





Orang Rimba Language Speech Recognition with XLS-R

Azinurrachman Maulana^{*}, Amalia Zahra[†]

Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Corresponding Author Email: azinurrachman.maulana@binus.ac.id

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.290604>

ABSTRACT

Received: 18 July 2024

Revised: 13 November 2024

Accepted: 28 November 2024

Available online: 25 December 2024

Keywords:

speech recognition, Cross-lingual Speech Representation (XLS-R), Orang Rimba language, n-gram language model

The Orang Rimba, an indigenous tribe in Indonesia, has a unique language central to their cultural identity. However, like many indigenous languages globally, the Orang Rimba language faces challenges from deforestation, economic development, and cultural assimilation, leading to its declining use. This study proposes the use of speech recognition, specifically the XLS-R model as a solution. The XLS-R model, designed for effectiveness even with limited linguistic resources, can help recognize and transcribing the Orang Rimba language, aiding its preservation and documentation for future generations. The initial phase involved training the XLS-R model with 3 hours and 40 minutes of Orang Rimba audio data, resulting in a baseline performance with a Word Error Rate (WER) of 22.59%. To improve accuracy, data augmentation was employed to expand the training dataset to 22 hours and 3 minutes, reducing the WER to 18.39%. Furthermore, the integration of an n-gram language model significantly enhanced the model's performance, further lowering the WER to 9.66%.

1. INTRODUCTION

The Orang Rimba is a unique indigenous group inhabiting the dense rainforests of Jambi, Riau, and South Sumatra in Indonesia. Living a nomadic lifestyle deeply entwined with the forest, they depend on hunting, gathering, and trading forest products for their livelihood. This connection with nature is not just a matter of survival but is intrinsic to their cultural identity and way of life [1].

The Orang Rimba has their own distinctive language, which is an essential part of their cultural identity. This language has unique vocabulary and grammar that reflect their close relationship with the forest environment. It serves as a crucial communication tool within the community and is a repository of their knowledge, history, and cultural practices.

However, like many tribal languages around the world, the Orang Rimba language faces threats. Jambi, Indonesia, the region where the Orang Rimba resides, has experienced extensive deforestation. According to Global Forest Watch, from 2001 to 2023, Jambi lost around 1.91 million hectares of tree cover, representing a 43% decrease in forested area since 2000. This loss, primarily due to logging and agricultural expansion, disrupts the Orang Rimba's traditional lifestyle and often forces them into urban areas or plantations. Such displacement diminishes their ability to use and pass down the Orang Rimba language within their community. Figure 1 shows Jambi's annual tree cover loss from 2001 to 2023, illustrating the significant deforestation trends impacting the region.

In schools, the children encounter educational policies that prioritize Bahasa Indonesia as the primary medium of

instruction, with little to no support for indigenous languages [2]. This lack of institutional support means that Orang Rimba children have limited opportunities to learn and practice their native language in formal educational settings, further diminishing their proficiency and connection to their cultural heritage. Preserving and documenting the Orang Rimba language is therefore essential—not only for maintaining the language itself but also for safeguarding the cultural heritage and identity of the Orang Rimba people.

Speech recognition can act as a tool to preserve the Orang Rimba language. Speech recognition is a field within computer science aimed at developing systems that can understand spoken language. These systems can convert spoken language into written text, allowing for the creation of written records of predominantly spoken and unwritten languages, like the Orang Rimba language [3].

However, the development of speech recognition systems for such unique languages presents its own set of challenges. The primary obstacle is the scarcity of audio and textual data in the Orang Rimba language, which is necessary for training accurate speech recognition models. Most existing models are trained on languages with abundant resources, leaving underrepresented languages like the Orang Rimba's at a disadvantage.

XLS-R is a speech recognition model capable of understanding multiple languages, developed by Facebook's AI team. An advantage of XLS-R, especially in the context of the Orang Rimba language, is its exposure to a vast range of languages during training. The representations learned from languages like Indonesian and Malay, which are geographically and linguistically close to the Orang Rimba,

can potentially enhance the model's performance for the Orang Rimba language. XLS-R is designed to operate effectively even with limited resources. The model employs a technique known as self-supervised learning, allowing it to learn useful representations from unlabelled data [4].

To enhance the effectiveness of the model, this study incorporates the strategy of data augmentation. Data augmentation is a crucial technique in machine learning, particularly valuable when dealing with limited datasets, as is the case with the Orang Rimba language [5]. It involves artificially expanding the dataset by creating modified versions of the existing data, thereby increasing the diversity and quantity of training samples. This process can be achieved through various methods, such as adding noise, varying pitch, and speed, or even simulating different acoustic environments [6].

Additionally, integrating a language model could significantly augment the capabilities of the model in processing the language. A language model would greatly enhance the speech recognition system by improving its

context understanding and predictive capabilities. This enhancement is crucial for accurately transcribing continuous speech and managing the complex syntax and vocabulary that are typical in natural language processing tasks. Through the analysis of text data, the language model learns the probabilities of word sequences, thereby serving as a powerful complement to the speech recognition model [7].

Given its capability across various languages, including those with scarce resources, XLS-R was selected as the foundation for the Orang Rimba language speech recognition system in this research. This study utilizes a dataset of self-recorded Orang Rimba audio, which was augmented to enhance the model's capabilities. Additionally, the integration of a language model can significantly improve the system's ability to understand context and predict speech patterns, thereby increasing transcription accuracy.

The structure of the remainder of this paper is organized as follows: Section 2 discusses related works, Section 3 focuses on the methodology employed, Section 4 examines the results and discussion, and Section 5 concludes the paper.

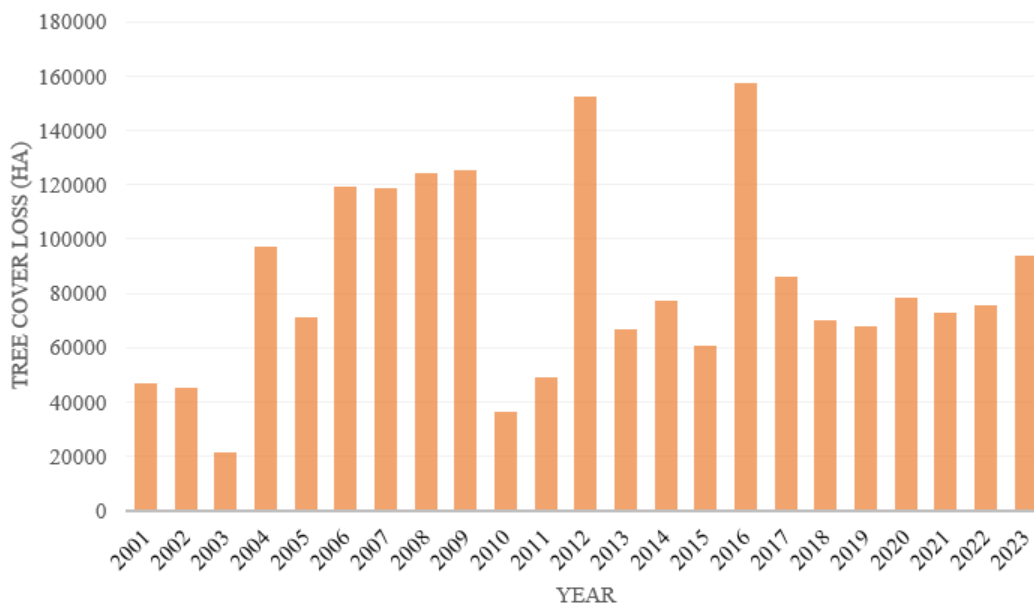


Figure 1. Tree cover loss in Jambi, Indonesia (2001-2023)

2. RELATED WORKS

Various advancements and applications of speech recognition models have been discovered. The wav2vec 2.0 model, introduced by Baevski et al. [8], has demonstrated its ability to deliver superior performance in speech recognition, even with limited resources. This capability is further showcased in the research conducted by Lehečka et al. [9], where wav2vec 2.0 is utilised for Czech language ASR, indicating the model's potential across various language contexts.

Building on these advancements, the XLSR model, introduced by Conneau et al. [10], marks a progression in the field of multilingual learning. Its successful implementation in various contexts, such as in the research conducted by Arisaputra and Zahra [11] for the Indonesian language, demonstrates its effectiveness across languages. The XLSR model's ability to accurately detect linguistic information and handle variations and abnormalities in speech is also

evidenced in Bartelds and Wieling [12] and Hernandez et al. [13].

However, the XLS-R model introduced by Babu et al. [4] signifies an enhancement in cross-language voice representation learning. The model's performance results in an average 7.4 BLEU score improvement in voice translation and a 14-34% average decrease in voice recognition errors. Moreover, improvements have been demonstrated in languages with limited resources, as tested on the BABEL dataset.

Because of this, the XLS-R model has been chosen for the development of a speech recognition system for the "Bahasa Orang Rimba" (Orang Rimba language). The development process is carried out by implementing transfer learning, where the XLS-R model is fine-tuned with collected Orang Rimba speech data. This choice is based on the model's superior performance, its effectiveness across various languages, including those with limited resources, and its advancements in voice representation learning.

Some research also shows the impact of data augmentation on improving model performance, especially in the context of limited training samples. The research by Nugroho and Noersasongko [14] demonstrates this effect. By incorporating techniques such as white noise addition, pitch shifting, and time stretching, they achieved a remarkable 99.76% accuracy in ethnic speaker recognition using a deep neural network architecture. Similarly, Amjad et al. [15] in their work on Pakistani racial speaker recognition applied the same augmentation strategies, resulting in significant improvements in speech emotion recognition accuracy.

The integration of the n-gram language model into the speech recognition model has also been proven to significantly improve performance. In the study by Baller et al. [16] the use of the n-gram model in the speech recognition system for low-resource languages successfully reduced the Word Error Rate (WER) to 42.27%, marking a significant improvement from the previous model. Meanwhile, research by Arisaputra and Zahra [11] also resulted in a decrease in WER from 20% to 12% when using the n-gram language model.

Therefore, data augmentation techniques such as adding noise, time stretching, and pitch shifting, along with the integration of the n-gram language model, were applied in the development process to artificially create a broader variation of data and enhance the linguistic understanding of the model. The use of the XLS-R model, combined with data that has been enriched through augmentation techniques and strengthened by the adaptation of the n-gram language model, has the potential to produce a speech recognition system that is more adaptive to voice input variations and more effective in interpreting linguistic contexts. This integration helps the model manage a wide range of linguistic variations, enhancing overall effectiveness in speech recognition.

3. METHODOLOGY

3.1 Data collection

Data collection for this study was conducted within the Bukit Duabelas National Park in Jambi, Indonesia, from September 2023 to November 2023. A total of 5 hours of audio recordings in the Orang Rimba language were amassed. These recordings were divided into various segments of differing durations to facilitate model training and evaluation. Participants for the recordings were selected based on gender (male and female), age, and reading proficiency, aiming for a representative and diverse sample of the Orang Rimba community. The sample consisted of three male and four female speakers, aged between 21 to 35 years.

The recording sessions took place in rooms where others were also working, which resulted in occasional background noise. Despite this, efforts were made to minimize disturbances by selecting the quietest possible times and locations. The recordings were conducted using smartphones, saved in .m4a format.

During each recording session, participants were provided with prepared transcripts to read aloud. These transcripts were also used as labels for model training. This approach was chosen to enhance efficiency, as spontaneous speech transcription would have been more time-consuming and inconsistent. Considering that the Orang Rimba language is predominantly oral, some participants were not accustomed to

reading. To address this, participants were introduced to the process and given opportunities to practice until they felt comfortable and fluent, minimizing hesitations and stutters in the recordings.

After each recording session, the audio files were meticulously reviewed to ensure quality. The review involved assessing the clarity of speech, the presence of background noise, and the overall consistency across all recordings. Sections with unclear speech or disruptive noise were edited, retaining only corrected and clear segments. For instance, if a speaker stumbled over a word but immediately corrected it, the unclear portion was removed, leaving only the corrected version. Table 1 contains several examples of sentences in the Orang Rimba language.

Table 1. Example sentences in the Orang Rimba language

Orang Rimba Language	Indonesian
Saloh tijak salah langkoh, Saloh pandong saloh pengoli.	Salah pijak salah langkah, Salah pandang salah penglihatan.
Sajang saji masok matah, cibuk aik, tikor bentol gawe betina. Kerayat belanjo, louk, maniy rapa gawe jenton.	Memasak, mengambil air, membuat tikar dan bantal merupakan pekerjaan perempuan. Memenuhi kebutuhan sehari-hari, mencari lauk, dan mengambil madu merupakan pekerjaan laki-laki.

From the examples, it can be seen that the Orang Rimba language has a vocabulary that is different from Indonesian, but there are words that are similar. The sentence structure of the Orang Rimba language also seems simpler compared to Indonesian.

3.2 Preprocessing data

Once the data collection is complete, the next step involves converting the gathered audio files from the .m4a format to .wav. Additionally, the sample rate of these files is adjusted to 16kHz. This conversion and sampling rate adjustment are crucial because the XLS-R model has been trained specifically on audio data in this format and at this sample rate.

Following the format conversion, the data is then split into distinct sets for training and testing purposes. The data is divided following a 70:10:20 ratio. Accordingly, 70% of the data is allocated for training the model, 10% for validation purposes, and the remaining 20% is reserved for testing. The training set is fundamental in training the model, the validation set is used to monitor the model's progress during training, and the testing set is key to assessing the overall performance of the model once the training is complete. Before being inputted into the model, the data is normalized, a step that is essential for improving the model's future performance [17].

In parallel with audio processing, the transcriptions undergo their preparatory steps. This includes converting all text to lowercase and removing any special characters. The conversion to lowercase is done to ensure uniformity, so the model does not differentiate between capitalized and non-capitalized words. Similarly, the elimination of special characters simplifies the textual data, thereby reducing the complexity and the number of unique tokens that the model needs to recognize and learn.

3.3 Augmenting data

The development of the speech recognition system also involved data augmentation. This process was focused on enhancing the training dataset through three specific techniques: adding white noise, time stretching, and pitch shifting. These methods were selected to introduce a variety of challenging auditory conditions to the model.

White noise was introduced to the audio samples at a level of approximately 10% of the original volume. This technique simulates ambient background sounds that are commonplace in natural environments where the Orang Rimba typically communicate, such as forests. Incorporating white noise trains the model to distinguish speech from environmental sounds, thereby improving its ability to accurately recognize spoken words in noisy settings [18].

Time stretching was applied by altering the playback speed to between 90% (slower) and 120% (faster) of the original duration. This technique accounts for variations in speech tempo among different speakers. The Orang Rimba community, like any group, includes individuals who naturally speak at different rates. By training the model on time-stretched audio, it becomes adept at handling both slower and faster speech rates, ensuring more consistent performance across varied speaking styles. Moreover, time stretching helps the model generalize better by exposing it to a wider range of temporal variations [19].

Lastly, the research incorporated pitch shifting. Pitch shifting involved adjusting the pitch of the audio samples up and down by two semitones. This augmentation technique addresses the diversity in speaker vocal characteristics, such as age-related vocal changes and individual pitch variations. The Orang Rimba community encompasses a range of ages and genders, each potentially exhibiting distinct pitch patterns. By incorporating pitch-shifted audio, the model learns to recognize speech across different pitch levels, thereby enhancing its ability to accurately transcribe speech from speakers with varying vocal pitches. Pitch shifting also aids in mitigating the effects of tonal variations that may be present in the Orang Rimba language, ensuring that the model remains effective despite subtle changes in pitch [20].

The selection of white noise, time stretching, and pitch shifting was driven by the specific acoustic and linguistic characteristics of the Orang Rimba language and its speakers. Given that the Orang Rimba primarily uses their language in oral communication within dynamic and sometimes noisy environments, these augmentation techniques address the challenges posed by such settings. Furthermore, the computational efficiency of these methods ensures that the augmentation process remains feasible without requiring extensive processing power or specialized hardware. Additionally, the limited availability of native language data necessitates techniques that maximize the diversity and quantity of training samples without requiring additional data collection.

3.4 Tokenization data

In the development of the speech recognition system, a tokenizer plays a crucial role by transforming the input data into a format that the model can comprehend. This process starts with establishing a vocabulary, which is essentially a collection of unique elements that the model recognizes.

However, since the model being developed is used for

speech recognition, the vocabulary consists of unique characters rather than words. This character-based approach includes all lowercase letters that have been preprocessed, spaces, and special characters designated for representing unknown elements. Each character in this vocabulary is assigned a specific numerical label for identification. Moreover, given the use of Connectionist Temporal Classification (CTC) in the fine-tuning process, a padding token is also incorporated into the vocabulary.

CTC is a technique in neural networks and training algorithms that aids in training sequence data. This technique was introduced to address sequence problems where the alignment between input and target is unknown. In many sequential tasks, such as speech recognition, the length of the sequential input (e.g., audio) does not always match the length of the sequential output (e.g., transcription text).

To address this issue, CTC introduces a special character called the 'blank' symbol into the output vocabulary. When CTC processes an input sequence, it generates a probability distribution for all possible output symbols, including the 'blank' symbol. The CTC algorithm then sums the probabilities of all possible alignments from input to output, resulting in a differentiable loss function. The model is then trained by optimizing weights to minimize this loss function [21].

Once the vocabulary is defined, the next step is to construct the tokenizer itself. This tokenizer converts transcription text into labels that match the established vocabulary. Additionally, the tokenizer is saved for future use, ensuring consistency and efficiency in subsequent tasks.

3.5 Fine-tuning model

Before commencing the fine-tuning process for the speech recognition model, it's essential to define the data collator. The data collator's function is to aggregate the processed data into a single batch. It also manages padding, ensuring uniformity in the length of all data within a batch for consistent processing.

The next step involves loading the pre-trained XLS-R model along with its configuration. This configuration details the model's architecture and the settings used in its initial training phase. An important modification made to the model is the replacement of its output layer with a new one tailored to the specific requirements of the Orang Rimba language speech recognition task. For this task, the model recognizes 26 classes, corresponding to each letter of the alphabet except "x", plus a space. It's set to be trained with the CTC loss function, which is vital for handling sequence alignment. The previous tokenizer was also recalled and integrated with the feature extractor from XLS-R itself to form a processor, which will subsequently be used for audio inference to generate transcriptions.

In the fine-tuning process, not all layers of the model are adjusted. The weights of the initial layers are kept frozen, meaning they won't be updated during the training. This is because these layers capture general features, such as basic phonetic elements, which are broadly applicable across languages. In contrast, the weights of the later layers, which detect more specific features like distinct sounds or words in the Orang Rimba language, are fine-tuned.

After setting up the model, the training phase begins. The training configuration, data collator, preprocessed dataset, and tokenizer are all integrated into a training class, forming the foundation for fine-tuning the model.

During training, the preprocessed audio data are first

converted into a latent sound representation through a convolutional feature encoder. This representation is then processed by the transformer within the model, creating a contextual understanding of the audio input. Concurrently, the latent representation undergoes discretization, resulting in a new form that the model can more easily interpret. Initially, the model undergoes training on a contrastive task, which aims to maximize the similarity between the latent and discrete representations. This step is crucial for the model to learn effective language representations. Following this, the model undergoes further training specifically for speech recognition, focusing on the linear layer and optimizing it by minimizing the CTC loss.

3.6 Integrating language model

After the fine-tuning process, the language model was constructed and then integrated into the speech recognition model. This process begins with the creation of an n-gram language model from the transcription of training data using KenLM. KenLM, developed by Heafield [22], is a language modeling library focused on optimizing speed and memory efficiency during language model query operations. This is important for speech recognition applications, where the speed of accessing the language model significantly impacts the system's ability to transcribe speech into text quickly and accurately. The library is specifically designed to address the computational challenges faced by large language models used in various natural language processing applications.

In this study, a 5-gram model was used, which considers word sequences up to five words long. A 5-gram model considers all word sequences from unigrams (1-grams) up to five-word sequences (5-grams). This inclusion allows the model to capture a wide range of contextual information, providing a deeper understanding of the syntactic and semantic nuances of the Orang Rimba language.

Higher-order n-gram models (e.g., 7-gram or 10-gram) offer more detailed contextual information but require more computational resources and memory. Given the limited dataset size and resource constraints of this study, a 5-gram model strikes an optimal balance between contextual depth and computational feasibility. Also, the limited nature of the Orang Rimba dataset necessitates a language model that can be effectively trained without overfitting. A 5-gram model is sufficiently complex to leverage the available data without demanding the extensive corpora typically required for higher-order models.

KenLM processes the transcription text and calculates the probabilities of up to five-word sequences appearing together. The outcome of this process is an ARPA file, which is the standard format for storing n-gram models. This file contains the n-grams and associated log probabilities, which are used to determine the likelihood of word sequences during the decoding process. The file is then modified by adding an end-of-sentence marker (</s>). This modification is crucial for helping the language model recognize when a sentence ends, which is important for predicting accurate structure and flow in continuous speech.

Once the ARPA file is prepared, it is utilized to build a CTC (Connectionist Temporal Classification) decoder. This decoder provides a probabilistic context that helps to refine and perfect the raw predictions of the speech recognition model. The decoder is integrated by incorporating it into the previously constructed processor.

3.7 Model evaluation

The primary evaluation metric used in this study is the Word Error Rate, also known as WER. WER is a standard metric in automatic speech recognition (ASR) systems, measuring the discrepancy between the transcribed text produced by the model and the reference (ground truth) text. It is calculated as the sum of substitutions, deletions, and insertions required to transform the model's output into the reference text, divided by the total number of words in the reference.

WER is widely adopted in the ASR community due to its straightforward interpretation and ability to provide a quantitative measure of transcription accuracy. It effectively captures the types of errors that are most impactful in practical applications, such as misheard words (substitutions), omitted words (deletions), and added words (insertions). By focusing on word-level accuracy, WER aligns closely with user-perceived performance, making it a relevant and meaningful metric for evaluating speech recognition systems.

The dataset was divided into three subsets to facilitate evaluation: 70% of the data, approximately 3 hours and 40 minutes was used for the training set; 10% of the data, about 32 minutes, was allocated to the validation set; and the remaining 20%, roughly 1 hour and 3 minutes was used for the testing set.

This division ensures that the model is trained on a substantial portion of the data while reserving separate sets for tuning (validation) and performance assessment (testing). The validation set was used to monitor the model's progress during training, enabling the implementation of early stopping to prevent overfitting. The testing set remained unseen during training and validation, providing an objective measure of the model's generalization capabilities.

4. RESULT AND DISCUSSION

Following the completion of data collection for the Orang Rimba language speech recognition, a corpus of audio recordings was compiled featuring seven individuals from the community, providing a balanced representation of gender with recordings from three male and four female speakers. Each participant contributed approximately 45 minutes of speech, resulting in a total of 5 hours and 15 minutes of raw audio data. The list of the speakers' gender and age can be seen in Table 2.

Table 2. List of speakers' gender and age

No.	Gender	Age
1	Male	21
2	Male	23
3	Male	31
4	Female	25
5	Female	22
6	Female	29
7	Female	35

One of the challenges faced during the data collection process is The Orang Rimba language is predominantly used in oral communication, and as such, its native speakers are not accustomed to reading. Consequently, some speakers experienced difficulties and tended to hesitate and stutter in their speech. This condition may add complexity to the data collected and, as a result, presents additional challenges for the

model during training. The inconsistent quality of this data could become one of the primary factors affecting the outcome of the model’s performance.

This data was then processed and prepared for training and evaluation of the speech recognition models. The recordings were segmented into smaller utterances and then divided into training, validation, and test sets. The distribution of the data consisted of 3 hours and 40 minutes allocated for training, 32 minutes dedicated to validation, and 1 hour and 3 minutes reserved for testing purposes.

In line with the methodology outlined in the previous section, the training data underwent augmentation to enhance the robustness and generalizability of the speech recognition model. Once augmented, the dataset expanded from 3 hours and 40 minutes to 22 hours and 3 minutes, making it a

substantial foundation for training the XLS-R model. Figure 2 illustrates the original audio waveform (top left) and its augmented versions: with white noise (top right), time stretched to 90% (middle left), and time stretched to 120% (middle right), pitch-shifted +2 (bottom left), and pitch-shifted -2 (bottom right).

Compared to the original audio, the audio with noise has a higher amplitude during quieter sections. The time-stretched and pitch-shifted waveforms generally have lower overall amplitudes due to the effects of the processing. Time-stretching the audio to 120% compresses the waveform while time-stretching it to 90% expands it. Since pitch shifting changes frequency, both pitch-shifted versions maintain similar amplitude characteristics, with the version shifted up showing slightly higher amplitudes.

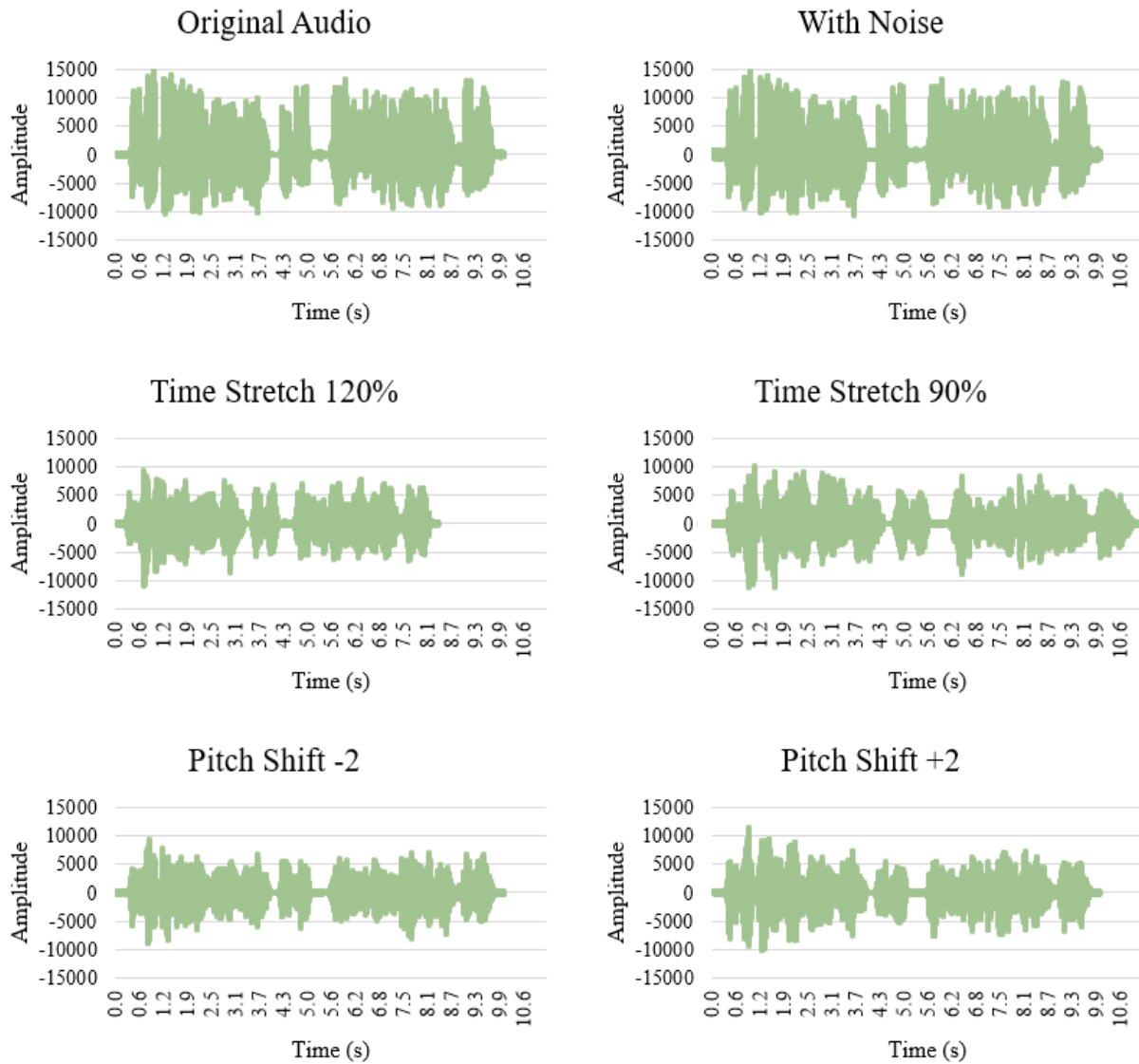


Figure 2. Waveform comparison of original and augmented audio samples

Table 3. Word error rate comparison of the models trained on original and augmented Orang Rimba datasets, with and without language model integration

Model	Training Data	Validation WER	Test WER	
			Without Language Model	With Language Model
XLSR-53	Original Data (3.675h)	20.83%	23.99%	12.73%
XLS-R		18.67%	22.59%	12.10%
XLSR-53	Augmented Data (22.05h)	16.14%	20.53%	11.02%
XLS-R		15.56%	18.39%	9.66%

The model was then fine-tuned using both just the original data and augmented data, and then integrated with the 5-gram language model. The experiment set out to evaluate the performance of the model with XLS-R as the primary system for speech recognition of the Orang Rimba language, with the XLSR-53 model serving as a reference point for comparison.

The primary reason for selecting XLSR-53 as the baseline model is that the dataset used in this study is unique, and no existing speech recognition models for the Orang Rimba language were available at the time of research. Therefore, the previously developed Wav2Vec 2.0 model, specifically XLSR-53, was chosen for training and used as a benchmark for comparison in this study. Table 3 shows the performance of all models.

For the original dataset of 3 hours and 40 minutes, the XLS-R model demonstrated superior performance compared to XLSR-53. Specifically, the XLS-R model achieved a validation WER of 18.67%, which is lower than the 20.83% recorded by XLSR-53. In the testing set, XLS-R also outperformed XLSR-53, with a WER of 22.59% compared to 23.99%. Upon integrating the language model, the WER for XLS-R decreased further to 12.10%, while XLSR-53 improved to 12.73%. These results indicate that XLS-R not only possesses a superior foundational capability but also maintains its advantage when augmented with a language model, especially in scenarios with limited training data.

When both models were trained using the augmented dataset, the positive impact of data augmentation became more pronounced. The XLS-R model achieved a validation WER of 15.56% and a testing WER of 18.39%, both outperforming XLSR-53, which recorded a validation WER of 16.14% and a testing WER of 20.53%. Although the validation WER for both models is relatively close, the more substantial difference in testing WER scores suggests that XLS-R possesses better generalization capabilities. This implies that XLS-R not only starts with a stronger baseline but also benefits more effectively from the increased and varied data introduced through augmentation, leading to enhanced performance on unseen data.

With the integration of the 5-gram language model, the performance of both models improved significantly. The XLS-R model's WER decreased to 9.66%, while XLSR-53's WER reduced to 11.02%. These outcomes consistently demonstrate that the use of a language model significantly enhances speech recognition accuracy, underscoring the language model's role in improving the model's ability to handle the diversity and complexity of speech data. Table 4 shows a few examples of the transcription created by this model.

Figure 3 illustrates the differences in performance between the XLS-R model when trained on original versus augmented data. Starting from step 1000th, the model trained with augmented data consistently exhibits a lower WER compared to its counterpart trained on original data.

Table 4. Model transcription compared to reference text

Model Transcript	Reference Text
delom rimba sen hopi klaku hakeh teringon waktu lagi moin ke rombong orang rimba ado di sungo sari	di delom rimba sen hopi laku akeh teringot waktu lagi mion ke rombong orang rimba nang ado di singosari
kalau ndo makon bison delok beno louk di delok hopi pakoi osen mumpa diluarr nio uje muans	kalau ndok makon bison ndelok benor louk dicari hopi pakoi sen mumpa di luar nio uje muas

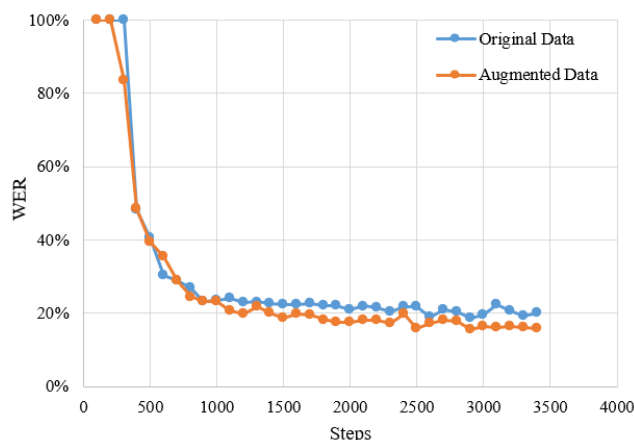


Figure 3. Comparison of validation WER on original vs augmented data on raining

The combined application of all three augmentation techniques leads to a comprehensive enhancement of the training data, introducing diverse speech variations that enrich model training. This strategy not only expands the dataset size but also boosts its variety, allowing the models to achieve a more robust and generalized grasp of the Orang Rimba language. The effects are evident in the WER reductions observed, particularly in the XLS-R model.

Moreover, integrating the 5-gram language model markedly improves both models' performance. The observed reduction in WER following this integration emphasizes the critical role of contextual information in transcription accuracy. By accounting for sequences of up to five words, the language model helps to resolve homophones, predict likely word patterns, and maintain syntactic coherence, effectively reducing substitution, insertion, and deletion errors.

The Orang Rimba community demonstrates rich dialectal diversity, with variations influenced by factors such as age, and gender. Recognizing and accounting for this linguistic variety was essential during data collection to capture a comprehensive spectrum of speech patterns. Efforts were made to include recordings from a wide range of speakers, ensuring that the speech recognition model could generalize effectively across dialects and minimize biases toward any single subgroup. Additionally, data augmentation techniques, such as time stretching and pitch shifting, were applied to replicate variations in speech tempo and pitch. This combination of inclusivity and augmentation ultimately strengthens the model's adaptability and accuracy, ensuring more reliable performance in real-world applications.

The Orang Rimba's lifestyle and varied environments, such as dense forest, introduce unique acoustic challenges, such as background noise and fluctuating speech tempos. Addressing these factors during data augmentation, through techniques like adding white noise and time stretching, was vital to simulate realistic scenarios. This approach ensures that the speech recognition system performs reliably even amid the dynamic, often noisy environments typical of the Orang Rimba's daily lives.

5. CONCLUSION

In this study, the capabilities of the XLS-R model have been harnessed to develop a speech recognition system for the Orang Rimba language, addressing the crucial need for

linguistic preservation amidst limited data resources. The application of data augmentation and language model emerged as a pivotal strategy, significantly enhancing the model's performance.

Initially, the XLS-R model, trained on 3 hours and 40 minutes of data, achieved a WER of 22.59%. This performance established a baseline and represented a substantial achievement given the constraints of available data. However, when the training data was augmented to a total of 22 hours and 3 minutes by incorporating noise, stretching time, and shifting pitch, the performance improved markedly, with the WER dropping to 18.39%. These results from the augmented dataset underscore the effectiveness of data augmentation in enhancing the accuracy of speech recognition models for languages with limited linguistic data.

Further significant advancements were observed with the integration of a 5-gram language model tailored to the Orang Rimba language. With this integration, the WER for the XLS-R model decreased further to 9.66%. This integration has demonstrated its effectiveness in improving language context and structure by predicting the next word based on previous ones. This significantly enhances the accuracy of the model's transcriptions.

The application of the XLS-R model, combined with data augmentation and language model, has proven to be a successful approach in mitigating the challenges presented by data scarcity for the Orang Rimba language. Despite this success, there remains significant potential for further improvement. A key challenge faced in this research was the limited dataset size; the current dataset is only 5 hours and 15 minutes. Expanding the dataset would greatly enhance the model's quality.

Collaborative data collection could be employed to engage more extensively with the Orang Rimba community. Crowdsourcing initiatives could involve community members directly in data collection, allowing the dataset to grow organically and better reflect the linguistic diversity of the community.

The dataset collection method could also be improved. To address the issue of speakers who struggle with reading, audio recordings could be collected first and then transcribed. This approach allows speakers to communicate more naturally without reading text, enhancing the quality of the dataset used for model training.

To enhance the data further, other data augmentation techniques could be applied, such as time and frequency masking. These improvements could create a more comprehensive and representative dataset, ultimately strengthening the model's performance and its ability to accurately transcribe the Orang Rimba language.

REFERENCES

- [1] Prasetijo, A. (2017). Living without the forest: Adaptive strategy of Orang Rimba. *Senri Ethnological Studies*, 95: 255-278. <https://doi.org/10.15021/00008586>
- [2] Hamied, F.A., Musthafa, B. (2019). Policies on language education in Indonesia. *Indonesian Journal of Applied Linguistics*, 9(2): 308-315. <https://doi.org/10.17509/ijal.v9i2.20279>
- [3] Nassif, A.B., Shahin, I., Attili, I., Azzeh, M., Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7: 19143-19165. <https://doi.org/10.1109/ACCESS.2019.2896880>
- [4] Babu, A., Wang, C., Tjandra, A., Lakhota, K., Xu, Q., Goyal, N., Auli, M. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*. <https://doi.org/10.21437/Interspeech.2022-143>
- [5] Gokay, R., Yalcin, H. (2019). Improving low resource Turkish speech recognition with data augmentation and TTS. In *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*, Istanbul, Turkey, pp. 357-360. <https://doi.org/10.1109/SSD.2019.8893184>
- [6] Kipyatkova, I., Kagirow, I. (2022). Analytical review of methods for solving data scarcity issues regarding elaboration of automatic speech recognition systems for low-resource languages. *Informatics and Automation*, 21(4): 678-709. <https://doi.org/10.15622/ia.21.4.2>
- [7] Pakoci, E., Pekar, D., Popović, B., Sečujski, M., Delić, V. (2022). Overcoming data sparsity in automatic transcription of dictated medical findings. In *2022 30th European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, pp. 454-458. <https://doi.org/10.23919/eusipco55093.2022.9909893>
- [8] Baevski, A., Zhou, Y., Mohamed, A., Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33: 12449-12460.
- [9] Lehečka, J., Švec, J., Pražák, A., Psutka, J.V. (2022). Exploring capabilities of monolingual audio transformers using large datasets in automatic speech recognition of Czech. *arXiv preprint arXiv:2206.07627*. <https://doi.org/10.21437/Interspeech.2022-10439>
- [10] Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*. <https://doi.org/10.21437/Interspeech.2021-329>
- [11] Arisaputra, P., Zahra, A. (2023). Indonesian Automatic Speech Recognition with XLSR-53. *arXiv preprint arXiv:2308.11589*. <https://doi.org/10.18280/isi.270614>
- [12] Bartelds, M., Wieling, M. (2022). Quantifying language variation acoustically with few resources. *arXiv preprint arXiv:2205.02694*. <https://doi.org/10.18653/v1/2022.naacl-main.273>
- [13] Hernandez, A., Pérez-Toro, P.A., Nöth, E., Orozco-Aroyave, J.R., Maier, A., Yang, S.H. (2022). Cross-lingual self-supervised speech representations for improved dysarthric speech recognition. *arXiv preprint arXiv:2204.01670*. <https://doi.org/10.21437/Interspeech.2022-10674>
- [14] Nugroho, K., Noersasongko, E. (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>
- [15] Amjad, A., Khan, L., Chang, H.T. (2022). Data augmentation and deep neural networks for the classification of Pakistani racial speakers recognition. *PeerJ Computer Science*, 8: e1053. <https://doi.org/10.7717/PEERJ-CS.1053>
- [16] Baller, T., Bennett, K., Hamilton, H.J. (2021). Transfer Learning and language model adaption for low resource speech recognition. *Canadian Artificial Intelligence Association*. <https://doi.org/10.21428/594757db.d3394351>

- [17] Labied, M., Belangour, A., Banane, M., Erraissi, A. (2022). An overview of automatic speech recognition preprocessing techniques. In 2022 International Conference on Decision Aid Sciences and Applications (DASA), Chiangrai, Thailand, pp. 804-809. <https://doi.org/10.1109/DASA54658.2022.9765043>
- [18] Ma, D., Li, G., Xu, H., Chng, E.S. (2019). Improving code-switching speech recognition with data augmentation and system combination. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, pp. 1308-1312. <https://doi.org/10.1109/APSIPAASC47483.2019.9023316>
- [19] Lounnas, K., Lichouri, M., Abbas, M. (2022). Analysis of the effect of audio data augmentation techniques on phone digit recognition for Algerian Arabic dialect. In 2022 International Conference on Advanced Aspects of Software Engineering (ICAASE), Constantine, Algeria, pp. 1-5. <https://doi.org/10.1109/ICAASE56196.2022.9931574>
- [20] Slizovskaia, O., Janer, J., Chandna, P., Mayor, O. (2022). Voice conversion with limited data and limitless data augmentations. arXiv preprint arXiv:2212.13581. <https://doi.org/10.48550/arXiv.2212.13581>
- [21] Graves, A., Fernández, S., Gomez, F., Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, New York, United States, pp. 369-376. <https://doi.org/10.1145/1143844.1143891>
- [22] Heafield, K. (2011). KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation, UK, pp. 187-197.