MALAYSIAN JOURNAL OF INNOVATION IN ENGINEERING AND APPLIED SOCIAL SCIENCE (MyJIEAS)
e-ISSN : 2811-3780
VOLUME 04, ISSUE 01

Deep Learning-Based Speech Recognition of Malay Dialect Instructions using Edge Impulse

S.A Muhamed[1], S.N Makhtar[2,3]

[1]Department of Electrical and Electronics Engineering, Politeknik Sultan Salahuddin Abdul Aziz Shah, Shah Alam, Selangor, Malaysia
[2]Department of Electrical and Electronics Engineering, National Defence University of Malaysia, Kuala Lumpur, Malaysia
[3]Cyber Security and Digital Industrial Revolution Centre, National Defence University of Malaysia, Kuala Lumpur, Malaysia

Corresponding Author's Email: sitianizahmuhamed@gmail.com

Article History: Received 04092024; Revised 02102024; Accepted 30102024;

ABSTRACT – The increasing demand for voice-controlled systems in Malaysia for assistive technology, smart home appliances, and navigation purposes is typically served with an English-only solution. This reduces the accessibility of such technologies within a non-native English-speaking community. Given Malaysia's rich linguistic diversity, this study focuses on dialect speech recognition, specifically targeting various Malay dialects. This research uses the Edge Impulse platform to present a deep learning-based methodology for recognising speech in Malay dialects. The presented system collects and pre-processes audio data using a smartphone, while the deep learning algorithms perform efficient training and classification. Our model showed that robust and dialect-inclusive Malay voice recognition is possible, with an accuracy of around 80% for the four classes of instructions tested in four different dialects of Kelantan, Terengganu, Kedah and Standard Malay. This study therefore provides more foundation towards accessibility and usability of creating more inclusive speech recognition systems which can be tuned to regional linguistic variabilities.

KEYWORDS: Deep Learning, Speech Recognition, Malay Dialects, Edge Impulse

1.0 INTRODUCTION

Interest in systems operating in languages other than English is growing with the rise of voice-operated applications and devices. Most speech recognition technologies today are developed to take the English language as the main input language [1]. This puts limitations on the access of the system to users in multilingual regions like Malaysia. In Malaysia, for example, where the majority are not native English speakers, there is an urgent need to have speech recognition systems that take instructions in Bahasa Melayu, the national language. This would upgrade their usability and accessibility, especially in mundane usages from simple smart home controls to assistive devices that are used by the elderly and also persons with disabilities.

However, creating a speech recognition system for Bahasa Melayu poses special challenges. The Malaysian Malay dialect varies widely, and each of the variations possesses distinct phonetic, lexical, and syntactic features [2], [3]. While the standard Malay language, also well known as Bahasa Melayu Baku, is the national and formal medium of communication, in Peninsular Malaysia alone, many other dialects like Kedah in the North, Kelantan and Terengganu of East Coast, and Johorean Malay in the South have equal usage in everyday life. These dialects reflect the cultural and geographical diversity within Malaysia and may differ significantly in pronunciation, vocabulary, and intonation. Even the same instructions can be uttered in many different ways, depending on the regional background of the speaker. For example, the word "buka" which means open, if pronounced in the northern dialect is "bʊ kak"; " bʊ koʊ " in Kelantan and " bʊ kə" for the South. Furthermore, the instruction may even take a different form depending on the regional background of the speaker, hence building a uniformly operating speech recognition system is a difficult task.

Machine learning, such as deep learning has great potential to overcome such complexity [4], [5], [6], [7], [8]. While rule-based systems intrinsically depend on predefined linguistic rules, machine learning models can be trained on diverse datasets; hence, enabling them to learn the pattern and details of languages independently. By generally training machine learning models across variations, classifiers can be trained on speech data associated with many different

dialects, recognising the commands with high accuracy regardless of dialect. Due to this adaptability, for any Malay dialect spoken by those giving instructions, voice-controlled systems will be able to respond dependably, hence increasing accessibility and inclusivity for Malaysian users.

This study explored the Edge Impulse platform in designing a speech recognition system using a machine learning-based approach that would classify utterances from the major dialects of Malay. By using deep learning models trained on audio data in various dialectical pronunciations, this work looks forward to presenting a system that can handle the linguistic diversity of Bahasa Melayu. The results will therefore contribute to the basis on which further applications may be founded later, particularly in assistive tools developed to promote independent living among elderly and disabled persons who might benefit from voice-controlled technologies in their native language.

2.0 DIALECT-INPUT SPEECH RECOGNITION USING MACHINE LEARNING (ML)

In recent years, speech recognition technology has started to be part of modern devices and systems, enabling speech-driven operation and thereby facilitating human-computer interaction. This section gives a simplified review of speech recognition and the application of machine learning in handling the challenges caused by non-standard language such as dialect input recognition.

2.1 Automatic Speech Recognition (ASR)

While it is easy for a human to recognise words and sentences, it's a huge task for a computer to understand what the speaker is trying to communicate. This has been a problem for a very long time, where over time different solutions have been suggested, but realistic and accurate ones have been found only in recent years through ASR [9].

Speech recognition is an interdisciplinary subfield of computer science and computational linguistics that designs methodologies and technologies that allow computers to recognise audio-spoken language into text. It is more commonly known as Automatic Speech Recognition (ASR), computer speech recognition, or speech-to-text. It draws upon knowledge and research in computer science, linguistics, and computer engineering. Speech recognition is often mistaken for voice recognition; however, speech recognition is concerned with converting spoken words into text, while voice recognition is focused on identifying an individual user's voice. Early speech technology was focused on a minimal vocabulary, yet today it is used in almost any sector that involves constant interaction, such as automotive and technology. The applications have grown rapidly in recent years due to the development of deep learning and big data [10], [11]

Generally, today there are two main approaches to ASR: (a) traditional handcrafted pipelines, which use separate modules for preprocessing, postprocessing, and speech recognition, and (b) end-to-end (E2E) deep learning or neural network-based approaches as shown in Figure 1 [12]. The traditional approach divides speech recognition into subproblems, each solved by specific algorithms, where improving accuracy and performance requires adjusting each component. However, this method is complex and often reaches a local rather than global optimum. On the other hand, contemporary methods utilise multi-layered deep neural networks that can process inputs and outputs with varying lengths. Even though the E2E methods require much larger datasets, these newer models eliminate the need for complex pipelines.

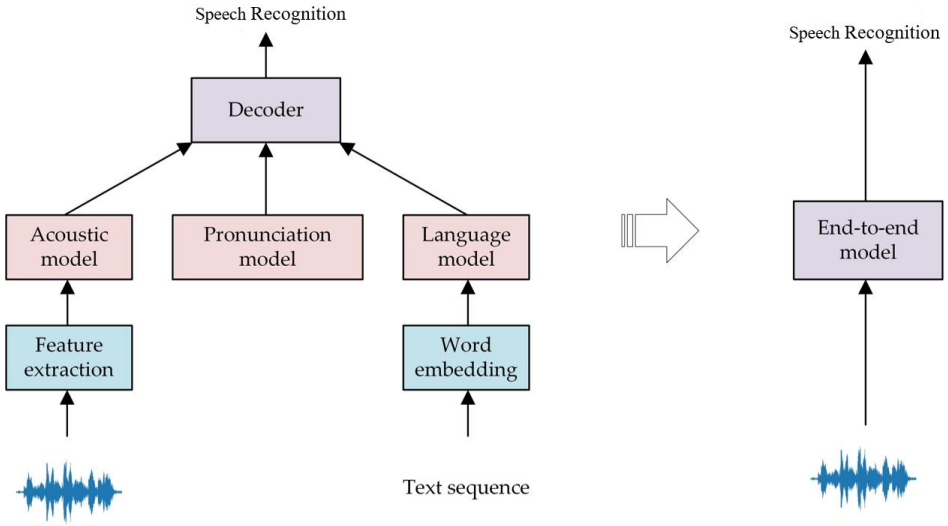


Figure 1. Comparison between traditional and E2E speech recognition [12]

2.2 Speech Recognition with Machine Learning (ML)

ML is a subset of artificial intelligence where computers learn from example inputs to attain a pattern. The ML model will, through various algorithms, learn patterns, predict, and thereby improve from experience to effectively mimic certain aspects of human learning. It uses various data types and has applications in image processing, natural language processing, health, finance, and recommendation systems. The most important recent development within ML is deep learning or DL—a subfield of ML inspired by the structure and function of the human brain [13]. Deep learning uses Artificial Neural Networks, with deep neural networks running several layers capable of modelling complex, abstract patterns in data. Its layers hierarchically process data to turn these DL models into learning sophisticated features—from simple edges in images to higher-order language structures.

Speech recognition is one of the fields that benefitted the most from the advancement of ML and DL in handling the complexities of human speech [14], [15], [16]. Traditional rule-based approaches were relatively limited in adapting to various accents, speech speeds, and background noise conditions. However, ML algorithms, especially deep neural networks, that can handle large and varied speech datasets, allowing systems to identify spoken language patterns with high accuracy despite factors of variation and noise. These have led to the development of speech recognition systems that are not only highly accurate but also capable of adapting to different languages, dialects, and accents.

2.4 Dialect-Specific Speech Recognition

In the domain of linguistic technology, considerable attention is given to developing speech recognition systems with dialectical options. This is because dialects allow for phonetic, lexical, and even syntactic aspects that differ from standard languages. In regions with many dialects, the average speech recognition system using trained language data finds it difficult to interpret data from dialects. This barrier has led researchers to design speech recognition models in order to fit and conform to dialects thereby increasing accessibility and usability for real-life applications**.**

One prominent area of research focuses on Arabic dialects, which differ significantly from Modern Standard Arabic and vary widely across regions [5]. Projects such as the MGB Challenge in Egypt [17] and the Multi Arabic Dialect Applications and Resources (MADAR) [18] have produced datasets and models specifically tailored to Egyptian, Levantine, Gulf, and North African Arabic dialects. Using deep learning techniques, these studies have shown that dialectal data enhances recognition accuracy and adaptability, underlining the importance of training models on language-specific variations.

In China, dialectal variation in Mandarin and Cantonese also poses challenges for speech recognition. Studies in [19], [20], [21] and [22] have focused on collecting regional dialect data, applying machine learning models to recognise dialect-specific pronunciations, and incorporating this into mainstream Mandarin-based systems. These studies highlight that integrating dialectal data with standard language models improves user experience and accuracy, especially in large-scale applications like automated customer service and smart assistants.

For English, accents and regional dialects, particularly from countries like the UK, Australia, and India, have also been the subject of recent research. Techniques such as transfer learning have proven effective, enabling speech models to generalise across accent variations by fine-tuning on smaller dialect-specific datasets. The use of transfer learning and fine-tuning in models like Google's Wav2Vec [23] and OpenAI's Whisper [24] has led to improved recognition of English dialects and accents, highlighting the efficacy of machine learning in accommodating linguistic diversity.

3.0 METHODOLOGY

Edge Impulse is a pioneering platform for developing machine learning models tailored for edge devices, which are computing devices capable of processing data locally, such as smartphones, microcontrollers, and Internet of Things (IoT) sensors. Designed to facilitate the deployment of ML models on low-power, resource-constrained devices, Edge Impulse enables developers to create, train, and deploy machine learning models that operate efficiently in real-time, close to where data is generated. This capability is particularly valuable in applications requiring immediate responses, such as real-time monitoring, environmental sensing, and health diagnostics [25].

In this study, Edge Impulse's deep learning capabilities are utilized to develop a speech recognition model that accurately interprets instructions across multiple Malay dialects. Audio data were collected from smartphone recordings, preprocessed, and used to train a model that can generalize across dialectal variations. By utilizing Edge Impulse, this study demonstrates the platform's potential to enhance accessibility and inclusivity in voice recognition applications.

3.1 Data Collection and Preprocessing

The Malay dialect speech commands were collected using the digital microphones of smartphones, allowing for a natural and accessible data capture process. The data set includes four classes of instructions: "Depan" (Front), "Belakang" (Back), "Berhenti" (Stop), and "Unknown"; which serves as a category for non-command sounds or irrelevant speech. The speech commands were recorded across four dialects representing dialects that are prominently different from each other: Kelantan, Kedah, Terengganu, and Standard Bahasa Melayu, as summarised in Table 1. Each dialect provided a different number of recordings, with a total of 1,533 samples distributed across the four instruction classes. To ensure data quality, the recordings were manually cleaned. This involved removing irrelevant sounds, handling background noise, and ensuring each recording contained only the intended command. This manual cleaning step was crucial for improving the accuracy of the subsequent machine learning (ML) model by providing it with a high-quality, well-labelled data set.

Table 1. Speech recording data

Class of Data	Number of Data	Dialect (n)			
		Kelantan	Kedah	Terengganu	Bahasa Melayu
Depan	336	61	80	80	115
Belakang	437	88	97	102	150
Berhenti	303	48	52	75	128
Unknown	457	-	-	-	-
Total	1533	197	229	257	393

The example of raw data utilised in this research is shown in Figure 2. This data represents the results obtained after recording participants' speech. Each recording session lasts for 10 seconds, during which participants are allowed to repeat the target word as many times as possible within the given time. As shown in Figure 3, one participant repeated the target word eight times within the 10-second recording period. After collecting the raw data, it undergoes a splitting process to extract the specific segments required for analysis. From the segmented data, Edge Impulse selects the instances of the target word that match the assigned label. This segmentation facilitates easier access to individual data points, enabling the extraction of speech samples on a per-second basis if needed. Additionally, this preparation process simplifies the training and testing phases for machine learning models, enhancing their efficiency and performance.

Figure 2. Raw data recording participant saying “belakang” as many times as possible with normal speaking speed within 10 seconds

Figure 3. Raw data were split into individual samples of the word

3.3 Data Processing and Model Development

The cleaned data set underwent standard machine learning preprocessing and classification steps on the Edge Impulse platform.

Building the Dataset

The dataset was balanced to ensure even representation across the command classes and dialects. This was followed by an 80-20 split, where 80% of the data was used for training and 20% for testing. The total duration of the entire data collection process can be calculated after

organizing the data by labels. Figure 4 illustrates the total duration of data collected and labelled for the word "depan"

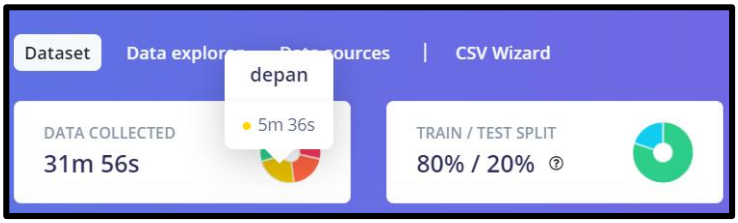


Figure 4. Total data collected for the word "depan".

Impulse Design

The design of the ML model, termed "Impulse" on the Edge Impulse platform, was configured to process time-series audio data. The impulse design involved adding an Audio MFCC (Mel Frequency Cepstral Coefficients) block to extract relevant features from the recordings, followed by a Classification ANN (Artificial Neural Network) for command classification. Neural networks are algorithms that are loosely based on the human brain and can recognize patterns in training data. The features extracted by using MFCC were used as input for the network training. Figure 5 shows the MFCC block and ANN block as illustrated in Edge Impulse.

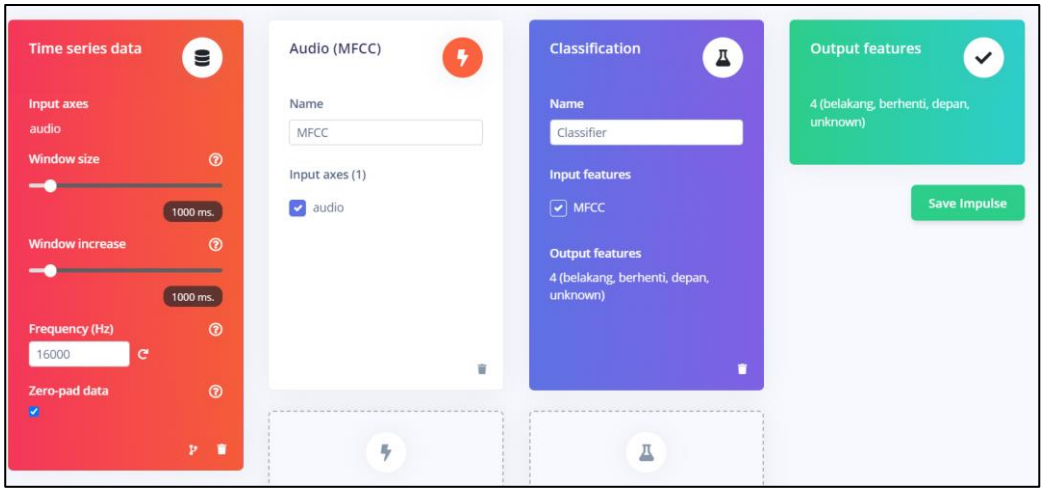


Figure 5. The setting of MFCC and Neural Network blocks in Edge Impulse

Data Training and Testing

The training process used the collected dataset to train the ANN model. This phase helped assess the model's performance on the provided data and ensured that it learned the correct patterns associated with each command and dialect. The internal states of the neurons are progressively adjusted and refined during training, enabling the network to transform its inputs in the exact manner required to produce the desired output. This process involves feeding a sample of training data into the network, evaluating its output for accuracy, and modifying the neurons' internal states to enhance the probability of delivering the correct output in subsequent attempts. The trained model was then tested on the 20% of data that was withheld from training. Testing on unseen data ensures that the model performs robustly across new inputs.

4.0 RESULTS AND DISCUSSION

There are several popular ways to evaluate the performance of the ability to map the input to the right class such as Accuracy, Precision, Recall, F1 score, Area under Curve, Confusion Matrix, and Mean Square Error.

A confusion matrix is a performance evaluation tool used to assess the accuracy of a classification model by comparing predicted results to actual outcomes. It is a table that summarizes the model's predictions into four categories: True Positives (TP), where the model correctly predicts the positive class; True Negatives (TN), where it correctly predicts the negative class; False Positives (FP), where it incorrectly predicts a positive result for a negative case (Type I error); and False Negatives (FN), where it fails to predict a positive result for a true positive case (Type II error).

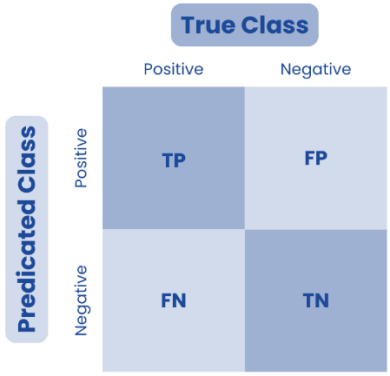


Figure 6. The basic structure of a Confusion Matrix

In this study, accuracy is used to measure the percentage of audio windows that were correctly classified. It is given as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Predictions } (\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

For example, if the model makes 100 predictions and gets 90 correct, the accuracy is 90%. Accuracy is straightforward but may not be reliable if the data is imbalanced

In the training phase, the model achieved an accuracy of 81.2% with a loss of 0.63. The class-wise performance, as seen in the confusion matrix, reveals insights into the model's ability to distinguish between different commands across dialects (Figure 7)

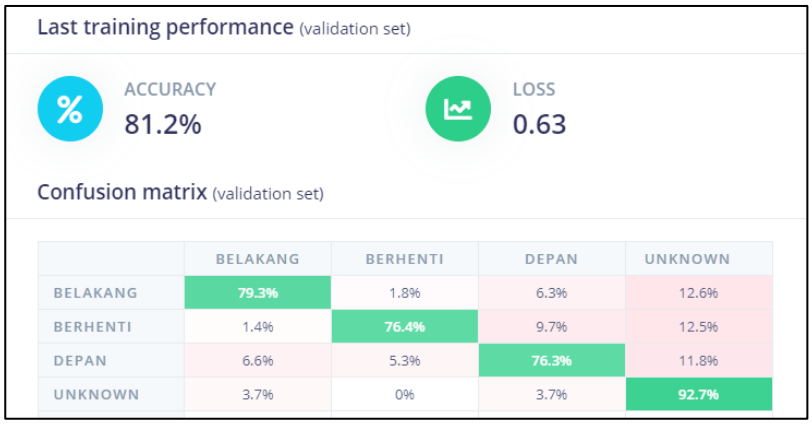
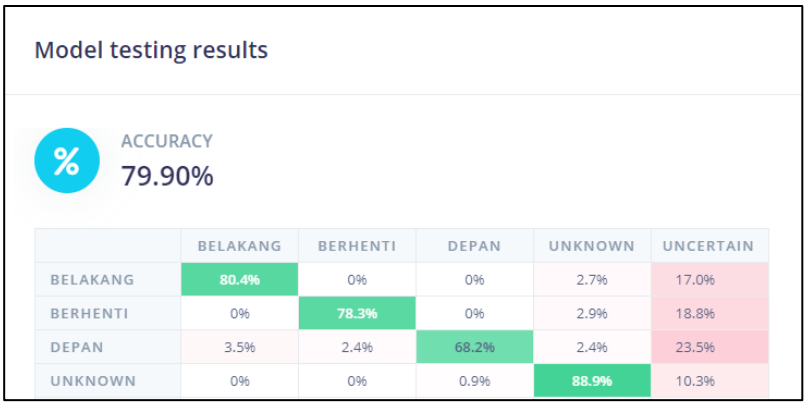


Figure 7. Accuracy of the training data

In the testing phase, the model achieved an accuracy of 79.9%, slightly lower than the training accuracy, which is common and indicates that the model generalizes well but encounters new challenges with unseen data (Figure 8). For class-specific performance, the model classified 80.4% of "Belakang" commands correctly, which is consistent with the training performance. A minor proportion was misclassified as "Unknown" (3.7%) and "Uncertain" (17%). For the "Berhenti" command, the model achieved 78.3% accuracy, with slight misclassifications into "Unknown" (2.9%) and "Uncertain" (18.8%). The accuracy for "Depan" decreased slightly to 68.2% on the test set. This class saw more confusion with "Uncertain" (23.5%) and "Unknown" (2.4%), suggesting that the model may have difficulties distinguishing "Depan" from other commands under certain dialectal variations or acoustic conditions. The model showed strong performance in the "Unknown" category, with an accuracy of 88.9%.



	BELAKANG	BERHENTI	DEPAN	UNKNOWN	UNCERTAIN
BELAKANG	80.4%	0%	0%	2.7%	17.0%
BERHENTI	0%	78.3%	0%	2.9%	18.8%
DEPAN	3.5%	2.4%	68.2%	2.4%	23.5%
UNKNOWN	0%	0%	0.9%	88.9%	10.3%

Figure 8. Accuracy of testing 20% remaining unseen data on the trained model

The model's performance, with approximately 80% accuracy in both training and testing phases, demonstrates its ability to effectively recognize Malay dialect commands across different classes. The slight decrease in testing accuracy is within acceptable limits and indicates a reasonable generalization capacity. However, there are a few notable limitations and areas for improvement. The higher rates of misclassification for "Depan" and "Berhenti" suggest that these commands may have similar acoustic features or may be more susceptible to dialectal variations that confuse the model. Addressing this may require augmenting the data for these commands or incorporating additional preprocessing techniques to highlight distinguishing features. In the test set, a considerable percentage of "Depan" and "Belakang" commands were marked as "Uncertain." This could indicate that the model faces difficulties when new data exhibits variations in pronunciation or background noise does not present in the training set. Further training with more diverse data could help reduce this issue. The model's high accuracy for the "Unknown" class across both training and testing sets demonstrates its strength in filtering out irrelevant sounds, an essential aspect for practical voice command applications. This suggests that the model can perform effectively in real-world scenarios where non-command sounds are prevalent.

Overall, these results indicate a promising foundation for a dialect-specific Malay speech recognition system. The model’s high performance in detecting command classes and filtering non-commands provides a solid base, though refinements could further improve recognition accuracy, particularly in more challenging dialectal variations.

5.0 CONCLUSION AND FUTURE WORKS

This exploratory study aimed to address the need for a speech recognition system capable of understanding Malay dialect instructions, specifically as a step toward developing assistive devices for disabled and elderly individuals living alone. Using the collected data, which included four classes of instructions from various Malay dialects, we developed a deep learning model on the Edge Impulse platform. The methodology involved data collection via smartphone

recordings, manual data cleaning, and classification using an artificial neural network with MFCC feature extraction. The model achieved a training accuracy of 81.2% and a testing accuracy of 79.9 %, demonstrating promising performance in recognizing commands across different Malay dialects. The system showed particular strength in identifying the "Unknown" class, which is essential for filtering out irrelevant audio in real-world applications. However, some challenges were observed with specific commands like "Depan" and "Berhenti," which had higher misclassification rates, highlighting the need for a more diverse data set to improve the model's robustness and generalization.

As an exploratory study, this research has provided valuable insights into the tools and methods needed to build effective speech recognition systems for assistive devices that cater to the Malay-speaking population, including those who use regional dialects. The results indicate that the Edge Impulse platform, combined with deep learning models, is a viable approach for developing voice-controlled systems that can enhance accessibility for disabled and elderly individuals.

Future work could involve expanding the dataset to include additional dialectal variations and speaker profiles, further enhancing the model's accuracy and robustness. Additionally, collecting data in more natural environments could improve the model's performance under real-world conditions, preparing it for deployment in assistive devices that aid disabled and elderly individuals. Exploring additional tools and architectures to handle complex variations in dialect and pronunciation will also be crucial in building reliable and inclusive voice-controlled assistive systems for this demographic. Ultimately, this study lays the groundwork for creating accessible and adaptive technologies that respond accurately to Malay dialect instructions, supporting independent living for those who may benefit most from assistive technology.

REFERENCES

[1] M. Anwar *et al.*, "MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation," 2023. [Online]. Available: https://github.com/

[2] A. H. Omar, "Standard Language and the Standardization of Malay," *Source: Anthropological Linguistics*, vol. 13, no. 2, pp. 75–89, 1971, [Online]. Available: http://www.jstor.org

[3] I. Aman and R. Mustaffa, "Social Variation Of Malay Language In Kuching, Sarawak, Malaysia: A Study On Accent, Identity And Integration," vol. 9, no. 1, 2009.

[4] R. Imaizumi, R. Masumura, S. Shiota, and H. Kiya, "End-to-end Japanese Multi-dialect Speech Recognition and Dialect Identification with Multi-task Learning," *APSIPA Trans Signal Inf Process*, vol. 11, no. 1, 2022, doi: 10.1561/116.00000045.

[5] H. A. Alsayadi, A. A. Abdelhamid, I. Hegazy, B. Alotaibi, and Z. T. Fayed, "Deep Investigation of the Recent Advances in Dialectal Arabic Speech Recognition," *IEEE Access*, vol. 10, pp. 57063–57079, 2022, doi: 10.1109/ACCESS.2022.3177191.

[6] A. Das, K. Kumar, and J. Wu, "Multi-dialect speech recognition in English using attention on ensemble of experts," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 6244–6248. doi: 10.1109/ICASSP39728.2021.9413952.

[7] A. Yadavalli, G. S. Mirishkar, and A. K. Vuppala, "Multi-Task End-to-End Model for Telugu Dialect and Speech Recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2022, pp. 1387–1391. doi: 10.21437/Interspeech.2022-10739.

[8] J. Zhang, Y. Peng, P. Van Tung, H. Xu, H. Huang, and E. S. Chng, "E2E-based Multi-task Learning Approach to Joint Speech and Accent Recognition." [Online]. Available: https://www.datatang.com/INTERSPEECH2020

[9] Vasile-Florian Păiș, *Computer Science, Technology and Applications*. New York: Nova Science Publishers., 2022.

[10] W. Lee, J. J. Seong, B. Ozlu, B. S. Shim, A. Marakhimov, and S. Lee, "Biosignal sensors and deep learning-based speech recognition: A review," Feb. 02, 2021, *MDPI AG*. doi: 10.3390/s21041399.

[11] D. Al-Fraihat, Y. Sharrab, F. Alzyoud, A. Qahmash, M. Tarawneh, and A. Maaita, "Speech Recognition Utilizing Deep Learning: A Systematic Review of the Latest Developments," 2024, *Korea Information Processing Society*. doi: 10.22967/HCIS.2024.14.015.

[12] S. Zhang, J. Kong, C. Chen, Y. Li, and H. Liang, "Speech GAU: A Single Head Attention for Mandarin Speech Recognition for Air Traffic Control," *Aerospace*, vol. 9, no. 8, Aug. 2022, doi: 10.3390/aerospace9080395.

[13] K. Sharifani and M. Amini, “Machine Learning and Deep Learning: A Review of Methods and Applications,” *World Information Technology and Engineering Journal*, vol. 10, 2023, [Online]. Available: https://ssrn.com/abstract=4458723

[14] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, “Deep Learning for Intelligent Human–Computer Interaction,” Nov. 01, 2022, *MDPI*. doi: 10.3390/app122211457.

[15] K. Ismael Taher and A. Mohsin Abdulazeez, “Deep Learning Convolutional Neural Network for Speech Recognition: A Review,” 2021, doi: 10.5281/zenodo.4475361.

[16] C. Yu, M. Kang, Y. Chen, J. Wu, and X. Zhao, “Acoustic modeling based on deep learning for low-resource speech recognition: An overview,” 2020, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2020.3020421.

[17] A. Ali, S. Vogel, and S. Renals, “Speech Recognition Challenge in the Wild: Arabic MGB-3,” Sep. 2017, [Online]. Available: http://arxiv.org/abs/1709.07276

[18] H. Bouamor *et al.*, “The MADAR Arabic Dialect Corpus and Lexicon”, [Online]. Available: http://www.ustar-consortium.com/

[19] D. Wang, S. Ye, X. Hu, S. Li, and X. Xu, “An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2021, pp. 1590–1594. doi: 10.21437/Interspeech.2021-374.

[20] Z. Tang *et al.*, “KeSpeech: An Open Source Speech Dataset of Mandarin and Its Eight Subdialects”, [Online]. Available: https://github.com/KeSpeech/KeSpeech/blob/main/dataset_license.md

[21] F. Xu, J. Luo, M. Wang, and G. Zhou, “Speech-driven end-to-end language discrimination toward Chinese dialects,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 19, no. 5, Aug. 2020, doi: 10.1145/3389021.

[22] Q. Li, Q. Mai, M. Wang, and M. Ma, “Chinese dialect speech recognition: a comprehensive survey,” *Artif Intell Rev*, vol. 57, no. 2, Feb. 2024, doi: 10.1007/s10462-023-10668-0.

[23] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.11477

[24] C. Graham and N. Roll, “Evaluating OpenAI’s Whisper ASR: Performance analysis across diverse accents and speaker traits,” *JASA Express Lett*, vol. 4, no. 2, Feb. 2024, doi: 10.1121/10.0024876.

[25] S. Hymel *et al.*, “Edge Impulse: An MLOps Platform for Tiny Machine Learning,” Nov. 2022, [Online]. Available: http://arxiv.org/abs/2212.03332