A Noise-Resilient Voice Command System for Smart Wheelchairs Using Gammatone Frequency Cepstral Coefficients and ResNet50

Fitri Utaminingrum
Department of Informatics Engineering, Brawijaya University, Indonesia
f3_ningrum@ub.ac.id (corresponding author)

Aulia Riza Mufita
Department of Informatics Engineering, Brawijaya University, Indonesia
auliariza@student.ub.ac.id

Aldiansyah Satrio Kabisat
Department of Informatics Engineering, Brawijaya University, Indonesia
aldisk@student.ub.ac.id

I Komang Somawirata
Department of Electrical Engineering, National Institute of Technology, Indonesia
kmgsomawirata@lecturer.itn.ac.id

Received: 16 November 2025 | Revised: 29 December 2025, 19 January 2026, and 7 February 2026 | Accepted: 8 February 2026

 DOI: https://doi.org/10.48084/etasr.16301

ABSTRACT

This study introduces a voice-activated smart wheelchair system engineered to support individuals with physical disabilities, especially in noisy environments. The proposed system employs GFCC for noise-resistant feature extraction and utilizes a ResNet50 deep learning architecture for command classification, implemented on an NVIDIA Jetson TX2 embedded platform. The model is designed to accurately identify Indonesian vocal commands related to wheelchair movement directions. Experimental evaluation encompasses epoch-wise performance analysis, confusion matrix evaluation, computational time measurement, and comprehensive testing within real-world environments under both calm and noisy conditions. The best model was found at epoch 72, when the validation accuracy was 94.6%, the validation loss was 0.221, and the macro-averaged precision, recall, and F1-score values were 0.955, 0.957, and 0.956, respectively. The average GFCC extraction and inference durations are 0.089 and 0.578 s, respectively, culminating in a total system latency of 0.667 s, thereby meeting real-time control specifications. Integrated testing shows that the proposed system works 88% of the time in quiet settings and 73.33% of the time in noisy ones. These findings demonstrate that the proposed GFCC–ResNet50 framework exhibits robust noise resistance and dependable real-time performance, rendering it appropriate for practical assistive mobility applications.

Keywords-smart wheelchair; voice command recognition; GFCC; ResNet50; Jetson TX2; noisy environments

I. INTRODUCTION

Individuals with disabilities consistently encounter substantial limitations in mobility and independence, which affect their ability to fully participate in daily activities. In 2022, around 1.3 billion people, or 16% of the world's population, were identified as having a disability, demonstrating the extent of the problem [1]. Individuals with physical disabilities, particularly those with limited hand or limb mobility, face significant challenges when utilizing traditional wheelchairs that rely exclusively on manual propulsion. This condition emphasizes the need to quickly develop alternative mobility solutions that enable autonomous navigation for individuals with limited upper limb mobility.

Recent developments in assistive technology have explored voice-controlled systems as a promising approach to improving wheelchair accessibility. In [2], a voice-based command classification system was developed for opening and closing doors using Mel Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Networks (CNN). Similar MFCC-based voice command systems have also been implemented for

navigation control in embedded robotic platforms using artificial neural networks [3]. However, the system in [3] exhibited reduced efficacy under chaotic conditions, thus significantly limiting its applicability in real-world environments. This limitation points out the importance of noise-robust speech processing methods in assistive mobility systems. Consequently, the effectiveness of various feature extraction methods, including MFCC and its alternatives, has become a central focus of research in the development of resilient speech recognition systems [4]. Subsequent research has shown that Gammatone Frequency Cepstral Coefficients (GFCC) are effective in capturing detailed spectral and temporal characteristics of speech signals and exhibit improved robustness in noisy environments compared to conventional MFCC features [5-6].

At the same time, advances in deep learning have shown that deeper convolutional architectures, such as ResNet50, achieve improved classification accuracy due to their residual learning mechanisms, which effectively mitigate gradient degradation issues in deep networks [7-8]. The NVIDIA Jetson TX2 platform is widely recognized as a suitable embedded computing solution for real-time applications in robotics and autonomous systems, due to its optimal combination of computational performance and energy efficiency [9]. These developments suggest that combining noise-resistant feature extraction with deep residual learning on an embedded platform presents significant opportunities for real-time assistive applications.

Several research studies have further illustrated the importance of noise-robust speech recognition for navigation and assistive systems. In [10], deep convolutional models augmented with noise-enhanced data were shown to achieve improved recognition accuracy in complex acoustic environments. In [5], it was stated that GFCC improves speech recognition performance in noisy environments. Furthermore, in [11], it was verified that NVIDIA Jetson-enabled peripheral AI systems can deliver effective real-time performance for speech-controlled navigation interfaces. This research confirmed the efficacy of individual components. However, the primary focus of contemporary studies continues to be on discrete improvements in feature extraction, classification, or hardware implementation.

Despite these advances, a clear gap remains. Few research studies have demonstrated a fully integrated voice-controlled wheelchair system that simultaneously incorporates noise-robust GFCC feature extraction, deep residual learning with ResNet50, and real-time embedded deployment on a compact edge platform, particularly for Indonesian language voice commands. In addition, existing research rarely evaluates system performance in real-world outdoor environments where wheelchair users operate predominantly. This deficiency in thorough integration and contextual evaluation limits the practical utility of prior solutions.

The primary innovation of this research lies in the design of a fully integrated voice-controlled wheelchair system that incorporates GFCC-based feature extraction, deep residual learning through a ResNet50 architecture, and real-time embedded deployment on an NVIDIA Jetson TX2 platform for Indonesian language voice commands under both silent and authentic noisy outdoor environments, offering tangible evidence of its robustness and practical applicability. By incorporating noise-resistant speech representation, deep residual learning, and real-time edge computing into a cohesive framework, this research presents a practical and scalable assistive mobility solution designed to improve the autonomy, safety, and independence of individuals with physical disabilities.

II. PROPOSED METHOD

A. System Architecture and Edge Deployment

The proposed approach combines GFCC with ResNet50 classification to develop a reliable voice-controlled smart wheelchair system capable of functioning efficiently in noisy environments. Figure 1 offers a comprehensive overview of the system workflow, depicting the architecture of the proposed smart wheelchair in a system design block diagram.

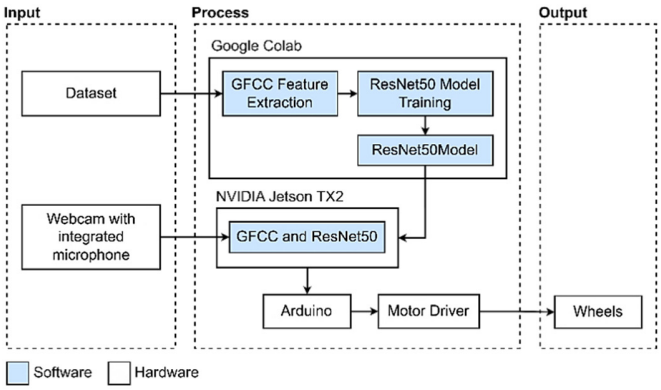

Fig. 1. Diagram of the proposed system.

In this architecture, the dataset is initially processed in Google Colab, where GFCC is conducted to derive noise-robust spectral representations. Subsequently, these features are utilized to train the ResNet50 model, resulting in a streamlined classifier appropriate for deployment on edge hardware. The trained model is then deployed on the NVIDIA Jetson TX2, which processes real-time audio input from an integrated webcam microphone. This device was chosen for its optimal combination of computational performance, energy efficiency, and appropriateness for deep learning-based audio processing on edge devices. Table I provides a summary of the hardware specifications employed.

TABLE I. HARDWARE SPECIFICATIONS OF NVIDIA JETSON TX2 USED FOR EDGE INFERENCE

Component	Specification
CPU	Dual-core NVIDIA Denver 2 + Quad-core ARM Cortex-A57
GPU	NVIDIA Pascal GPU, 256 CUDA cores
Software stack	JetPack 4.6 (CUDA 10.2, cuDNN, TensorRT)
RAM	8 GB LPDDR4, 128 bit
Storage	32 GB eMMC 5.1
Peak performance	1.3 TFLOPS (FP16)

During operation, the NVIDIA Jetson TX2 executes GFCC extraction and ResNet50 inference to identify the user command, subsequently transmitting control signals to the Arduino. The Arduino communicates with the motor driver to control the wheelchair wheels according to the voice command received. This modular architecture ensures effective segregation between cloud-based model training and edge-based inference, while ensuring dependable real-time command execution.

B. Dataset Preparation

The dataset employed in this research comprises recorded Indonesian voice commands gathered in an authentic outdoor setting. A total of 5,413 audio recordings were employed, comprising 1,266 samples of "Maju" (Forward), 1,076 samples of "Mundur" (Backward), 1,028 samples of "Kanan" (Right), 1,136 samples of "Kiri" (Left), and 907 samples of "Berhenti" (Stop). All recordings were archived in .WAV format at a sampling rate of 16 kHz, with an approximate duration of 1 s per sample.

The audio data was recorded using the integrated microphone of a Logitech C270 webcam. Data collection was carried out in an outdoor roadside pedestrian zone, where ambient noise emerged from actual traffic and surrounding environmental sounds (approximately 70 dB). Although sound pressure levels were not directly measured, the recordings reflect natural and unregulated noise conditions. The dataset was compiled from multiple individuals, whereas the evaluation was performed with three subjects to determine practical usability. The dataset was partitioned into training, validation, and testing subsets with proportions of 70%, 20%, and 10%, respectively.

C. Gammatone-Frequency Cepstral Coefficients (GFCC)

GFCC was selected as the primary feature extraction technique due to its enhanced robustness to noise from a signal processing perspective compared to MFCC [6]. Inspired by the human auditory system, GFCC leverages cochlear filtering properties and utilizes a Gammatone filterbank instead of the Mel filterbank employed in MFCC [12-13]. Although MFCC encodes frequency perception through the Mel scale, GFCC employs the Equivalent Rectangular Bandwidth (ERB) scale, offering a more physiologically precise representation of human auditory frequency resolution, especially in the low and high frequency ranges [12]. This ERB-based frequency spacing allows GFCC to detect more detailed spectral features and enhances its robustness in noisy environments relative to MFCC [13]. The GFCC extraction procedure comprises multiple stages [14]. Figure 2 depicts the GFCC architecture employed in this research.

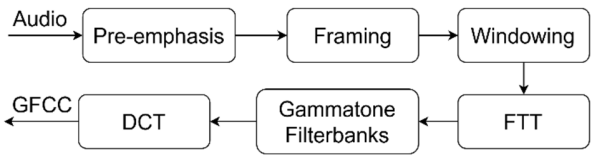

Fig. 2. Architecture of the GFCC feature extraction process.

- Pre-emphasis: High-frequency components are enhanced through a pre-emphasis filter to improve the signal-to-noise ratio.
- Framing and Windowing: The speech signal is segmented into overlapping frames, with each frame multiplied by a Hamming window to reduce spectral leakage.
- Fast Fourier Transform (FFT): The time-domain signal is transformed into the frequency domain.
- Gammatone Filterbank: The frequency spectrum is analyzed using a bank of Gammatone filters, whose impulse responses replicate the band-pass filtering characteristics of the human cochlea. These filters are arranged according to the ERB scale, which fundamentally differs from the triangular Mel filterbank used in MFCC and enables more precise auditory modeling [13].
- Non-linear Compression and Discrete Cosine Transform (DCT): The output of the Gammatone filterbank is subjected to non-linear compression, typically utilizing cubic root compression rather than the logarithmic compression conventionally employed in MFCC [15]. The compressed features are subsequently transformed via DCT to generate the final cepstral coefficients.

To further demonstrate the attributes of the extracted features, Figure 3 provides an example visualization of the GFCC representation.

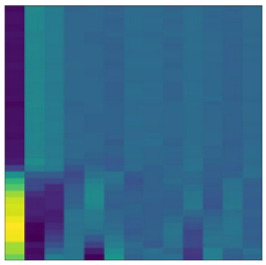

Fig. 3. Example of a GFCC spectrogram.

The figure displays a GFCC spectrogram, or cepstral feature map, derived from a single spoken Indonesian voice command. This visualization illustrates the time-frequency architecture generated by the Gammatone filterbank and ERB-based frequency scaling, demonstrating how prominent speech features are maintained while noise-related fluctuations are minimized. The specific parameters employed for GFCC extraction in this research are outlined in Table II.

TABLE II. GFCC FEATURE EXTRACTION PARAMETERS

Parameter	Value
Sampling rate	16 kHz
Pre-emphasis coefficient	0.97
Frame length	30 ms
Frame shift	15 ms
Window type	Hamming
FFT size	2048
Number of Gammatone filters	128
Frequency range	0 to 8000 Hz

D. ResNet50

The classification stage employs the ResNet50 architecture, a 50-layer deep CNN that incorporates residual connections to address the vanishing gradient and degradation issues typically encountered in deep networks. Figure 4 illustrates the structural overview of the ResNet50 network employed in this research. ResNet50 was selected due to its proven effectiveness in deep feature learning and classification tasks, outperforming conventional CNN architectures such as VGG and AlexNet through residual learning mechanisms [7-8]. The network is trained utilizing GFCC feature maps as input, with each frame annotated in accordance with the respective command class.

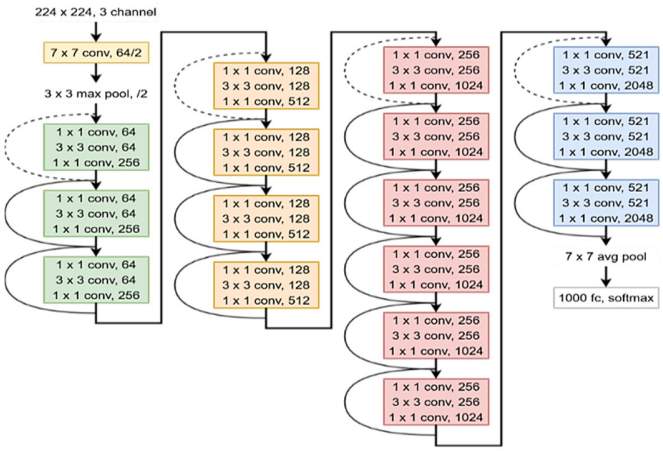

Fig. 4. Block diagram of the ResNet50 architecture.

Before being fed into the ResNet50 network, the GFCC spectrograms are treated as image-like feature maps and resized to a fixed spatial resolution of 224×224 pixels. To satisfy the three-channel input requirement of ResNet50, the single-channel GFCC spectrogram is replicated across three channels, resulting in an input tensor of size 224×224×3.

The ResNet50 model is trained utilizing GFCC feature maps as input, with each sample annotated according to the respective voice command class. The training configuration and hyperparameters are meticulously chosen to optimize both classification accuracy and computational efficiency. Table II presents a comprehensive overview of the training hyperparameters employed in this research.

TABLE III. RESNET50 TRAINING HYPERPARAMETERS

Hyperparameter	Value
Optimizer	Adam
Learning rate	0.0001
Batch size	16
Number of epochs	75
Loss function	Categorical cross entropy
Data augmentation	White noise injection
Noise amplitude factor	0.02

III. RESULTS AND DISCUSSION

A. Impact of Epoch Variation on Model Performance

The initial experiment sought to determine the ideal number of training epochs that produces the most favorable model performance with respect to accuracy, loss, precision, recall, and F1-score. Table IV encapsulates the quantitative findings of this experiment, showcasing the five most effective epochs. The table demonstrates that an increase in the number of epochs typically enhances classification performance, evidenced by elevated accuracy values and decreased loss values during both the training and validation phases. This trend indicates that the model acquires proficiency in distinguishing vocal characteristics as training progresses.

TABLE IV. FIVE BEST EPOCHS BASED ON F1-SCORE FROM THE RESNET50 MODEL

Epoch	Training		Validation		Macro	
	Acc.	Loss	Acc.	Loss	Avg Prec.	Avg Rec.
72	**0.871**	**0.368**	**0.946**	**0.221**	**0.955**	**0.957**
63	0.869	0.378	0.936	0.246	0.955	0.955
70	0.870	0.372	0.936	0.229	0.952	0.953
69	0.868	0.370	0.942	0.229	0.949	0.952
59	0.857	0.399	0.932	0.264	0.949	0.951

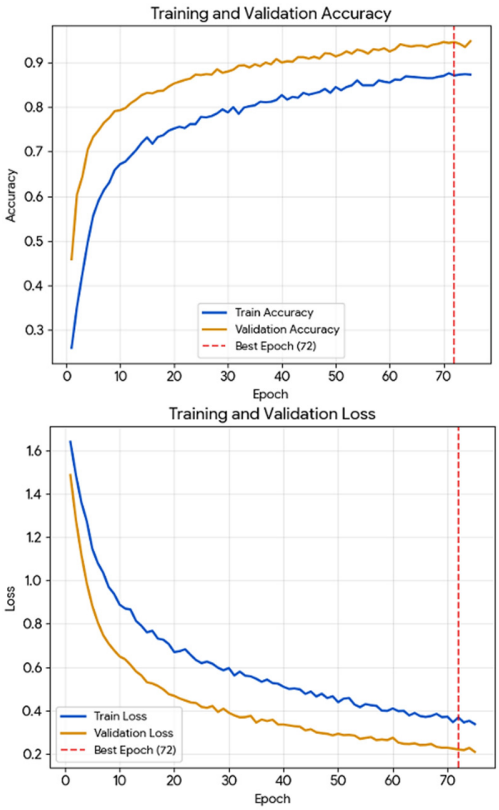

Fig. 5. Training and validation accuracy and loss curves.

The model trained for 72 epochs achieved the highest overall performance. At this epoch, the validation accuracy was 94.6% with a validation loss of 0.221, whereas the training accuracy reached 87.1%. The macro-averaged precision and recall values were 0.955 and 0.957, respectively, signifying a well-balanced classification performance across all command classes. Figure 5 shows the training and validation accuracy and loss curves to enhance the selection of the optimal epoch and to examine the model's learning behavior. The accuracy curves exhibit a consistent rise for both training and validation

sets until epoch 72, with a negligible disparity between them. This behavior indicates that the model effectively generalizes to novel data at this phase of training. The loss curves demonstrate a steady decline for both training and validation losses until they attain their minimum values around epoch 72. Past this juncture, the validation loss exhibits minimal variations, whereas the training loss persists in a slight decline. This divergence indicates the onset of overfitting, a phenomenon where further training does not improve generalization performance and may lead to model degradation.

Epoch 72 was identified as the optimal training configuration based on the numerical results in Table IV and the learning trends depicted in Figure 5. This epoch offers an advantageous equilibrium between increased classification accuracy and reduced risk of overfitting. The findings validate that the ResNet50 model can achieve effective feature generalization when trained with GFCC features, aligning with previous research on the efficacy of residual learning architectures for speech classification tasks.

B. Confusion Matrix Analysis

The confusion matrix assesses the classification accuracy of the optimal model (epoch 72) when evaluated on 543 novel voice commands. Figure 6 depicts the confusion matrix outcomes of the ResNet50 model, demonstrating substantial classification consistency among all five classes. The matrix indicates that most samples were accurately classified, with minimal misclassifications occurring between acoustically analogous commands ("Maju" vs. "Mundur"). The performance metrics for each class, presented in Table V, indicate that the model achieved a macro-average precision of 0.955, a recall of 0.957, and an F1-score of 0.956.

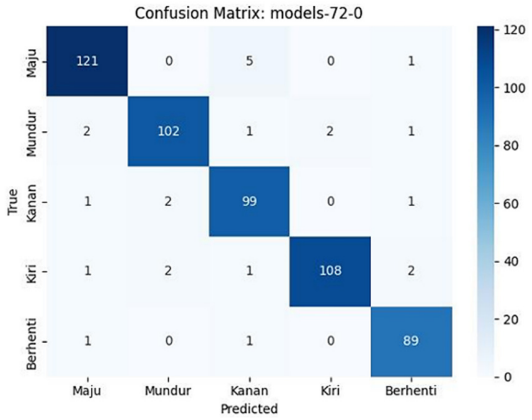

Fig. 6. Confusion matrix of the ResNet50 model (Epoch 72).

TABLE V. PRECISION, RECALL, AND F1-SCORE RESULTS

Class	Precision	Recall	F1-Score
Maju (Forward)	0.95	0.96	0.96
Mundur (Backward)	0.94	0.96	0.95
Kanan (Right)	0.96	0.93	0.94
Kiri (Left)	0.95	0.98	0.96
Berhenti (Stop)	0.95	0.95	0.95
Average	**0.95**	**0.96**	**0.95**

The findings validate that the integration of GFCC and ResNet50 facilitates precise voice command recognition, efficiently distinguishing analogous acoustic features due to GFCC's noise-resistant spectral representation and ResNet50's deep residual learning architecture.

C. Computation Duration for GFCC Extraction and Model Prediction

The average processing time for each voice command was assessed on the Jetson TX2 to validate real-time functionality on embedded hardware. The measurement encompassed the duration of GFCC feature extraction and ResNet50 inference. Table VI displays the computation duration for each command, while Table VII illustrates the average processing analysis. The average latency of 0.667 s indicates that the proposed system meets real-time response criteria for assistive mobility control. The GPU acceleration of the Jetson TX2 facilitated stable processing within these temporal constraints.

TABLE VI. AVERAGE COMPUTATION TIME OF GFCC EXTRACTION AND RESNET50 PREDICTION

Command	GFCC Extraction (s)	Prediction (s)
Maju (Forward)	0.092	0.599
Mundur (Backward)	0.084	0.590
Kanan (Right)	0.086	0.545
Kiri (Left)	0.095	0.554
Berhenti (Stop)	0.087	0.600
Average	**0.089**	**0.578**

TABLE VII. ANALYSIS OF AVERAGE COMPUTATION TIME FOR GFCC EXTRACTION AND RESNET50 PREDICTION

Process Type	Average Time (s)	Observation
GFCC Extraction	0.089	Efficient (<0.1s)
ResNet50 Prediction	0.578	Near real-time inference
Total Average Time	**0.667**	**Real-time response**

D. Integrated Accuracy of the ResNet50 Model under Quiet and Noisy Environments

The conclusive experiment assessed the comprehensive system performance when implemented in authentic environments. Fifteen subjects executed the five commands under both quiet and noisy (±70 dB) conditions. To provide a consolidated view while maintaining the integrity of the data, the results from the quiet and noisy tests were merged into Table VIII without any numerical modification.

TABLE VIII. INTEGRATED ACCURACY OF RESNET50 MODEL IN QUIET AND NOISY ENVIRONMENTS

Env.	Maju (Forward)	Mundur (Backward)	Kanan (Right)	Kiri (Left)	Berhenti (Stop)	Avg. Acc.
Quiet Avg.	86.67	60.00	100.00	93.33	100.00	88.00
Noisy Avg.	66.67	46.67	100.00	73.33	80.00	73.33

The consolidated data indicate that the system attained an average accuracy of 88% in tranquil settings and 73.33% in auditory-challenged environments. The "Kanan" (Right) command consistently achieved 100% accuracy in both conditions, demonstrating robust acoustic differentiation. Performance declines for "Mundur" (Backward) and "Kiri" (Left) in noisy environments can be ascribed to spectral overlap

and masking effects caused by background noise. Despite this, the system maintained robust overall accuracy, making it suitable for practical wheelchair operation.

IV. CONCLUSION

This research successfully designed a voice command system for a smart wheelchair that can recognize Indonesian commands. The NVIDIA Jetson TX2 runs the system, which uses GFCC for feature extraction and ResNet50 for classification. The integration of GFCC and ResNet50 exhibited robust noise resistance, high precision, and real-time capabilities appropriate for assistive mobility.

The optimal training configuration was achieved at epoch 72, resulting in a validation accuracy of 94.6%, a loss of 0.221, and balanced precision, recall, and F1-score of approximately 0.956. Additional epochs resulted in overfitting, affirming that this point represented the optimal balance between learning complexity and generalization. The confusion matrix analysis corroborated these results, with all command classes attaining F1-scores exceeding 0.94 and exhibiting minimal confusion between acoustically similar commands such as "Maju" (Forward) and "Mundur" (Backward). This demonstrates that GFCC effectively retained spectral features, while ResNet50 preserved robust feature discriminability.

The system also exhibited real-time functionality, with an average GFCC extraction time of 0.089 s, an inference time of 0.578 s, and a total latency of 0.667 s per command, which satisfied the criteria for responsive wheelchair control on the Jetson TX2 platform. When tested in real-world settings, the model attained an accuracy of 88% in quiet conditions and 73.33% in noisy environments, demonstrating strong adaptability despite some decline in performance for the Backward and Left commands caused by noise interference. The Right command achieved 100% accuracy in both environments, validating the model's robustness against interference.

In conclusion, the combination of GFCC and ResNet50 provides a noise-resistant and computationally efficient approach to speech-controlled wheelchair operation. The equilibrium among accuracy, inference speed, and environmental robustness achieved underscores the readiness of the system for practical deployment. Future research will investigate adaptive noise suppression and multimodal control to further improve reliability, safety, and accessibility for users with physical disabilities.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Penelitian Guru Besar 2025 program, organized by Direktorat Riset dan Pengabdian kepada Masyarakat (DRPM), Brawijaya University, for its financial support and facilitation throughout this research. This research was conducted under contract number 08470/UN10.F1501/B/PT/2025.

REFERENCES

[1] *Global Report on Health Equity for Persons with Disabilities*, 1st ed. World Health Organization, 2022.

[2] B. S. P. Laksono, T. Syaifuddin, and F. Utaminingrum, "Voice Recognition to Classify 'Buka' and 'Tutup' Sound to Open and Closes Door Using Mel Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Network (CNN)," *Journal of Information Technology and Computer Science*, vol. 9, no. 1, pp. 58–66, Apr. 2024, https://doi.org/10.25126/jitecs.202491579.

[3] M. Z. Abbiyansyah and F. Utaminingrum, "Voice Recognition on Humanoid Robot Darwin OP Using Mel Frequency Cepstrum Coefficients (MFCC) Feature and Artificial Neural Networks (ANN) Method," in *2022 2nd International Conference on Information Technology and Education (ICIT&E)*, Jan. 2022, pp. 251–256, https://doi.org/10.1109/ICITE54466.2022.9759883.

[4] A. A. Alasadi, T. H. Aldhayni, R. R. Deshmukh, A. H. Alahmadi, and A. S. Alshebami, "Efficient Feature Extraction Algorithms to Develop an Arabic Speech Recognition System," *Engineering, Technology & Applied Science Research*, vol. 10, no. 2, pp. 5547–5553, Apr. 2020, https://doi.org/10.48084/etasr.3465.

[5] P. Bawa, V. Kadyan, and G. Chhabra, "A Multifaceted Feature Extraction Approach for Noise-Robust Punjabi Spoken Digit Recognition System Under Low-Resource Conditions," in *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Mar. 2024, pp. 1–6, https://doi.org/10.1109/ICRITO61523.2024.10522268.

[6] N. Boualoulou, T. Belhoussine Drissi, and B. Nsiri, "Comparison of Feature Extraction Methods Between MFCC, BFCC, and GFCC with SVM Classifier for Parkinson's Disease Diagnosis," in *IoT Based Control Networks and Intelligent Systems*, 2024, pp. 231–247, https://doi.org/10.1007/978-981-99-6586-1_16.

[7] L. Borawar and R. Kaur, "ResNet: Solving Vanishing Gradient in Deep Networks," in *Proceedings of International Conference on Recent Trends in Computing*, 2023, pp. 235–247, https://doi.org/10.1007/978-981-19-8825-7_21.

[8] P. Nagpal, S. A. Bhinge, and A. Shitole, "A Comparative Analysis of ResNet Architectures," in *2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, Dec. 2022, pp. 1–8, https://doi.org/10.1109/SMARTGENCON56628.2022.10083966.

[9] B. Bangennavar, S. Patil, S. Kudal, B. D. Parmeshachari, and R. Latti, "People Tracking and Counting using Jetson TX2 Kit with Tracking Algorithm," in *2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon)*, Nov. 2022, pp. 1–5, https://doi.org/10.1109/NKCon56289.2022.10126790.

[10] N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool, "Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Recognition," in *Interspeech 2016*, Sept. 2016, pp. 2982–2986, https://doi.org/10.21437/Interspeech.2016-805.

[11] S. Gondi and V. Pratap, "Performance Evaluation of Offline Speech Recognition on Edge Devices," *Electronics*, vol. 10, no. 21, Nov. 2021, Art. no. 2697, https://doi.org/10.3390/electronics10212697.

[12] N. M. Sharma, V. Kumar, P. K. Mahapatra, and V. Gandhi, "Comparative analysis of various feature extraction techniques for classification of speech disfluencies," *Speech Communication*, vol. 150, pp. 23–31, May 2023, https://doi.org/10.1016/j.specom.2023.04.003.

[13] Z. Mengxi and T. Zhiguo, "Research on Failure Identification of Partial Discharge Ultrasonic Signal Based on GFCC," in *2020 IEEE Electrical Insulation Conference (EIC)*, June 2020, pp. 412–416, https://doi.org/10.1109/EIC47619.2020.9158683.

[14] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, Jan. 2020, Art. no. 107020, https://doi.org/10.1016/j.apacoust.2019.107020.

[15] M. A. Islam, "Non-linear Power Exponent Effect in GFCC for Bangla and Malay speech Separation," in *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, Feb. 2022, pp. 206–211, https://doi.org/10.1109/ICISET54810.2022.9775917.