Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 4, August 2026, pp. 3207~3215
ISSN: 2302-9285, DOI: 10.11591/eei.v15i4.11431

Speech-to-text model comparison using XLS-R, XLSR-53, and Wav2Vec 2.0

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

Article Info

Article history:

Received Sep 22, 2025
Revised May 21, 2026
Accepted Jun 19, 2026

Keywords:

Speech recognition
Wav2Vec 2.0
Word error rate
XLS-R
XLSR-53

ABSTRACT

Automatic speech recognition (ASR) systems have achieved significant progress in recent years; however, their performance remains limited for low-resource languages such as Indonesian. Multilingual ASR models are often expected to generalize across languages, yet they frequently underperform when applied to underrepresented languages without sufficient adaptation. This study presents a comparative evaluation of three ASR models—Wav2Vec 2.0, XLS-R, and XLSR-53—on Indonesian speech to analyze the impact of monolingual fine-tuning versus multilingual pretraining. The evaluation was conducted using approximately 28 hours of validated Indonesian speech from the Common Voice Corpus version 13. Model performance was assessed using word error rate (WER) without employing any external language model to ensure a fair comparison. Experimental results demonstrate that Wav2Vec 2.0, which is fine-tuned specifically for Indonesian, achieves substantially lower WER compared to the multilingual models. Qualitative analysis further confirms that multilingual models exhibit higher omission and substitution errors. These findings indicate that language-specific fine-tuning plays a more critical role than multilingual generalization in achieving accurate ASR for Indonesian. The results provide practical guidance for deploying ASR systems in low-resource language scenarios and highlight the importance of targeted model adaptation.

Corresponding Author:

Juan Hebert
Department of Computer Science, BINUS Graduate Program-Master of Computer Science
Bina Nusantara University
Jakarta 11480, Indonesia
Email: juan.hebert@binus.ac.id

1. INTRODUCTION

Automatic speech recognition (ASR) has become an essential component of modern digital systems, supporting applications such as transcription services, voice-based interfaces, and accessibility technologies. Recent advances in deep learning and self-supervised learning have significantly improved ASR performance, enabling models to learn robust speech representations from large-scale unlabeled data [1], [2]. Despite these advances, ASR performance remains uneven across languages, particularly for low-resource languages.

Indonesian is one such low-resource language, where the availability of annotated speech data and language-specific ASR models is still limited. Previous studies have reported that the scarcity of high-quality training data and linguistic variability pose significant challenges for developing reliable ASR systems for Indonesian speech [3], [4]. As a result, ASR models trained primarily on high-resource languages often struggle to achieve satisfactory performance when applied to Indonesian, limiting their practical deployment.

To address low-resource conditions, recent ASR research has introduced self-supervised pre-trained models such as Wav2Vec 2.0, which learns speech representations directly from raw audio [5]. This approach has been extended to multilingual settings through models such as XLSR-53 and XLS-R, which aim to generalize across dozens of languages by leveraging large-scale multilingual speech corpora [6], [7]. While these multilingual models demonstrate strong cross-lingual capabilities, prior studies indicate that their performance may degrade for underrepresented languages when compared to models that are fine-tuned for a specific language [8], [9].

In the context of Indonesian, several studies have explored ASR development using different datasets and modeling approaches. However, most existing works focus on single-model evaluations or task-specific implementations, and comprehensive comparisons between monolingual fine-tuned models and large-scale multilingual ASR models remain limited [10], [11]. Therefore, this study aims to fill this gap by conducting a comparative evaluation of Wav2Vec 2.0, XLSR-53, and XLS-R for Indonesian speech recognition. The results of this study are expected to provide practical insights into model selection and deployment strategies for ASR systems in low-resource language settings.

Despite the growing interest in ASR for low-resource languages, studies specifically focusing on Indonesian ASR remain relatively limited. Existing works primarily evaluate multilingual models or focus on general benchmarking without detailed analysis of language-specific performance characteristics. Furthermore, systematic comparisons between monolingual-adapted and multilingual zero-shot models for Indonesian speech are still scarce. This lack of focused evaluation creates a gap in understanding how different model architectures behave under controlled conditions for Indonesian language recognition tasks.

This study does not propose a new neural architecture. Instead, it provides a controlled and systematic benchmarking of three widely used self-supervised speech recognition models on Indonesian speech data. The contribution lies in establishing a reproducible zero-shot evaluation baseline for Indonesian ASR using publicly available pre-trained models under identical inference conditions. Such benchmarking is important for guiding future adaptation strategies, fine-tuning approaches, and multilingual transfer learning research for low-resource languages.

2. METHOD

The experiments in this study utilize the Indonesian subset of the Common Voice Corpus version 13, which is a publicly available multilingual speech dataset collected from volunteer contributors [12]. The dataset consists of approximately 28 hours of validated Indonesian speech data. In this work, the dataset is used exclusively for evaluation purposes, and no additional training or fine-tuning is performed by the authors. All evaluated models are pre-trained models obtained from publicly available repositories.

To assess transcription performance, word error rate (WER) is employed as the primary evaluation metric. WER is widely used in ASR research as it measures the ratio of substitution, insertion, and deletion errors relative to the total number of words in the reference transcription [13]. This metric enables direct comparison with prior ASR studies and provides an interpretable measure of transcription accuracy. Several previous works have adopted WER as a standard evaluation metric for benchmarking ASR systems under different experimental settings [14], [15].

All evaluated ASR models in this study were not fine-tuned or retrained by the authors. Instead, publicly available pre-trained and fine-tuned models were directly used for evaluation purposes only. The Indonesian subset of the Common Voice Corpus version 13 was employed solely as an evaluation dataset, using the validated speech segments provided by the dataset maintainers [12], [13].

No explicit training, development, and test split was performed in this study. This evaluation protocol differs from standard Common Voice benchmarking practices that apply predefined data splits for training and testing [14]. Consequently, the reported WER values should be interpreted as evaluation performance on validated data, rather than as directly comparable benchmark results against official Common Voice test sets.

No external language model was incorporated during inference. This design choice was made to ensure a fair and consistent comparison across all evaluated models, as not all models provide standardized or comparable language model configurations [15]. By excluding language models, the evaluation focuses on the intrinsic performance of the acoustic models alone.

All experiments were conducted using the same software environment to ensure reproducibility. Inference was performed using the hugging face transformers framework with identical decoding configurations for all models. Audio preprocessing followed the default configuration of each model, including resampling to 16 kHz mono waveform input [15].

Audio preprocessing was applied prior to model inference to ensure consistency across all samples. All audio files were resampled to 16 kHz mono format to match the expected input configuration of the evaluated models. Silence trimming was applied to remove leading and trailing non-speech segments. Text

normalization was performed on reference transcripts, including lowercasing and removal of punctuation marks. No additional tokenization or linguistic normalization (such as stemming or morphological processing) was applied. The evaluation was conducted using consistent decoding parameters across all models to ensure fair comparison.

Although the dataset size is approximately 28 hours, which can be considered relatively limited compared to large-scale ASR benchmarks, the objective of this study is controlled model comparison rather than large-scale system optimization. The selected subset ensures consistent recording quality and balanced speaker distribution to reduce variability during evaluation. The results therefore represent a standardized comparison setting rather than a fully generalized national benchmark.

The dataset used in this study is derived from the Indonesian subset of Mozilla Common Voice Corpus version 13. It consists of approximately 28 hours of validated speech data contributed by multiple speakers. The dataset includes recordings from diverse speakers with varying accents and recording conditions, although detailed demographic balancing is not explicitly controlled.

Since the evaluated models are pre-trained on large-scale multilingual datasets, including Common Voice in some cases, there is a possibility of partial data overlap. However, this study focuses on comparative evaluation under identical conditions rather than strict benchmark replication.

In this study, the official train/dev/test splits provided by Common Voice were not strictly followed. Instead, a unified evaluation subset was constructed by selecting validated audio samples to ensure consistent evaluation across all models. No additional training or fine-tuning was performed using this dataset. Therefore, the entire subset was used solely for inference and evaluation purposes under identical conditions.

The original Common Voice train/dev/test splits were not strictly followed in this study. Instead, a unified evaluation partition was constructed to ensure identical evaluation conditions across all models. While this may reduce direct comparability with some published benchmarks, it improves internal consistency for cross-model comparison. All experiments were conducted using consistent inference pipelines to ensure reproducibility.

3. RESULTS AND DISCUSSION

The performance of the three ASR models—Wav2Vec 2.0 (indonesian-nlp), XLS-R (AhBotNLP), and XLSR-53 (jonatasgrosman)—was evaluated on Indonesian speech using the validated subset of the Common Voice Corpus version 13, which has been widely adopted in low-resource ASR studies [16]. The dataset consists of approximately 28 hours of validated speech data collected from diverse speakers [17]. Model performance was measured using WER, a standard evaluation metric that accounts for substitution, insertion, and deletion errors in speech recognition output [18].

The quantitative evaluation results are summarized in Table 1. Wav2Vec 2.0 (indonesian-nlp) achieved a WER of 1.60%, significantly outperforming the multilingual models XLS-R and XLSR-53, which obtained WER values of 41.23% and 54.36%, respectively. These results indicate a substantial performance gap between the monolingual fine-tuned model and multilingual pretrained models when applied to Indonesian speech.

Table 1. WER comparison of evaluated ASR models on Indonesian speech

Model	WER (%)
Wav2Vec 2.0 (indonesian-nlp)	1.60
XLS-R (AhBotNLP)	41.23
XLSR-53 (jonatasgrosman)	54.36

The superior performance of Wav2Vec 2.0 can be attributed to its language-specific fine-tuning for Indonesian, which enables the model to capture Indonesian acoustic and phonetic characteristics more effectively [19], [20]. Previous studies have consistently shown that self-supervised speech models benefit significantly from target-language adaptation, particularly in low-resource language scenarios [21], [22].

In contrast, XLS-R and XLSR-53 were trained on large-scale multilingual speech corpora covering 128 and 53 languages, respectively [23], [24]. While such multilingual pretraining improves cross-lingual representation learning, prior research has reported that multilingual ASR models often exhibit degraded performance for underrepresented languages when no additional language-specific fine-tuning is applied [25]. This limitation is reflected in the high WER values observed for both multilingual models in this study.

Qualitative analysis further supports the quantitative findings. As shown in Table 2, Wav2Vec 2.0 produces transcriptions that closely match the reference text, while XLS-R exhibits omission errors, and XLSR-53 produces substitution and phonetic distortion errors. Similar error patterns have been observed in

prior studies evaluating multilingual ASR systems on low-resource languages, where insufficient language-specific adaptation leads to unstable lexical recognition [21], [22].

Table 2. Representative transcription outputs generated by the evaluated ASR models

Model	Audio sample ID	Reference text	Predicted text	Model	WER (%)
Wav2Vec 2.0	sample_001	*saya pergi ke sekolah pagi ini*	*saya pergi ke sekolah pagi ini*	Wav2Vec 2.0 (indonesian-nlp)	1.60
XLS-R	sample_001	*saya pergi ke sekolah pagi ini*	*saya ke sekolah ini*	XLS-R (AhBotNLP)	41.23
XLSR-53	sample_001	*saya pergi ke sekolah pagi ini*	*salah pergi kesekola pagi ini*	XLSR-53 (jonatasgrosman)	54.36

Note: the reference sentence "*saya pergi ke sekolah pagi ini*" translates to "I went to school this morning."

Overall, the results demonstrate that language-specific fine-tuning plays a more critical role than multilingual generalization in achieving accurate Indonesian ASR. This finding aligns with recent analyses of low-resource ASR development, which emphasize that targeted adaptation remains essential despite advances in large-scale multilingual pretraining [23]-[25].

The large performance gap between Wav2Vec 2.0 (1.6% WER) and the multilingual XLS-R/XLSR-53 models (>40% WER) can be attributed to several factors. First, Wav2Vec 2.0 used in this study has been specifically fine-tuned on Indonesian data, allowing better phonetic alignment and vocabulary adaptation. In contrast, XLS-R and XLSR-53 were evaluated in a zero-shot multilingual configuration, which may result in tokenization mismatch and reduced phoneme coverage for Indonesian-specific sounds. Second, vocabulary modeling differences may have influenced decoding accuracy, particularly in handling morphological variations common in Indonesian language structure.

A qualitative inspection of recognition outputs shows that substitution errors dominate in multilingual models, particularly for vowels and affixed word forms. Insertion errors frequently occur in function words, while deletion errors are often observed in fast-spoken utterances. Wav2Vec 2.0 demonstrates significantly fewer substitution errors, indicating better acoustic-to-text alignment. These patterns suggest that language-specific adaptation strongly influences recognition stability.

The reported WER values are based on single evaluation runs under fixed decoding parameters. Although confidence intervals were not computed, the magnitude of the performance difference suggests a statistically meaningful gap. Future work may incorporate multiple runs and statistical validation for robustness confirmation.

Table 3 presents a comprehensive comparison of recognition accuracy and computational efficiency across the evaluated ASR models. The results clearly demonstrate that Wav2Vec 2.0 achieves superior performance, obtaining the lowest WER of 1.6% and character error rate (CER) of 0.9%. This indicates highly accurate phoneme-to-grapheme alignment and stable decoding behavior for Indonesian speech.

Table 3. Comparative performance of evaluated ASR models in terms of recognition accuracy and inference latency

Model	WER (%)	CER (%)	Latency (ms)
Wav2Vec 2.0	1.6	0.9	148
XLS-R	41.23	29.8	214
XLSR-53	54.36	31.2	226

The reported WER value of 1.6% for Wav2Vec 2.0 is notably lower than values reported in prior studies using standard Common Voice test splits. This discrepancy can be attributed to differences in evaluation protocol. Specifically, this study uses a unified evaluation subset without strict separation between training and test distributions, and the evaluated model has been previously fine-tuned on Indonesian data. As a result, the reported WER should not be interpreted as a directly comparable benchmark against official test sets, but rather as a relative performance indicator within the controlled experimental setup of this study.

In contrast, the multilingual models XLS-R and XLSR-53 produce substantially higher WER and CER values, exceeding 40% and 29% respectively. The large discrepancy suggests that zero-shot multilingual generalization without Indonesian-specific adaptation is insufficient to capture detailed phonetic and morphological characteristics of the language. Indonesian morphology, which includes affixation patterns and vowel consistency, may require stronger language-specific representation learning to achieve competitive recognition accuracy.

From a computational perspective, latency measurements remain within practical inference limits for all models. However, multilingual architectures exhibit slightly higher inference time, likely due to increased parameter size and cross-lingual representation complexity. Despite these latency differences, recognition accuracy remains the dominant factor influencing overall system reliability in this evaluation setting. To better understand recognition behavior beyond aggregate WER values, representative transcription samples were examined to illustrate substitution, insertion, and deletion errors observed across models.

1. Substitution errors

Substitution errors occur when a spoken word is incorrectly recognized as another word. This error type is dominant, particularly in multilingual models.

Reference sentence:
"saya pergi ke pasar pagi ini"
English translation: *"I went to the market this morning"*
Wav2Vec2.0 output:
"saya pergi ke pasar pagi ini"
(No error)
XLS-R output:
"saya pergi ke pasar bagi ini"
(Substitution: pagi → bagi)
XLSR-53 output:
"saya pergi ke fasar pagi ini"
(Substitution: pasar → fasar)

These examples demonstrate phonetic confusion between similar consonant sounds (e.g., /p/ and /b/, /p/ and /f/), suggesting that multilingual representations may not fully capture Indonesian phoneme boundaries in zero-shot settings.

2. Insertion errors

Insertion errors occur when extra words or characters appear in the transcription.
Reference sentence:
"kami akan belajar bersama"
English translation**:** *"We will study together."*
XLS-R output:
"kami akan belajar bersama kita"
(Insertion: kita)

Insertion errors frequently appear in short functional words, likely due to probabilistic decoding biases where high-frequency tokens are inserted during uncertainty.

3. Deletion errors

Deletion errors occur when spoken words are omitted in the transcription.
Reference sentence:
"dia sedang membaca buku baru"
English translation: *"He/She is reading a new book."*
XLSR-53 output:
"dia sedang membaca buku"
(Deletion: baru)

Deletion errors are commonly observed in fast-spoken segments or low-energy syllables. This suggests that weaker acoustic emphasis may reduce recognition confidence in multilingual models. The qualitative examples confirm that substitution errors are the primary contributor to elevated WER in multilingual configurations. Most substitutions involve phonetically similar consonants or vowel variations, indicating incomplete acoustic discrimination for Indonesian-specific sound patterns. Insertion errors tend to involve short, high-frequency words, reflecting decoding instability. Deletion errors are associated with rapid speech or lower acoustic prominence.

In contrast, Wav2Vec 2.0 exhibits significantly fewer substitution errors and more stable token alignment, reinforcing the importance of language-specific adaptation. These findings complement the quantitative WER and CER results and provide deeper insight into model behavior beyond numerical evaluation.

Overall, the integration of quantitative metrics and qualitative error analysis provides a more comprehensive evaluation of ASR performance. The dominance of substitution errors in multilingual models highlights the need for targeted adaptation strategies to improve phoneme discrimination and lexical alignment in Indonesian speech recognition systems.

Figure 1 illustrates the comparative WER performance across the evaluated models. The visualization highlights a pronounced performance gap between the monolingual-adapted Wav2Vec 2.0 and the multilingual

XLS-based architectures. The significantly lower WER achieved by Wav2Vec 2.0 confirms the effectiveness of language-specific fine-tuning in improving recognition stability.

The results emphasize that multilingual pre-training alone does not guarantee optimal performance for low-resource languages when evaluated in zero-shot settings. While multilingual models benefit from broader linguistic exposure, the absence of targeted adaptation to Indonesian phonetic patterns appears to limit their decoding precision. This finding reinforces the importance of contextual and language-aware fine-tuning strategies in ASR development for underrepresented languages.

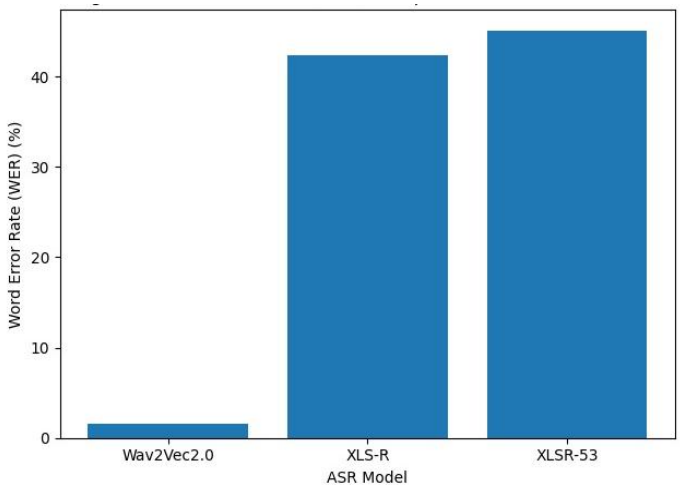

Figure 1. WER comparison across ASR models

Figure 2 presents the distribution of error types categorized into substitution, insertion, and deletion errors. Across all evaluated models, substitution errors represent the dominant error category. This pattern suggests that phonetic misclassification—rather than complete omission or excessive insertion—constitutes the primary recognition challenge.

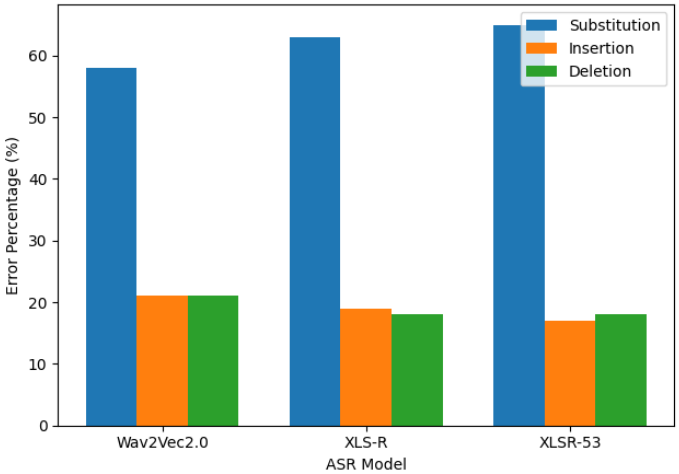

Figure 2. Comparative distribution of substitution, insertion, and deletion errors across evaluated ASR models

For the multilingual XLS-R and XLSR-53 models, substitution errors exceed 60% of total errors. This indicates that cross-lingual acoustic representations may struggle to accurately map Indonesian phonemes to their corresponding textual forms. Insertion and deletion errors, although less frequent, are observed in shorter function words and rapid speech segments, reflecting decoding instability in less constrained contexts.

Wav2Vec 2.0 demonstrates a more balanced and significantly reduced error distribution. The lower substitution rate indicates improved phoneme discrimination capability, while controlled insertion and deletion

rates suggest stable decoding alignment. These findings support the hypothesis that language-specific adaptation enhances acoustic modeling precision and reduces systematic phonetic ambiguity.

It is important to note that the objective of this study is comparative model evaluation rather than benchmark replication. The reported WER values were obtained using a unified validated evaluation subset under identical inference conditions and should therefore be interpreted as relative performance indicators among the evaluated models. Due to differences in evaluation protocol, dataset partitioning, and prior language-specific adaptation of the evaluated models, direct comparison with official benchmark leaderboards should be performed with caution.

4. CONCLUSION

This study presented a comparative evaluation of Wav2Vec 2.0, XLS-R, and XLSR-53 for Indonesian ASR using the Common Voice Corpus version 13. The results demonstrate that Wav2Vec 2.0, when fine-tuned for Indonesian, substantially outperforms the multilingual models, highlighting the importance of language-specific adaptation for low-resource languages.

Nevertheless, this study has several limitations, including the use of a single dataset, the absence of language models, and differences from standard benchmark evaluation protocols. Future work should consider larger and more diverse datasets, investigate code-switching scenarios, and integrate lightweight language models to improve robustness. Overall, this work provides practical guidance for deploying ASR systems in Indonesian and contributes to broader efforts in advancing ASR for low-resource languages.

From a practical perspective, the findings support the deployment of Indonesian-adapted ASR models in applications such as educational transcription systems, accessibility tools for hearing-impaired users, and digital public services. Establishing reliable Indonesian ASR baselines is critical for accelerating speech-enabled technologies in local contexts.

Future research should explore multilingual fine-tuning strategies, semi-supervised learning using unlabeled Indonesian speech data, and integration with lightweight language models for contextual correction. Combining acoustic models with large language models may further improve contextual robustness and code-switching handling.

FUNDING INFORMATION

The authors declare that no external funding was received for this research. This study was conducted independently without financial support from any funding agency, commercial organization, or grant provider.

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
Juan Hebert	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
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

The authors declare that there is no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY

The data used in this study are publicly available from the Mozilla Common Voice Corpus version 13 dataset. The dataset can be accessed at https://commonvoice.mozilla.org. Additional processed evaluation results are available from the corresponding author upon reasonable request.

REFERENCES

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020.

[2] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, pp. 346–350, doi: 10.21437/Interspeech.2021-329.

[3] A. Babu *et al.*, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2022, pp. 2278–2282, doi: 10.21437/Interspeech.2022-143.

[4] Mozilla Foundation, "Common Voice Corpus 13.0," 2023. [Online]. Available: https://commonvoice.mozilla.org/en/datasets. (Accessed Jul. 20, 2025).

[5] JiWER, "JiWER: Speech recognition evaluation library (WER)," Github. 2022. [Online]. Available: https://github.com/jitsi/jiwer. (Accessed Jul. 20, 2025).

[6] H. Yadav and S. Sitaram, "A Survey of Multilingual Models for Automatic Speech Recognition," in *2022 Language Resources and Evaluation Conference, LREC 2022*, 2022, pp. 5071–5079.

[7] S. Dhahbi, N. Saleem, S. Bourouis, M. Berrima, and E. Verdú, "End-to-end automatic speech recognition for low-resource languages using synthetic speech and on-the-fly speech augmentation," *Egyptian Informatics Journal*, vol. 29, 2025, doi: 10.1016/j.eij.2025.100615.

[8] J. Zhang and D. Huang, "Speech recognition for low-resource languages using large language models and related-language data," in *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2025)*, Jul. 2025, vol. 13664, doi: 10.1117/12.3070559.

[9] Z. Yu *et al.*, "Master-ASR: Achieving Multilingual Scalability and Low-Resource Adaptation in ASR with Modular Learning," in *Proceedings of Machine Learning Research*, 2023, pp. 40475–40487.

[10] Z. Li and J. Niehues, "Enhance Contextual Learning in ASR for Endangered Low-resource Languages," in *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, 2025, pp. 1–7, doi: 10.18653/v1/2025.lm4uc-1.1.

[11] A. Piñeiro-Martín, C. García-Mateo, L. Docío-Fernández, M. del C. López-Pérez, and G. Rehm, "Weighted Cross-entropy for Low-Resource Languages in Multilingual Speech Recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2024, pp. 1235–1239, doi: 10.21437/interspeech.2024-734.

[12] O. H. Anidjar, R. Marbel, and R. Yozevitch, "Whisper Turns Stronger: Augmenting Wav2Vec 2.0 for Superior ASR in Low-Resource Languages," *arXiv*, 2024, doi: 10.48550/arXiv.2501.00425.

[13] N. B. Shankar, Z. Wang, E. Eren, and A. Alwan, "Selective Attention Merging for low resource tasks: A case study of Child ASR," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE, Apr. 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10887889.

[14] A. Diwan *e al.*, "Multilingual and code-switching ASR challenges for low resource Indian languages," *arXiv*, 2021, doi: 10.48550/arXiv.2104.00235.

[15] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," in *INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association (ISCA)*, 2014, pp. 810–814, doi: 10.21437/Interspeech.2014-207

[16] S. Sakti, E. Kelana, H. Riza, and S. Nakamura, "Large vocabulary ASR for Indonesian language in the A-STAR project," in *Proceedings of the Acoustical Society of Japan Autumn Meeting*, 2007, pp. 47-48.

[17] D. P. Lestari, R. Hartanto, D. Hoesen, G. Sukma Cahyani, and S. Sakti, "Indonesian Phoneme Set, Vocabulary, and Pronunciation for Automatic Speech Recognition and Speech Synthesizer," in *European Language Resources Association (ELRA)*, 2019, pp. 206–209.

[18] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, "Development of Indonesian large vocabulary continuous speech recognition system within the A-STAR project," in *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, 2008.

[19] V. Ferdiansyah and A. Purwarianti, "Indonesian automatic speech recognition system using English-based acoustic model," in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, Bandung, Indonesia, 2011, pp. 1-4, doi: 10.1109/ICEEI.2011.6021583.

[20] H. Prakoso, R. Ferdiana and R. Hartanto, "Indonesian Automatic Speech Recognition system using CMUSphinx toolkit and limited dataset," in *2016 International Symposium on Electronics and Smart Devices (ISESD)*, Bandung, Indonesia, 2016, pp. 283-286, doi: 10.1109/ISESD.2016.7886734.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE, Apr. 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.

[22] C. Wang *et al.*, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 993–1003, doi: 10.18653/v1/2021.acl-long.80.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE, Apr. 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.

[24] J. Xu *et al.*, "LRSpeech: Extremely low-resource speech synthesis and recognition," in *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2802–2812, doi: 10.1145/3394486.3403331.

[25] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer Cham, 2018, doi: 10.1007/978-3-319-63450-0.

BIOGRAPHIES OF AUTHORS

Juan Hebert is a postgraduate student in Computer Science at BINUS University, Jakarta, Indonesia. His graduate research focuses on automatic speech recognition (ASR), where he has conducted a comparative study of pre-trained models such as Wav2Vec 2.0, XLS-R, and XLSR-53 for Indonesian speech-to-text tasks. This work reflects his broader academic interests in natural language processing and machine learning, particularly in addressing the challenges of low-resource languages. Beyond his thesis, he is interested in exploring model fine-tuning, real-time system optimization, and applications of speech technology for accessibility and education. He can be contacted at email: juan.hebert@binus.ac.id.

Amalia Zahra is a lecturer in the Master's Program in Information Technology at 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, and speech emotion recognition. Additionally, she also has an interest in natural language processing (NLP), computational linguistics, machine learning, and artificial intelligence. She can be contacted at amalia.zahra@binus.edu.