Computer Science and Information Technologies
Vol. 6, No. 3, November 2025, pp. 253~261
ISSN: 2722-3221, DOI: 10.11591/csit.v6i3.p253-261

Javanese and Sundanese speech recognition using Whisper

Alim Raharjo, Amalia Zahra
Department of Computer Science, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Nov 12, 2024
Revised May 27, 2025
Accepted Jun 13, 2025

Keywords:

Fine tune
Javanese
Speech recognition
Sundanese
Whisper

ABSTRACT

Automatic speech recognition (ASR) technology is essential for advancing human-computer interaction, particularly in a linguistically diverse country like Indonesia, where approximately 700 native languages are spoken, including widely used languages like Javanese and Sundanese. This study leverages the pre-trained Whisper Small model an end-to-end transformer pretrained on 680,000 hours of multilingual speech, fine tuning it specifically to improve ASR performance for these low resource languages. The primary goal is to increase transcription accuracy and reliability for Javanese and Sundanese, which have historically had limited ASR resources. Approximately 100 hours of speech from OpenSLR were selected, covering both reading and conversational prompts, the data exhibited dialectal variation, ambient noise, and incomplete demographic metadata, necessitating normalization and fixed-length padding. with model evaluation based on the word error rate (WER) metric. Unlike approaches that combine separate acoustic encoders with external language models, Whisper unified architecture streamlines adaptation for low-resource settings. Evaluated on held-out test sets, the fine-tuned models achieved Word Error Rates of 14.97% for Javanese and 2.03% for Sundanese, substantially outperforming baseline systems. These results demonstrate Whisper effectiveness in low-resource ASR and highlight its potential to enhance transcription accuracy, support language preservation, and broaden digital access for underrepresented speech communities.

Corresponding Author:

Alim Raharjo
Department of Computer Science, BINUS Graduate Program-Master of Computer Science
Bina Nusantara University
DKI Jakarta, Indonesia, 11480
Email: alim.raharjo@binus.ac.id

1. INTRODUCTION

Automatic speech recognition (ASR) is a pivotal technology in human-computer interaction, allowing seamless communication between humans and machines through speech [1]. Over time, ASR has seen significant advancements, leading to the integration of speech interfaces in a wide range of applications, such as transcription services, virtual assistants, and voice-controlled devices [2]–[8]. The global ASR market is expected to grow substantially, driven by the increasing demand for efficient and accurate speech recognition technologies [9]. Language diversity is one of the core characteristics of human civilization, with over 7,000 languages spoken worldwide, showcasing the complexity of human communication [10]. Language goes beyond simple communication; it encapsulates worldviews, cultural expressions, and unique grammatical and phonetic systems. Indonesia, with its population of approximately 280 million, is home to around 700 native languages, in addition to the national language, Bahasa Indonesia [10]. This makes Indonesia one of the most linguistically diverse nations in the world, accounting for roughly 10% of the

global languages. Javanese and Sundanese, spoken by approximately 90 million and 40 million people respectively, are the second and third most widely spoken languages in Indonesia [11].

Recent advancements in ASR for Indonesian languages have been driven by self-supervised learning models, notably Wav2Vec2 [12], which has shown promise in low-resource language settings. Studies have reported word error rates (WER) as low as 22% in some applications [13]–[22]. Notably, research on Sundanese ASR utilizing the OpenSLR dataset achieved a WER of 23.5% through fine-tuning of the Wav2Vec2 model [23]. Similarly, another study on Javanese ASR, leveraging the XLS-R variant of Wav2Vec2, achieved a WER of 17.95% [24]. These results highlight the effectiveness of self-supervised models in tackling the challenges posed by low-resource languages with limited training data. This study focuses on fine-tuning the Whisper model [7], a state-of-the-art ASR model, to evaluate its performance for low-resource languages, specifically Javanese and Sundanese. The datasets used are sourced from OpenSLR [11], a repository offering valuable resources for languages with limited speech data. Whisper's extensive pre-training on a dataset of over 680,000 hours of audio-transcription data, including 117,000 multilingual hours covering more than 96 languages [7], positions it as a strong candidate for this task. A recent study [25]–[27] comparing Whisper to Wav2Vec2.0 found that Whisper consistently outperforms Wav2Vec2.0 across multiple languages, achieving lower WERs, especially in noisy environments such as the GRACE [25] and CORAA [28] corpora. For instance, Whisper yielded an average WER range of 11.3% to 24.9%, notably outperforming Wav2Vec2.0, which ranged from 13.1% to 34.8% in similar conditions [25]. The statistical significance of Whisper's performance, as verified by a Mann-Whitney test, underlines its adaptability and efficiency in diverse multilingual scenarios. Unlike Wav2Vec 2.0, which rely on unsupervised pre-training on unlabelled data, Whisper benefits from its pre-training directly on speech-to-text tasks, enabling it to generalize well across diverse languages, domains, and datasets.

Whisper use of labeled data allows it to perform ASR tasks with minimal fine-tuning compared to models that require extensive fine-tuning for optimal performance. Its pre-training on a diverse multilingual dataset makes it particularly suited for underrepresented languages like Javanese and Sundanese, which lack sufficient ASR models. Furthermore, Whisper architecture a transformer based encoder-decoder model—facilitates accurate transcriptions by converting raw audio into log-Mel spectrograms, which are encoded into hidden states and autoregressively decoded into text transcriptions [7]. The model's deep fusion language model integration provides superior performance compared to shallow fusion approaches.

Whisper exceptional performance in multilingual ASR tasks, with WERs of 3% on the LibriSpeech test-clean subset and 4.7% on the TED-LIUM corpus [7], underscores its potential to generalize across various languages and domains. These qualities make it an ideal candidate for improving ASR for Javanese and Sundanese, where data scarcity and linguistic variability present significant challenges. Sundanese and Javanese were chosen for this study due to their widespread use in Indonesia, despite the limited availability of high-quality, annotated speech data. Sundanese, spoken by approximately 40 million people, represents one of Indonesia's largest linguistic groups but has received relatively little attention in ASR research [11]. The success of fine-tuning Wav2Vec2 for Sundanese ASR, as shown by [23], suggests that further improvements may be achievable, particularly with Whisper's comprehensive multilingual capabilities.

In Indonesia's diverse linguistic landscape, Sundanese, one of the most spoken regional languages, illustrates the challenges of language complexity and variation. As a primary language for communities in West Java, Sundanese serves as a key cultural and communicative tool. However, the adoption of Indonesian as the national language and the close proximity of Sundanese-speaking areas to Javanese-speaking regions have introduced additional layers of variation within Sundanese. In areas like Banjar City, where Sundanese and Javanese communities intersect, distinct sub-dialects have emerged, including Java-influenced and Java-dominated Sundanese varieties [29]. A detailed study of Banjar Sundanese identified three primary sub-dialect: standard Sundanese, Java-influenced Sundanese, and Java-dominated Sundanese. Each reflect different levels of Javanese integration observable in lexical and phonological shifts [29].

While speech recognition technology has seen rapid development in recent years, its applications remain largely concentrated on major national and international languages such as English, Mandarin Chinese, and Indonesian. Consequently, many regional and indigenous languages like Javanese and Sundanese have received limited attention in technological research and innovation. This lack of representation not only reflects a digital divide but also contributes to the ongoing threat of language endangerment. To address this, integrating speech technology into language preservation efforts has become increasingly important. For instance, implementing ASR systems for under-represented languages can encourage broader use among speakers and foster linguistic pride within communities. In the case of Sundanese, the availability of tools such as speech-to-text applications can raise awareness about the language declining use, promote everyday communication in Sundanese, and inspire younger generations to learn and engage with their linguistic heritage. Furthermore, such systems can bridge communication gaps and make digital content more accessible in local languages, supporting both cultural preservation and inclusivity in digital environments.

Beyond its role in language preservation, speech recognition technology also offers significant potential in educational contexts, particularly for supporting the learning of regional languages such as Sundanese and Javanese [30], [31]. By enabling the development of language learning systems that capture and transcribe native speech, ASR tools can help everyday users become more familiar with the pronunciation, structure, and use of these languages in real-life scenarios. This is especially valuable for languages like Sundanese and Javanese, which feature complex speech levels or registers ranging from informal forms used among peers to more refined and polite forms used when addressing elders or in formal settings.

These linguistic nuances are often difficult for learners or non-native speakers to grasp, especially without guided exposure to native usage. A speech-to-text or audio-based learning system powered by ASR can provide real-time feedback and structured language materials that incorporate these variations, making the learning process more interactive and culturally informed. Through such applications, ASR technology not only helps bridge the accessibility gap in language learning but also strengthens community engagement with local languages, ensuring that these cultural identities remain vibrant in the digital age. For Javanese, spoken by around 90 million people, additional challenges stem from its linguistic complexity and dialectal variations. Previous research using the XLS-R variant of Wav2Vec2 achieved a WER of 17.95% [24], but there remains significant room for improvement. Whisper, with its broader pre-training and multilingual architecture, is expected to enhance ASR performance for Javanese. Studies have also shown that fine-tuning Whisper for low-resource child speech yields promising results compared to non-finetuned models [32]–[34], further supporting its potential to improve ASR for underrepresented languages like Javanese and Sundanese.

2. METHOD

2.1. Research stages

The first stage of this research involved gathering the relevant datasets from publicly available sources, as illustrated in Figure 1. We utilized two datasets: OpenSLR (SLR35) for Javanese and OpenSLR (SLR36) for Sundanese [11]. These datasets were selected due to their comprehensiveness, consisting of thousands of hours of transcribed speech. After dataset collection, we proceeded with the pre-processing stage. This phase involved organizing the data into training and validation sets, ensuring it was appropriately prepared for the Whisper model's requirements. Part of this preparation included extracting audio features by converting the raw sound files into log-Mel spectrograms, the input format Whisper uses. This transformation captures essential frequency details in the speech, making the data well-suited for the fine-tuning process. Notably, the testing dataset was kept separate to ensure an unbiased final evaluation.

With pre-processing complete, the next stage involved fine-tuning the Whisper model, a state-of-the-art sequence-to-sequence ASR model originally trained on over 680,000 hours of multilingual speech data [7]. This extensive pre-training provided a strong foundation, but due to the unique characteristics of Javanese and Sundanese, specific adaptations were necessary. We fine-tuned Whisper using the training subset, modifying model weights to enhance its recognition accuracy for these languages. Throughout the training, we periodically assessed model performance with the validation dataset to track its progress and mitigate overfitting. Key performance metrics, including WER, were calculated at various intervals to monitor transcription accuracy, and training continued until predefined stopping criteria were met.

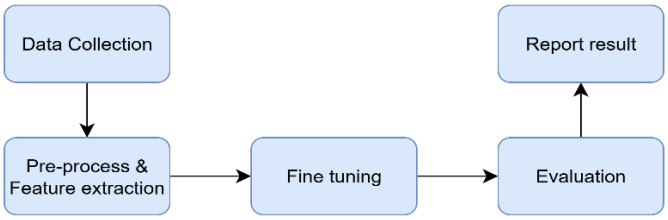

Figure 1. Research stages

In the final evaluation stage, illustrated in Figure 1, the model was tested on the held-out datasets for both Javanese (SLR35) and Sundanese (SLR36). This stage was essential for assessing how well the model could generalize to new, unseen data. Performance was measured using WER, providing a quantitative assessment of transcription accuracy in both languages. The outcomes were then compared with previous studies on Javanese and Sundanese ASR [23], [24] to gauge the improvements Whisper's fine-tuning brought to ASR accuracy for low-resource languages. These findings offer insights into the potential of Whisper in advancing ASR performance for underrepresented languages like Javanese and Sundanese, supporting broader efforts to make ASR more inclusive.

2.2. Data collection

The datasets employed in this study were sourced from OpenSLR, specifically the SLR35 corpus for Javanese and SLR36 for Sundanese [11]. SLR35 consists of 185,076 utterances amounting to approximately 296 hours of audio, contributed by 1,019 native Javanese speakers. SLR36 contains 219,156 utterances totaling approximately 333 hours of Sundanese speech from 542 speakers. Both datasets were recorded at 16 kHz, and all speakers were between the ages of 18 and 35. However, gender-specific metadata was not provided in the dataset documentation. The speech data was collected in collaboration with the Department of Javanese Literature at Universitas Gadjah Mada (UGM) in Yogyakarta and Universitas Pendidikan Indonesia (UPI) in Bandung. The recordings capture a variety of speaking styles, including reading speech and conversational phrases, and are accompanied by manually transcribed text. Nevertheless, challenges such as inconsistent spelling, informal punctuation, and occasional transcription errors were identified during preprocessing. These were addressed through a normalization process, including lowercasing and removal of special characters.

To facilitate controlled evaluation and to replicate prior experimental setups [24], only a portion of the total data was utilized. As shown in Table 1, training and validation used content from zip archives 0 through 2, while testing was conducted solely on archive number 3. The resulting data split, as shown in Table 2, comprised 44 hours and 33 minutes for Javanese training and 11 hours and 9 minutes for validation. For Sundanese, the training dataset totaled 49 hours and 52 minutes, while the validation dataset covered 12 hours and 28 minutes. The testing dataset involved 18 hours and 28 minutes of Javanese data and 20 hours and 55 minutes of Sundanese. Although the corpora are extensive, several limitations were noted. These include the lack of speaker gender metadata, variations in audio quality due to differing recording environments, and the wide range of utterance durations. Since Whisper requires fixed-length input sequences of 30 seconds, audio clips were either truncated or zero-padded as needed. These preprocessing adjustments ensured uniformity and model compatibility throughout the fine-tuning process.

Table 1. Dataset splitting

Sundanese	Javanese	Used in
Sundanese-asr-1...2	Javanese-asr-1...2	Training & validation
Sundanese-asr-3	Javanese-asr-3	Testing
Sundanese-asr-4…15	Javanese-asr-4…15	Not used

Table 2. Dataset duration and partitioning

Language	Training duration	Validation duration	Testing duration	Total used
Javanese	44 h 33 m	11 h 09 m	18 h 28 m	74 h 10 m
Sundanese	49 h 52 m	12 h 28 m	20 h 55 m	83 h 15 m

2.3. Pre-processs and feature extraction

The OpenSLR datasets were divided into two main subsets, 80% allocated to training and 20% reserved for validation. This division maximized the amount of data available for training while providing a validation set for tuning and monitoring model performance during training. Each audio file was paired with its respective transcriptions, forming organized and consistent subsets. To ensure uniformity across the transcriptions, a text normalization process was applied, which involved removing any special characters and retaining only lowercase alphabetic characters. Simplifying the text in this way minimized potential inconsistencies, helping to align the transcriptions with the model's expectations.

To fine-tune the Whisper model with the OpenSLR data, an important step was ensuring compatibility with Whisper's input format requirements. The model does not directly accept audio in ".flac" format. Therefore, files needed to be converted into log-Mel spectrograms, which is the format Whisper processes. Converting the audio data requires an understanding of how digital systems interpret sound: audio is represented as a one-dimensional array of amplitude values over time. Since continuous audio has an infinite range of values, it must be discretized by sampling amplitude at fixed intervals, a process known as sampling. Sampling rate, typically measured in Hertz (Hz), defines these intervals.

Maintaining consistent sampling rates is essential to prevent errors during ASR processing. Mismatched sampling rates between the audio inputs and Whisper model requirements could result in unexpected outputs. For instance, playing a 16 kHz audio file at an 8 kHz rate distorts playback speed and quality. For this study, Whisper's feature extractor required a 16 kHz sampling rate, so audio data was up-sampled or down-sampled as necessary to ensure compatibility. Ensuring correct sampling rates allowed the model to process the data accurately and reduced the risk of misinterpretations during ASR tasks. Ensuring correct sampling rates allowed the model to process the data accurately and reduced the risk of misinterpretations during ASR tasks.

In terms of data augmentation, no advanced techniques such as noise injection, pitch shifting, or time stretching were applied in this study. While these methods have been shown to enhance generalization in ASR models by simulating various acoustic conditions and speaker variability [5], [35], the focus of this research was to evaluate the effectiveness of the Whisper model with minimal intervention. Nonetheless, simple padding and truncation techniques were used to standardize audio sample lengths to 30 seconds, as required by the Whisper architecture. Shorter clips were zero-padded (representing silence), and longer clips were truncated to maintain uniform input dimensions. These steps ensure input compatibility and stability during model training, although future work may explore more sophisticated augmentation strategies to further improve robustness.

2.4. Fine tuning

In this stage, the fine-tuning process began by loading the pre-trained Whisper-Small model from the Hugging Face Hub, a popular repository for machine learning models. Whisper, pre-trained on a largescale multilingual dataset, includes automatic language detection. However, due to the linguistic similarities between regional languages, particularly in Indonesia, it was necessary to explicitly define the target language. To ensure that transcriptions were generated in either Javanese or Sundanese, the language and task arguments were specified in the generation configuration prior to training.

For fine-tuning, the model was trained using language-specific datasets with input features (log-Mel spectrograms) and corresponding transcription labels. The training was performed using the trainer application programming interface (API) provided by Hugging Face, which allowed control over various hyperparameters. The model was trained with a batch size of 16, a learning rate of 1e-5, and a total of 5,000 training steps, using the AdamW optimizer under the hood. A warmup of 500 steps was applied to gradually ramp up the learning rate during the early phase of training. Evaluation was performed every 1,000 steps, and checkpoint saving occurred at the same interval. To enhance memory efficiency, gradient checkpointing was enabled, and the model was trained using mixed precision (fp16) for faster computation. The best model was selected automatically based on the lowest WER observed on the validation set. All training progress and evaluation metrics were logged through TensorBoard.

Although the focus of evaluation was on WER, which is widely used in ASR tasks, we also considered other potential metrics such as sentence error rate (SER) and word recognition accuracy (WRA). However, due to the limited availability of sentence-segmented labels in the dataset, SER was not calculated. WRA, the percentage of correctly recognized words out of all actual words, was examined during internal evaluation to support WER results. Including WRA, it helped provide additional insight into how often the model accurately recognized full words, particularly in noisy or dialect-influenced samples. Future studies could benefit from more extensive use of these complementary metrics to offer a broader evaluation of ASR performance.

2.5. Evaluation

The evaluation process was done in two main stages: an initial stage during training to fine-tune the model, and a final evaluation after the fine-tuning was complete. In the first stage, evaluations were performed within the training cycle using the validation data split. The primary metric was WER, where lower values indicated better model accuracy. These evaluations were performed at each checkpoint specified in the training configuration, using the validation dataset to assess model performance. This cycle continued until the stopping criteria were met, with the model saving its best version based on the lowest recorded WER. In the second stage, the best-performing model from fine-tuning was reloaded for testing on a separate set of audio data. In this evaluation, the model transcribed each audio sample directly, and WER was calculated by comparing the predictions to the provided transcription files.

To ensure accurate WER calculation, a normalization process was applied to both the predicted and true transcriptions. This process included converting all text to lowercase, removing punctuation, trimming spaces, eliminating multiple spaces, and filtering out empty strings. These adjustments standardized the text for consistent WER evaluation. The overall model performance was then assessed by averaging the WER across the dataset, with all individual WER scores summed and divided by the total number of audio files. This evaluation process was carried out separately for Javanese and Sundanese, with each language tested using its respective fine-tuned model.

3. RESULTS AND DISCUSSION

In this study, all processes including data pre-processing, model fine tuning, and evaluation were implemented using Python within the Kaggle environment, leveraging 4 cores of an NVIDIA Tesla P100 GPU with 29 GB of RAM, and an Intel Xeon 2.20 GHz CPU with 30 GB of RAM. Libraries primarily available in the Kaggle notebook environment, such as transformers, huggingface_hub, jiwer, torch,

torchaudio, and sklearn, were used extensively throughout the research. For the Whisper model generation, transformers and huggingface_hub libraries were utilized. Before fine tuning, the Tokenizer and FeatureExtractor from the transformers library were imported; the FeatureExtractor processed the raw audio inputs, while the Tokenizer converted model outputs into readable text.

The use of the Whisper Small model represents a novel contribution to ASR research for Javanese and Sundanese. Previous studies, such as those by Cryssiover and Zahra [23] and Arisaputra *et al*. [24], primarily relied on Wav2Vec2 Base, Wav2Vec2 large, or XLS-R models, often requiring additional N-gram language models to boost performance. In contrast, Whisper is pre-trained on a massive, supervised dataset and integrates both acoustic modeling and language generation into a single transformer-based architecture. This end-to-end approach eliminates the need for post-processing with external decoders, which is especially advantageous for low-resource languages where text corpora are scarce. Moreover, unlike models like BERT or mBART that focus on textual language understanding, Whisper is optimized for speech-to-text transcription tasks and does not require separate tokenizers or ASR-specific architectures. This makes it inherently better suited for direct transcription in languages with limited digital resources. To the best of our knowledge, this study represents the first application of Whisper Small to Javanese and Sundanese ASR, establishing a new benchmark for these languages in terms of simplicity, performance, and scalability.

The Javanese and Sundanese datasets, collected from the OpenSLR website, were downloaded as files numbered zero to three along with their respective transcription files [11]. The audio files were in .flac format, and the transcription files included columns such as FileID, UserID, and Transcription. Since the UserID column was irrelevant to this study, it was removed, and the FileID column was renamed to "Audio" to clarify its function. To ensure each entry linked correctly to its corresponding audio file, complete file paths were added to the Audio column. The dataset was split by allocating folders zero to two for training (80%) and validation (20%), as described in Table 1, while the files in folder three comprised the testing set (100%).

Using the modified transcription files, feature extraction was performed with the WhisperProcessor library, which supports both audio pre-processing and tokenization. Audio files were transformed into log-Mel spectrograms, capturing frequency data over time to better prepare the data for model fine-tuning. The transformed datasets consisted of two main components: input_features and Labels. Input_features contained the log-Mel spectrograms derived from the one-dimensional audio arrays, which represent the frequency content of audio signals. This transformation provided the model with information that spans both frequency and time, crucial for accurate speech recognition. Labels, generated by the Tokenizer function, were tokenized representations of the transcription text, encoding each transcription as a sequence of integer IDs (LabelIDs) that the model could interpret. With input_features and Labels in memory, the models were ready for fine-tuning. The same pre-trained Whisper base model was used for both languages, but each model was trained separately with language-specific data: Javanese data for the "Whisper-small-jv" model and Sundanese data for the "Whisper-small-su" model.

After fine-tuning, the evaluation phase commenced. This phase involved testing the models with distinct testing datasets specific to each language. During this phase, each model transcribed its testing dataset, with transcription accuracy measured by WER. For example, the model fine-tuned on Javanese data transcribed the Javanese test set, while the model fine-tuned on Sundanese data handled the Sundanese test set. After transcription, a normalization step was applied to both predicted and true transcriptions. This included removing empty strings, converting all text to lowercase, eliminating extra spaces, trimming whitespace, and removing punctuation, ensuring consistency across WER calculations. The WER for each model was calculated by averaging individual WERs across all transcribed files. The Javanese dataset contained 11,574 audio files, while the Sundanese dataset had 13,820 audio files. The average WER results, as shown in Table 3, revealed that the "Whisper-small-jv" model achieved a WER of 14.97% on the Javanese test set, while the "Whisper-small-su" model achieved a lower WER of 2.03% on the Sundanese test set, demonstrating Whisper's effectiveness in low-resource ASR for these languages.

The performance of the Whisper models was then compared to previous research [23], [24], as shown in Table 4. On the Sundanese dataset, the Whisper model demonstrated significant improvements in WER over the Wav2Vec2 Base and Wav2Vec2 Large models, which achieved WERs of 23.5% and 24%, respectively. However, despite these gains, the Whisper model's performance on the Javanese dataset fell short compared to the XLS-R model with an N-gram language model, which reported a lower WER of 5.4%. While the Whisper model's outcomes on Sundanese data mark a meaningful advancement in low-resource ASR, the findings indicate that additional improvement could be achieved by integrating language models with ASR models. As detailed in Table 3, the Whisper models achieved training WERs of 21.4% for Javanese and 2.1% for Sundanese, with testing WERs of 14.97% and 2.03%, respectively. The significant difference between WERs for Sundanese and Javanese is believed to stem, in part, from inconsistencies in the transcription quality of the Javanese dataset. Although the dataset's creators manually reviewed transcriptions, some errors may still be present [11].

Table 3. Evaluation result

Model	Language	Validation WER (%)	Testing WER (%)
Whisper-small-jv	Javanese	21.4	14.97
Whisper-small-su	Sundanese	2.1	2.03

Table 4. Previous work comparison

Previous work	Model	Testing WER	
		Javanese (%)	Sundanese (%)
[24]	XLSR-300 m + N-gram	10.1	5.4
[23]	Wav2Vec2 Base	-	23.5
	Wav2Vec2 Large	-	24
This study	Fine tuned whisper small	14.97	2.03

The noticeable gap in WER between the two languages 2.03% for Sundanese versus 14.97% for Javanese can be attributed to several factors. First, transcription quality plays a critical role in supervised learning, and the Javanese dataset is known to contain more inconsistent annotations and orthographic variability, as acknowledged by the dataset's creators [11]. These inconsistencies may have introduced noise during training, making it harder for the model to learn reliable mappings between speech and text. In contrast, the Sundanese dataset appears to be more standardized, which likely improved model convergence and recognition accuracy. Second, linguistic complexity and variability may contribute to the difference. Javanese exhibits a more intricate speech level system, including multiple registers such as Ngoko, Madya, and Krama, each with its own vocabulary and usage rules. These honorific distinctions introduce vocabulary diversity and contextual ambiguity, which can increase model confusion during decoding. Sundanese, while it also has polite forms, generally features less morphological variation and a flatter structure in spoken form, which may simplify the learning task for the model.

Third, the difference may also reflect pronunciation variability and dialectal influences. Javanese is spoken across a broader geographical region with stronger dialectal variation such as Solo, Yogyakarta, Banyumas, and Eastern Javanese varieties, while the Sundanese dataset is more regionally concentrated around West Java, particularly Bandung. This regional focus may have reduced variability in speaker accents and phonetic realizations, resulting in a cleaner training signal. Lastly, the distribution of utterance lengths and speaking rates might have played a role. Preliminary inspection suggests that Javanese recordings include a wider range of speaking speeds and more complex sentence constructions, while Sundanese recordings tend to be more concise and consistent. These subtle differences in speaking style could affect how well the model generalizes across utterances during decoding.

4. CONCLUSION

This study set out to investigate the effectiveness of the Whisper ASR model for two low resource languages, Javanese and Sundanese, aiming to improve transcription accuracy for underrepresented languages in Indonesia. By fine tuning Whisper a model pre trained on over 680,000 hours of multilingual audio data this study evaluated the model's capacity to recognize and transcribe speech in these languages accurately. The datasets which were obtained from the openslr (SLR35 for Javanese and SLR36 for Sundanese) were used for the training and testing of the Whisper model on large collections of audio data, where the performance evaluation was based on the WER metric. The results demonstrated substantial improvements in WER for Sundanese, where Whisper outperformed previous models like Wav2Vec2 Base and Wav2Vec2 Large, achieving a significantly lower WER on the testing set. However, while the Whisper model's performance on Sundanese data was noteworthy, the outcomes for Javanese highlighted areas for further enhancement. Compared to the XLS-R model, which used an N-gram language model and achieved a lower WER on Javanese, the Whisper model showed slightly higher WER scores, suggesting that additional language-specific modeling strategies might enhance Whisper's performance for Javanese. Future improvement for Javanese language specific could be done with more hyperparameter tuning and more testing. This discrepancy between Sundanese and Javanese WERs appears to stem partly from transcription inconsistencies within the Javanese dataset, as noted by the dataset's creators. These findings underscore the importance of high-quality, consistent transcriptions in ASR training data, especially for languages with limited resources. Improvements to transcription accuracy in training datasets could further reduce WER in future ASR models for low-resource languages. The performance disparity between the two languages underscores the importance of both transcription quality and linguistic complexity in ASR model training. Future work may benefit from addressing these issues through improved annotation practices and more dialect-aware modeling strategies, particularly for Javanese.

FUNDING INFORMATION

No funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Alim Raharjo	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	
Amalia Zahra				✓	✓			✓		✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

No conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in [OpenSLR] at https://www.openslr.org/resources.php, reference number [11].

REFERENCES

[1] D. Yu and L. Deng, "Automatic speech recognition," *Springer London*, 2015.
[2] Y. Zhang *et al.*, "Google USM: scaling automatic speech recognition beyond 100 languages," *arXiv Computer Science*, 2023.
[3] J. Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022, doi: https://doi.org/10.1561/116.00000050.
[4] D. Amodei *et al.*, "Deep speech 2: end-to-end speech recognition in English and Mandarin," *33rd International Conference on Machine Learning, ICML 2016*, vol. 1, pp. 312–321, 2016.
[5] A. Gulati *et al.*, "Conformer: convolution-augmented transformer for speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 5036–5040, 2020, doi: 10.21437/Interspeech.2020-3015.
[6] S. Kriman *et al.*, "Quartznet: deep automatic speech recognition with 1D time-channel separable convolutions," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2020, pp. 6124–6128, doi: 10.1109/ICASSP40776.2020.9053889.
[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *Proceedings of Machine Learning Research*, vol. 202, pp. 28492–28518, 2023.
[8] Z. Zhao and W. Q. Zhang, "End-to-end keyword search based on attention and energy scorer for low resource languages," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2587–2591, 2020, doi: 10.21437/Interspeech.2020-2613.
[9] C. M. R. K. Raje, "Automatic speech recognition - ASR software market report 2024," *cognitivemarketresearch.com*, 2025. [Online]. Available: https://www.cognitivemarketresearch.com/automatic-speech-recognition-%28asr%29-software-market-report. [accessed Oct. 25, 2024].
[10] D. M. Eberhard, G. F. Simons, and C. D. Fennig, "Ethnologue: languages of the world. twenty-seventh edition," *Languages of the World*, 2024. [Online]. Available: https://www.ethnologue.com.
[11] O. Kjartansson, S. Sarin, K. Pipatsrisawat, M. Jansche, and L. Ha, "Crowd-sourced speech corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali," *6th Workshop on Spoken Language Technologies for Under-Resourced Languages, SLTU 2018*, pp. 52–55, 2018, doi: 10.21437/SLTU.2018-11.
[12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: a framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, 2020.
[13] Z. Maxwell-Smith and B. Foley, "Automated speech recognition of Indonesian-English language lessons on YouTube using transfer learning," *FieldMatters 2023 - 2nd Workshop on NLP Applications to Field Linguistics, Proceedings*, pp. 1–16, 2023, doi: 10.18653/v1/2023.fieldmatters-1.1.
[14] K. Azizah and M. Adriani, "Hierarchical transfer learning for text-to-speech in Indonesian, Javanese, and Sundanese languages," *2020 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020*, pp. 421–428, 2020, doi: 10.1109/ICACSIS51025.2020.9263086.
[15] A. Adila, D. Lestari, A. Purwarianti, D. Tanaya, K. Azizah, and S. Sakti, "Enhancing Indonesian automatic speech recognition: evaluating multilingual models with diverse speech variabilities," *2024 27th Conference on the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques, O-COCOSDA 2024 - Proceedings*, 2024, doi: 10.1109/O-COCOSDA64382.2024.10800336.
[16] O. H. Anidjar, R. Marbel, and R. Yozevitch, "Whisper turns stronger: augmenting Wav2Vec 2.0 for superior ASR in low-resource languages," a*rXiv Computer Science*, 2024.
[17] M.-H. Hsu and H. Lee, "SMILE: speech meta in-context learning for low-resource language automatic speech recognition," *arXiv Electrical Engineering and Systems Science*, 2025.

[18] Y. Liu, X. Yang, and D. Qu, "Exploration of whisper fine-tuning strategies for low-resource ASR," *Eurasip Journal on Audio, Speech, and Music Processing*, no. 1, 2024, doi: 10.1186/s13636-024-00349-3.
[19] L. Zhang, N. Jiang, Q. Wang, Y. Li, Q. Lu, and L. Xie, "Whisper-SV: adapting whisper for low-data-resource speaker verification," *Speech Communication*, vol. 163, 2024, doi: 10.1016/j.specom.2024.103103.
[20] V. Timmel, C. Paonessa, R. Kakooee, M. Vogel, and D. Perruchoud, "Fine-tuning whisper on low-resource languages for real-world applications," *arXiv Computer Science*, 2024.
[21] D. K. Gete *et al*., "Whispering in Amharic: fine-tuning whisper for low-resource language," *arXiv Computer Science*, 2025.
[22] X. de Zuazo, E. Navas, I. Saratxaga, and I. H. Rioja, "Whisper-LM: improving ASR models with language models for low-resource languages," *arXiv Computer Science*, 2025.
[23] A. Cryssiover and A. Zahra, "Speech recognition model design for Sundanese language using WAV2VEC 2.0," *International Journal of Speech Technology*, vol. 27, no. 1, pp. 171–177, 2024, doi: 10.1007/s10772-023-10066-5.
[24] P. Arisaputra, A. T. Handoyo, and A. Zahra, "XLS-R deep learning model for multilingual ASR on low-resource languages: Indonesian, Javanese, and Sundanese," *ICIC Express Letters, Part B: Applications*, vol. 15, no. 6, pp. 551–559, 2024.
[25] J. C. Vásquez-Correa and A. Álvarez Muniain, "Novel speech recognition systems applied to forensics within child exploitation: Wav2vec2.0 vs. whisper," *Sensors*, vol. 23, no. 4, 2023, doi: 10.3390/s23041843.
[26] D. R. Yerramreddy, J. Marasani, P. S. V. Gowtham, G. Harshit, and Anjali, "Speech recognition paradigms: a comparative evaluation of speechbrain, whisper and Wav2Vec2 models," *2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024*, 2024, doi: 10.1109/I2CT61223.2024.10544133.
[27] A. Barcovschi, R. Jain, and P. Corcoran, "A comparative analysis between conformer-transducer, whisper, and wav2vec2 for improving the child speech recognition," in *2023 International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2023*, 2023, pp. 42–47, doi: 10.1109/SpeD59241.2023.10314867.
[28] A. Candido Junior *et al*., "CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese," *Language Resources and Evaluation*, vol. 57, no. 3, pp. 1139–1171, 2023, doi: 10.1007/s10579-022-09621-4.
[29] Wagiati, N. Darmayanti, Y. Yohanarisagarniwa, and D. Zein, "Mapping the dimensions of linguistic distance: a study on quantitative and qualitative geolinguistics of Banjar Sundanese dialect," *European Journal of Language and Culture Studies*, vol. 2, no. 4, pp. 8–17, 2023, doi: 10.24018/ejlang.2023.2.4.87.
[30] W. Udasmoro *et al*., "The preservation of the Javanese language in the Special Region of Yogyakarta," *Indonesian Journal of Geography*, vol. 55, no. 1, pp. 59–59, Feb. 2023, doi: https://doi.org/10.22146/ijg.68183.
[31] R. Alhammad, "The phonology, morphology, and syntax of Sundanese," *Forum for Linguistic Studies*, vol. 5, no. 3, Dec. 2023, doi: https://doi.org/10.59400/fls.v5i3.1945.
[32] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of whisper models to child speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 5242–5246, 2023, doi: 10.21437/Interspeech.2023-935.
[33] R. Jain, A. Barcovschi, M. Y. Yiwere, P. Corcoran, and H. Cucu, "Exploring native and non-native English child speech recognition with whisper," *IEEE Access*, vol. 12, pp. 41601–41610, 2024, doi: 10.1109/ACCESS.2024.3378738.
[34] W. Liu, Y. Qin, Z. Peng, and T. Lee, "Sparsely shared lora on whisper for child speech recognition," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 11751–11755, 2024, doi: 10.1109/ICASSP48485.2024.10447004.
[35] D. S. Park *et al*., "Specaugment: a simple data augmentation method for automatic speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2613–2617, 2019, doi: 10.21437/Interspeech.2019-2680.

BIOGRAPHIES OF AUTHORS

Alim Raharjo Received a bachelor's degree in computer science from the Faculty of Computer Science, Bina Nusantara University, Indonesia in 2023. Currently getting his master's degree at Computer Science in Bina Nusantara University, Indonesia. His main interests include speech recognition, software development, and machine learning. He can be contacted at email: alim.raharjo@binus.ac.id.

Amalia Zahra is a lecturer at the Master of Information Technology, Bina Nusantara University, Indonesia. She received her bachelor's degree in computer science from the Faculty of Computer Science, University of Indonesia (UI) in 2008. She does not have a master's degree. Her Ph.D. was obtained from the School of Computer Science and Informatics, University College Dublin (UCD), Ireland in 2014. Her research interests cover various fields in speech technology, such as speech recognition, spoken language identification, speaker verification, speech emotion recognition, and so on. Additionally, she also has an interest in natural language processing (NLP), computational linguistics, machine learning, and artificial intelligence. She can be contacted at email: amalia.zahra@binus.edu.