Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14257
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Leveraging Label Preprocessing for Effective End-to-End Indonesian Automatic Speech Recognition

Mohammad Noval Althoff[1)*], Affandy[2)], Ardytha Luthfiarta[3)], Mohammad Wahyu Bagus Dwi Satya[4)], Halizah Basiron[5)]
[1,2,3,4,5)] Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia,
[5)]Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia
[1)]novalalthoff@gmail.com, [2)]affandy@dsn.dinus.ac.id, [3)]ardytha.luthfiarta@dsn.dinus.ac.id,
[4)]bagus8454@gmail.com , [5)]halizah@utem.edu.my

Submitted : Nov 25, 2024 | **Accepted** : Dec 9, 2024 | **Published** : Jan 8, 2025

Abstract: This research explores the potential of improving low-resource Automatic Speech Recognition (ASR) performance by leveraging label preprocessing techniques in conjunction with the wav2vec2-large Self-Supervised Learning (SSL) model. ASR technology plays a critical role in enhancing educational accessibility for children with disabilities in Indonesia, yet its development faces challenges due to limited labeled datasets. SSL models like wav2vec 2.0 have shown promise by learning rich speech representations from raw audio with minimal labeled data. Still, their dependence on large datasets and significant computational resources limits their application in low-resource settings. This study introduces a label preprocessing technique to address these limitations, comparing three scenarios: training without preprocessing, with the proposed preprocessing method, and with previous method. Using only 16 hours of labeled data, the proposed preprocessing approach achieves a Word Error Rate (WER) of 15.83%, significantly outperforming the baseline scenario (33.45% WER) and the previous preprocessing method (19.62% WER). Further training using the proposed preprocessing technique with increased epochs reduces the WER to 14.00%. These results highlight the effectiveness of label preprocessing in reducing data dependency while enhancing model performance. The findings demonstrate the feasibility of developing robust ASR models for low-resource languages, offering a scalable solution for advancing ASR technology and improving educational accessibility, particularly for underrepresented languages.

Keywords: Automatic Speech Recognition (ASR); Label Preprocessing; Low-Resource Language; Self-Supervised Speech Representation Learning; wav2vec 2.0

INTRODUCTION

Access to quality education remains a significant challenge for many children in Indonesia, with the issue being particularly pronounced for children with disabilities. According to the (WHO & UNICEF, 2023), approximately 30% of children with disabilities in Indonesia have never attended school, compared to nearly universal access among children without disabilities. This stark disparity underscores the pressing need for innovative strategies to promote equitable learning opportunities. One potential solution lies in leveraging Automatic Speech Recognition (ASR) technology, which can transcribe spoken language into text in real time (Oh & Song, 2021). This capability has demonstrated potential in enhancing educational accessibility for children with hearing impairments and learning difficulties. For instance, ASR has been found to improve reading comprehension in children with dyslexia and foster more inclusive educational experiences in diverse learning contexts (Ashshidiqi & Wijiastuti, 2020).

Despite its promise, development of effective ASR systems for low-resource languages like Indonesian poses significant challenges. ASR performance depends heavily on high-quality, labeled speech data, which is limited for the Indonesian language (Aji et al., 2022). Furthermore, Indonesian is characterized by linguistic diversity, including numerous dialects, tonal variations, and spontaneous speech patterns, making it particularly complex for model training and deployment. Consequently, existing ASR systems for Indonesian often exhibit suboptimal accuracy, especially in practical applications. To mitigate these challenges, researchers are exploring advanced

*Mohammad Noval Althoff

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14257
e-ISSN : 2541-2019
p-ISSN : 2541-044X

methodologies such as data augmentation (Nugroho et al., 2022), synthetic data generation (Rossenbach et al., 2020), and transfer learning techniques (Mamyrbayev et al., 2022). These approaches aim to enhance model performance in resource-constrained settings by maximizing the utility of available data.

The advent of Self-Supervised Learning (SSL) in speech representation has further revolutionized ASR development. Models such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and data2vec (Baevski et al., 2022) have demonstrated remarkable success by extracting rich representations directly from raw audio, reducing dependence on large labeled datasets. These SSL models rely on extensive pre-training with unlabeled audio to uncover intricate acoustic and linguistic patterns, which can subsequently be fine-tuned using smaller labeled datasets for specific ASR tasks. However, implementing SSL-based approaches often demands significant computational resources, particularly when processing large audio datasets, posing a considerable challenge in low-resource environments.

Building on recent advancements in SSL, this study explores the use of an End-to-End (E2E) Automatic Speech Recognition (ASR) model, specifically wav2vec2-large, to address the challenges of developing ASR systems for low-resource languages such as Indonesian. Recognizing the constraints posed by limited high-quality labeled datasets, the study introduces a novel transcription preprocessing technique designed to enhance the quality and reduce noise in training data, thereby optimizing the fine-tuning process. This method enables the model to more effectively learn linguistic and acoustic patterns from limited resources, significantly improving transcription accuracy and robustness, as supported by findings in prior research (Dubey & Shah, 2022).

A key feature of the proposed preprocessing method is its ability to refine the alignment between raw audio inputs and their corresponding textual labels. This alignment not only enhances the efficiency of SSL models in resource-constrained environments but also creates a scalable framework for improving accessibility in education and other critical applications. By addressing the limitations of existing ASR systems, particularly in low-resource settings, this approach highlights the transformative potential of combining innovative data preprocessing with cutting-edge SSL technologies.

The findings of this research emphasize the critical role of high-quality data preparation in maximizing the capabilities of modern ASR models. Moreover, the study provides actionable insights into the development of scalable, robust solutions to support equitable access to education and information in Indonesia. This contribution represents a significant step forward in overcoming the barriers to effective ASR implementation in low-resource languages, with implications for broader applications in education and beyond.

LITERATURE REVIEW

The evolution of ASR research has been marked by significant technological advancements over the years. In its early popular stages, ASR systems relied heavily on statistical methods, with the Acoustic Model (AM) and Hidden Markov Models (HMM) forming the backbone of speech-to-text conversion. These systems paired the AM with a Language Model (LM) to map speech sounds to words, achieving notable success but struggling with complex languages and spontaneous speech. The advent of deep learning brought a paradigm shift, as neural networks like Time Delay Neural Networks (TDNN) and Recurrent Neural Networks (RNN) improved the ability to model acoustic and linguistic patterns.

(Abidin et al., 2020) compared traditional statistical models, such as the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM), with modern approaches like Time Delay Neural Network Factorization (TDNNF). Using a dataset collected from YouTube channels, they demonstrated that the GMM-HMM model achieved its best performance at 19.28% WER, while the TDNNF model trained on a validated dataset outperformed it, achieving 11.35% WER. This highlights the importance of dataset quality and the effectiveness of deep learning for ASR. Building on this, (Abidin et al., 2022) extended their work by comparing GMM-HMM, TDNNF, and CNN-TDNNF-augmented models using a larger dataset of 181 hours of Indonesian audio and 215,291 utterances. The CNN-TDNNF model achieved the lowest WER of 19.03%, compared to 24.96% for TDNNF and 40.85% for GMM-HMM, showcasing the potential of deep learning in capturing complex acoustic and linguistic features.

Dialectal variation presents additional challenges for ASR systems, requiring models to account for diverse linguistic patterns. (Tawaqal & Suyanto, 2021) addressed this by utilizing Mel Frequency Cepstral Coefficients (MFCC) for feature extraction and Deep Recurrent Neural Networks (DRNN) for classifying five major Indonesian dialects: Javanese, Sundanese, Banjar, Bugis, and Malay. The dataset includes recordings from 500 speakers, evenly distributed across gender and age groups, with each speaker contributing 100 sentences. Experimental results show that the optimized configuration of 30 epochs and a batch size of 30 achieves the highest training accuracy of 89% and testing accuracy of 85.4%. These findings demonstrate the potential of DRNN in handling multi-dialect classification tasks.

More recently, End-to-End approaches have gained traction, where a single model replaces the traditional AM-LM pipeline. Innovations like the Transformer architecture and Self-Supervised Speech Representation Learning models, such as wav2vec 2.0 and HuBERT, have further revolutionized ASR by learning speech patterns from raw audio with minimal labeled data. For instance, Transformers have been effectively applied to improve

*Mohammad Noval Althoff

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14257
e-ISSN : 2541-2019
p-ISSN : 2541-044X

the readability of ASR outputs by restoring punctuation and capitalization, achieving high performance for language like Vietnamese (Nguyen et al., 2021). Similarly, Transformers have demonstrated success in language understanding tasks, as highlighted by (Devlin et al., 2019).

Despite these advancements, significant challenges remain, particularly for low-resource languages. According to (Chen et al., 2019), over 95% of the world's languages are low-resource, with insufficient annotated speech corpora to train effective ASR models. E2E ASR systems, while powerful, often require vast datasets to train their acoustic and language components. As a result, developing robust ASR systems for low-resource languages remains a considerable hurdle, with models frequently underperforming due to the scarcity of training data.

Label preprocessing has emerged as an effective strategy to address these challenges by optimizing the quality of training datasets. Inconsistent transcription labels—caused by variations in special characters, diacritical marks, capitalization, and spacing—introduce noise into the training process, reducing ASR accuracy. Prior study (Dubey & Shah, 2022) have demonstrated that preprocessing transcription data to normalize and simplify labels can enhance model performance.

In the context of Indonesian ASR, previous research (Suyanto et al., 2020) employed an E2E approach using 33 hours of labeled Indonesian audio, achieving an average WER of 22% after extensive hyperparameter tuning. While this demonstrates the potential of E2E ASR, it also underscores the practical challenges of data scarcity in low-resource environments, where acquiring sufficient labeled datasets is often difficult and resource-intensive. Similarly, (Arisaputra & Zahra, 2022) demonstrated the benefit of integrating a language model with the XLSR-53, an E2E cross-lingual wav2vec 2.0 model pre-trained on 53 languages. By fine-tuning 24 hours of labeled Indonesian audio and incorporating a 5-gram KenLM language model, they reduced WER from 20% to 12%.

These studies collectively highlight the need for innovative techniques to address data scarcity in low-resource settings. Therefore, this research explores the use of a self-supervised ASR model, *wav2vec2-large*, coupled with a novel label preprocessing method. While the specifics of this preprocessing method are described in the chapter below, its theoretical basis lies in its ability to reduce noise, enhance label consistency, and improve the generalization capability of ASR models. By comparing models trained with and without label preprocessing, this study aims to evaluate the impact of this proposed technique on ASR performance.

METHOD

The research methodology is shown in Figure 1. The process begins with data collection, followed by preprocessing to ensure data quality. The dataset is then divided into three distinct sets: training, validation, and testing. During the training stage, the study examines the impact of different label preprocessing techniques on ASR performance. Finally, model evaluation is conducted to quantify transcription accuracy and compare the results across preprocessing scenarios.

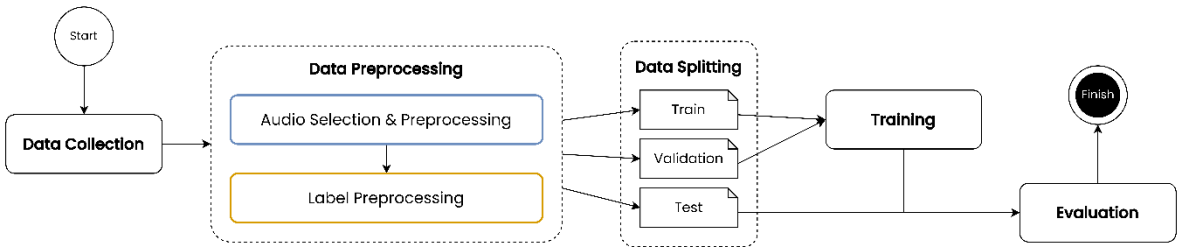

Figure 1. Research Workflow

Data Collection

In this research, we collected the Indonesian audio datasets from Mozilla Common Voice (MCV) (Ardila et al., 2020) and Google Fleurs (GF) (Conneau et al., 2023) as primary resources for fine-tuning our Self-Supervised Speech Representation Learning (SSL) model. MCV provides diverse audio samples contributed by volunteers, representing a wide range of demographic backgrounds and regional accents, which helps capture the phonetic richness of the Indonesian language. GF, a multilingual dataset created by Google, further supplements this variety with high-quality recordings and consistent transcription accuracy, essential for building a robust and adaptable ASR model.

Audio Selection and Preprocessing

The audio selection process in this study involved two main steps to ensure quality and representation within the dataset. First, audio samples with loudness levels below -40 LKFS (Loudness K-weighted Full Scale) were discarded, as these recordings tend to have lower clarity and can introduce noise that may hinder model

*Mohammad Noval Althoff

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14257
e-ISSN : 2541-2019
p-ISSN : 2541-044X

performance. Second, to achieve gender balance among speakers, the dataset was adjusted to include an equal number of samples from male and female speakers, which is crucial for reducing gender bias in model recognition accuracy. These steps resulted in a more balanced, high-quality dataset, enhancing the model's robustness and fairness across different speaker profiles.

After selection processes were performed, all audio data underwent a frequency resampling step to ensure consistency with the input requirements of the ASR model. Each audio sample was resampled to a specific frequency using *librosa: 0.10.2.post1* (McFee et al., 2024), the most recent update of a widely used tool for audio and music analysis. Resampling to a uniform frequency is needed, as it aligns all audio inputs to the model's expected format, improving processing efficiency and recognition accuracy.

Label Preprocessing

Label preprocessing is the highlighted step in this study, as it aims to enhance the ASR model's performance by refining transcription labels to create cleaner, more consistent data for fine-tuning. It plays a central role in comparing the model's effectiveness with different preprocessing techniques, as the study explores three scenarios: (S1) fine-tuning the model without any label preprocessing, (S2) using 3-steps previous label preprocessing technique used by (Dubey & Shah, 2022), and (S3) applying 5-steps label preprocessing technique as proposed. Given that end-to-end ASR models typically rely on self-supervised learning, they often require vast amounts of unlabeled data, leading to high resource demands during training. By integrating label preprocessing, this study proposes a more efficient approach that allows robust ASR model development with less data, albeit labeled.

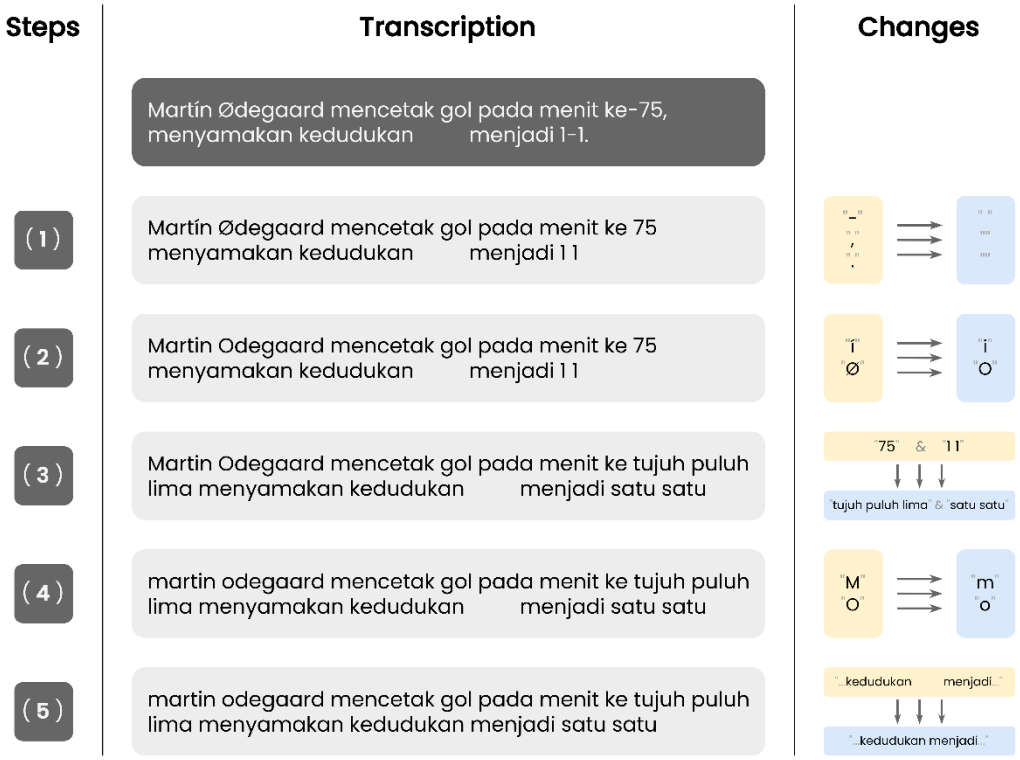

Figure 2. Illustration of Step-by-Step Proposed Technique Results

As shown in Figure 2, the proposed label preprocessing technique itself comprises five key steps: (1) removing special characters to eliminate foreign symbols that might confuse the model, (2) mapping letters with diacritical marks (e.g., converting "ã" to "a") to simplify character variations, (3) converting numbers into their string representations (e.g., "11" to "eleven") for consistent interpretation, (4) converting all characters to lowercase to maintain uniformity, and (5) removing extra spaces to avoid processing errors. Table 1 below describes the steps comparison between the Dubey & Shah (previous) technique and the proposed technique.

Table 1. Comparison of Technique Steps

Previous Technique	Proposed Technique
(1) Remove special characters	(1) Remove special characters

*Mohammad Noval Althoff

(2) Remove extra spaces	(2) Remap letters marked with diacritics
(3) Lowercase characters	(3) Convert numbers into their string representations
	(4) Lowercase characters
	(5) Remove extra spaces

Data Splitting

The final data is divided into three distinct sets to facilitate effective model development and evaluation: a training set (80%), a validation set (10%), and a test set (10%). Additionally, each set is balanced to ensure fair distribution of speaker gender, with approximately 50% male and 50% female representation. The training set is used to adjust the model, allowing it to learn patterns and features from the data. The validation set is used to fine-tune model parameters and prevent overfitting by evaluating performance during the training process. Finally, the test set provides an unbiased assessment of the model's accuracy on unseen data, ensuring a reliable measure of its generalization ability in real-world applications.

Training

This research utilizes a derivative E2E ASR model from the wav2vec 2.0 framework, specifically *wav2vec2-large*, to analyze the impact of label preprocessing on model performance. Transformer-based models, like the wav2vec 2.0, are used in automatic speech recognition due to their ability to efficiently handle sequential data, such as audio recordings, with variable lengths. The transformer architecture, introduced in "Attention Is All You Need," enables the model to access and emphasize different parts of the input sequence through a self-attention mechanism. This mechanism allows the model to selectively focus on certain parts of the input, assigning attention weights based on the relevance of each segment. Multi-headed attention further enhances this by enabling parallel focus on multiple parts of the input, capturing complex relationships within the data.

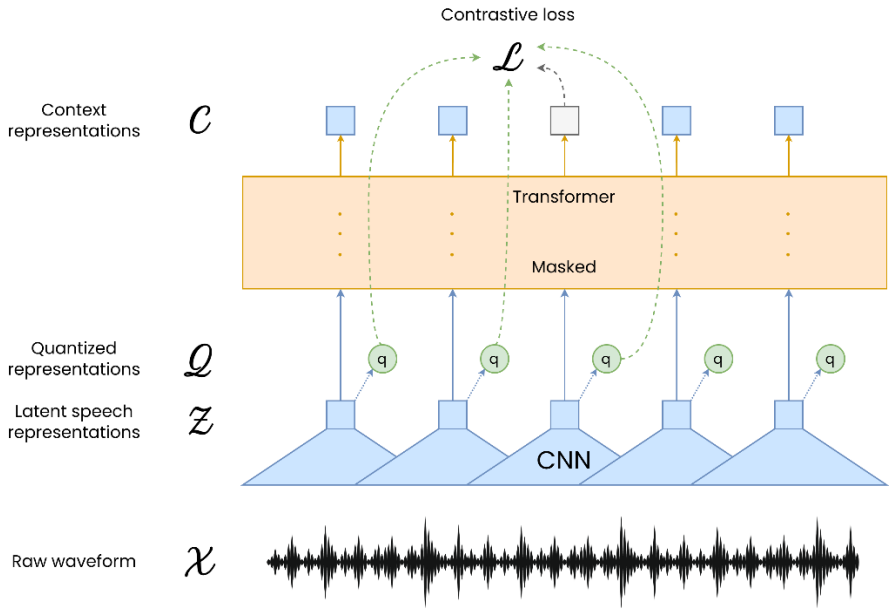

Figure 3. Illustration of the wav2vec 2.0 architecture

As also introduced in (Baevski et al., 2020), The wav2vec 2.0 architecture processes raw audio in three main stages: a feature encoder, a context network, and a quantization module. The feature encoder takes raw audio, normalizes it, and passes it through multiple convolutional layers with layer normalization and GELU activation to produce compact latent speech representations. These representations are fed into a context network based on a Transformer, which uses self-attention to capture dependencies across the entire audio sequence while incorporating relative positional information through a convolutional layer. For self-supervised learning, the model uses a quantization module to discretize the encoder's output into a set of discrete units. This is achieved via product quantization, where the output is divided into groups (codebooks), and discrete entries are selected using Gumbel-Softmax. Together, these components allow wav2vec 2.0 to learn powerful, contextualized speech representations end-to-end, excelling in tasks with limited labeled data.

*Mohammad Noval Althoff

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14257
e-ISSN : 2541-2019
p-ISSN : 2541-044X

To further process the sequential outputs and map them to corresponding text transcriptions, the model utilizes Connectionist Temporal Classification (CTC) as a loss function and decoding method. CTC aligns the input audio features with the output text sequence without requiring pre-aligned data, making it ideal for speech recognition tasks. By introducing a special blank token, CTC enables the model to handle variable-length inputs and outputs, collapsing repeated predictions and blanks into meaningful sequences. For instance, the model's output sequence "$aa\emptyset b\emptyset bccc\emptyset$" is simplified to "$abc$". Interestingly, in contrast to other label topologies, CTC does not require alignments with single label emissions per label position during training. However, throughout training, CTC models exhibit a property known as peaky behavior (Zeyer, Schlüter, & Ney, 2021), where the model eventually emits single-labels per hypothesized position. (Laptev, Majumdar, & Ginsburg, 2022) even introduces a CTC variant that eliminates non-blank loop transitions.

Evaluation

In the evaluation stage, Word Error Rate (WER) is used as the primary metric to assess the ASR model's performance. WER is a widely recognized metric in speech recognition (Malik et al., 2021) that measures the accuracy of transcribed text by calculating the number of errors relative to the total number of words in the reference transcription.

$$WER = \frac{S+I+D}{N} \quad (1)$$

On formula (1) above: ***S*** *(Substitutions)* refers to instances where an incorrect word replaces a correct word in the transcription, ***I*** *(Insertions)* are extra words inserted in the transcription that are not present in the reference, ***D*** *(Deletions)* are words missing from the transcription that are present in the reference, and ***N*** is the total number of words in the reference. WER enables a quantitative evaluation of the model's transcription accuracy, with lower WER scores indicating higher accuracy and more reliable ASR performance.

RESULT

Data Preparation

Data was collected from two primary sources: Mozilla Common Voice (MCV) and Google Fleurs (GF). These datasets have distinct data structures and characteristics. MCV provides a larger set of 11,960 audio samples with an average duration of 6.39 seconds, amounting to a cumulative total of 21 hours and 13 minutes. In contrast, GF consists of 3,616 utterances with an average duration of 12.56 seconds, resulting in a total of 12 hours and 37 minutes of audio data. Given these structural differences, a series of selection and preprocessing steps were carried out to prepare a well-balanced, high-quality dataset for training. This involved ensuring uniform loudness, balancing the dataset's duration, gender distribution, and harmonizing the audio sample rate to match the input requirements of the model.

To ensure audio quality, samples with low loudness levels were removed. Any audio with a loudness below -40 LKFS was discarded, resulting in the removal of 111 samples from MCV and 466 from GF. After filtering, further steps were taken to balance both duration and speaker gender across the datasets. Since the initial data structures were uneven—particularly in duration—the MCV and GF datasets were each down-sampled to 8 hours, bringing the total combined duration to 16 hours of Indonesian audio. Additionally, to ensure gender balance, the final dataset retained an equal distribution of male and female speakers from each source, with 1,830 males and 1,873 females in MCV and 1,150 males and 1,107 females in GF. The final combined dataset, therefore, included 2,980 male and 2,980 female speakers, ensuring an equitable representation. Since the Mozilla Common Voice dataset originally used a sample rate of 48 kHz, it was necessary to resample this audio to 16 kHz to meet the input requirements of the wav2vec 2.0 model used in this research.

These preprocessing steps resulted in a dataset that is balanced in both duration and gender representation, with audio quality tailored to the requirements of the ASR model, providing a strong foundation for the model's development. The final data structures are presented in Table 2.

Table 2. Post-Preparation Data Structures

Attribute	Dataset		
	MCV	GF	Combined
Utterance count	3703	2257	5960
Averege record length *(s)*	7.78	12.77	9.67
Total duration *(HH:mm:ss)*	07:59:52	08:00:14	16:00:06
Gender representation	M: 1830	M: 1150	M: 2980

*Mohammad Noval Althoff

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14257
e-ISSN : 2541-2019
p-ISSN : 2541-044X

	F: 1873	F: 1107	F: 2980

M (Male)
F (Female)

Label Preprocessing Overview

The label preprocessing step is central to this study, as it directly impacts the ASR model's ability to achieve optimal performance in low-resource settings. By refining transcription labels, this process ensures cleaner and more consistent data for fine-tuning. The following table showcases a step-by-step comparison of sample results between the previous technique (Scenario 2) and the proposed technique (Scenario 3).

Table 3. Step-by-Step Label Preprocessing Results

Raw Text			
"Perangko ke-1000-nya adalah "Karya Besar Raja Swedia" oleh David KlÃ¶cker Ehrenstrahl di tahun 2000, yang terdaftar di Buku Rekor Dunia Guinness."			
Previous Technique (S2)		**Proposed Technique** (S3)	
Step	**Result**	**Step**	**Result**
(1)	Perangko ke 1000 nya adalah Karya Besar Raja Swedia oleh David KlÃcker Ehrenstrahl di tahun 2000 yang terdaftar di Buku Rekor Dunia Guinness	(1)	Perangko ke 1000 nya adalah Karya Besar Raja Swedia oleh David KlÃcker Ehrenstrahl di tahun 2000 yang terdaftar di Buku Rekor Dunia Guinness
(2)	Perangko ke 1000 nya adalah Karya Besar Raja Swedia oleh David KlÃcker Ehrenstrahl di tahun 2000 yang terdaftar di Buku Rekor Dunia Guinness	(2)	Perangko ke 1000 nya adalah Karya Besar Raja Swedia oleh David KlAcker Ehrenstrahl di tahun 2000 yang terdaftar di Buku Rekor Dunia Guinness
(3)	perangko ke 1000 nya adalah karya besar raja swedia oleh david klãcker ehrenstrahl di tahun 2000 yang terdaftar di buku rekor dunia guinness	(3)	Perangko ke seribu nya adalah Karya Besar Raja Swedia oleh David KlAcker Ehrenstrahl di tahun dua ribu yang terdaftar di Buku Rekor Dunia Guinness
		(4)	perangko ke seribu nya adalah karya besar raja swedia oleh david klacker ehrenstrahl di tahun dua ribu yang terdaftar di buku rekor dunia guinness
		(5)	perangko ke seribu nya adalah karya besar raja swedia oleh david klacker ehrenstrahl di tahun dua ribu yang terdaftar di buku rekor dunia guinness
Clean Text Comparison			
perangko ke 1000 nya adalah karya besar raja swedia oleh david klãcker ehrenstrahl di tahun 2000 yang terdaftar di buku rekor dunia guinness		perangko ke seribu nya adalah karya besar raja swedia oleh david klacker ehrenstrahl di tahun dua ribu yang terdaftar di buku rekor dunia guinness	

Training and Evaluation

wav2vec2-large is utilized during the training stage to evaluate the impact of label preprocessing techniques on Indonesian speech recognition. Three distinct training scenarios are implemented: (S1) training without any label preprocessing, (S2) training with previous label preprocessing method done by (Dubey & Shah, 2022), and (S3) training with the proposed label preprocessing technique.

Each of these scenarios is evaluated by measuring their performance using Word Error Rate (WER) and the results are then compared to determine the relative effectiveness of each training approach. (S3), incorporating the proposed label preprocessing technique, is expected to significantly improve the WER compared to (S1), where no preprocessing is applied. Moreover, the proposed method aims to outperform the previous preprocessing approach in (S2), demonstrating its superior capability to optimize transcription quality and model learning in low-resource settings.

By systematically comparing these scenarios, this evaluation seeks to validate the effectiveness of the proposed label preprocessing technique in enhancing ASR model performance while highlighting its potential as a more robust and efficient solution for low-resource speech recognition tasks.

*Mohammad Noval Althoff

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14257
e-ISSN : 2541-2019
p-ISSN : 2541-044X

DISCUSSIONS

As previously mentioned, this study evaluated three training scenarios to assess the impact of label preprocessing on the performance of SSL ASR model: (S1) training without label preprocessing, (S2) training with previous preprocessing method conducted by (Dubey & Shah, 2022), and (S3) training with the proposed label preprocessing technique. The evaluation results for each scenario are presented in the table below.

Table 4. Evaluation of The Training Scenarios

Scenario	**WER** (%)
(S1)	33.45
(S2)	19.62
(S3)	**15.83**

The results demonstrate that the proposed label preprocessing method significantly enhances performance compared to the baseline scenario (S1), reducing the WER from 33.45% to 15.83%. This improvement highlights the effectiveness of the preprocessing technique in optimizing the quality of training data and improving the ASR model's ability to generalize. Furthermore, the proposed method outperforms the previous label preprocessing in scenario (S2), which achieved a WER of 19.62%. These findings underscore the superiority of the proposed preprocessing technique in handling low-resource settings for ASR development.

The differences in WER between scenarios can be attributed to several key factors. First, the proposed label preprocessing method focuses on reducing inconsistencies and noise in the training data, enabling the model to better capture linguistic patterns and align with audio features. In contrast, the baseline scenario (S1) lacks any preprocessing, leaving errors and inconsistencies unaddressed, which hinders the model's learning process. The prior method used in scenario (S2), while partially effective, relies on basic filtering and normalization steps that fail to address more nuanced linguistic irregularities.

The findings of this study have important implications for advancing ASR technologies in low-resource languages like Indonesian. The demonstrated reduction in WER through label preprocessing highlights the potential for improved ASR performance even when large-scale labeled datasets are unavailable. For low-resource languages, where collecting extensive annotated data is both costly and time-intensive, label preprocessing offers a practical solution to enhance the utility of existing datasets. Furthermore, this research underscores the importance of tailoring preprocessing techniques to the linguistic characteristics of the target language. For example, the Indonesian language's relatively simple morphology and phoneme structure benefited from preprocessing steps that minimized transcription inconsistencies and improved phoneme alignment.

Table 5. Impact of Increased Epochs on ASR Performance

Epoch	**WER** (%)	**Training Time**
50	15.83	6 hours, 24 minutes
100	**14.00**	13 hours, 24 minutes

After training with an increased number of epochs, the ASR model further demonstrated performance improvements, with the WER reducing from 15.83% at 50 epochs to 14.00% at 100 epochs. However, this improvement came at the cost of nearly doubling the training time. This suggests that while additional training epochs can further refine model performance, there is a trade-off with computational efficiency that must be considered in low-resource environments.

The primary limitation of this study lies in its focus on the Indonesian language, which may limit the generalizability of the findings to other low-resource languages. Each language has unique linguistic and phonetic characteristics that could influence the effectiveness of the proposed label preprocessing technique. For example, tonal languages, agglutinative languages, or languages with more complex morphologies may require different preprocessing strategies to achieve similar improvements. While the results demonstrate significant gains in ASR performance for Indonesian, future research should expand the scope to include multiple languages with diverse linguistic and structural properties. Doing so would validate the versatility of the proposed method and identify any language-specific adaptations that may be needed. This line of inquiry is critical for the broader application of ASR technology in global low-resource language contexts.

CONCLUSION

This study demonstrates the effectiveness of leveraging label preprocessing to enhance the performance of low-resource self-supervised speech representation learning for Automatic Speech Recognition (ASR). By comparing three training scenarios—without preprocessing, with previous preprocessing method, and with the proposed preprocessing technique—the proposed approach significantly reduced the Word Error Rate (WER) and

*Mohammad Noval Althoff

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14257
e-ISSN : 2541-2019
p-ISSN : 2541-044X

outperformed previous methods in similar low-resource contexts. Notably, the proposed technique achieved a WER of 14.00% while utilizing 16 hours of labeled data. These findings highlight the potential of label preprocessing to optimize ASR model performance in low-resource settings by improving data efficiency and reducing the reliance on extensive labeled datasets. This approach offers a promising pathway for developing robust ASR models for low-resource languages, contributing to more accessible and equitable language technology solutions.

REFERENCES

Abidin, T. F., Misbullah, A., Ferdhiana, R., Aksana, M. Z., & Farsiah, L. (2020). Deep Neural Network for Automatic Speech Recognition from Indonesian Audio using Several Lexicon Types. *2020 International Conference on Electrical Engineering and Informatics (ICELTICs)*, *IEEE*, 1–5. https://doi.org/10.1109/ICELTICs50595.2020.9315538

Abidin, T. F., Misbullah, A., Ferdhiana, R., Farsiah, L., Aksana, M. Z., & Riza, H. (2022). Acoustic Model with Multiple Lexicon Types for Indonesian Speech Recognition. *Applied Computational Intelligence and Soft Computing*. https://doi.org/10.1155/2022/3227828

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., … Ruder, S. (2022). One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*, 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., … Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of the 12th International Language Resources and Evaluation Conference (LREC)*, 4218–4222. https://doi.org/10.48550/arXiv.1912.06670

Arisaputra, P., & Zahra, A. (2022). Indonesian Automatic Speech Recognition with XLSR-53. *Ingenierie Des Systemes d'Information*, *27*(6), 973–982. https://doi.org/10.18280/isi.270614

Ashshidiqi, M. H., & Wijiastuti, A. (2020). Teknologi Asistif Text To Speech (TTS) Pada Kemapuan Membaca pemahaman Anak Disleksia. *Jurnal Pendidikan Khusus*, *15*(1).

Baevski, A., Hsu, W. N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. *Proceedings of the 39th International Conference on Machine Learning Research (PMLR)*, *162*, 1298–1312. https://doi.org/10.48550/arXiv.2202.03555

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*, 12449-12460. https://doi.org/10.48550/arXiv.2006.11477

Chen, K., Tsai, C., Liu, D., Lee, H., & Lee, L. (2019). Completely Unsupervised Speech Recognition By A Generative Adversarial Network Harmonized With Iteratively Refined Hidden Markov Models. *arXiv e-print*, arXiv-1904. https://doi.org/10.48550/arXiv.1904.04100

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., … Bapna, A. (2023). FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech. *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, 798–805. https://doi.org/10.1109/SLT54892.2023.10023141

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dubey, P., & Shah, B. (2022). Deep Speech Based End-to-End Automated Speech Recognition (ASR) for Indian-English Accents. *arXiv e-print*, arXiv-2204. https://doi.org/10.48550/arXiv.2204.00977

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *29*, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

Laptev, A., Majumdar, S., & Ginsburg, B. (2022). CTC Variations Through New WFST Topologies. *Proceedings*

*Mohammad Noval Althoff

of the Annual Conference of the International Speech Communication Association (Interspeech), 1041–1045. https://doi.org/10.21437/Interspeech.2022-10854

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, *80*(6), 9411–9457. https://doi.org/10.1007/s11042-020-10073-7

Mamyrbayev, O., Alimhan, K., Oralbekova, D., Bekarystankyzy, A., & Zhumazhanov, B. (2022). Identifying The Influence Of Transfer Learning Method In Developing An End-To-End Automatic Speech Recognition System With A Low Data Level. *Eastern-European Journal of Enterprise Technologies*, *1*(9 (115)), 84–92. https://doi.org/10.15587/1729-4061.2022.252801

McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., … Pimenta, W. (2024). *librosa/librosa: 0.10.2.post1*. Zenodo. https://doi.org/10.5281/zenodo.11192913

Nguyen, T. T. H., Nguyen, T. B., Pham, N. P., Do, Q. T., Le, T. L., & Luong, C. M. (2021). Toward Human-Friendly ASR Systems: Recovering Capitalization and Punctuation for Vietnamese Text. *IEICE TRANSACTIONS on Information and Systems*, *104*(8), 1195–1203. https://doi.org/10.1587/transinf.2020BDP0005

Nugroho, K., Noersasongko, E., Ignatius, D. R., & Setiadi, M. (2022). Enhanced Indonesian Ethnic Speaker Recognition using Data Augmentation Deep Neural Network. *Journal of King Saud University - Computer and Information Sciences*, *34*(7), 4375–4384. https://doi.org/10.1016/j.jksuci.2021.04.002

Oh, E. Y., & Song, D. (2021). Developmental research on an interactive application for language speaking practice using speech recognition technology. *Educational Technology Research and Development*, *69*(2), 861–884. https://doi.org/10.1007/s11423-020-09910-1

Rossenbach, N., Zeyer, A., Schlüter, R., & Ney, H. (2020). Generating Synthetic Audio Data for Attention-Based Speech Recognition Systems. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7069–7073. https://doi.org/10.1109/ICASSP40776.2020.9053008

Suyanto, S., Arifianto, A., Sirwan, A., & Rizaendra, A. P. (2020). End-to-End Speech Recognition Models for a Low-Resourced Indonesian Language. *2020 8th International Conference on Information and Communication Technology (ICoICT), IEEE*, 1-6. https://doi.org/10.1109/ICoICT49345.2020.9166346

Tawaqal, B., & Suyanto, S. (2021). Recognizing Five Major Dialects in Indonesia Based on MFCC and DRNN. *Journal of Physics: Conference Series*, *IOP Publishing*, *1844*(1). https://doi.org/10.1088/1742-6596/1844/1/012003

WHO, & UNICEF. (2023). *From the margins to the mainstream Executive summary Global report on children with developmental disabilities Executive summary*.

Zeyer, A., Schlüter, R., & Ney, H. (2021). Why does CTC result in peaky behavior? *arXiv e-print*, arXiv-2105. https://doi.org/10.48550/arXiv.2105.14849

*Mohammad Noval Althoff