsymbol{H} \in R^{T\times d_{model}}$ are mean-pooled over time and passed to a classification head (LayerNorm → Dropout → Linear) to produce logits over the target command set. This mirrors common "encoder → global pooling → linear" heads in audio Transformers and PD-voice classification works using ViT/AST-style encoders on time–frequency inputs [15]. The network is trained end-to-end with cross-entropy over class labels. Linear layers use Xavier uniform initialization (as in _init_weights) to stabilize optimization in attention stacks, a standard good practice in recent small-scale ASR/Transformer implementations [8].

E. BiLSTM and Transformer

We classify short speech commands from MFCC sequences using a sequential BiLSTM→Transformer encoder. Given $X \in \mathbb{R}^{B\times T\times F}$with $F = n_{\mathrm{mfcc}}30$, the model accepts either $[B, T, F]$ or $[B, F, T]$; if needed, inputs are transposed inside forward to $[B, T, F]$. A 2-layer bidirectional LSTM (128 hidden per direction; inter-layer dropout 0.2 applied only when layers > 1) encodes local/phoneme-scale dynamics and outputs $H_{\mathrm{BiLSTM}} \in \mathbb{R}^{B\times T\times 256}$. A linear projection maps these features to the Transformer width $Z \in \mathbb{R}^{B\times T\times 128}$. Because self-attention is permutation-invariant, we add sinusoidal positional encodings to Z before attention so temporal order remains explicit for encoder, consistent with audio/sequence Transformer practice [4].

The sequence is refined by a Transformer encoder with two encoder layers (nn. TransformerEncoderLayer, batch_first=True), model width $d_{\mathrm{model}} = 128$, four heads of scaled dot-product attention, and a position-wise FFN of size 256 (ReLU). Each layer performs Query (Q), Key (K), and Value (V) projections, multi-head attention softmax $\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V$, where d_k is the dimension of the key vectors (used for scaling), residual connections, and LayerNorm around both attention and FFN; dropout 0.3 is applied inside attention and FFN sublayers. This follows standard, survey-summarized encoder designs for audio detection[4].

After the encoder aggregates global context, we apply temporal mean pooling over frames to get an utterance-level vector $\bar{h} \in \mathbb{R}^{B\times 128}$, and a lightweight head (LayerNorm → Linear) produces logits for the target classes.

This local-then-global design lets the BiLSTM capture ordered, short-range temporal structure, while the Transformer models long-range dependencies and cross-frame interactions an approach that has been shown effective by hybrid LSTM↔Transformer models on speech MFCCs (SER) and in broader time-series settings [16], [17], [18].

V. RESULTS

A. Overall Model Performance

Analysis of the grand mean accuracy Table 1, averaged across all seeds and sampling rates, reveals significant performance differences. The hybrid BiLSTM+Transformer (Hybrid) model achieved the highest accuracy at 0.975250 ≈ 97.53%. This performance substantially surpassed the pure BiLSTM architecture (0.958889, ≈ 95.89%) and the pure Transformer architecture (0.948146, ≈ 94.81%). The observed absolute accuracy improvements were 1.64 percentage points over BiLSTM and 2.71 percentage points over the Transformer. Furthermore, the BiLSTM outperformed the Transformer by 1.07 percentage points. These results indicate that the sequential integration of local temporal modeling (BiLSTM) with global context (Transformer encoder) yields consistent and significant accuracy improvements compared to isolated architecture.

Table 1. Overall Model Performance Comparison (Accuracy)

Model	Grand_MeanAcc	Weighted_MeanAcc
Hybrid	**0.9752503394**	**0.9752503394**
BiLSTM	0.9588890165	0.9588890165
Transformer	0.9481456727	0.9481456727

B. Sampling Rate Effect

Evaluation of the sampling rate effect Table 2 demonstrated that increasing from 8 kHz to 16 kHz provided consistent but marginal accuracy improvements across all models. The observed absolute delta (gain) was +0.07 percentage points (pp) for the BiLSTM+Transformer (Hybrid) (0.974900 vs. 0.975601), +0.21 pp for the BiLSTM (0.957860 vs. 0.959919), and +0.15 pp for the Transformer (0.947368 vs. 0.948923). This limited gain (<0.25 pp) suggests that the additional temporal resolution provides stable but limited benefits for this task, which involves only four simple voice commands and short utterance durations. The BiLSTM+Transformer model showed the smallest improvement (0.07 pp), suggesting the hybrid architecture is already well-optimized for lower-resolution inputs.

Table 2. Impact of Sampling Rate and Comparative Accuracy Gap

Model	Sample Rates		Accuracy Gap (vs Hybrid)	
	8kHz	16kHz	8kHz	16kHz
Hybrid	**0.974899**	**0.975600**	**0**	**0**
BiLSTM	0.957859	0.959918	-0.0170	-0.0156
Transformer	0.947368	0.948923	-0.0275	-0.0266

C. Parameter Complexity and Computational Efficiency

In Table 3, the terms of model complexity, BiLSTM baseline is significantly smaller (560,644 parameters), representing approximately 33% of the parameter count of the hybrid (1,687,303) and Transformer (1,687,303) models. Despite its smaller size, the BiLSTM exhibited higher computational cost during training, evidenced by lower throughput. Both the Transformer and hybrid models share identical parameter counts, yet the sequential fusion (BiLSTM → projection → Transformer) delivered superior performance. Regarding training efficiency, higher sampling rates (16 kHz) accelerated throughput (e.g., 356.21 samples/sec for hybrid @16kHz vs. 302.11 @8kHz), likely due to improved batch utilization and I/O pipeline efficiency. The BiLSTM+Transformer achieved an ≈ 2.2 × throughput advantage over the pure BiLSTM (356.21 vs. 183.21 samples/sec @16kHz), representing an optimal trade-off between the highest accuracy and practical computational efficiency during training.

Table 3. Model Complexity and Training Efficiency

Model	Condition	Parameter
Hybrid	All SR	1,687,303
LSTM	~	560,644
Transformer	~	1,687,303

D. Robustness to Noise

Model robustness was evaluated using 30% noise augmentation (white/pink, SNR 5-20 dB) across two noise variants. Table 4 per-fold data analysis revealed perfect symmetry between performance on pink and white noise; the resulting accuracy difference was ≈0.0 pp for all model and sampling rate combinations (e.g., BiLSTM+Transformer 16kHz: Pink = 97.56%, White = 97.56%). This finding indicates the augmentation scheme is unbiased toward specific noise characteristics. Consistently high accuracy (≥ 94.8%) across all noise variants confirms that MFCC features with log-power normalization effectively isolate perceptually-relevant acoustic content independent of noise type. The models successfully learn robust representations that generalize across diverse augmentation without preference for the pink or white spectrum.

Table 4. Noise Robustness Analysis (Accuracy under Augmentation)

Model	Noise	SR	Accuracy				Avg Times	Total Times
			Mean	Std	Min	Max		
bilstm	pink	8,000	95.79%	0.30%	95.11%	96.12%	88.52	26.556.80
	pink	16,000	95.99%	0.46%	95.34%	97.14%	66.56	19.966.73
	white	8,000	95.79%	0.30%	95.11%	96.12%	88.52	26.556.80
	white	16,000	95.99%	0.46%	95.34%	97.14%	66.56	19.966.73
transformer	pink	8,000	94.74%	0.34%	94.22%	95.30%	37.15	371.53
	pink	16,000	94.89%	0.54%	94.02%	96.29%	31.68	316.76
	white	8,000	94.74%	0.34%	94.22%	95.30%	37.15	371.53
	white	16,000	94.89%	0.54%	94.02%	96.29%	31.68	316.76
bilstm_transformer	pink	8,000	97.49%	0.19%	97.14%	97.90%	40.54	12.161.74
	pink	16,000	97.56%	0.19%	97.21%	97.83%	34.3	10.289.55
	white	8,000	97.49%	0.19%	97.14%	97.90%	40.54	12.161.74
	white	16,000	97.56%	0.19%	97.21%	97.83%	34.3	10.289.55

E. Cross Seed Stability

Training stability was measured using the standard deviation of accuracy across seed means Table 5. The Transformer architecture at 16 kHz exhibited the lowest cross-seed variability (sd ≈0.000122, stability_score=8167.53), indicating high training consistency across random initializations. However, its accuracy remained below that of the hybrid model. The BiLSTM+Transformer maintained the highest accuracy while demonstrating excellent cross-seed stability (sd $\approx 2.56 \times 10^{\{-4\}}$ at 8kHz and $\approx 3.59 \times 10^{\{-4\}}$ at 16kHz). This offers the optimal trade-off between high accuracy and generalization consistency.

Table 5. Training Stability Analysis (Cross-Seed Variance)

Model	SR	Stability Score	Sd_Seed Means
Hybrid	8000	**3905.671509**	**0.0002560379176**
	16000	**2784.223238**	**0.0003591666021**
LSTM	8000	677.6650569	0.001475655252
	16000	466.2354928	0.002144838854
Transformer	8000	1096.160545	0.000912275127
	16000	8167.534563	0.0001224359679

F. Statistical Significance

Statistical significance testing using Holm-Bonferroni correction (applied to both t-tests and Wilcoxon signed-rank tests) confirmed that in Table 6 all architectural accuracy differences are statistically significant ($p < 0.001$) for both sampling rates. For example, the corrected $p - value$ $(p_{\{holm\}})$ for the 8 kHz t-test between BiLSTM+Transformer and Transformer was $\approx 2.87 \times 10^{\{-24\}}$, and between BiLSTM+Transformer and BiLSTM was $\approx 2.28 \times 10^{\{-20\}}$. This concludes that the observed accuracy differences are not due to random chance and that the practical effects are consistent across sampling rates and random seeds.

G. Per-Class Performance and ROC-AUC

Per-class accuracy analysis demonstrated robust and balanced performance across all command classes, with accuracy values consistently ranging from 0.94-0.99 regardless of model architecture or sampling rate configuration. BiLSTM+Transformer achieved the highest per-class accuracy at 97.62% (±0.64%) at 16 kHz, while BiLSTM and Transformer variants maintained competitive performance at 95.99% (±0.94%) and 94.89% (±1.27%), respectively, with no individual class exhibiting systematic degradation. ROC-AUC analysis revealed exceptionally strong class separability across all conditions, with macro-ROC-AUC values consistently ≈ 0.9989-0.9992 for all model sampling rate combinations. As illustrated in Figure 2, the ROC-AUC heatmaps demonstrate uniformly high discrimination capability across all four command classes (Down, Left, Right, Up) and both sampling rates (8 kHz and 16 kHz), with values ranging from 0.9960-1.0000. This exceptional and uniform performance across all classes indicates that errors are primarily stochastic misclassifications arising from inherent acoustic variability under noisy conditions, rather than systematic failures on specific command classes, demonstrating the model's robust generalization capability.

Table 6. Statistical Significance Tests (Holm-Bonferroni)

SR	Model a	Model b	Test	P_Raw	P_Holm
8kHz	BiLSTM+Transformer	LSTM	ttest_p	1.14E-20	2.28E-20
	BiLSTM+Transformer	Transformer	ttest_p	9.57E-25	2.87E-24
	BiLSTM	Transformer	ttest_p	5.67E-18	5.67E-18
	BiLSTM+Transformer	LSTM	wilcoxon_p	1.72E-06	3.44E-06
	BiLSTM+Transformer	Transformer	wilcoxon_p	1.71E-06	5.13E-06
	BiLSTM	Transformer	wilcoxon_p	1.72E-06	1.72E-06
16kHz	BiLSTM+Transformer	LSTM	ttest_p	1.31E-15	2.61E-15
	BiLSTM+Transformer	Transformer	ttest_p	8.73E-21	2.62E-20
	BiLSTM	Transformer	ttest_p	2.09E-10	2.09E-10
	BiLSTM+Transformer	LSTM	wilcoxon_p	1.71E-06	5.13E-06
	BiLSTM+Transformer	Transformer	wilcoxon_p	1.72E-06	3.44E-06
	BiLSTM	Transformer	wilcoxon_p	3.47E-06	3.47E-06

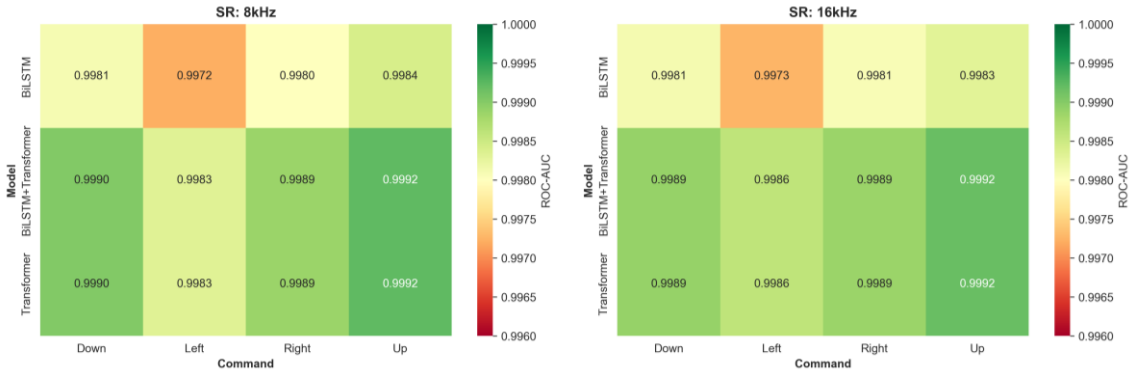

Figure 2. ROC-AUC Heatmaps

Figure 2, Heatmaps showing mean ROC-AUC values by Model × Command for both sampling rates (8 kHz and 16 kHz). The uniformly high and consistent values (predominantly 0.9980–0.9992) across all model–command combinations indicate exceptional class separability and balanced discrimination performance with minimal variance.

H. Real-Time Integration Performance

To validate the system's feasibility for interactive applications, we integrated the Hybrid model (16 kHz, Pink Noise) with a 3D Endless Runner game via a local TCP bridge. Performance metrics were logged over a continuous 5-minute gameplay session (N=50 commands). As shown in Table 7, the system achieved an average Total Response Time of 5.23 ms. The decoupled architecture (Python Inference + C# TCP Client) introduced negligible overhead, with an average transport latency of just 0.15 ms. These results confirm that the system operates well within the real-time constraints (<100 ms) required for fluid game control.

Table 7. Real-Time Performance Assessment (Endless Runner Integration)

Metric	Mean (ms)	Min (ms)	Max (ms)	Std Dev (ms)
Inference Time	5.08	3.51	9.31	1.12
Transport Latency	0.15	0.12	0.28	0.04
Total Response Time	5.23	3.68	9.52	1.15

I. Optimal Configuration and Recommendations

The strictly optimal configuration, defined by the highest Grand Mean Accuracy, is the BiLSTM+Transformer at 16 kHz, which achieved 0.9756 (97.56%). However, the BiLSTM+Transformer at 8 kHz (0.9749, 97.49%) performed practically identically, with a negligible difference of less than 0.07 percentage points. Given this minimal gap, we recommend the 8 kHz configuration for resource-constrained edge deployments (e.g., mobile devices) as it offers significant computational savings with virtually no loss in accuracy. The 16 kHz configuration is recommended only for scenarios where absolute peak accuracy is paramount and computational resources are unconstrained.

J. Summary of Key Findings

Key findings indicate a clear accuracy hierarchy: BiLSTM+Transformer dominates at 97.53%, followed by BiLSTM (95.89%) and Transformer (94.81%); all differences are statistically significant ($p < 0.001$ Holm-Bonferroni). Notably, the small-parameter BiLSTM (560.6K) outperformed the 3x-larger pure Transformer (1.69M, suggesting Transformers may be less data-efficient in this low-data, from-scratch training regime. The hybrid model also provided ≈2.2× faster training throughput than the BiLSTM (356 vs. 183 smp/sec at 16kHz). The sampling rate increase (8-16 kHz) yielded consistent but marginal improvements (+0.07–0.21pp). Finally, all models demonstrated high noise robustness (accuracy ≥ 94.8% under 30% augmentation) and excellent per-class performance (ROC-AUC≈0.998–0.999).

VI. DISCUSSION

This section analyzes the experimental results to validate our hypotheses, interpret key findings, and derive practical and theoretical implications. Our study demonstrates that integrating BiLSTM and Transformer architectures captures complementary temporal information, improves classification accuracy, and enhances model stability for speech command recognition[6], [16], [18], [19].

A. Empirical Validation of Architectural Complementarity

Our results highlight a critical finding regarding data efficiency. The BiLSTM baseline (560.6K parameters) achieved an accuracy of 95.89%, outperforming the pure Transformer baseline (94.81%) by 1.07 percentage points, despite the Transformer utilizing three times the parameter count (1.69M). This suggests that for this small-scale, from-scratch training task, the Transformer's self-attention mechanism may be less data-efficient than the recurrent inductive bias of the BiLSTM. These findings are consistent with research on small-scale ASRs, which note that large transformer architectures often require significantly more

parameters and training data to outperform traditional models, especially in a limited corpus [8].

The core hypothesis that these architectures are complementary is strongly supported by the superior performance of hybrid models. BiLSTM+Transformer hybrid model achieves 97.53% accuracy, (these results are comparable to other hybrid models, such as the 97.04% accuracy achieved by the Transformer-BiLSTM model of Mou et al.), an absolute increase of 1.64 percentage points (pp) over BiLSTM and 2.71 pp over Transformer. This shows that for a fixed large parameter budget (1.69M), augmenting a Transformer with a BiLSTM (hybrid model) is a more effective use of parameters than a pure Transformer architecture. This confirms findings from other ablation studies in which hybrid models outperformed stand-alone BiLSTM and Transformer models [18]. This integration is effective because the architecture captures complementary information: LSTM handles sequential dependencies while Transformer manages global contextual interactions [6]. The 2.2x training throughput advantage of the hybrid model over pure BiLSTM (356 vs 183 smp/sec at 16 kHz) further demonstrates that the attention-based component efficiently parallelizes computing during training.

B. Robustness and Noise Invariance

Noise augmentation robustness is exceptional: pink and white noise variants produce identical accuracy profiles ($\Delta \approx$ 0.0 pp difference across all architectures). This symmetry indicates that MFCCs with log-scale power spectra effectively isolate the perceptually relevant acoustic content, as this feature is designed to mimic the human auditory system [16]. Meanwhile, our augmentation scheme (30% probability, SNR 5–20 dB) provides uniform coverage without bias to specific noise characteristics. All models maintained an accuracy of >94% under augmentation, suggesting that robust feature learning is architecture-agnostic when proper input pre-processing is applied.

C. Sampling Rate and Signal-Processing Trade-offs

The marginal accuracy gain from 8-16 kHz (0.07-0.21 pp) reflects the simplified task structure: four distinct, monosyllabic commands. The BiLSTM+Transformer model exhibited the smallest accuracy gain (0.07 pp) from this change, suggesting its hybrid architecture is optimized for lower-resolution inputs, while the pure Transformer benefits slightly more (0.15 pp). Training throughput increases substantially at higher SR (17.9 - 32.9% improvement), likely due to improved I/O pipeline utilization and batch packing efficiency. For voice control applications, the 8 kHz setting offers a practical operating point, balancing minimal accuracy loss against a reduced resource footprint.

D. Parameter Efficiency and Deployment Implications

The BiLSTM's 560.6K parameters (approximately 33% of the hybrid size) illustrate the recurrent advantage for parameter-efficient models. However, its 2.2 × lower training throughput (183 vs. 356 smp/sec at 16 kHz) and lower accuracy (95.89% vs. 97.53%) suggest that pure LSTM optimization requires careful trade-off assessment. While the memory footprint is smaller, recurrent computation can be a bottleneck. Conversely, the hybrid and Transformer share parameter counts (1.69M) but differ in accuracy and robustness. The hybrid architecture demonstrates clear superiority for real-time deployment, achieving an average total response time of 5.23 ms (including 0.15 ms transport latency) in live game testing. This ultra-low latency, combined with high accuracy and parallelizable training, presents a compelling case for interactive applications. The optimal choice favors the hybrid model where both accuracy and response speed are critical.

E. Stability and Generalization

Cross-seed variance (standard deviation of seed means) is low for the BiLSTM+Transformer (sd $\approx 2.56 \times 10^{\{-4\}}$ at 8kHz). Higher variance in the pure Transformer at 8 kHz (sd $\approx 9.12 \times 10^{\{-4\}}$) suggests greater sensitivity to initialization, a known challenge for attention-based models on small datasets[8]. The stratified 5-fold cross-validation protocol, combined with repeated seeds (180 trained models total), yields highly reliable performance metrics [20].

F. Real-Time Feasibility in Interactive Gaming

The integration with a 3D Endless Runner game serves as a critical stress test for the system's real-time capabilities. In this genre, player reaction times are measured in milliseconds; a delay exceeding 100 ms can result in a "Game Over" due to ignored inputs. Our system's verified average total response time of 5.23 ms provides a safety margin of nearly an order of magnitude relative to this threshold, ensuring that voice commands feel instantaneous to the player. Furthermore, the 0.15 ms transport latency via TCP confirms that the decoupled architecture (Python inference engine communicating with a C# game client) introduces negligible overhead. This finding is significant for game developers, as it validates that complex hybrid deep learning models can be deployed in interactive loops without requiring monolithic integration into the game engine itself, preserving development modularity while maintaining ultra-low latency performance.

G. Limitations and Future Directions

This study evaluated voice command recognition on a small-scale, four-class task with fewer than 10,000 training samples using standardized MFCC features and controlled noise augmentation (pink and white noise, SNR 5–20 dB). While this controlled setting enables fair architectural comparison, the findings may not directly generalize to larger speech recognition tasks with extended vocabularies or more complex utterances. Real-time game control feasibility was verified on a standard PC; future work should validate performance on constrained edge devices (e.g., mobile phones) to confirm sub-50 ms latency in strictly low-power environments. The study employed a fixed MFCC front-end; alternative feature representations (e.g., log-mel spectrograms or self-supervised learned features) were not explored[4]. Finally, the dataset does not distinguish between speaker-dependent and speaker-independent scenarios, leaving open

the question of generalization to unseen speakers[20]. Future work should prioritize three practical directions: (1) Knowledge Distillation for Mobile Deployment to compress the hybrid model into a lightweight student network (50–70% parameter reduction) while maintaining 95%+ accuracy, (2) Transfer Learning from Pre-Trained Speech Datasets to fine-tune the hybrid architecture on large public speech command datasets (e.g., Google Speech Commands, 65K samples) before adapting to the four-class gaming task[8], leveraging external acoustic knowledge to improve robustness and reduce training variance[4], (3) Adaptive Augmentation and Speaker Generalization to extend robustness evaluation beyond fixed pink/white noise to include time-domain augmentations (pitch shifting, time stretching, random cropping) and evaluate performance across multiple speakers or implement speaker adaptation mechanisms using few-shot learning to enable multi-user gaming systems [4], [20]. These directions maintain practical feasibility while addressing real-world deployment constraints and improving generalization to diverse acoustic environments and user populations.

VII. CONCLUSION

This study provides strong empirical evidence that hybrid BiLSTM-Transformer architectures can outperform isolated recurrent and attention-based models for small-scale voice command recognition when trained from scratch. The hybrid model achieved a superior accuracy of 97.53%, significantly outperforming the BiLSTM (95.89%) and Transformer (94.81%) baselines. Beyond accuracy, the proposed architecture demonstrated exceptional quantitative advantages: (1) Robustness: It maintained consistent high performance (>97.5%) across both pink and white noise conditions, verifying unbiased feature learning; (2) Efficiency: It delivered a 2.2 × training throughput advantage over the BiLSTM baseline (356 vs. 183 samples/sec), optimizing the accuracy-efficiency trade-off; and (3) Real-Time Feasibility: Empirical integration with a 3D Endless Runner game confirmed an ultra-low total response time of 5.23 ms (consisting of 5.08 ms inference and 0.15 ms transport latency). These findings validate the hybrid architecture as a robust, data-efficient, and computationally practical solution for low-latency interactive applications, setting a reproducible baseline for future research in resource-constrained ASR.

REFERENCES

[1] D. M. Waqar, T. S. Gunawan, M. Kartiwi, and R. Ahmad, "Real-Time Voice-Controlled Game Interaction using Convolutional Neural Networks," in *2021 IEEE 7th International Conference on Smart Instrumentation, Measurement and Applications, ICSIMA 2021*, Institute of Electrical and Electronics Engineers Inc., Aug. 2021, pp. 76–81. doi: 10.1109/ICSIMA50015.2021.9526318.

[2] F. Adnan, I. Amelia, D. Sayyid ', and U. Shiddiq, "Implementasi Voice Recognition Berbasis Machine Learning," *Edu Elektrika Journal*, vol. 11, no. 1, 2022.

[3] X. Wang, P. Zhang, and C. Liu, "Acoustic signal-based identification of pipeline defects using optimized MFCC and LSTM," *Journal of Pipeline Science and Engineering*, p. 100355, Sep. 2025, doi: 10.1016/j.jpse.2025.100355.

[4] K. Zaman, K. Li, M. Sah, C. Direkoglu, S. Okada, and M. Unoki, "Transformers and audio detection tasks: An overview," Mar. 01, 2025, *Elsevier Inc.* doi: 10.1016/j.dsp.2024.104956.

[5] S. Ünalan, O. Günay, I. Akkurt, K. Gunoglu, and H. O. Tekin, "A comparative study on breast cancer classification with stratified shuffle split and K-fold cross validation via ensembled machine learning," *J Radiat Res Appl Sci*, vol. 17, no. 4, p. 101080, Dec. 2024, doi: 10.1016/j.jrras.2024.101080.

[6] S. Riaz, A. Saghir, M. J. Khan, H. Khan, H. S. Khan, and M. J. Khan, "TransLSTM: A hybrid LSTM-Transformer model for fine-grained suggestion mining," *Natural Language Processing Journal*, vol. 8, p. 100089, Sep. 2024, doi: 10.1016/j.nlp.2024.100089.

[7] A. A. Alsuwaylimi, "Arabic dialect identification in social media: A hybrid model with transformer models and BiLSTM," *Heliyon*, vol. 10, no. 17, Sep. 2024, doi: 10.1016/j.heliyon.2024.e36280.

[8] A. Loubser, P. De Villiers, and A. De Freitas, "End-to-end automated speech recognition using a character based small scale transformer architecture," *Expert Syst Appl*, vol. 252, Oct. 2024, doi: 10.1016/j.eswa.2024.124119.

[9] Y. Yan, S. O. Simons, L. van Bemmel, L. G. Reinders, F. M. E. Franssen, and V. Urovi, "Optimizing MFCC parameters for the automatic detection of respiratory diseases," *Applied Acoustics*, vol. 228, Jan. 2025, doi: 10.1016/j.apacoust.2024.110299.

[10] A. Sabha and A. Selwal, "A novel Approach for Audio-based Video Analysis via MFCC Features," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1512–1521. doi: 10.1016/j.procs.2024.04.142.

[11] K. Pizzi, M. Pizarro, and A. Fischer, "Comparative study on noise-augmented training and its effect on adversarial robustness in ASR systems," *Comput Speech Lang*, vol. 96, Feb. 2026, doi: 10.1016/j.csl.2025.101869.

[12] S. Tirronen, S. R. Kadiri, and P. Alku, "The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection," *Journal of Voice*, vol. 38, no. 5, pp. 975–982, Sep. 2024, doi: 10.1016/j.jvoice.2022.03.021.

[13] J. Tayebi, A. Rezaie, M. Rezaie, M. Hassanpour, and M. R. I. Faruque, "Bi-LSTM neural network for enhanced radiotherapy: Detection of tissue parameters," *Nuclear Engineering and Technology*, vol. 58, no. 1, p. 103906, Jan. 2026, doi: 10.1016/j.net.2025.103906.

[14] P. Calle *et al.*, "Integration of nested cross-validation, automated hyperparameter optimization, high-performance computing to reduce and quantify the

variance of test performance estimation of deep learning models," *Comput Methods Programs Biomed*, vol. 272, Dec. 2025, doi: 10.1016/j.cmpb.2025.109063.

[15] B. Perrone, F. Amato, and G. Olmo, "Voice classification in Parkinson's disease: A deep learning approach using transformers and error rate metrics," *Biomed Signal Process Control*, vol. 113, Mar. 2026, doi: 10.1016/j.bspc.2025.108954.

[16] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," *IEEE Access*, vol. 10, pp. 36018–36027, 2022, doi: 10.1109/ACCESS.2022.3163856.

[17] J. Zhao *et al.*, "Advancing calving time prediction using tail acceleration data: Comparative evaluation of transformer, hybrid CNN-transformer and LSTM-transformer models," *Smart Agricultural Technology*, vol. 12, p. 101531, Dec. 2025, doi: 10.1016/j.atech.2025.101531.

[18] H. Mou, H. Rong, and A. P. Teixeira, "Detecting abnormal ship trajectory to avoid bridge collisions via a Transformer-BiLSTM model," *Ocean Engineering*, vol. 343, Jan. 2026, doi: 10.1016/j.oceaneng.2025.123232.

[19] Z. Jamshidzadeh, M. Ehteram, and H. Shabanian, "Bidirectional Long Short-Term Memory (BILSTM) - Support Vector Machine: A new machine learning model for predicting water quality parameters," *Ain Shams Engineering Journal*, vol. 15, no. 3, Mar. 2024, doi: 10.1016/j.asej.2023.102510.

[20] T. Leinonen *et al.*, "Empirical investigation of multi-source cross-validation in clinical ECG classification," *Comput Biol Med*, vol. 183, Dec. 2024, doi: 10.1016/j.compbiomed.2024.109271.