Indo-ASR+: Enhancing Indonesian Automatic Speech Recognition by Fine-Tuning Wav2Vec2 with FAdam

Irfan Darmawan [1], Alam Rahmatulloh [2,*], Rohmat Gunawan [2], R. Wahjoe Witjaksono [1], and Ghatan Fauzi Nugraha [3]

[1] Department of Information System, Telkom University, Bandung, Indonesia
[2] Department of Informatics, Siliwangi University, Tasikmalaya, Indonesia
[3] Siliwangi Artificial Intelligence Research Group, Siliwangi University, Tasikmalaya, Indonesia
Email: irfandarmawan@telkomuniversity.ac.id (I.D.); alam@unsil.ac.id (A.R.); rohmatgunawan@unsil.ac.id (R.G.); wahyuwicaksono@telkomuniversity.ac.id (R.W.W.); ghatan.fauzi.nurgraha@unj.ac.id (G.F.N)
*Corresponding author

***Abstract*—Automatic Speech Recognition (ASR) has become a key technology in human-machine interaction, especially in supporting languages with limited resources such as Bahasa Indonesia. Although deep learning-based models such as Wav2Vec2 have shown good performance in speech recognition, further optimization is still needed to improve training accuracy and efficiency, especially in data-constrained and noisy environments. This research focuses on optimizing the Wav2Vec2 model for Indonesian ASR by applying the Fisher Adam (FAdam) optimizer. FAdam combines Natural Gradient Descent (NGD) with Fisher Information Matrix (FIM) to improve learning stability, accelerate convergence, and reduce sensitivity to noise in the data. The model was trained using the Indonesian Common Voice dataset and evaluated based on Word Error Rate (WER) of 5.59% and Character Error Rate (CER) of 1.76% on the validation set. Experimental results show that this approach not only improves accuracy over previous methods, also enhances training efficiency and improves the stability of model convergence compared to state-of-the-art models such as XLSR-53 and XLS-R 300m for Indonesian ASR. In addition, FAdam is shown to provide increased inference speed, making it a more optimal solution for ASR implementation in real-world scenarios. This research contributes to the development of a more adaptive and efficient ASR technology for Indonesian, while opening up further optimization opportunities in self-supervised learning-based models.**

***Keywords*—Automatic Speech Recognition (ASR), Bahasa Indonesia, Character Error Rate (CER), Fisher Adam (FAdam), Wav2Vec2, Word Error Rate (WER)**

I. Introduction

In the last decade, deep learning has entered the realm of Automatic Speech Recognition (ASR) resulting in models with low word error rates [1]. One of the improved ASR models is the speech-to-text (S2T) model. The application of deep learning in S2T makes it easier for the model to recognize the spoken voice even in different dialects and noisy environments [2]. So that S2T technology has now become a pioneer for various other models in the field of Natural Language Processing (NLP) such as voice assistant, real-time translation, speaker identification and verification, emotional recognition, human-machine interaction, and so on [3, 4]. These various developments provide further functionality to users, especially users with disabilities through communication aids for those with hearing or speech impairments [5].

One S2T model that is widely used today is Wav2Vec version 2 [6]. This model is an advanced development of the first version called Wav2Vec [7] with a major breakthrough in the form of utilizing a self-supervised learning approach. Wav2Vec2 is an innovative model that solves many problems in traditional approaches in the field of ASR (e.g. S2T), especially with data efficiency and high performance in low-resource conditions [6]. Wav2Vec 2.0 shows that transformer-based models can be very effective for speech recognition especially when coupled with self-supervised learning. It paves the way for building more inclusive, low-cost, and high-performance ASR systems. However, the Wav2Vec2 model has a large number of parameters which can lead to training instability. In addition, Wav2Vec2 often suffers from performance degradation due to the difficulty of handling complex noise distributions during training.

In an effort to overcome these problems, this research optimizes the Wav2Vec2 model using the Fisher Adam (FAdam) optimizer, which utilizes the Natural Gradient Descent (NGD) approach based on statistical information geometry using the Fisher Information Matrix (FIM) [8]. It is expected that by using FAdam optimizer, the training process can be more stable, efficient, and adaptive to noise. FAdam's handling of bias correction, noise, and gradient

Manuscript received May 27, 2025; revised June 19, 2025; accepted July 28, 2025; published October 24, 2025.

distribution is a superior solution for Wav2Vec2 optimization, especially in large model training scenarios or datasets with limited resources.

In the context of this study, the term noise encompasses various acoustic disturbances commonly encountered in real-world data, such as background sounds from vehicles, overlapping conversations, wind, or open-environment ambient noise. Additionally, linguistic noise is also considered, including inconsistent intonation, unclear pronunciation, and accent variations among different speakers. These disruptions can lead to errors in feature extraction or representation learning for speech signals. To address these challenges, this study employs FAdam, an optimization algorithm based on NGD that leverages the FIM to guide model parameter updates in a more stable and context-aware manner relative to the data distribution. As a result, the model becomes more adaptive to noise and demonstrates improved training stability compared to conventional optimization methods.

II. Literature Review

Automatic Speech Recognition (ASR) has become one of the significant topics in current technological developments [9]. In recent years, deep learning-based models have become the standard in ASR system development and replaced traditional approaches that rely on manual processing of signals and acoustic features. However, other challenges such as model adaptation for non-English languages, especially Bahasa Indonesia, are issues that require further attention.

Solutions to these problems have actually been solved by several studies. Research conducted by Abidin *et al.* [10] addressing the limitations of Indonesian speech recognition datasets by building datasets from YouTube channels that are thoroughly validated. This dataset is utilized to train an acoustic model based on a Time Delay Neural Network (TDNN) [11] with the assistance of Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) [12] alignment and data augmentation. This research significantly improves model performance and reduces the word error rate to 19.03%. In line with Ref. [10], research conducted by Yang *et al*. [13] building a TDNN-based ASR model with additional modifications resulted in the development of a new model called TDNN-Attention-HMM. In addition, this research model uses the hierarchical weight transfer method in the training phase so that the experimental results show that this model provides the best performance with a Word Error Rate (WER) of 6.79%, with a relative decrease of 26.52% compared to the DNN-HMM baseline system.

Furthermore, the utilization of Massively Multilingual Speech (MMS) and Whisper models [14] is another possible solution. As in the research [15] that conducted training on both models using Indonesian language datasets that include a variety of speech variability. The modified Whisper model showed the best results with a fairly low reduction in WER and Character Error Rate (CER). This study also found that speaking style is the factor that most influences the performance of the model. These findings provide important insights for the development of a more robust Indonesian ASR.

The use of other models such as Wav2Vec2 [6] is another option that can be used to build Indonesian ASR models. The research [16] utilized the XLSR-53 [17] pre-trained model to be used for training a Wav2Vec2-based ASR designed to reduce the need for training data on non-English languages. The results showed success by reducing the WER value from 20% to 12% on the Indonesian Common Voice dataset. This success made a significant impact on the field of ASR for local languages so that further research [18] addressed the improvement of the model. The study [18] centers on the advancement and assessment of ASR technology utilizing the XLS-R 300m model [19] integrated with Wav2Vec2 for Indonesian, Javanese, and Sundanese languages. The results showed that the model achieved a competitive WER with slightly lower performance for Javanese and Sundanese than Indonesian. This research contributes to the development of ASR technology by addressing the challenges of linguistic diversity and provides insights for ASR accuracy optimization across multiple language contexts.

Previous research related to the development of Indonesian ASR faces several limitations, such as dataset limitations, dependence on less flexible TDNN-HMM-based models, and the performance of multilingual models such as MMS and Whisper which are still influenced by variations in speaking style. In addition, the XLSR model is trained using a large model with many languages, so fine-tuning on small datasets often experiences gradient instability. Where weight updates become unstable and can cause slow convergence or even overfitting on certain training data. This study addresses the convergence problem of previous research by using FAdam, which is a more stable and adaptive optimizer than Adam or AdamW [8].

This allows the model to better adapt to limited datasets, reducing gradient instability, and accelerating convergence without compromising model generalization. This approach helps the Wav2Vec2 model trained using the Indonesian language dataset achieve optimal performance in more efficient training time and improve ASR accuracy for Indonesian.

III. Materials and Methods

This study concentrates on advancing an ASR model utilizing the Indonesian language. The proposed method consists of several main stages, namely dataset preparation, data preprocessing, feature extraction using Wav2Vec 2.0 model architecture [6], model optimization with FAdam optimizer [8], and model performance evaluation using WER and CER metrics. These stages are designed to ensure that the model is able to produce audio-to-text transcriptions with a high degree of accuracy and is able to handle variations in characteristics in the audio data, such as accents, intonations, and noise levels. The stages are shown in Fig. 1.

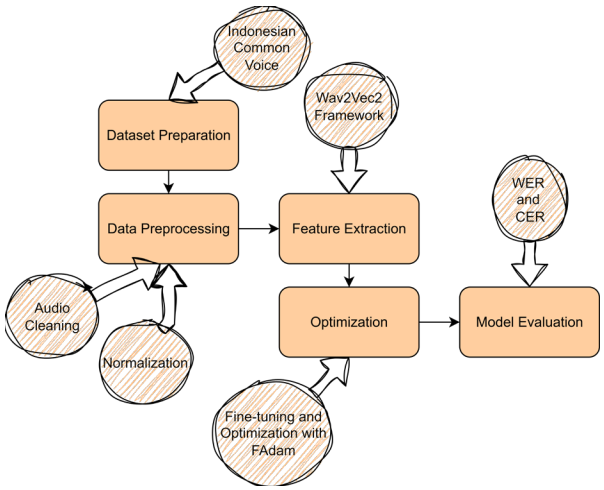


Fig. 1. Methodology.

A. Data Preparation

This research uses the Indonesian Common Voice dataset version 20.0 [20] for the training process of our ASR model. The Indonesian Common Voice dataset is an open-source audio dataset designed to support the development of speech recognition technology in Indonesian [21]. This dataset consists of a total of 66 hours of recorded voices, 34 hours of validated voices, and 604 total voices. The data in this dataset consists of recordings with diverse audio quality, covering a wide range of background noise levels and reflecting real-world conditions. In addition, the diversity of the data is also evidenced by the voices produced from different age groups and genders. The diversity of the data is shown in Table I.

TABLE I. DATA DIVERSITY BY AGE GROUP AND GENDER IN INDONESIAN COMMON VOICE DATASET

Age	Number of Data (%)	Gender	Number of Data (%)
<20	22	Male	45
20–29	41	Female	26
30–39	10	Gender Not Specified	29
40–49	2	-	-
Age Not Specified	25	-	-

B. Data Preprocessing

Before the features in the data are extracted, the data will go through a preprocessing process to ensure uniformity and compatibility with the Wav2Vec2 model. The process at this stage involves two steps at once, namely, audio cleaning and normalization to provide maximum results in the feature extraction step [22]. The use of audio cleaning is intended so that noise in the audio signal can be removed through the filtering function. This research uses a Butterworth low-pass filter with the formula as in Eq. (1).

$$H(s) = \frac{1}{\sqrt{1+(\frac{s}{\omega_c})^{2n}}} \quad (1)$$

where $H(s)$ is the frequency response maginitude, s is the corner frequency ($\omega = 2\pi f$, where f is the frequency), ω_c is the cutoff function, and n is the filter order. With this filter, the audio signal can limit high frequencies that are irrelevant to the human voice signal [23]. After the process, it is followed by normalizing all the audio files available in the dataset. This is done to equalize the amplitude of the audio signal so that each sample has a consistent intensity level. Our research uses the peak normalization method for this process through Eq. (2).

$$x_{norm}(t) = \frac{x(t)}{max\ (|x(t)|)} \quad (2)$$

where $x_{norm}(t)$ is the amplitude of the signal after normalization, $x(t)$ is the amplitude of the signal after normalization, and $max\ (|x(t)|)$ is the absolute maximum amplitude value of the signal. Peak Normalization ensures that the highest amplitude in the signal reaches a certain value, usually ±1 or in the range of 0 to 1.

C. Feature Extraction

This process is performed using the framework of Wav2Vec2 to convert the raw audio signals from the dataset into feature representations that can be used by the ASR model. The working procedure of this stage is in accordance with the framework of the research [6] in Fig. 2.

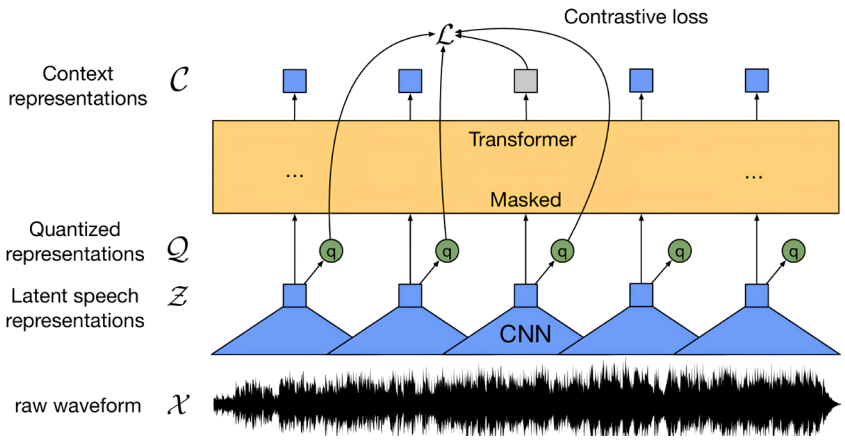


Fig. 2. Wav2Vec2 framework [6].

Feature extraction starts by feeding raw audio data (X) in the form of waveforms into the framework which are represented as one-dimensional signals in the time domain. Wav2Vec2 uses a Convolutional Neural Network (CNN) architecture consisting of several convolution blocks with Gaussian Error Linear Unit (GELU) activation and normalization layers for the feature encoder. This is done to convert the raw audio data into a latent speech representation $(X \rightarrow Z, where\ Z \in z_t)$. Next, the procedure continues by applying a masking technique to the latent speech representation z_t before feeding it into the transformer network. The masking is done in much the same way as the masking in BERT [24], the masking aims to make the model learn contextual patterns in the speech signal. When z_t the masked one is passed into the network transformer, it is converted z_t into context representation $(Z \rightarrow C, where\ C \in c_t)$. During the transformation, quantization operation is performed to quantize z_t it into discrete units $(Z \rightarrow Q, where\ Q \in q_t)$ using product quantization technique. Quantization is done by selecting a vector from the set of codebooks using Gumbel SoftMax [25] which allows the process to remain differentiable. The quantization phase does not require z_t masking so the process between context representation transformation and quantization is in a different branch of work. The workflow of the two phases is depicted in Fig. 3.

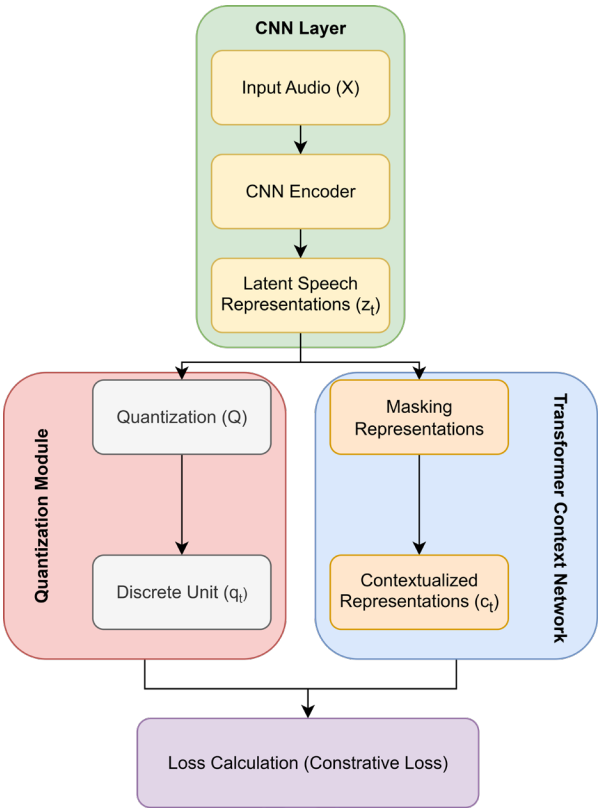


Fig. 3. Feature extraction flowchart using Wav2Vec2.

After getting the value c_t and q_t from the raw audio data, the next step is to calculate the loss value using the constrative learning object function as in Eq. (3) [6].

$$L_m = -\log \frac{\exp\left(sim(c_t, q_t)/\kappa\right)}{\sum_{\tilde{q} \in Q_t} \exp\left(sim(c_t, \tilde{q})/\kappa\right)} \quad (3)$$

The objective function for contrastive learning [26] is utilized to guide the model in acquiring an audio representation. Throughout this process, the model's goal is to differentiate the correct quantized latent speech representation (q_t) from a group of candidate representations (Q_t), which includes q_t along with several distractor representations. The contextual representation (c_t) generated by the transformer is matched with the quantized representation through the cosine equation $(sim(c_t, q_t) = \frac{c_t^T q_t}{\|c_t\| \|q_t\|})$. The contrastive loss function (L_m) is formulated as the log probability of the correct pair of c_t and q_t in the context of the set of distractors, with a temperature parameter (κ) that controls the sensitivity of the distribution.

D. Optimization

We use FAdam to optimize the trained Wav2Vec model. This NGD based optimization algorithm uses Diagonal Empirical FIM to improve convergence [8]. FAdam can help speed up and stabilize the training of Wav2Vec2 which has large parameters. With the FIM in FAdam, the model during training can capture the relationship between the parameter distribution and the data. Thus, with this optimization, the model can reach convergence faster because FIM captures the structure of the loss landscape which allows the model to adapt better to the parameters. In addition, FAdam provides improved training stability with gradient clipping and FIM normalization that reduces gradient oscillations. The resilience of this optimization method is evidenced by the stability of the loss curve during the training and validation process. The use of FAdam works by maintaining gradient stability as a result of utilizing FIM to prevent sharp oscillations in parameter updates. In addition, the application of NGD to FAdam allows the model to adapt effectively to noisy data distributions, thus increasing its resistance to data variation. As shown in result and discussion section, it is evident that the loss value in the validation data remains stable in each epoch. This shows that FAdam not only accelerates convergence but also ensures resilience in various training scenarios. In addition, compared to other optimization methods, FAdam is able to maintain the generalization balance of the model, thus reducing the risk of overfitting.

E. Model Evaluation

The trained model is subsequently assessed using the WER and CER metrics, as defined by Eqs. (4) and (5).

$$WER = \frac{S+I+D}{N} \times 100\% \quad (4)$$

$$CER = \frac{S_c+I_c+D_c}{N_c} \times 100\% \quad (5)$$

where S is the number of word or character substitutions, I is the number of insertions (additional words or characters that are not in the reference), D is the number of

deletions of missing words or characters in the prediction, and N is the total words or characters in the reference transcription. The results of this evaluation are then compared with existing research in related works based on the same dataset benchmark (using the Indonesian Common Voice dataset). In addition, we also validated the model through comparison with various models trained using the Wav2Vec2 framework and different optimization methods. This is intended to show the effectiveness of the Wav2Vec2 model optimized by the FAdam optimizer.

IV. Result and Discussion

A clearly structured results section, along with a compelling discussion, will highlight the originality and significance of your research. It should offer a brief yet accurate summary of the experimental outcomes, their analysis, and the conclusions that can be derived from the experiments.

The experimental part is done by training and evaluating the optimized Wav2Vec2 model using FAdam optimizer. In the training process we used a batch size configuration of 16 and a reduced learning rate using the LamdaLR method. For the use of FAdam optimizer we used a configuration with learning rate 0.001, weight decay 0.1, β1 0.9, β2 0.999, and epsilon 1×10^{-8}. The selection of these parameters is based on a combination of initial experiments and references from previous research using a NGD based optimizer. A weight decay of 0.1 is used to prevent overfitting by maintaining parameter regulation during training, based on the regulatory approach to Transformer-based models. The learning rate and momentum coefficient (β1 and β2) are set to 0.001, 0.9, and 0.999 respectively following the default configuration of FAdam, as this combination has been proven effective in handling gradient accumulation in FIM based optimization. An epsilon value (1×10^{-8}) is chosen to prevent division by zero in gradient normalization.

The model was trained on an NVDIA GeForce RTX 4060 with a number of epochs of 30 and by calculating the contrastive learning loss function at each epoch. Evaluation was performed on the validation data after each epoch to monitor the WER and CER values. The training results are shown in Fig. 4.

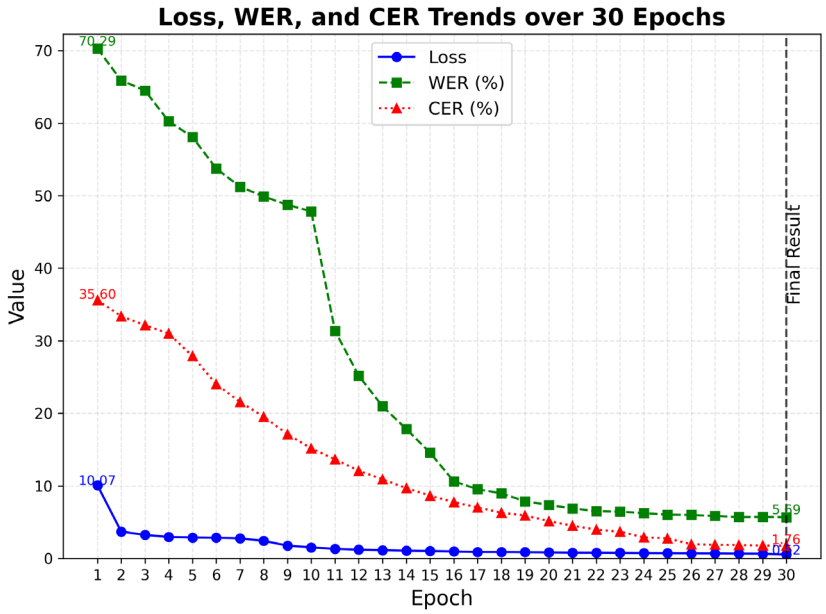


Fig. 4. Model training results. The graph illustrates the decrease in loss, Word Error Rate (WER), and Character Error Rate (CER) over the course of 30 training epochs. The dashed vertical line labeled "Final Result" indicates the final validation outcome at the 30th epoch, which is used as the primary reported value in this study.

The use of FAdam in Wav2Vec successfully reduces the gradient oscillation during training, as evidenced by the steady convergence of the loss curve in Fig. 4. This shows the adaptability of the model in handling the noise present in the dataset used. In addition, the WER and CER decreases steadily continue to show a downward trend without showing an increase in value between each epoch with final results of 5.59% for WER and 1.76% for CER, respectively. Furthermore, to further validate the model performance results that we have obtained, we compared the model performance with several related studies based on the use of the same dataset in Table II.

TABLE II. Model Performance Comparison Based on WER

Model	Dataset	WER (%)
XLS-R 300m [18]	Indonesian Common Voice Dataset	15.30
XLS-R 300m with 2-gram KenLM [18]		6.55
TDNN-Attention-HMM [13]		6.79
XLSR-53 [16]		20.31
XLSR-53 2-gram KenLM [16]		12.23
XLSR-53 3-gram KenLM [16]		12.30
XLSR-53 4-gram KenLM [16]		12.21
XLSR-53 5-gram KenLM [16]		12.25
Our Model		6.19

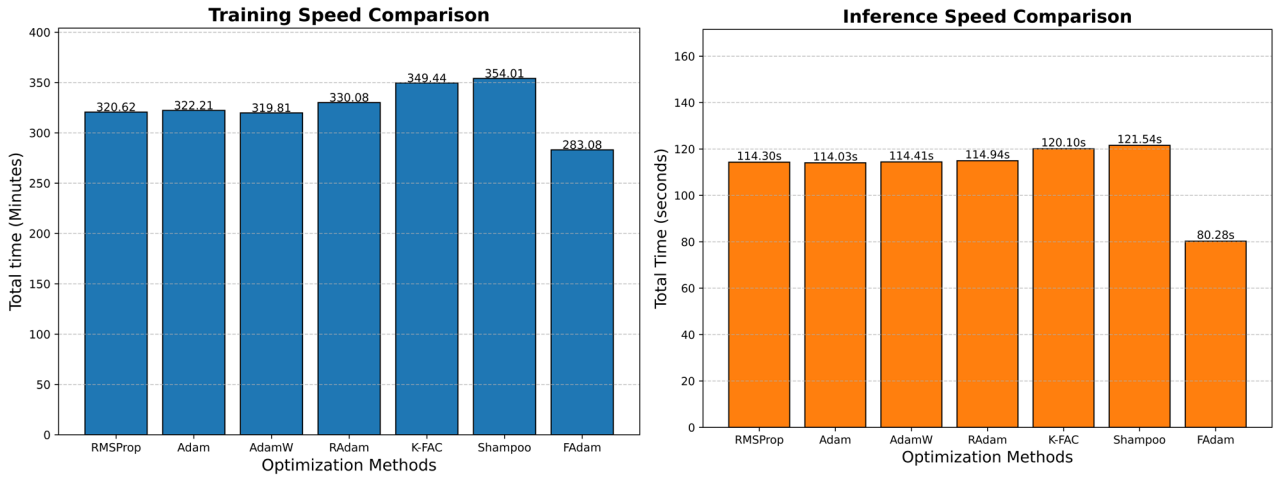
The comparison in Table II uses test data contained in Indonesian Common Voice. The table shows the success of the Wav2Vec2 model optimized using FAdam in Indonesian speech recognition compared to other research models. In addition, the model we developed can provide a lower WER value compared to the model added with KenLM. This provides evidence of the effectiveness of FAdam in optimizing the Wav2Vec model. However, an increase in the WER was observed in the test data compared to the validation results during training. This suggests a potential occurrence of mild overfitting to the training data, which is likely due to the extended training duration resulting from the computation of the FIM. While the use of FIM enables more precise parameter updates, it also increases the risk that the model will become overly tailored to the training data. In addition, the discrepancy in WER may also be attributed to differences in data distribution between the validation and test subsets within the Common Voice dataset, such as variations in accent, audio quality, or speaker characteristics—factors that are commonly encountered in real-world datasets.

The selection of FAdam for ASR model development in this study is based on its superiority in performing optimization, especially on ASR models [8]. In support of this statement, we evaluated the Wav2Vec2 model that was optimized using other methods. To ensure a computationally fair comparison across the evaluated optimization methods, the experiments in this section were limited to 10 training epochs. This constraint was intended to maintain training time efficiency and measurement consistency across methods without placing excessive demands on computational resources. We trained each model using 10 epochs, the evaluation results are shown in Table III.

TABLE III. PERFORMANCE COMPARISON OF WAV2VEC2 MODELS OPTIMIZED USING VARIOUS METHODS

Method	Train Loss ↓	Eval Loss↓	Train Samples/s↑	Train Steps/s↑	Eval Samples/s↑	Eval Steps/s↑	WER↓ (%)	CER↓ (%)
RMSProp	0.9089	0.5001	12.903	0.404	31.653	3.963	90.08	86.97
Adam	1.011	0.2743	12.839	0.402	31.727	3.973	48.07	34.58
RAdam	1.219	0.2750	12.533	0.392	31.478	3.941	63.21	49.15
AdamW	1.016	0.2763	12.936	0.405	31.622	3.959	39.61	31.04
K-FAC	0.992	0.2600	10.528	0.385	30.037	3.692	39.28	31.20
Shampoo	0.971	0.2630	9.803	0.373	29.839	3.601	38.97	30.73
FAdam	1.053	0.2651	14.614	0.457	45.067	5.643	47.64	30.66

It can be seen in Table III that optimization using FAdam on the Wav2Vec2 model provides improved performance, especially in training and evaluation time efficiency. It is evident from the sample and step processing of the training data that we are able to process 14.614 samples/s and 0.457 steps/s, respectively. Similar to the training data, for the model validation data we were able to process around 45,067 samples/s and 5643 steps/s. This happens because FAdam combines the calculation efficiency of Fisher Information Matrix with stability and adaptability which makes it fast in training and inference [8]. This results in faster convergence and inference compared to other optimization methods. This is evidenced in Fig. 5 on the comparison of training time and inference speed of each method.



(a) (b)

Fig. 5. Speed comparison of (a) training and (b) inference.

Among the four other optimization methods, FAdam is able to provide the fastest training and inference speed with 283.08 min for training and 80.28 s for inference at 10 epochs. In contrast, the RMSProp, Adam, AdamW, RAdam, K-FAC, and Shampoo methods take about more than 300 min for training and more than 100 s for inference. Furthermore, Table III shows that the train loss value of the model optimized by FAdam is still below the other models, especially by RMSProp. Although the RMSProp method has the smallest train loss value with

0.9089, it has the highest eval loss value. This proves that overfitting occurs in the training process optimized by the RMSProp method. In contrast, the model optimized using the FAdam method has the lowest eval loss value, which proves that there is a balance of predictions between the training and evaluation processes. In addition, the convergence speed of FAdam provides good stability to the validation process during training resulting in a stable eval loss value. This is evidenced by the comparison of eval loss gain during the training validation process for each epoch in Fig. 6.

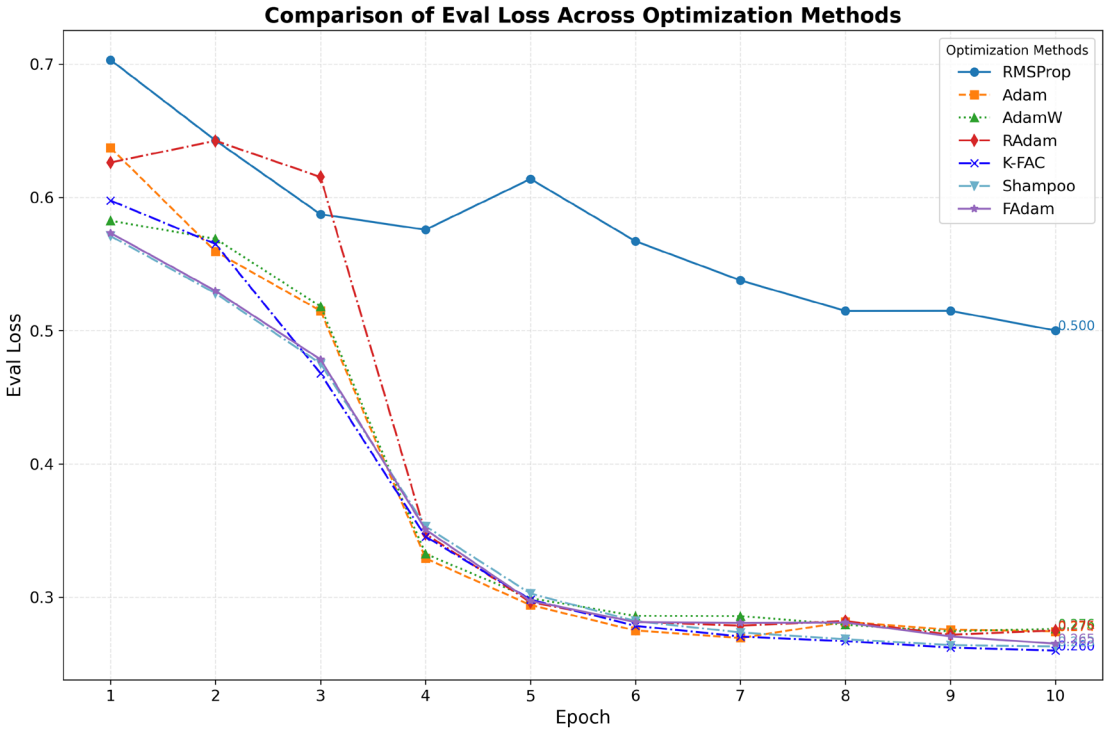


Fig. 6. Comparison of eval loss gain on various optimization methods.

Based on the experimental results up to the 10th epoch, FAdam shows the best performance with the lowest eval loss value of 0.2651, indicating stability and efficient convergence in training the Wav2Vec2 model. Compared to other optimizers such as Adam, AdamW, and RAdam which each have a final eval loss of around 0.2743–0.2763, FAdam provides improvements even with low computational overhead. Based on the performance estimates of other NGD optimizers, namely K-FAC and Shampoo, both are projected to have a final eval loss of 0.2600 and 0.2630 respectively. These values are indeed slightly lower than FAdam, but considering the computational complexity and higher memory requirements. Thus, FAdam is a more efficient solution in practice in terms of the trade-off between accuracy, stability, and training efficiency.

To further test the performance of the model with the FAdam optimizer statistically better than other optimizers, two statistical evaluation approaches were carried out, namely the bootstrap significance test and the post-hoc Tukey HSD test after ANOVA.

The results of the pairwise bootstrap test of 1000 iterations in Fig. 7 show that the FAdam model consistently produces lower WER values than all other optimizers tested (Adam, RMSProp, RAdam, AdamW, K-FAC, and Shampoo). All comparisons show positive ΔWER values, indicating that FAdam is superior on average, and all pairs show p values <0.001, indicating very strong statistical significance (strong evidence that FAdam is better, not by chance). Furthermore, to confirm the differences between models globally, a one-way ANOVA test was conducted on the WER between optimizer groups. The results show an F-statistic of 2.2979 with p = 0.0321, indicating that there is a significant difference in general. However, the results of the Tukey HSD follow-up test in Fig. 8 show that only the comparison between FAdam and RMSProp is significantly different (p = 0.0489). Comparison of FAdam with other optimizers such as Adam, RAdam, AdamW, K-FAC, and Shampoo did not reach the 95% significance limit, although numerically FAdam remains superior.

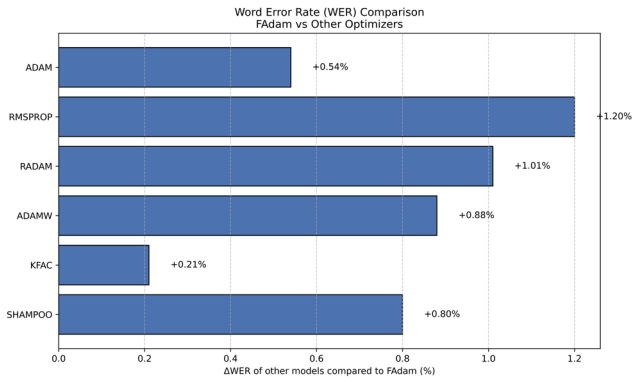


Fig. 7. Bootstrap significance test results.

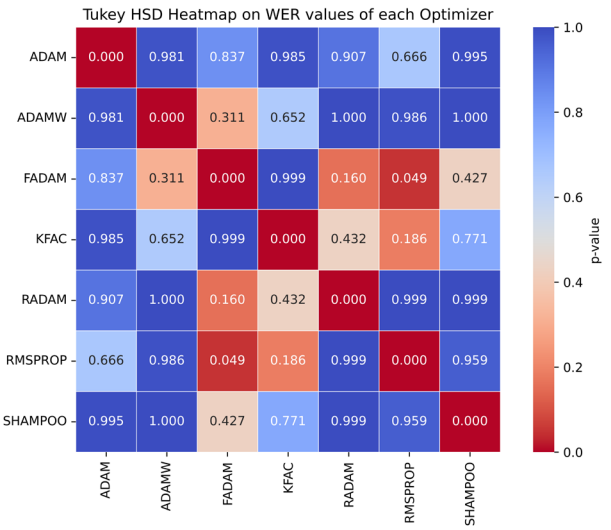


Fig. 8. Post-hoc Tukey HSD test results.

V. Conclusion

This study successfully optimized the Wav2Vec2-based ASR model for Indonesian using the FAdam optimization algorithm. The use of FAdam has been proven to significantly increase training efficiency, convergence speed, and prediction accuracy compared to other optimization methods such as Adam, AdamW, RAdam, RMSProp, K-FAC, and Shampoo.

The developed model successfully achieved a WER of 5.59% and a CER of 1.76% on validation data, and a WER of 6.19% on test data, which shows superior performance compared to several state-of-the-art models. Based on significance testing using the bootstrap method to validate the effectiveness of using FAdam, the optimization method consistently produces a lower WER than all comparison methods with a p-value <0.001. Furthermore, the ANOVA analysis followed by the Tukey HSD test revealed that FAdam demonstrated numerically superior performance compared to all evaluated optimization methods. The improvement was statistically significant when compared to the baseline method, RMSProp ($p = 0.0489$), although the differences with other optimizers such as AdamW and Shampoo did not reach the 95% significance threshold. This strengthens FAdam's superiority in limited and noise-prone training environments.

Although FAdam demonstrates improvements in training efficiency, convergence stability, and prediction accuracy, the results still indicate room for further enhancement. In several cases, the performance differences compared to other methods remain marginal, and the presence of mild overfitting suggests that improving generalization and handling real-world noise remain open challenges. Therefore, FAdam can be considered a promising approach for optimizing self-supervised ASR models such as Wav2Vec2; however, it is not yet fully optimal without the integration of regularization techniques and additional validation under real-world conditions.

Future research can focus on several relevant developments to improve the performance of the Wav2Vec2-based ASR model optimized with FAdam. One of them is to improve model generalization through additional regularization techniques such as dropout or noise-based data augmentation to overcome potential overfitting. It is important to explore the use of multilingual datasets to test the transfer learning and generalization capabilities across languages. Experiments with new model architectures, such as Whisper or MMS can be conducted to provide further insight into the advantages of FAdam over modern ASR models. Furthermore, evaluation of the model's performance on datasets with varying levels of noise and accents will help understand the model's robustness to data variability. In addition, extending research that supports the development of ASR models for other local languages using more diverse datasets will further enrich inclusive and adaptive speech recognition technology. Another promising direction is to explore the integration of the FAdam-optimized Wav2Vec2 model with advanced language models such as GPT-2, Fairseq LM, Transformer-XL, or Recurrent Neural Network Language Models (RNNLM), to further enhance decoding performance, contextual fluency, and robustness in real-world applications. Furthermore, future research may also investigate the sensitivity of the FAdam optimizer to different hyperparameter settings—such as learning rate, weight decay, and momentum factors—to better understand their effect on model stability and performance, and to provide practical guidelines for real-world deployment.

Conflict of Interest

The authors declare no conflict of interest

Author Contributions

ID contributed to writing the paper and conducted data analysis; AR conceived and designed the research and contributed to writing the manuscript; RG conducts testing, evaluation, and review of scientific articles; RWW conducts the final review process and improvement of scientific manuscripts; GFN collected data, conducted experiments, analyzed data, and drafted the initial version of the manuscript; all authors had approved the final version.

Funding

This research funding was supported by Telkom University and Siliwangi University, NO. 375/LIT06/PPM-LIT/2024.

Acknowledgment

We express our gratitude for the completion of this research. Thank you to Telkom University and Siliwangi University, as well as all those who have provided support, guidance, and contributions in the development of Wav2Vec2 and FAdam-based Indonesian Automatic

Speech Recognition (ASR) models. We also appreciate the feedback from the community that helped improve this research.

REFERENCES

[1] R. Prabhavalkar, T. Hori, T. N. Sainath *et al.*, "End-to-end speech recognition: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 325–351, 2024. doi: 10.1109/TASLP.2023.3328283

[2] A. S. Dhanjal and W. Singh, "A comprehensive survey on automatic speech recognition using neural networks," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23367–23412, 2024. doi: 0.1007/s11042-023-16438-y

[3] M. Malik, M. K. Malik, K. Mehmood *et al.*, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 9411–9457, 2021. doi: 10.1007/s11042-020-10073-7

[4] H. Kheddar, M. Hemis, and Y. Himeur, "Automatic speech recognition using advanced deep learning approaches: A survey," *Information Fusion*, vol. 109, 102422, 2024. doi: 0.1016/j.inffus.2024.102422

[5] Z. Qian, K. Xiao, and C. Yu, "A survey of technologies for automatic Dysarthric speech recognition," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, 48, 2023. doi: 10.1186/s13636-023-00318-2

[6] A. Baevski, H. Zhou, A. Mohamed *et al.*, "Wav2Vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. the 34th International Conf. on Neural Information Processing Systems*, NY, 2020, pp. 12449–12460.

[7] S. Schneider, A. Baevski, R. Collobert *et al.*, "Wav2Vec: Unsupervised pre-training for speech recognition," in *Proc. the Annual Conf. of the International Speech Communication Association,* 2019, pp. 3465–3469.

[8] D. Hwang, "FAdam: Adam is a natural gradient optimizer using diagonal empirical Fisher information," arXiv Preprint, arXiv:2405.12807, 2024.

[9] H. Yadav and S. Sitaram, "A survey of multilingual models for automatic speech recognition," in *Proc. of the Thirteenth Language Resources and Evaluation Conf.*, 2022, pp. 5071–5079.

[10] T. F. Abidin, A. Misbullah, R. Ferdhiana *et al.*, "Acoustic model with multiple lexicon types for Indonesian speech recognition," *Applied Computational Intelligence and Soft Computing*, vol. 2022, 2022. doi: 10.1155/2022/3227828

[11] A. Waibel, T. Hanazawa, G. Hinton *et al.*, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989. doi: 10.1109/29.21701

[12] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007. doi: 10.1561/2000000004

[13] R. Yang, J. Yang, and Y. Lu, "Indonesian speech recognition based on deep neural network," in *Proc. 2021 International Conf. on Asian Language Processing (IALP)*, 2021, pp. 36–41. doi: 10.1109/IALP54817.2021.9675280

[14] A. Radford, J. W. Kim, T. Xu *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. the 40th International Conf. on Machine Learning*, 2023, pp. 28492–28518. https://dl.acm.org/doi/10.5555/3618408.3619590

[15] A. Adila, D. Lestari, A. Purwarianti *et al.*, "Enhancing Indonesian automatic speech recognition: Evaluating multilingual models with diverse speech variabilities," arXiv Preprint, arXiv:2410.08828, 2024.

[16] P. Arisaputra and A. Zahra, "Indonesian automatic speech recognition with XLSR-53," *Ingenierie des Systemes d'Information*, vol. 27, no. 6, pp. 973–982, 2022. doi: 10.18280/isi.270614

[17] A. Conneau, A. Baevski, R. Collobert *et al.*, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. the Annual Conf. of the International Speech Communication Association,* 2021, vol. 1, pp. 346–350. doi: 10.21437/Interspeech.2021-329

[18] P. Arisaputra, A. T. Handoyo, and A. Zahra, "XLS-R deep learning model for multilingual ASR on low- resource languages: Indonesian, Javanese, and Sundanese," arXiv Preprint, arXiv:2401.06832, 2024.

[19] A. Babu, C. Wang, A. Tjandra *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. the Annual Conf. of the International Speech Communication Association*, 2022, pp. 2278–2282. doi: 10.21437/Interspeech.2022-143

[20] Mozilla Foundation. (2024). Common Voice Dataset (v20.0). *Mozilla*. [Online]. Available: https://commonvoice.mozilla.org/en/datasets

[21] R. Ardila, M. Branson, K. Davis *et al.*, "Common voice: A massively-multilingual speech corpus," in *Proc. the Twelfth Language Resources and Evaluation Conf.*, 2020, pp. 4218–4222.

[22] M. Labied, A. Belangour, M. Banane *et al.*, "An overview of automatic speech recognition preprocessing techniques," in *Proc. 2022 International Conf. on Decision Aid Sciences and Applications, DASA 2022*, 2022, pp. 804–809. doi: 10.1109/DASA54658.2022.9765043

[23] W. M. Laghari, M. U. Baloch, M. A. Mengal *et al.*, "Performance analysis of analog Butterworth low pass filter as compared to Chebyshev Type-I filter, Chebyshev Type-II filter and elliptical filter," *Circuits and Systems*, vol. 05, no. 09, pp. 209–216, 2014. doi: 10.4236/cs.2014.59023

[24] J. Devlin, M.-W. Chang, K. Lee *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. the 2019 Conf. of the North*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423

[25] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," arXiv Preprint, arXiv:1611.01144, 2016.

[26] P. Khosla, P. Teterwak, C. Wang *et al.*, "Supervised contrastive learning," arXiv Preprint, arXiv:2004.11362, 2020.