Enhancing Low-Resource Dialectal ASR in Indonesian Using Speech-Transformer Models and Data Augmentation

Sukmawati Nur Endah
Informatics Department, Universitas Diponegoro, Indonesia | Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia
sukmane@lecturer.undip.ac.id

Suprapto
Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia
sprapto@ugm.ac.id (corresponding author)

Yohanes Suyanto
Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia
yanto@ugm.ac.id

Received: 17 June 2025 | Revised: 25 July 2025 | Accepted: 14 August 2025

 DOI: https://doi.org/10.48084/etasr.12734

ABSTRACT

One of the main challenges faced by researchers in speech recognition is the limitation of data, especially for low-resource languages. A common strategy to improve a model's performance is to expand the data space through data augmentation techniques. Data augmentation has proven effective in increasing the amount of training data and reducing the mismatch between training and testing data. Furthermore, data augmentation is essential for improving the performance of deep neural networks by mitigating overfitting and enhancing the models' generalization capabilities. This study compares the impact of several standard augmentation techniques applied to low-resource dialect speech (time stretching, pitch shifting, noise addition, and gain) on speech recognition performance using a Speech-Transformer architecture. The dataset used consists of Indonesian dialectal speech. The results indicate that the average accuracy improvement in recognition was 57.6%, 57.9%, and 59.3% for Character Error Rate (CER), Word Error Rate (WER), and Sentence Error Rate (SER), respectively, compared to speech recognition without any data augmentation.

Keywords-augmentation; dialectal speech recognition; low-resource

I. INTRODUCTION

In several countries, such as Indonesia, various ethnic groups speak in different dialects. Speech recognition for dialects requires more effort than for standard pronunciations. A further challenge is the limited availability of data for dialectal speech, which represents one of the main obstacles in speech recognition research: either insufficient data [1] or the presence of low-resource data [2].

On the other hand, modern Automatic Speech Recognition (ASR) models such as Transformer, Speech-Transformer, and Conformer typically require extensive speech datasets for effective training. However, the development and annotation of such datasets entail substantial time, expertise, and resources. In practical applications, obtaining natural and representative speech data for training is often challenging or unfeasible [3]. One way to address this issue is by expanding the data space through data augmentation [4]. Data augmentation is an effective approach to enlarge the training corpus and minimize the discrepancy between training and testing conditions [5]. Additionally, data augmentation is essential for improving the performance of deep neural networks by mitigating overfitting and enhancing the models' generalization capabilities [6-8].

This study examines the impact of several standard augmentation techniques that can be implemented for dialectal speech with limited resources, including time stretching, pitch shifting, noise addition, and gain adjustment, on speech recognition performance using a Speech-Transformer architecture. The dataset used consists of Indonesian dialectal speech. The primary contribution of this study is the development of an audio data augmentation software, along with an in-depth evaluation of the effectiveness of conventional audio augmentation techniques in enhancing the accuracy of dialectal speech recognition. The results indicate that data

augmentation significantly enhances the performance of dialectal speech recognition, with a 57.6% reduction in Character Error Rate (CER), 57.9% reduction in Word Error Rate (WER), and 59.3% reduction in Sentence Error Rate (SER), compared to speech recognition without any data augmentation.

II. RELATED WORK

In recent years, the research community has shown increasing attention to augmentation techniques due to their promising benefits for various ASR applications, particularly for languages with limited resources, including minority, regional, and dialectal varieties. For example, employing Cycle-consistent Generative Adversarial Networks (CycleGAN) for data augmentation resulted in a 5.58% reduction in WER [9], and Generative Adversarial Networks (GANs) showed a relative WER reduction of more than 20 % for end-to-end children ASR [10]. Speaker augmentation using SpecAugment resulted in a substantial relative reduction in WER, achieving a 30% improvement over systems without data augmentation and an approximately 18% improvement compared to systems using SpecAugment only in low-resource speech recognition tasks [11]. In the speech emotion recognition task, augmentation with CycleGAN [12] and GAN [13] also showed improvement. Furthermore, authors in [14] showed that selecting suitable data augmentation techniques tailored to specific conditions is essential for enhancing performance in speech emotion recognition tasks using Support Vector Machines (SVM). CycleGAN has also been used as a data augmentation technique to generate whispers from original speech [15].

Several novel data augmentation techniques have been explored to enhance ASR performance for low-resource languages [16]. A systematic review by authors in [17] highlighted various augmentation methods used in deep learning-based classifiers for the classification of audio signals, including speech. Furthermore, authors in [18] investigated the effectiveness of four specific augmentation strategies for environmental sound prediction: specified independent, random independent, specified sequential, and random sequential.

The influence of different augmentation techniques (time stretching, pitch shifting, background noise, and combining time stretching and background noise) was assessed on a mechanical noise dataset using classification accuracy as the evaluation criterion. The study showed that the combined augmentation approach yielded the highest precision and the lowest error rates for both squeak and rattle datasets [19]. These three augmentation techniques (pitch shifting, noise addition, time stretching) have also been used to improve system robustness in the speaker verification task for Persian speech [20]. Furthermore, the effect of data augmentation techniques on ASR system development indicates that pitch shifting can reduce the WER for Convolutional Neural Network (CNN)-based recognizers [3].

ASR technology allows machines to transcribe spoken language into text or executable commands [21]. It is often used in robots and intelligent machines and as a communication tool between humans and computers [22]. Deep learning approaches have recently become increasingly prominent in ASR systems development. Given that speech data are inherently sequential, deep learning models like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Bidirectional-LSTM (Bi-LSTM) networks were initially considered more appropriate [23]. However, recent advancements have increasingly shifted towards Transformer-based architectures due to their superior performance and flexibility.

The Transformer architecture was proposed as a solution to challenges encountered in machine translation tasks [24]. Since then, it has been widely adopted across various domains within Natural Language Processing (NLP), including speech synthesis, ASR, computer vision, and others. The original Transformer architecture consists of an encoder-decoder structure based on self-attention mechanisms, where each component is composed of a stack of identical blocks.

The first application of a Transformer for ASR was a model called Speech-Transformer, which introduced convolutional layers preceding the Transformer blocks to better capture local acoustic features [25]. A well-known advancement of this model is the Conformer architecture [26], which integrates convolutional modules into the Transformer to enhance the modeling of both local and global dependencies in speech.

However, the attention mechanism in the Transformer and Conformer models incurs quadratic computational complexity concerning input length. To address this limitation, authors in [27] proposed a Transformer with linear attention and authors in [28] proposed the Linear Attention-based Conformer (LAC), which leverages linear attention mechanisms to improve computational efficiency while maintaining recognition performance.

III. METHODOLOGY

This research was conducted through four main stages: data preparation, data augmentation, model development (training) and model evaluation, as illustrated in Figure 1. The data preparation phase consists of data collection, audio segmentation, and data splitting. The data augmentation phase utilizes four augmentation techniques: time stretching, pitch shifting, noise addition, and gain adjustment. The model development (training) process consists of two phases: feature extraction and speech recognition with the Speech-Transformer. The subsequent sections offer a detailed explanation of each phase.

A. Data Preparation

1) Data Collection

This study uses two types of datasets: Indonesian Dataset1 (ID1) [29] and Indonesian Dataset2 (ID2) [30]. All 54 speakers (23 in ID1 and 31 in ID2) represent different regions of Indonesia, capturing the dialects of their respective ethnic groups. The dialects used in this study originate from the five major ethnic groups in Indonesia: Javanese, Balinese, Batak, Sundanese, and Minang. The speakers' ages range from 17 to 25 years. The distribution of dialect data and speaker gender for each dataset is shown in Table I.

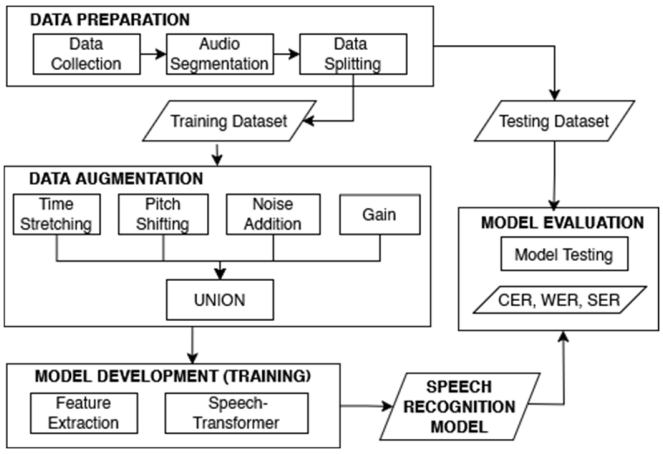

Fig. 1. Overview of the research methodology.

TABLE I. DISTRIBUTION OF DIALECTS AND SPEAKER GENDER

Dialect of ethnic group	ID1		ID2	
	Male	Female	Male	Female
Javanese	7	2	2	3
Sundanese	0	2	5	3
Batak	3	3	2	2
Balinese	1	0	4	5.
Minang	2	3	3	2
Total	13	10	16	15

The recording process was conducted using the Audacity software with a sampling rate of 44,100 Hz, 32-bit float type, and a mono channel. Each audio file was saved in Waveform Audio File Format (WAV). The recordings were performed in a closed, soundproof studio to minimize noise interference.

The ID1 dataset contains 70 sentences, whereas ID2 contains 330 sentences, including the 70 sentences from ID1. The total number of sentences and the duration of speech recordings are shown in Table II. This study also uses a combination of the ID1 and ID2 datasets, referred to as the ID12 dataset. The ID12 dataset represents the total number of sentences and total duration from both ID1 and ID2.

TABLE II. TOTAL NUMBER OF SENTENCES AND DURATION

Dataset	Number of sentences	Duration
ID1	1,410	53 m 49 s
ID2	10,230	7 h 40 m 48 s
Total	11,640	8 h 34 m 37 s

2) Audio Segmentation

The recordings from each speaker consist of continuous speech, where the recording stops after several sentences are spoken. Audio segmentation involves extracting individual sentence processes from each speaker's speech. Each segmented sentence is considered a single data instance.

The segmentation process was performed manually using Audacity software. Before segmentation, the signal undergoes visual inspection to identify any noise that could disrupt the recognition process. If noise is present in the signal, it is reduced manually using Audacity's noise reduction feature.

3) Dataset Splitting

Each dataset (ID1, ID2, and ID12) was split into two subsets, allocating 80% for training and 20% for testing. The details of this division are presented in Table III.

TABLE III. DATA SPLITTING OF TRAINING AND TESTING SETS

Dataset	Training data		Testing data	
	NS	Duration	NS	Duration
ID1	1,128	43m 14s	282	10m 35s
ID2	8,184	6h 9m 1s	2,046	1h 31m 47s
ID12	9,312	6h 52m 15s	2,328	1h 42m 22s

NS: number of sentences.

B. Data Augmentation

We augmented only the training dataset using four augmentation techniques: time stretching, pitch shifting, noise addition, and gain adjustment. Moreover, we integrated all four augmentation techniques to evaluate their effectiveness in dialectal speech recognition. The following sections provide a detailed explanation of each technique.

1) Time Stretching

Time stretching is an audio processing technique that changes the length of an audio signal while preserving its original pitch. This approach enables audio to be slowed or speeded up, without causing substantial distortion. Time stretching is beneficial for training speech recognition models to improve their robustness against variations in speech speed.

In this study, the length of the audio signal was modified using the time-stretch function from Librosa with a speed variation factor randomly selected between 0.9 and 1.1. If the factor is less than 1, the speech becomes longer (slower); if it is greater than 1, the speech becomes shorter (faster). This range was chosen to preserve the dialect in each utterance.

2) Pitch Shifting

Pitch shifting is a technique that alters the pitch of an audio signal while maintaining its original duration. This technique can increase or decrease the frequency of a sound to simulate voice variations among different individuals, such as differences between male and female voices. Pitch shifting enhances model robustness to natural frequency variations in speech.

In this study, the pitch of the audio signal was adjusted using a randomly generated value between -1 and 1 semitones. This range was chosen to preserve the dialect in each utterance. If the generated value is greater than 0, the pitch is raised (higher, resembling a female voice); if the value is less than 0, the pitch is lowered (deeper, resembling a male voice). If the value is exactly 0, the pitch remains unchanged.

3) Noise Addition

Noise addition is the process of adding background noise to an audio signal to enhance the model's robustness against noisy environments. Common types of noise used in augmentation include white noise, pink noise, or background sounds such as traffic or office noise. White noise is a type of noise with an

evenly distributed energy across the whole audible frequency spectrum (typically 20 Hz – 20 kHz). In contrast, pink noise has more energy decreases as the frequency increases (each octave contains the same amount of energy). In other words, pink noise has more energy in lower frequencies compared to higher frequencies. This technique improves the model's robustness to real-world conditions, which often do not involve ideal recording environments.

In this study, the noise used was white noise, specifically Gaussian noise, with a mean of 0, variance following the distribution of the original signal, and a length equal to that of the original signal. The generated noise was then added to the original signal at a certain scale. This scale was determined by generating a random value between 0.1 and 0.3.

4) Gain

Gain is a technique that modifies the amplitude of an audio signal, essentially adjusting the loudness (volume) of the sound. This approach is commonly used to enhance the robustness of speech recognition models against variations in volume that occur during recordings, such as differences in microphone distance or speaker intensity. In speech data augmentation, there are two types of gain modification: gain increase (amplification), which raises the signal amplitude to make the sound louder, and gain decrease (attenuation), which lowers the signal amplitude to make the sound quieter.

In this study, the signal amplitude was increased using a scaling factor randomly selected between 2 and 4. This range was selected to ensure that the scaling factor did not become excessively large, as an overly high amplitude may surpass normalization limits, resulting in audio distortion or clipping.

C. Model Development (Training)

1) Feature Extraction

After data augmentation, the feature extraction process was performed by creating a spectrogram from each speech sample using the Short-Time Fourier Transform (STFT). The steps include framing, windowing, computing the STFT, and spectrogram formation.

Framing refers to segmenting a continuous audio waveform into short, overlapping windows. In this study, each frame consists of 200 samples, with a hop size (stride) of 80 samples between successive frames. To reduce distortions at the beginning and end of each frame, the frame is multiplied by a window function. In this research, we use a Hamming window, as described in [31].

After the windowing process, the STFT is computed using (1):

$$\mathrm{STFT}_x(m,k) = \sum_{n=0}^{N_m-1} x(n+mH)w(n)\, e^{-j\frac{2\pi}{N_m}kn} \quad (1)$$

where $x(n)$ is the original signal in the time domain, m is the frame index (time segment), k is the frequency index, H is hop size (number of samples between frames), $w(n)$ is the Hamming window function, and N_m is the window length (frame length). The next step is to take the magnitude of $\mathrm{STFT}[m,k]$ and apply a power law transformation (root of $\mathrm{STFT}[m,k]$). After obtaining the spectrogram, it was normalized using mean-variance normalization and padded to 10 s.

2) Speech-Transformer

The extracted feature representations serve as the input to the speech recognition system, which employs The Speech-Transformer model [25]. This model is a development of the Transformer model [24], comprising an encoder block and a decoder block. Before entering the encoder block, the input undergoes convolution followed by a Rectified Linear Unit (ReLU) activation function.

The training process used the training dataset described previously. In this study, we employed four encoder layers and one decoder layer, with two heads in the multi-head attention mechanism. The result of the training process is a trained speech recognition model.

D. Model Evaluation

Model testing was conducted to evaluate the performance of the trained model. The test dataset, as described in Table III, was used for this purpose. The model was evaluated using CER, WER, and SER for each predefined testing scenario. The calculations of CER, WER, and SER are, respectively, performed using (2), (3), and (4):

$$\mathrm{CER} = \frac{S_c+D_c+I_c}{N_c} \quad (2)$$

where S_c is the number of substitutions (characters in the hypothesis that differ from the reference), D_c is the number of deletions (characters present in the reference but missing in the hypothesis), I_c is the number of insertions (extra characters in the hypothesis not in the reference), and N_c is the total number of characters in the reference transcription.

$$\mathrm{WER} = \frac{S_w+D_w+I_w}{N_w} \quad (3)$$

where S_w denotes the number of substitutions, D_w represents the number of deletions, I_w indicates the number of insertions, and N_w refers to the total number of words in the reference transcript.

$$\mathrm{SER} = \frac{E_S}{S} \quad (4)$$

where E_S represents the number of sentences with one or more errors (insertion, deletion, or substitution), and S represents the total number of reference sentences.

IV. EXPERIMENTAL RESULTS

A. Experiment Scenarios

As described above, the experiments were conducted using the datasets ID1, ID2 and ID12. Three testing scenarios for dialectal speech recognition were considered:

- Scenario 1: Testing ID1 without and with augmentation.
- Scenario 2: Testing ID2 without and with augmentation.
- Scenario 3: Testing ID12 without and with augmentation.

The best-performing test result can then be determined by comparing the three scenario results 1–3. In each scenario, the testing data were evaluated under the following conditions: augmentation using time stretch, augmentation using pitch shift, augmentation using noise addition, augmentation using gain adjustment, augmentation using the combination of all four techniques, and no augmentation. The hyperparameters used in this study were: initial and final learning rate = 10^{-5}, learning rate after warmup = 10^{-3} and 10^{-4}, batch size = 16, number of epochs =100, warmup epochs = 15, decay epochs = 85, Adam optimizer, and dropout rate = 0,1. All experiments across the three scenarios used a learning rate after warmup of 10^{-3}, except for Scenario 3 (testing ID12 using all four augmentation techniques), which employed a learning rate after warmup of 10^{-4}. This adjustment was made because using a learning rate of 10^{-3} in that scenario resulted in unsatisfactory recognition performance.

B. Experiment Results

Figure 2 presents the results of Scenario 1, which evaluates the ID1 dataset both with and without data augmentation, highlighting the effect of augmentation techniques on a smaller dataset. Figure 3 presents the results of Scenario 2 for the ID2 dataset, whereas Figure 4 illustrates the results of Scenario 3 for the combined ID12 dataset, providing insight into how augmentation performs with larger and more diverse data. By examining and comparing the WER values across these three scenarios, the best-performing results for each dataset are summarized in Table IV.

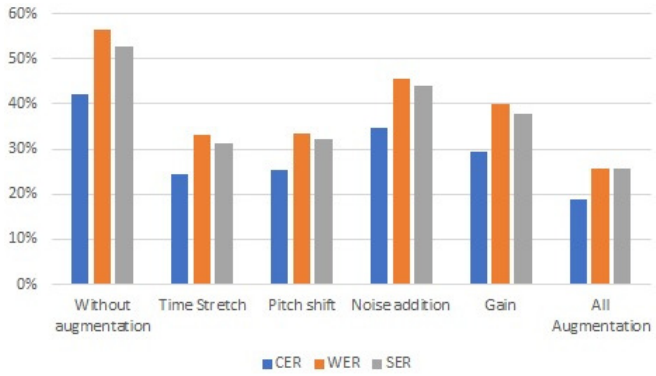

Fig. 2. Results of testing ID1 without and with augmentation.

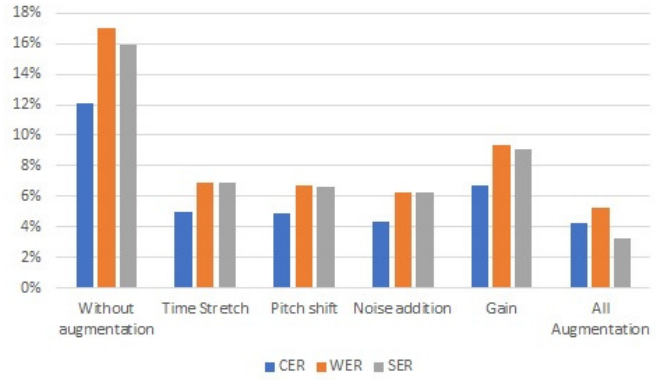

Fig. 3. Results of testing ID2 without and with augmentation.

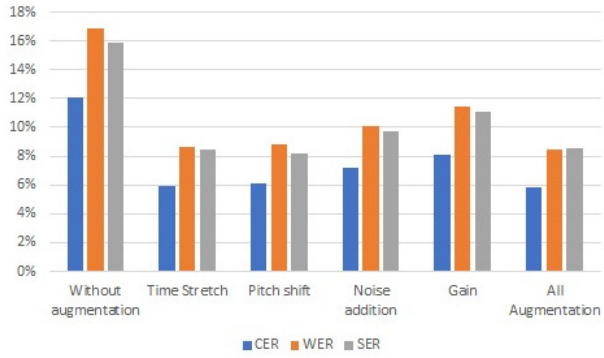

Fig. 4. Results of testing ID12 without and with augmentation.

TABLE IV. BEST RESULTS FROM EACH SCENARIO

Scenario	Dataset	Augmentation	WER (%)
1	ID1	Combination of all augmentation techniques	25.58
2	ID2	Combination of all augmentation techniques	5.29
3	ID12	Combination of all augmentation techniques	8.48

V. DISCUSSION

Based on the experimental results, it can be observed that adding augmented data consistently improves ASR performance for each dataset, whether evaluated using CER, WER, or SER. The extent of performance improvement for each dataset is presented in Table V.

TABLE V. IMPROVEMENT IN ASR PERFORMANCE

Dataset	Evaluation	Augmentation techniques (%)				
		TS	PS	NA	G	All
ID1	CER	42.3	40.3	17.6	30.4	55.7
	WER	41.5	40.8	19.0	29.5	54.7
	SER	40.8	38.9	16.8	28.2	51.7
ID2	CER	59.2	59.5	63.9	45.0	65.3
	WER	59.5	60.5	63.2	45.3	69.0
	SER	56.6	58.5	60.9	42.8	79.7
ID12	CER	50.6	49.6	40.8	33.1	51.9
	WER	49.0	47.7	40.6	32.1	49.9
	SER	46.9	48.5	38.8	30.2	46.4

TS = Time stretching; PS = Pitch shifting; NA = Noise addition; G = Gain; All = Combination of all augmentation techniques.

Table V shows that almost all experiments involving the combination of the four augmentation techniques (time stretching, pitch shifting, noise addition, and gain adjustment) resulted in better recognition accuracy compared to using any single augmentation technique alone. The average accuracy improvement in recognition was 57.6%, 57.9%, and 59.3% for CER, WER, and SER, respectively, compared to ASR without any data augmentation.

When considering the four techniques independently, no single augmentation method consistently outperformed the others. For example, in the case of ID1, which contains fewer data samples, time stretching produced better recognition results than the other three techniques. In contrast, for ID2, which has a larger number of data samples, the best recognition

performance was achieved using noise addition. For the combined dataset (ID12), examining CER and WER shows that time stretching outperformed the other three augmentation techniques.

Comparing the data volume between ID1 and ID2, the experimental results in Scenario 1 and 2 indicate that the more training data available, the better the recognition performance. However, when the datasets were combined (ID12), even though the total number of samples exceeded that of ID2, the recognition performance of ID12 was still inferior to that of ID2. This is likely due to the relatively high error rate in ID1, which negatively affected the combined model in ID12.

Authors in [32, 33] conducted speech recognition using the low-resource dataset ID1 with LSTM, Bi-LSTM, Convolutional Bi-LSTM (Conv-BiLSTM), and CNN-Bi-LSTM models. A comparison of their results with the best results obtained from the proposed method in this study is presented in Table VI.

TABLE VI. COMPARISON WITH OTHER WORKS

Method	Dataset	WER (%)
LSTM [32]	ID1	77.64
Bi-LSTM [32]	ID1	24.50
Bi-LSTM [33]	ID1	30.03
Conv-BiLSTM [33]	ID1	27.11
CNN-Bi-LSTM [32]	ID1	10.80
Speech-Transformer + augmentation (proposed)	ID1	25.58
Speech-Transformer + augmentation (proposed)	ID12	8.48

From Table VI, it can be observed that the proposed method achieves better results compared to other studies for the ID12 dataset. For the ID1 dataset, it still outperforms LSTM and is highly competitive with Bi-LSTM and Conv-BiLSTM models.

VI. CONCLUSION

This study investigated the effect of data augmentation on low-resource dialectal Automatic Speech Recognition (ASR) in Indonesian. Based on the experimental results, it can be concluded that adding augmented data enhances ASR performance. The average accuracy improvement in recognition was 57.6%, 57.9%, and 59.3% for Character Error Rate (CER), Word Error Rate (WER), and Sentence Error Rate (SER), respectively, compared to ASR using the Speech-Transformer model without any data augmentation. The combination of all augmented data produced by time stretching, pitch shifting, noise addition, and gain adjustment yields better recognition performance than using any single augmentation technique.

Increasing the amount of training data consistently improves recognition accuracy, either by adding original speech data or through data augmentation, especially for deep learning models. In this study, the Speech-Transformer model demonstrated that the proposed Speech-Transformer with augmentation method outperformed Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Convolutional Bi-LSTM (Conv-BiLSTM), and Convolutional Neural Network-LSTM (CNN-Bi-LSTM) for the ID12 dataset. Additionally, the proposed method was highly competitive with Bi-LSTM and Conv-BiLSTM for the low-resource ID1 dataset.

Future work will explore the integration of self-supervised pre-trained models and domain-specific transfer learning to achieve even greater performance on Indonesian dialects under limited data conditions.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest.

ACKNOWLEDGMENTS

The authors would like to acknowledge the research funding provided by the Indonesian Education Scholarship, the Center for Higher Education Funding and Assessment, and the Indonesian Endowment Fund for Education, Indonesia.

REFERENCES

[1] R. Gokay and H. Yalcin, "Improving Low Resource Turkish Speech Recognition with Data Augmentation and TTS," in *2019 16th International Multi-Conference on Systems, Signals & Devices*, Istanbul, Turkey, 2019, pp. 357–360, https://doi.org/10.1109/SSD.2019.8893184.

[2] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. Schwartz, "Two-Stage Data Augmentation for Low-Resourced Speech Recognition," in *Proc. Interspeech 2020*, San Francisco, CA, USA, 2016, pp. 2378–2382, https://doi.org/10.21437/Interspeech.2016-1386.

[3] J. Galic and D. Grozdic, "Exploring the Impact of Data Augmentation Techniques on Automatic Speech Recognition System Development: A Comparative Study," *Advances in Electrical and Computer Engineering*, vol. 23, no. 3, pp. 3–12, Aug. 2023, https://doi.org/10.4316/AECE.2023.03001.

[4] Z. Tu, J. Deadman, N. Ma, and J. Barker, "Auditory-Based Data Augmentation for end-to-end Automatic Speech Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, 2022, pp. 7447–7451, https://doi.org/10.1109/ICASSP43922.2022.9746252.

[5] M. Soleymanpour, M. T. Johnson, and J. Berry, "Dysarthric Speech Augmentation Using Prosodic Transformation and Masking for Subword End-to-end ASR," in *2021 International Conference on Speech Technology and Human-Computer Dialogue*, Bucharest, Romania, 2021, pp. 42–46, https://doi.org/10.1109/SpeD53181.2021.9587372.

[6] Y. Qian, H. Hu, and T. Tan, "Data augmentation using generative adversarial networks for robust speech recognition," *Speech Communication*, vol. 114, pp. 1–9, Nov. 2019, https://doi.org/10.1016/j.specom.2019.08.006.

[7] J. Wang, S. Kim, and Y. Lee, "Speech Augmentation Using Wavenet in Speech Recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, 2019, pp. 6770–6774, https://doi.org/10.1109/ICASSP.2019.8683388.

[8] X. Song, Z. Wu, Y. Huang, D. Su, and H. Meng, "SpecSwap: A Simple Data Augmentation Method for End-to-End Speech Recognition," in *Proc. Interspeech 2020*, Shanghai, China, 2020, pp. 581–585, https://doi.org/10.21437/Interspeech.2020-2275.

[9] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, "Data Augmentation Using CycleGAN for End-to-End Children ASR," in *2021 29th European Signal Processing Conference*, Dublin, Ireland, 2021, pp. 511–515, https://doi.org/10.23919/EUSIPCO54536.2021.9616228.

[10] P. Sheng, Z. Yang, and Y. Qian, "GANs for Children: A Generative Data Augmentation Strategy for Children Speech Recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop*, Singapore, Singapore, 2019, pp. 129–135, https://doi.org/10.1109/ASRU46091.2019.9003933.

[11] C. Du and K. Yu, "Speaker Augmentation for Low Resource Speech Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on*

Acoustics, Speech and Signal Processing, Barcelona, Spain, 2020, pp. 7719–7723, https://doi.org/10.1109/ICASSP40776.2020.9053139.

[12] F. Bao, M. Neumann, and N. T. Vu, "CycleGAN-Based Emotion Style Transfer as Data Augmentation for Speech Emotion Recognition," in *Proc. Interspeech 2019*, Graz, Austria, 2019, pp. 2828–2832, https://doi.org/10.21437/Interspeech.2019-2293.

[13] A. Chatziagapi *et al.*, "Data Augmentation Using GANs for Speech Emotion Recognition," in *Proc. Interspeech 2019*, Graz, Austria, 2019, pp. 171–175, https://doi.org/10.21437/Interspeech.2019-2561.

[14] B. T. Atmaja and A. Sasou, "Effects of Data Augmentations on Speech Emotion Recognition," *Sensors*, vol. 22, no. 16, Aug. 2022, Art. no. 5941, https://doi.org/10.3390/s22165941.

[15] P. R. R. Gudepu *et al.*, "Whisper Augmented End-to-End/Hybrid Speech Recognition System — CycleGAN Approach," in *Proc. Interspeech 2020*, Shanghai, China, 2020, pp. 2302–2306, https://doi.org/10.21437/Interspeech.2020-2639.

[16] R. Damania, "Data augmentation for automatic speech recognition for low resource languages," M.S. thesis, Department of Computer Science, Rochester Institute of Technology, Rochester, NY, USA, 2021.

[17] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, "Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review," *Electronics*, vol. 11, no. 22, Nov. 2022, Art. no. 3795, https://doi.org/10.3390/electronics11223795.

[18] M. Muthumari, C. A. Bhuvaneswari, J. E. N. S. Kumar Babu, and S. P. Raju, "Data Augmentation Model for Audio Signal Extraction," in *2022 3rd International Conference on Electronics and Sustainable Communication Systems*, Coimbatore, India, 2022, pp. 334–340, https://doi.org/10.1109/ICESC54411.2022.9885539.

[19] A. Abeysinghe, S. Tohmuang, J. L. Davy, and M. Fard, "Data augmentation on convolutional neural networks to classify mechanical noise," *Applied Acoustics*, vol. 203, Feb. 2023, Art. no. 109209, https://doi.org/10.1016/j.apacoust.2023.109209.

[20] N. Hosseini-Kivanani, H. Asadi, and C. Schommer, "Speaker Verification Enhancement via Speaking Rate Dynamics in Persian Speechprints," in *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods*, Porto, Portugal, 2025, pp. 665–672.

[21] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, Nov. 2023, Art. no. 101869, https://doi.org/10.1016/j.inffus.2023.101869.

[22] T. P. Rosin, J. Gachot, H.-L. Kordt, M. Kerzel, and S. Wermter, "Talking to Robots: A Practical Examination of Speech Foundation Models for HRI Applications." arXiv, Aug. 25, 2025, https://doi.org/10.48550/arXiv.2508.17753.

[23] A. Alahmadi, A. Alahmadi, E. Alduweib, W. Alromema, and B. Ahmed, "Development of a Deep Learning-based Arabic Speech Recognition System for Automatons," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18439–18446, Dec. 2024, https://doi.org/10.48084/etasr.8661.

[24] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[25] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, 2018, pp. 5884–5888, https://doi.org/10.1109/ICASSP.2018.8462506.

[26] A. Gulati *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, Shanghai, China, 2020, pp. 5036–5040, https://doi.org/10.21437/Interspeech.2020-3015.

[27] S. Yu and P. Li, "Transformer Based End-to-End Speech Recognition with Linear Attention," in *2022 8th International Conference on Control, Automation and Robotics*, Xiamen, China, 2022, pp. 340–344, https://doi.org/10.1109/ICCAR55106.2022.9782628.

[28] S. Li, M. Xu, and X.-L. Zhang, "Efficient conformer-based speech recognition with linear attention," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Tokyo, Japan, 2021, pp. 448–453.

[29] S. N. Endah, R. Kusumaningrum, and S. Adhy, "ID1 : Indonesian Dataset1." Zenodo, Jul. 11, 2025, https://doi.org/10.5281/zenodo.15862848.

[30] S. N. Endah, Suprapto, and Y. Suyanto, "ID2 : Indonesian Dataset2." Zenodo, Jul. 11, 2025, https://doi.org/10.5281/zenodo.15946165.

[31] S. Gupta, J. Jaafar, W. F. Wan Ahmad, and A. Bansal, "Feature Extraction using MFCC," *Signal & Image Processing : An International Journal*, vol. 4, no. 4, pp. 101–108, Aug. 2013, https://doi.org/10.5121/sipij.2013.4408.

[32] A. P. F. Naiborhu and S. N. Endah, "Indonesian Continuous Speech Recognition Using CNN and Bidirectional LSTM," in *2021 5th International Conference on Informatics and Computational Sciences*, Semarang, Indonesia, 2021, pp. 122–127, https://doi.org/10.1109/ICICoS53627.2021.9651902.

[33] S. N. Endah, R. Rismiyati, P. S. Sasongko, and A. P. F. Noiborhu, "Indonesian continuous speech recognition optimization with convolution bidirectional long short-term memory architecture," *Telecommunication Computing Electronics and Control*, vol. 23, no. 3, pp. 807–815, Jun. 2025, https://doi.org/10.12928/telkomnika.v23i3.24994.