Evaluasi Model Wavenet untuk Klasifikasi Perintah Suara Real-Time pada Kontrol Karakter Game

Nasyrun Adetiya[1], Oddy Virgantara Putra[2]
[12]Teknik Informatika, Universitas Darussalam Gontor
[12]Ponorogo, Jawa Timur - Indonesia
Email: [1]432022611073@stu.unida.gontor.ac.id, [2]oddy@unida.gontor.ac.id

Abstract

Traditional game control systems present accessibility barriers for players with motor impairments. This research evaluates the performance of the WaveNet deep learning architecture for real-time voice command classification in games. The WaveNetClassifier model was trained on a dataset of 1,200 audio samples consisting of four directional commands ("down", "up", "right", "left"). To improve generalization, the training data was augmented with background noise. Evaluation was conducted in three experimental runs (seeds 42, 123, 456). Results indicate the model achieves an average accuracy of 80.7% on clean data and 80.4% on noisy data. Statistical analysis shows high consistency with a standard deviation of only 3.66%. Latency testing resulted in an average inference time of 5.2 ms, demonstrating feasibility for real-time applications. However, a specific performance drop was observed in the "Left" command classification (Recall 51.9%) due to phonetic characteristics. Future work suggests expanding the dataset and implementing advanced augmentation techniques.

Keywords: *Deep Learning; Game Accessibility; Speech Recognition; Voice Control; WaveNet*

Abstraksi

Sistem kontrol game tradisional memiliki keterbatasan dalam aspek aksesibilitas, terutama bagi pemain dengan keterbatasan motorik. Penelitian ini mengevaluasi performa arsitektur deep learning WaveNet untuk klasifikasi perintah suara real-time dalam game. Model WaveNetClassifier dilatih menggunakan dataset 1.200 sampel perintah suara yang terdiri dari empat kelas arah: "down", "up", "right", dan "left". Untuk meningkatkan generalisasi, data latih diaugmentasi dengan background noise. Model dievaluasi dalam tiga eksperimen terpisah (seed 42, 123, dan 456). Hasil menunjukkan model mencapai rata-rata akurasi 80,7% pada data bersih dan 80,4% pada data bising. Analisis statistik menunjukkan konsistensi tinggi dengan standar deviasi hanya 3,66%. Pengujian latensi menghasilkan rata-rata waktu inferensi 5,2 ms, yang menunjukkan kelayakan untuk aplikasi real-time. Penurunan performa spesifik teridentifikasi pada klasifikasi perintah "Left" (Recall 51,9%) yang disebabkan oleh karakteristik fonetik. Penelitian selanjutnya disarankan untuk memperluas dataset dan menerapkan teknik augmentasi tingkat lanjut.

Kata Kunci: *Deep Learning; Aksesibilitas Game; Pengenalan Suara; Kendali Suara; WaveNet*

1. PENDAHULUAN

Perkembangan teknologi interaksi manusia-komputer telah membuka peluang baru dalam industri *gaming*. Meskipun sistem kontrol tradisional seperti *keyboard* dan *mouse* efektif untuk sebagian besar pengguna, sistem ini menciptakan hambatan bagi pemain dengan disabilitas motorik atau dalam skenario *Virtual Reality* (VR) yang membutuhkan interaksi *hands-free*. Oleh karena itu, pengembangan sistem kendali alternatif yang responsif menjadi kebutuhan mendesak.

Teknologi pengenalan suara (*speech recognition*) menawarkan solusi potensial untuk tantangan tersebut. Penelitian terkini oleh Zhang, dkk. [1] menegaskan bahwa *Deep Learning* telah menjadi metode standar dalam sistem pengenalan suara modern karena kemampuannya menangkap pola kompleks. Namun, implementasi dalam *game* menuntut latensi yang sangat rendah dan ketahanan terhadap gangguan suara latar (*noise*).

Sebagian besar penelitian sebelumnya menggunakan *Convolutional Neural Network* (CNN) yang memproses spektrogram (representasi visual suara). Meskipun metode ini mencapai akurasi tinggi seperti pada penelitian Waqar, dkk.[2] yang mencapai 96,5%, proses konversi audio ke spektrogram dapat menghilangkan detail temporal halus dan menambah latensi komputasi. Penelitian ini mengusulkan pendekatan berbeda dengan menggunakan arsitektur WaveNet. Berbeda dengan metode berbasis spektrogram, WaveNet memproses *raw audio waveform* (gelombang suara mentah) secara langsung menggunakan *dilated causal convolutions*. Pendekatan ini bertujuan menangkap fitur temporal yang lebih kaya untuk klasifikasi *real-time*.

Tujuan utama penelitian ini adalah mengevaluasi performa model WaveNet dalam mengklasifikasikan perintah suara navigasi (*up, down, left, right*) berdasarkan metrik akurasi, ketahanan terhadap *noise*, dan latensi pemrosesan. Evaluasi dilakukan untuk menentukan apakah arsitektur generatif ini layak digunakan sebagai kontrol alternatif yang efisien dalam *game*.

2. TINJAUAN PUSTAKA

2.1. Tinjauan Penelitian Terkait Kontrol Suara

Penggunaan suara sebagai kontrol game telah dieksplorasi dalam beberapa studi. Waqar, dkk. [2] mengembangkan sistem kendali suara untuk game *Snake* menggunakan CNN pada input MFCC (*Mel-Frequency Cepstral Coefficients*). Sementara itu, Vinayak, dkk. [3] mengimplementasikan kontrol suara untuk navigasi *maze* dengan fokus pada integrasi sistem. Studi komparatif oleh Li, dkk. [4] menunjukkan bahwa CNN ringan (*lightweight*) efektif untuk pengenalan perintah, namun masih bergantung pada pra-pemrosesan spektrogram. Penelitian ini melengkapi studi tersebut dengan mengevaluasi arsitektur yang memproses data mentah untuk meminimalkan latensi pra-pemrosesan.

2.2. Perbandingan Arsitektur Deep Learning

Pemilihan arsitektur model sangat krusial dalam pemrosesan suara. Tabel 1 menyajikan perbandingan karakteristik antara CNN, RNN, dan WaveNet dalam konteks klasifikasi audio.

Tabel 1. Perbandingan Arsitektur untuk Klasifikasi Audio

Fitur	CNN (Standard)	RNN (LSTM/GRU)	WaveNet (Dilated Conv)
Input Utama	Spektrogram (Citra 2D)	Sekuensial / Time-series	Raw Waveform / Sekuensial
Kelebihan	Efisien mengekstrak fitur spasial frekuensi	Baik menangkap konteks memori panjang	Menangkap dependensi panjang dengan resolusi temporal tinggi
Kekurangan	Kehilangan informasi fase/temporal halus	Training lambat (sulit diparalelkan)	Komputasi berat pada versi generatif asli
Relevansi Riset	Sering digunakan di penelitian terdahulu [2], [4]	Kurang efisien untuk real-time cepat	Potensial untuk presisi tinggi pada raw audio

2.3. Arsitektur WaveNet

WaveNet, yang diperkenalkan oleh Oord dkk. [5], menggunakan *dilated causal convolutions* untuk memperluas *receptive field* tanpa meningkatkan biaya komputasi secara eksponensial. Hal ini memungkinkan model untuk mempelajari pola frekuensi dan temporal langsung dari data mentah. Kemampuan ini dihipotesiskan dapat memberikan keunggulan dalam kecepatan inferensi untuk aplikasi *real-time*.

3. METODE PENELITIAN

Penelitian ini menggunakan metode eksperimental dengan tahapan pengumpulan dataset, pra-pemrosesan dan augmentasi, perancangan model, serta evaluasi performa.

3.1. Dataset

Dataset yang digunakan bersumber dari *Google Speech Commands Dataset* yang difokuskan pada empat kelas perintah navigasi inti, yaitu "down", "up", "right", dan "left". Dataset terdiri dari total 1.200 sampel audio dengan distribusi seimbang sebanyak 300 sampel untuk setiap kelas. Seluruh data memiliki format file .wav dengan spesifikasi *sample rate*16 kHz, *mono channel*, dan durasi rata-rata 1 detik per sampel. Untuk keperluan eksperimen, dataset dipartisi dengan proporsi 70% sebagai data pelatihan, 15% sebagai data validasi, dan 15% sebagai data pengujian.

3.2. Arsitektur Model

Model yang dibangun adalah WaveNetClassifier yang dimodifikasi untuk tugas klasifikasi. Model menerima input audio mentah (1 kanal) dan memprosesnya melalui 2 tumpukan (*stacks*) yang masing-masing terdiri dari 8 lapisan *dilated convolution*. *Residual*

channels dan *dilation channels* ditetapkan sebesar 32 untuk menyeimbangkan kompleksitas dan performa.

3.3. Prosedur Pelatihan

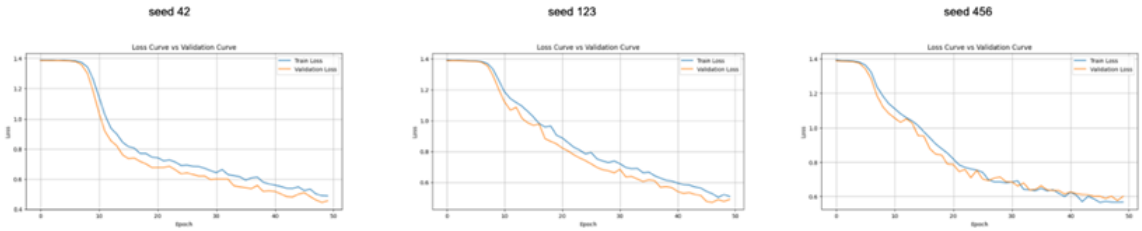
Model dilatih menggunakan *optimizer* Adam dengan *learning rate* 0.0001 selama 50 *epoch*. Untuk meningkatkan generalisasi dan ketahanan terhadap gangguan lingkungan, diterapkan dua strategi utama. Pertama, penggunaan *Dropout*dengan tingkat sebesar 0.3 pada lapisan akhir untuk mencegah *overfitting*. Kedua, penerapan Augmentasi *Noise* di mana data latih dicampur dengan *background noise* secara acak pada rasio *Signal-to-Noise* (SNR) antara 5dB hingga 20dB.

4. HASIL DAN PEMBAHASAN

Evaluasi dilakukan dalam tiga kali percobaan (*run*) dengan inisialisasi *seed* acak yang berbeda (42, 123, 456) untuk memastikan konsistensi hasil.

4.1. Analisis Loss dan Konvergensi

Selama proses pelatihan, diamati bahwa nilai *Validation Loss* secara konsisten lebih rendah dibandingkan *Training Loss*. Fenomena ini terjadi akibat penggunaan lapisan *dropout* yang aktif saat fase pelatihan (mematikan neuron secara acak untuk mempersulit pembelajaran), namun non-aktif saat fase validasi (menggunakan kapasitas penuh model).



Gambar 1. Kurva Loss vs Validasi (Seed 42, 123, 456)

Seperti terlihat pada Gambar 1, semua *run* mencapai konvergensi yang sangat baik ("Excellent") sebelum *epoch*maksimum, dengan rata-rata pengurangan *loss* sebesar 64,57%.

4.2. Performa Klasifikasi dan Analisis Statistik

Tabel 2 menunjukkan ringkasan performa model berdasarkan metrik akurasi, presisi, *recall*, dan F1-Score.

Tabel 2. Analisis Statistik Performa (Rata – rata 3 Seed)

Metrik	Mean	Std Dev	Min	Max	Range
Accuracy	8.074	366	7.556	8.389	833
Precision	8.119	416	7.543	8.518	975
Recall	8.074	366	7.556	8.389	833
F1-Score	8.001	340	7.512	8.251	739

Hasil pada Tabel 2 menunjukkan stabilitas model yang cukup baik. Standar deviasi yang rendah (3,66% untuk akurasi) mengindikasikan bahwa performa model konsisten meskipun dilatih ulang dengan inisialisasi (*seed*) yang berbeda. Model juga menunjukkan ketahanan (*robustness*) yang tinggi terhadap *noise*. Akurasi rata-rata pada data bersih adalah 80,74%, sedangkan pada data bising adalah 80,39%. Penurunan performa yang sangat kecil (kurang dari 0,4%) membuktikan efektivitas strategi augmentasi data.

4.3. Analisis Kesalahan (Error Analysis)

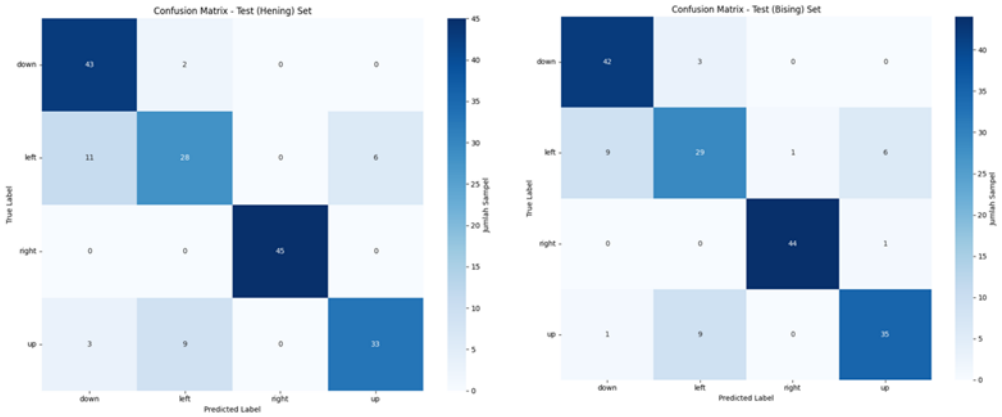
Analisis lebih mendalam dilakukan pada performa setiap kelas perintah, seperti disajikan pada Tabel 3.

Tabel 3. Performa per Kelas (Rata – rata)

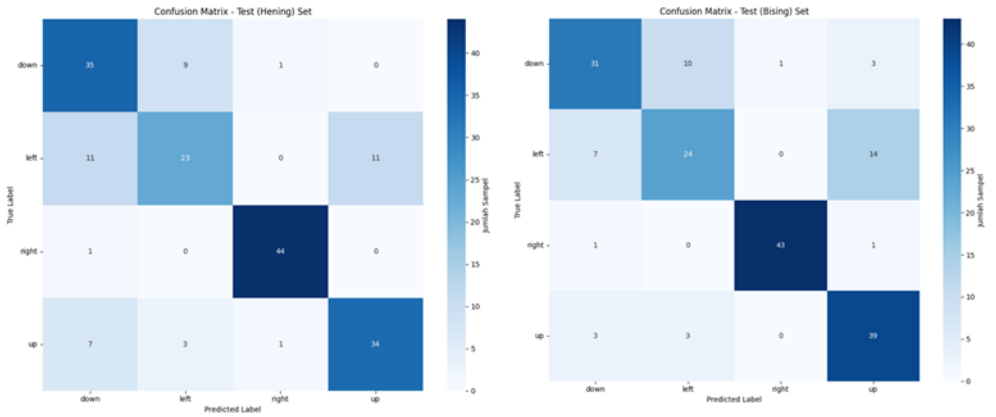
Voice Command	Precision	Recall	F1-Score	Kualitas Klasifikasi
Down	754	956	843	Baik
Left	835	519	639	Cukup
Right	993	1.000	996	Excellent
Up	788	852	819	Baik

Berdasarkan Tabel 3 dan *confusion matrix* (Gambar 2), terlihat ketimpangan performa. Perintah "Right" dikenali dengan sempurna (*Recall* 1.0), namun perintah "Left" memiliki *Recall* terendah (0,519). Artinya, hampir setengah dari perintah "Left" gagal dikenali atau salah diklasifikasikan sebagai "Down" atau "Up".

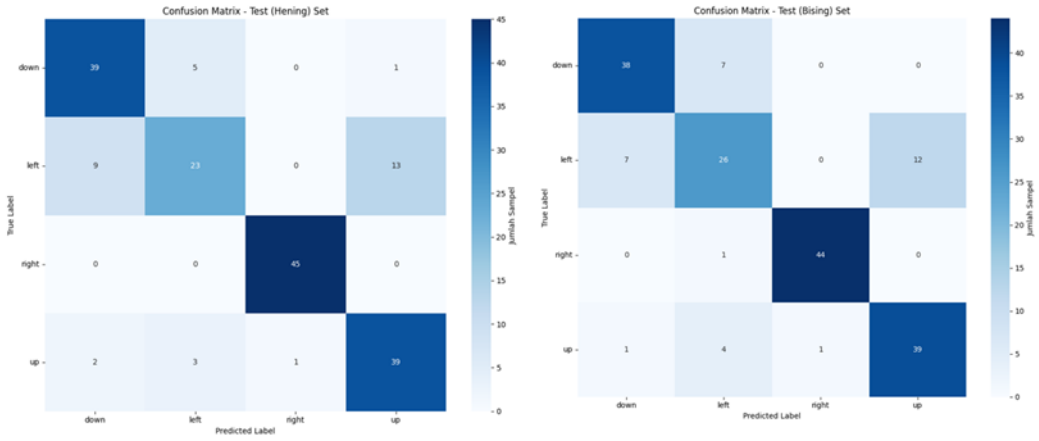
Hal ini mendukung hipotesis bahwa fitur fonetik pendek pada kata "Left" (/lɛft/)—yang diakhiri konsonan lemah—lebih rentan tertutup oleh *noise* atau mirip dengan fitur pendek kata lain dalam representasi *waveform* mentah, dibandingkan dengan diftong panjang dan kuat pada kata "Right" (/raɪt/).



Gambar 2. Confusion Matrix seed 42 (Hening dan Bising)



Gambar 3. Confusion Matrix seed 123 (Hening dan Bising)



Gambar 4. Confusion Matrix seed 456 (Hening dan Bising)

4.4. Evaluasi Latensi dan Real-Time Factor

Untuk memvalidasi kelayakan penggunaan dalam *game*, dilakukan pengujian latensi inferensi (Tabel 4).

Tabel 4. Hasil Pengujian Latensi dan RTF (Rata – Rata)

Seed	Latensi Rata - Rata	Letensi Maksimum	RTF Rata - Rata
42	5.19 ms	18.15 ms	0.0052
123	5.15 ms	18.21 ms	0.0051
456	5.26 ms	17.95 ms	0.0053

Nilai *Real-Time Factor* (RTF) rata-rata 0,0052 menunjukkan bahwa sistem dapat memproses 1 detik audio hanya dalam waktu sekitar 5,2 milidetik. Kecepatan ini sangat krusial untuk *game* berjenis aksi yang membutuhkan respons instan dan jauh di bawah batas toleransi latensi *game* pada umumnya.

4.5. Perbandingan dengan Metode Lain

Tabel 5 membandingkan hasil penelitian ini dengan metode *state-of-the-art* (SOTA) lainnya.

Tabel 5. Perbandingan dengan Metode Lain

Metode	Input	Akurasi Dilaporkan	Fokus Utama
CNN (Waqar et al. [2])	Spektrogram	96,5%	Game Sederhana
Broadcast Residual (Kim et al. [6])	Spektrogram	~94%	Efisiensi Keyword Spotting
WaveNet (Usulan)	Raw Waveform	80,7%	Robustness & Latensi

Meskipun akurasi WaveNet sedikit di bawah metode berbasis spektrogram, model ini menawarkan keunggulan signifikan dalam efisiensi pemrosesan data mentah untuk perangkat *edge*.

5. KESIMPULAN

Penelitian ini berhasil mengevaluasi penggunaan model WaveNet untuk kontrol game. Kesimpulan utama adalah Model WaveNet mampu memproses perintah suara *real-time* dengan latensi sangat rendah (5,2 ms), memenuhi syarat responsivitas game. Model menunjukkan ketahanan yang baik terhadap gangguan suara, dengan performa stabil di lingkungan bising (akurasi 80,4%). Keterbatasan penelitian ini terletak pada lingkup dataset yang hanya mencakup 4 perintah suara inti dan evaluasi yang masih dilakukan secara terisolasi, belum diuji coba dalam skenario permainan (*gameplay*) nyata yang kompleks. Selain itu, teridentifikasi kesulitan model dalam mengklasifikasikan kata berdurasi fonetik pendek seperti "Left" (*Recall* 51,9%).

Rekomendasi untuk penelitian selanjutnya adalah (1) memperluas dataset dengan variasi aksen dan jumlah perintah yang lebih banyak, (2) menerapkan teknik augmentasi spesifik seperti *time stretching* untuk memperbaiki akurasi kata pendek, dan (3) mengimplementasikan sistem ini ke dalam *game engine* (seperti Unity/Unreal) untuk pengujian pengalaman pengguna (*User Experience*) secara langsung.

DAFTAR PUSTAKA

[1] A. Li et al., “Learning Neural Vocoder from Range-Null Space Decomposition,” 2025. [Online]. Available: https://github.com/Andong-Li-speech/RNDVoC.

[2] 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.

[3] K. Vinayak, “Jatin Sharma 3rd , Shruti Gupta 4th 1st,2nd,3rd B. Tech Scholar Poornima Institute of Engineering & Technology,” 2024. [Online]. Available: www.ijnrd.org

[4] A. Bilal, S. Latif, S. A. Ghauri, O. Y. Song, A. A. Abbasi, and T. Karamat, “Modified Heuristic Computational Techniques for the Resource Optimization in Cognitive Radio Networks (CRNs),” Electronics (Switzerland), vol. 12, no. 4, Feb. 2023, doi: 10.3390/electronics12040973.

[5] A. van den Oord et al., “WaveNet: A Generative Model for Raw Audio,” Sep. 2016, [Online]. Available: http://arxiv.org/abs/1609.03499

[6] B. Kim, S. Chang, J. Lee, and D. Sung, “Broadcasted Residual Learning for Efficient Keyword Spotting,” Jul. 2023, [Online]. Available: http://arxiv.org/abs/2106.04140

[7] N. Zargham, et al., "Let's Talk Games: An Expert Exploration of Speech Interaction with NPCs," International Journal of Human-Computer Interaction, vol. 41, no. 5, pp. 3592-3612, 2025. doi: 10.1080/10447318.2024.2338666.

[8] J. Aguirre-Peralta, M. Rivas-Zavala, and W. Ugarte, "Speech to Text Recognition for Videogame Controlling with Convolutional Neural Networks," in International Conference on Pattern Recognition Applications and Methods, vol. 1, 2023, pp. 948–955. doi: 10.5220/0011782900003411.

[9] A. Copaceanu, J. Weinel, and S. Cunningham, "Using Voice Input to Control and Interact With a Narrative Video Game," BCS Learning & Development, 2024. doi: 10.14236/ewic/eva2024.66.

[10] M. R. Hasanabadi, "An Overview of Text-to-Speech Systems and Media Applications," Technical Review NO 294, pp. 7-13, 2023.

[11] J. Shen, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing: proceedings, 2018.

[12] S. Ma, D. Mcduff, and Y. Song, "NEURAL TTS STYLIZATION WITH ADVERSARIAL AND COLLABORATIVE GAMES," Tech. rep., 2019.