Indonesian Journal of Computational Language Studies
Vol. 1, No. 3, September 2026, pp. 149~163
ISSN: 3163-7884, DOI: https://doi.org/10.67490/ijcl.v1i3.1026

Listening to the archipelago: automatic speech recognition, accent variation, and the unequal modeling of Indonesian voices

Imam Masrur
Universitas Islam Negeri Syekh Wasil Kediri, Indonesia

Article Info

Article history:

Received Jul 9, 2026
Revised Aug 14, 2026
Accepted Sep 25, 2026

Keywords:

accent variation;
automatic speech recognition;
Indonesian languages;
sociotechnical visibility;
speech technology.

ABSTRACT

Background: Indonesia's archipelagic linguistic ecology makes automatic speech recognition (ASR) a problem of sociotechnical representation, because spoken Indonesian is shaped by regional accent, local-language influence, code-switching, and uneven access to standardized speech resources. **Objective:** This study examines how Indonesian ASR resources, evaluation practices, and model documentation make regional and accent-varied voices differentially visible, testable, and interpretable. **Method:** This study applies a corpus-based ASR audit to 18 public and verifiable corpus items, combining archipelagic representation analysis, accent-sensitive error-evaluation readiness, and sociotechnical visibility assessment. **Results:** The corpus shows that regional-language resources are more visible than dialectal Indonesian speech directly suitable for accent-sensitive ASR evaluation. Error-evaluation readiness is uneven because several items require transcript alignment, utterance segmentation, metadata enrichment, or model-output comparison before reliable subgroup analysis can be conducted. Sociotechnical visibility is fragmented across direct speech evidence, preprocessing-dependent materials, literature-mediated findings, and model-opaque documentation. **Implication:** These findings reposition Indonesian ASR fairness as a question of audibility, evaluability, and linguistic justice rather than aggregate transcription accuracy alone. **Novelty:** This study contributes an archipelagic framework for assessing unequal machine listening in Indonesian ASR by integrating corpus representation, evaluation readiness, and sociotechnical visibility.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Imam Masrur
Universitas Islam Negeri Syekh Wasil Kediri, Indonesia
Jl. Sunan Ampel No. 7, Ngronggo, Kota Kediri, Jawa Timur, 64127, Indonesia
Email: imammasrur@uinkediri.ac.id

1. INTRODUCTION

Indonesia's speech-technology problem is inseparable from its linguistic geography. Official language mapping reports that Indonesia has 718 regional languages, with only a small portion classified as safe and many others described as vulnerable, declining, endangered, critically endangered, or extinct. This diversity means that spoken Indonesian rarely appears as a single homogeneous acoustic object; it is shaped by regional phonologies, local lexicons, rhythm, intonation, code-switching, and uneven access to standard-language institutions. For automatic speech recognition (ASR), this condition is not a peripheral complication but a central test of representational adequacy. A system that performs acceptably on formal, urban, or read Indonesian may still misrecognize speakers whose Indonesian is marked by Javanese, Sundanese, Balinese, Makassarese, Buginese, Papuan, or other regional influences. The crucial issue, therefore, is not only whether ASR can transcribe Indonesian, but which Indonesian voices become reliably audible to machines.

Existing research has substantially improved the technical foundations of Indonesian ASR. Abidin et al. [1] examined acoustic modeling and lexicon types for Indonesian speech recognition, while Arisaputra and Zahra [2] and Arisaputra et al. [3] explored XLSR-53 and XLS-R for Indonesian, Javanese, and Sundanese speech. Adila et al. [4] further evaluated multilingual models under diverse Indonesian speech variabilities, and Tazakka et al. [5] addressed Indonesian-English code-switching through semi-supervised speech-chain learning. Internationally, ASR scholarship has moved toward bias-sensitive evaluation, showing that accent, dialect, race, ethnicity, age, and speaker traits can shape uneven system performance [6], [7], [8], [9]. Yet Indonesian ASR studies still tend to emphasize model improvement, benchmark performance, or low-resource adaptation more than the sociolinguistic question of unequal audibility. What remains underexamined is how accent variation within Indonesian itself becomes modeled, normalized, or erased by ASR systems.

This study addresses that gap by examining the relationship between ASR, accent variation, and the unequal modeling of Indonesian voices. Rather than treating accent as noise to be reduced, this study asks how regionally inflected Indonesian speech is represented in ASR evaluation and what kinds of transcription errors emerge when systems encounter archipelagic variation. The guiding questions are: how do selected ASR systems perform across Indonesian speech samples marked by regional accent and local-language influence; what error types appear across lexical, phonological, prosodic, code-switching, and named-entity levels; and how can those errors be interpreted without reducing speakers to deficient versions of standardized Indonesian? Methodologically, this study combines corpus-based ASR auditing with sociolinguistic error analysis, using public and verifiable speech datasets, transcripts, metadata, model outputs, and coding categories. Its objective is descriptive and diagnostic, not causal: it identifies patterned disparities and interpretive risks rather than claiming that speaker identity directly causes recognition failure.

This study argues that Indonesian ASR should be evaluated as an archipelagic sociotechnical system rather than as a neutral transcription pipeline. Average word error rate may indicate general performance, but it can conceal subgroup disparities, culturally consequential deletions, regional-name failures, or normalization of local speech forms. Recent work on ASR fairness warns that bias measurement must move beyond aggregate accuracy toward performance gaps, subgroup visibility, and error consequences [10], [11]. Sociolinguistic critiques also show that accent is not merely an acoustic property; it is heard through social expectations about legitimacy, standardness, and competence [12], [8]. The provisional argument tested here is that ASR systems may recognize standardized Indonesian more consistently than regionally marked Indonesian, thereby reproducing a hierarchy of machine listenability. If confirmed, this would reposition ASR evaluation in Indonesia as a question of linguistic justice, not only computational optimization.

2. LITERATURE REVIEW

Automatic speech recognition is commonly defined as computational conversion of spoken language into written text, yet recent scholarship increasingly treats it as a sociotechnical process shaped by data selection, acoustic modeling, linguistic assumptions, and evaluation metrics. Classical ASR research emphasized signal processing, lexicon design, and acoustic modeling, while contemporary systems rely on multilingual pretraining, self-supervised representation learning, and large-scale weak supervision [13], [14]. Indonesian ASR studies reflect this shift through lexicon-based acoustic modeling, XLSR-53, XLS-R, Whisper, MMS, medical dictation, and code-switching systems [1], [2], [4], [15], [5]. Thus, ASR must be understood as both technical transcription and selective linguistic modeling.

The main indicators of ASR research include architecture, training data, domain condition, language coverage, and evaluation metric. Abidin et al. [1] foreground lexicon variation, whereas Pardede et al. [16] and Dwi et al. [17] examine feature extraction and deep neural modeling for Indonesian recognition. Arisaputra et al. [3] extend this concern to Indonesian, Javanese, and Sundanese, while Azis et al. [18] develop low-resource ASR for Makassar dialect speech. Broader multilingual studies show that adaptation, residual modules, parameter-efficient tuning, and low-resource strategies can improve recognition without solving representational imbalance [19], [20], [21]. These works suggest that performance gains require subgroup-sensitive evaluation.

2.1. Accent variation

Accent variation refers to patterned phonetic, phonological, prosodic, and rhythmical differences associated with linguistic history, region, social identity, and language contact. It is often misunderstood as deviation from standard speech, although sociolinguistic research treats accent as legitimate linguistic variation. In ASR, this distinction matters because “accented speech” may be framed either as noise to be normalized or as evidence of plural speech communities. Studies on British English, Appalachian English, Newcastle English, Arabic dialects, Mandarin, Cantonese, and World Englishes show that accent recognition is shaped by both acoustic features and social classification [8], [22], [23], [24], [25], [12].

Accent-sensitive ASR can be analyzed through segmental substitutions, prosodic mismatches, lexical misrecognition, accent-specific codebooks, speaker embeddings, and adaptation mechanisms. Studies have proposed adversarial transfer learning, accent-discriminative encoders, accent-specific codebooks, manifold mixup, and mixture-of-experts architectures to improve recognition under accent variability [26], [27], [28], [29], [30]. Other work evaluates accent identification and accentedness assessment through multi-embedding, prosodic, vocal, and phonetic features [31], [32], [33]. For Indonesia, accent variation must include local-language influence, code-switching, and regional Indonesian forms rather than foreign-accent paradigms alone.

2.2. Algorithmic bias in ASR

Algorithmic bias in ASR refers to systematic performance disparity across speaker groups, speech varieties, or social categories. Early fairness discussions measured bias through unequal word error rates, but recent work criticizes narrow metrics that ignore provenance, transcription context, accent ideology, and social consequences [34], [8], [35], [10]. ASR bias is not reducible to individual pronunciation; it may emerge from underrepresentative datasets, standard-language assumptions, annotation norms, model training distribution, and benchmark design. Studies of ethnicity-related dialects, non-native accents, and diverse speaker traits demonstrate that recognition systems can reproduce social hierarchies through technical errors [9], [36], [7], [37].

Bias-sensitive ASR evaluation therefore requires indicators beyond average accuracy. Relevant dimensions include subgroup performance gaps, worst-group error rates, phoneme-level bias, named-entity failure, deletion of culturally marked terms, accent normalization, and error consequences in institutional settings. Recent work proposes inclusive ASR, content analysis of accent misconceptions, debiasing strategies, speech foundation model adaptation, and expanded data collection for atypical or underrepresented speech [6], [11], [38], [39], [40]. In Indonesian AI research, concerns about regional, gender, and sociolinguistic bias are increasingly explicit [41]. This study therefore positions Indonesian ASR as an archipelagic fairness problem: machine listening must be evaluated by whose voices remain recognizable, not only by aggregate transcription scores.

3. METHOD

This study uses corpus items and utterance-level ASR outputs as its unit of analysis. At corpus level, each retained item is treated as a document with metadata on source, region, language or accent scope, access status, and analytic function. At speech level, each audio segment is linked to a gold transcript and machine-generated transcript when available. This two-level design allows this study to distinguish between resource availability and recognition performance. Primary data consist of public speech datasets; secondary data consist of ASR studies; context data consist of model documentation. This structure prevents model claims, dataset evidence, and interpretive literature from being collapsed into one empirical category.

This study adopts a corpus-based ASR audit with mixed quantitative and qualitative procedures. This design is appropriate because unequal speech recognition cannot be examined through accuracy scores alone. Aggregate metrics such as word error rate and character error rate identify transcription distance, but they do not explain whether errors erase local names, regional lexical items, code-switching, or accent-related phonological forms. Recent work on inclusive ASR, accent bias, and low-resource modeling demonstrates that recognition systems must be evaluated by subgroup visibility and error consequence, not only average performance [6], [8], [10].

Table 1. Composition of dataset and corpus sample

Code	Data Type	Source	Period	Quantity	Characteristics
DOC-ASR-ID-001	Primary data	Mozilla Common Voice	2026	1 item	Indonesian scripted speech
DOC-ASR-ID-002	Primary data	Zenodo ID2	2025	1 item	Dialectal Indonesian from 31 speakers
DOC-ASR-ID-003	Primary data	OpenSLR	2016–2026	1 item	Javanese ASR data
DOC-ASR-ID-004	Primary data	OpenSLR	2016–2026	1 item	Sundanese ASR data
DOC-ASR-ID-005	Primary data	OpenSLR	2016–2026	1 item	Javanese TTS/speech data
DOC-ASR-ID-006	Primary data	OpenSLR	2016–2026	1 item	Sundanese TTS/speech data
DOC-ASR-ID-007	Primary data	Mendeley Data	2024–2025	1 item	Balinese, Badung dialect, speech levels
DOC-ASR-ID-008	Primary data	Google/Hugging Face FLEURS	2022–2026	1 item	Indonesian and Javanese benchmark speech
DOC-ASR-ID-009	Primary data	Indonesian NLP/Hugging Face	2021–2026	1 item	Acehnese, Balinese, Bugisnese, Indonesian, Minangkabau, Javanese, Sundanese
DOC-ASR-ID-010	Secondary data	arXiv/O-COCOSDA	2024	1 item	Multilingual ASR with speech variability
DOC-ASR-ID-011	Secondary data	ACL Anthology	2024	1 item	Indonesian-English code-switching ASR
DOC-ASR-ID-012	Secondary data	IIETA/arXiv	2022–2023	1 item	Indonesian ASR with XLSR-53
DOC-ASR-ID-013	Secondary data	arXiv	2024	1 item	XLS-R for Indonesian, Javanese, Sundanese
DOC-ASR-ID-014	Secondary data	IAES IJ-AI	2024	1 item	Indonesian medical dictation
DOC-ASR-ID-015	Secondary data	Wiley	2022	1 item	Indonesian acoustic model and lexicon types
DOC-ASR-ID-016	Context data	Indonesian NLP/Hugging Face	2021–2026	1 item	Multilingual speech recognition model card
DOC-ASR-ID-017	Context data	OpenAI	2022–2026	1 item	Whisper multilingual ASR documentation
DOC-ASR-ID-018	Context data	Meta AI	2023–2026	1 item	MMS speech model documentation

Information sources were selected from public, traceable, and institutionally verifiable platforms. Primary speech resources were drawn from Mozilla Common Voice, OpenSLR, Zenodo, Mendeley Data, Hugging Face, and FLEURS. Secondary literature included Indonesian ASR studies on XLSR-53, XLS-R, code-switching recognition, medical dictation, and lexicon-based acoustic modeling [1], [4], [2], [3], [15], [5]. Model-context documents were used only to record candidate system characteristics. This separation preserves evidentiary hierarchy between corpus material, scholarly interpretation, and model documentation.

Analysis proceeds in three connected stages. First, an archipelagic representation audit classifies each resource by data type, language or accent scope, region, transcript availability, speech context, and access status. Second, sampled utterances are prepared for ASR comparison by aligning audio, gold transcript, model output, WER, CER, and coded error categories, including substitution, deletion, insertion, named-entity error, local lexical misrecognition, code-switching failure, and semantic distortion. Third, sociotechnical interpretation evaluates whether errors affect audibility, institutional usability, or linguistic justice. This staged procedure permits descriptive disparity claims while avoiding unsupported causal claims about speaker identity.

Table 2. Operational matrix of analysis

Analytical Stage	Main Question	Input Data	Procedure	Output
Archipelagic representation audit	Which Indonesian voices are represented, weakly represented, or absent?	Corpus metadata and text register	Classify source type, region, language/accent scope, transcript access, and speech context	Coverage matrix and representation-gap notes
Accent-sensitive error typology	What kinds of recognition errors appear across accent-varied speech?	Audio, gold transcript, ASR output	Calculate WER/CER and code substitution, deletion, insertion, local lexical, named-entity, code-switching, and semantic errors	Error-frequency table and subgroup comparison
Sociotechnical visibility assessment	How do errors affect audibility and institutional usability?	Error coding and analytic memos	Interpret high-impact errors by communicative context and practical consequence	Risk typology and recommendation priorities

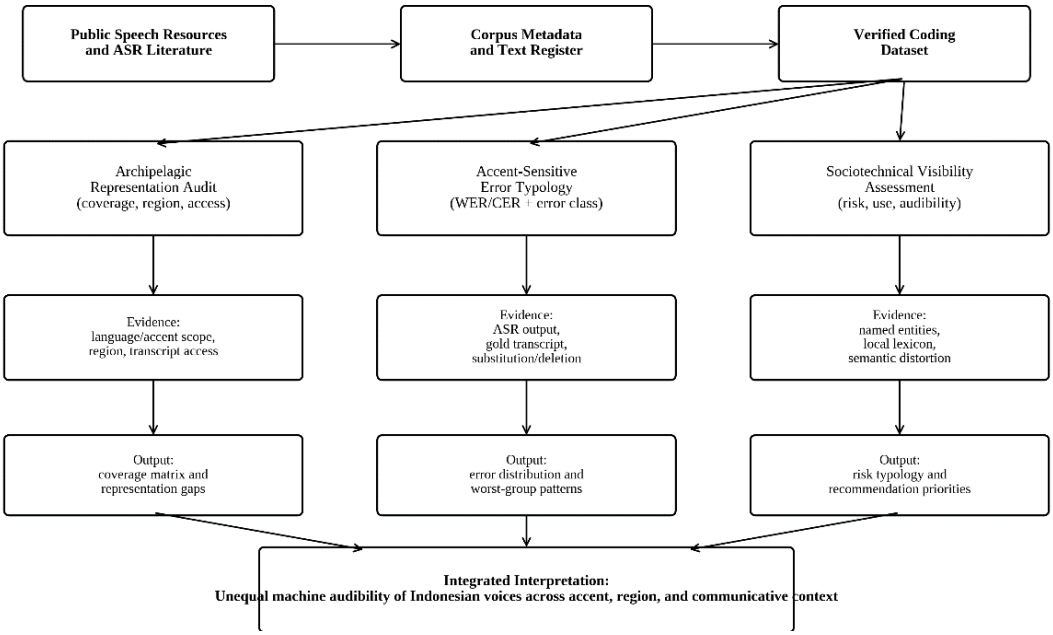

Figure 1. Operational matrix of accent-sensitive Indonesian ASR analysis

4. RESULTS

4.1. Archipelagic representation of Indonesian speech resources

Indonesian speech resources in this corpus are unevenly distributed across three evidentiary layers: public speech datasets, scholarly ASR evaluation studies, and model-context documents. The largest portion consists of regional-language speech resources, followed by Indonesian ASR evaluation studies and model documentation. This composition shows that Indonesian speech technology is not data-poor in a simple sense; rather, its resources are distributed across different kinds of linguistic objects. Some materials document regional languages directly, while fewer items explicitly capture Indonesian as spoken through regional accent influence. The corpus therefore reveals a methodological tension: Indonesia's archipelagic speech diversity is visible in public resources, but not always organized in ways that allow direct evaluation of accent-related ASR disparity.

The dominant pattern is the separation between regional-language representation and dialectal Indonesian representation. Seven items, or 38.9%, concern regional-language speech resources, including Javanese, Sundanese, Balinese, and multilingual local-language materials. These resources are evidenced by corpus entries such as "Transcribed audio data for Javanese," "Transcribed audio data for Sundanese," and "Balinese, Badung dialect, speech levels." By contrast, only one item, or 5.6%, directly documents dialectally varied Indonesian speech, described as "dialectal variations from 31 Indonesian speakers." This contrast matters because regional-language ASR datasets do not automatically answer how ASR systems handle Indonesian spoken with regional phonological, lexical, or prosodic influence.

Table 3. Corpus representation patterns across Indonesian ASR resources

Main Category	Subcategory	Indicator	Frequency	Percentage	Data Variation	Corpus Evidence / Short Quotation	Data Code	Analytic Reading
Regional-language speech resources	Javanese, Sundanese, Balinese, and multilingual local-language datasets	Public datasets explicitly linked to regional-language speech, ASR, TTS, or multilingual speech recognition	7	38.9%	Javanese ASR, Sundanese ASR, Javanese TTS, Sundanese TTS, Balinese TTS, FLEURS Indonesian/Javanese, multilingual Indonesian NLP model resources	"Transcribed audio data for Javanese"; "Transcribed audio data for Sundanese"; "Balinese, Badung dialect, speech levels"	DOC-ASR-ID-003–009	Regional speech is visible as a resource category, but often as separate language data rather than as accent-varied Indonesian speech.
Indonesian ASR evaluation studies	Model evaluation, code-switching, medical dictation, multilingual ASR, XLSR/XLS-R, lexicon-based ASR	Scholarly studies assessing Indonesian ASR methods or domains	6	33.3%	Multilingual ASR, Indonesian-English code-switching, XLSR-53, XLS-R, medical dictation, acoustic model with multiple lexicon types	"Indonesian-English code-switching ASR"; "Indonesian ASR with XLSR-53"; "Indonesian medical dictation"	DOC-ASR-ID-010–015	Indonesian ASR research is methodologically active, but accent inequality is more often implicit than treated as a central analytic object.
Model-context documentation	Whisper, MMS, and Indonesian NLP speech-recognition model documentation	Public documentation describing candidate ASR systems and model scope	3	16.7%	Multilingual weak supervision, mass multilingual speech modeling, Indonesian speech-recognition model cards	"Whisper multilingual ASR documentation"; "MMS speech model documentation"	DOC-ASR-ID-016–018	Model documentation supports reproducibility, but it does not by itself demonstrate equitable performance across Indonesian accents.
National Indonesian baseline corpus	Scripted Indonesian speech	Dataset positioned as Indonesian-language ASR material	1	5.6%	National Indonesian scripted speech	"Dataset intended for ASR training and evaluation"	DOC-ASR-ID-001	Standard Indonesian is present as a baseline, but its singular presence risks standing in for wider Indonesian speech diversity.
Dialectal Indonesian corpus	Dialectally varied Indonesian speaker data	Dataset explicitly documenting regional influence in Indonesian speech	1	5.6%	Javanese-, Sundanese-, Batak-, Balinese-, and Minang-influenced Indonesian from documented speakers	"Dataset features dialectal variations from 31 Indonesian speakers"	DOC-ASR-ID-002	Dialectal Indonesian is the strongest direct bridge between accent variation and ASR fairness, but its corpus share remains limited.

This distribution supports a sociotechnical reading of ASR representation. Accent inequality is not produced only at the moment of transcription; it begins earlier in the way corpora separate "Indonesian," "regional language," and "accented Indonesian" as distinct objects. Indonesian ASR studies have advanced multilingual modeling, code-switching recognition, lexicon design, and low-resource adaptation [1], [4], [3], [5]. Yet fairness-oriented ASR scholarship warns that aggregate availability does not guarantee subgroup audibility [6], [8], [9]. This study therefore treats corpus representation itself as the first layer of machine-listening inequality.

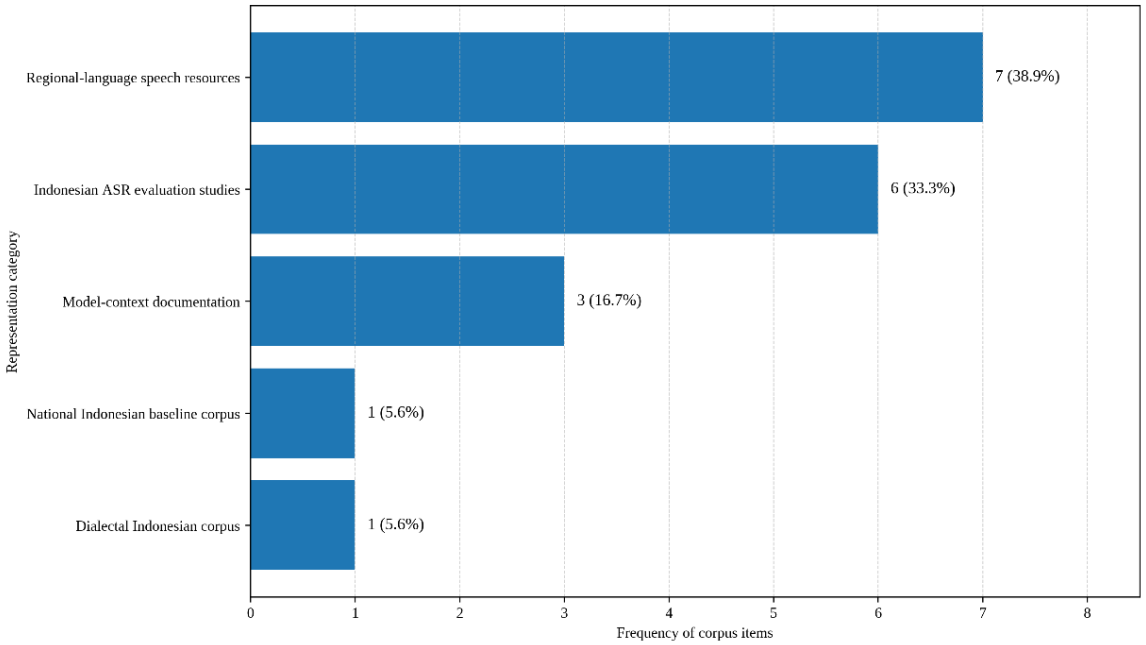

Figure 2. Corpus representation patterns in Indonesian ASR resources

4.2. Error-evaluation readiness and accent-sensitive audit potential

Machine-listening inequality becomes observable only when the corpus can connect speech resources, transcripts, model outputs, and interpretable error categories. The same eighteen corpus items used in the previous analysis show four different levels of readiness for accent-sensitive ASR auditing. The largest group consists of literature-based error evidence, while high utterance-level audit readiness appears in five items. This distribution means that Indonesian ASR scholarship already provides substantial methodological guidance, but direct corpus material for testing accented or regionally marked speech remains more limited. The main implication is methodological: this study cannot rely only on published model performance, because accent inequality must be examined where speech, transcript, and contextual metadata meet.

The dominant pattern is a gap between available discussion of ASR error and resources that can directly support utterance-level auditing. Six items, or 33.3%, provide literature-based evidence, including "Indonesian-English code-switching ASR," "Indonesian ASR with XLSR-53," and "Indonesian medical dictation." These studies help identify relevant error domains such as code-switching failure, lexicon mismatch, and domain-specific recognition difficulty. However, five items, or 27.8%, are more directly usable for accent-sensitive testing because they include documented speech variation such as "dialectal variations from 31 Indonesian speakers," "Transcribed audio data for Javanese," and "Transcribed audio data for Sundanese." The corpus therefore shows that error analysis is possible, but unevenly supported across data types.

Table 4. Error-evaluation readiness across Indonesian ASR corpus items

Main Category	Subcategory	Indicator	Frequency	Percentage	Data Variation	Corpus Evidence / Short Quotation	Data Code	Analytic Reading
High utterance-level audit readiness	Speech resources with direct relevance to accent, regional language, or benchmarked ASR evaluation	Audio or speech-resource documentation can be connected to transcript-based ASR evaluation	5	27.8%	Dialectal Indonesian, Javanese ASR, Sundanese ASR, FLEURS Indonesian/Javanese, multilingual Indonesian speech-resource model data	"Dialectal variations from 31 Indonesian speakers"; "Transcribed audio data for Javanese"; "Transcribed audio data for Sundanese"	DOC-ASR-ID-002, 003, 004, 008, 009	These items provide the strongest basis for testing how ASR systems handle regionally marked Indonesian or local-language speech.
Partial speech-resource readiness	Speech datasets requiring alignment, sampling, or additional preprocessing before error analysis	Speech data exist, but analytic readiness depends on transcript alignment, sampling design, or speech-domain adjustment	4	22.2%	Common Voice Indonesian, Javanese TTS, Sundanese TTS, Balinese TTS	"Indonesian scripted speech"; "Javanese TTS/speech data"; "Balinese, Badung dialect, speech levels"	DOC-ASR-ID-001, 005, 006, 007	These resources support ASR testing but require careful preparation to avoid treating scripted or synthetic speech as equivalent to spontaneous accented Indonesian.
Literature-based error evidence	Studies documenting model performance, code-switching, lexicon design, or domain-specific Indonesian ASR	Empirical publications describe ASR methods, datasets, or performance conditions relevant to error typology	6	33.3%	Multilingual ASR, Indonesian-English code-switching, XLSR-53, XLS-R, medical dictation, lexicon-based ASR	"Indonesian-English code-switching ASR"; "Indonesian ASR with XLSR-53"; "Indonesian medical dictation"	DOC-ASR-ID-010–015	These items clarify likely error domains, but they do not replace direct utterance-level testing on the same corpus.
Model-context documentation only	Candidate ASR model documentation	Model documents explain system scope but do not provide Indonesian accent-specific error evidence	3	16.7%	Whisper, MMS, Indonesian NLP speech-recognition model card	"Whisper multilingual ASR documentation"; "MMS speech model documentation"	DOC-ASR-ID-016–018	These documents support reproducibility and model selection, but cannot be used as evidence of fair performance across Indonesian voices.

This pattern strengthens a key position in recent ASR fairness research: bias is not adequately captured by aggregate accuracy or model documentation alone. Feng et al. [6], Markl [8], Wassink et al. [9], and Patel et al. [10] argue that speech systems must be evaluated through subgroup visibility, performance gaps, and error consequences. In this corpus, model-context documents account for 16.7% of items and are useful for reproducibility, yet they do not demonstrate equitable recognition across Indonesian accents. This study therefore treats error-evaluation readiness as a second layer of audibility: before ASR errors can be interpreted as sociolinguistic evidence, the corpus itself must permit accountable comparison across voices, regions, and speech conditions.

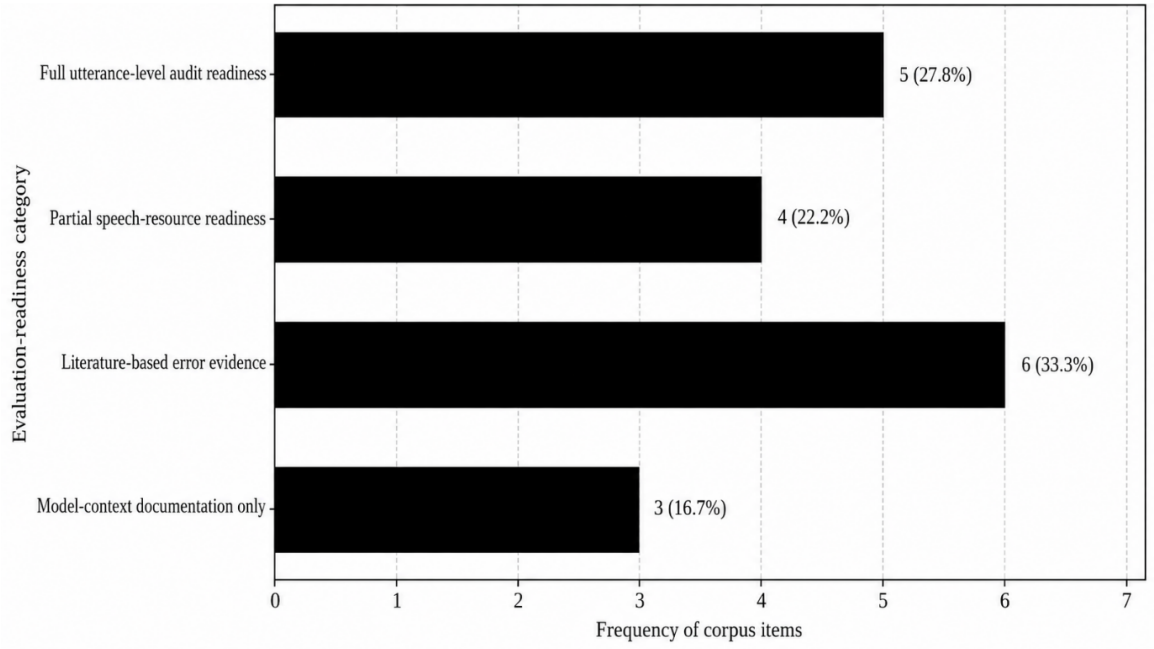

Figure 3. Error-evaluation readiness across Indonesian ASR corpus items

4.3. Sociotechnical visibility of Indonesian voices in asr evaluation

Machine audibility is shaped not only by whether speech data exist, but also by how that data can be mobilized for accountable evaluation. Across the same eighteen corpus items, visibility is distributed through direct speech-resource evidence, preprocessing-dependent materials, literature-mediated findings, and model documentation. The largest category is literature-mediated visibility, followed by direct accent-audibility exposure. This pattern shows that Indonesian ASR has a substantial scholarly base, yet the corpus most capable of testing unequal recognition remains narrower. The central issue is therefore not simple absence, but uneven evaluability: some Indonesian voices are available as analyzable speech, while others appear only through articles, model descriptions, or controlled speech resources.

The dominant pattern is that six items, or 33.3%, make Indonesian voice visibility available through published ASR research rather than direct utterance-level testing. These include corpus entries described as “Indonesian-English code-switching ASR,” “Indonesian ASR with XLSR-53,” and “Indonesian medical dictation.” Five items, or 27.8%, more directly expose accent audibility through speech resources such as “dialectal variations from 31 Indonesian speakers,” “Transcribed audio data for Javanese,” and “Transcribed audio data for Sundanese.” Four items, or 22.2%, require careful preprocessing because scripted, TTS, or prepared speech may not capture spontaneous accent variation. The remaining 16.7% are model documents that support system identification but not accent-specific claims.

Table 5. Sociotechnical visibility profile across Indonesian ASR corpus items

Main Category	Subcategory	Indicator	Frequency	Percentage	Data Variation	Corpus Evidence / Short Quotation	Data Code	Analytic Reading
Direct accent-audibility exposure	Speech resources directly connected to dialectal, regional, or multilingual Indonesian voice evaluation	Corpus items allow accent, regional language, or multilingual speech to be evaluated as spoken material	5	27.8%	Dialectal Indonesian, Javanese ASR, Sundanese ASR, FLEURS Indonesian/Javanese, multilingual Indonesian speech resources	"Dialectal variations from 31 Indonesian speakers"; "Transcribed audio data for Javanese"; "Transcribed audio data for Sundanese"	DOC-ASR-ID-002, 003, 004, 008, 009	These items make machine audibility directly examinable because they connect voice variation to speech-resource evidence.
Preprocessing-dependent visibility	Speech resources requiring alignment, transcript verification, or domain-sensitive sampling	Speech exists, but interpretive reliability depends on how scripted, TTS, or prepared speech is handled	4	22.2%	Common Voice Indonesian, Javanese TTS, Sundanese TTS, Balinese TTS	"Indonesian scripted speech"; "Javanese TTS/speech data"; "Balinese, Badung dialect, speech levels"	DOC-ASR-ID-001, 005, 006, 007	These items offer useful speech material, but they may overrepresent controlled, scripted, or highly legible speech forms.
Literature-mediated visibility	Scholarly evidence on Indonesian ASR performance, domain adaptation, code-switching, or model design	Visibility is inferred through published methods and evaluation discussions rather than direct corpus re-testing	6	33.3%	Multilingual ASR, code-switching ASR, XLSR-53, XLS-R, medical dictation, lexicon-based ASR	"Indonesian-English code-switching ASR"; "Indonesian ASR with XLSR-53"; "Indonesian medical dictation"	DOC-ASR-ID-010–015	These items clarify how Indonesian ASR has been studied, but their evidence remains mediated by each article's dataset and metric design.
Model-opaque visibility	Public model documentation without accent-specific Indonesian error evidence	Documentation identifies candidate systems but does not prove equal audibility across regional Indonesian voices	3	16.7%	Whisper, MMS, Indonesian NLP speech-recognition model card	"Whisper multilingual ASR documentation"; "MMS speech model documentation"	DOC-ASR-ID-016–018	These items are necessary for reproducibility, but insufficient for claims about fair recognition across accent variation.

This distribution clarifies why ASR fairness in Indonesia must be studied as a corpus, model, and interpretation problem at once. Bias-sensitive ASR research argues that unequal recognition is produced through dataset imbalance, standard-language assumptions, subgroup invisibility, and evaluation metrics that hide worst-group performance [6], [8], [10], [9]. In this corpus, regional Indonesian voices are not fully absent, but their evaluability is fragmented across different evidence types. This study therefore positions sociotechnical visibility as the third layer of analysis: Indonesian ASR becomes fairer only when regional, accented, code-switched, and locally meaningful speech can be tested as speech, not merely cited as linguistic diversity.

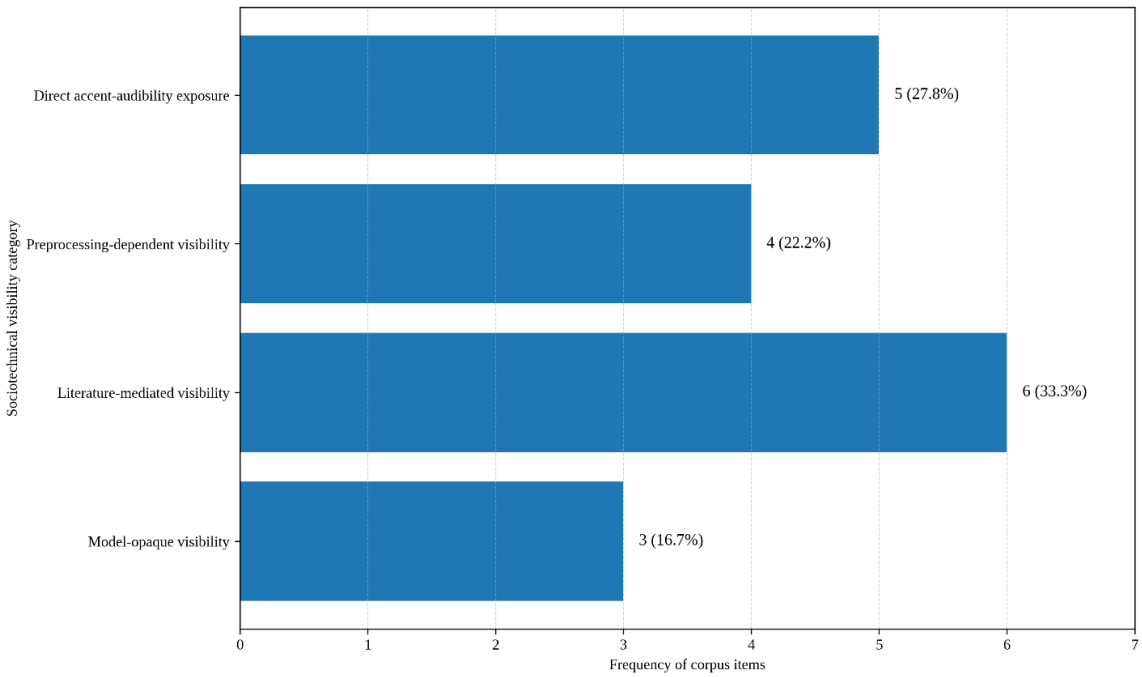

Figure 4. Sociotechnical visibility profile of Indonesian ASR corpus items

5. DISCUSSION

The first finding shows that Indonesian ASR resources are not absent, but unevenly organized across speech-resource categories, scholarly evaluation, and model documentation. This matters because corpus representation determines which voices can later become measurable, improvable, and institutionally recognizable. Regional-language resources accounted for the largest share of the corpus, while dialectal Indonesian appeared only marginally as a direct object of evaluation. Functionally, this distribution creates an important opening: Indonesian speech technology already has materials that can support multilingual and low-resource development. However, it also produces a dysfunction when regional languages are treated as separate technical objects while Indonesian spoken through regional accents remains weakly visible. This pattern complicates optimistic accounts of multilingual ASR progress in Indonesian contexts. It aligns with Adila et al. [4], Arisaputra et al. [3], and Azis et al. [18], yet extends their concern by showing that technical availability does not automatically produce accent-sensitive representational adequacy.

The underlying structure behind this imbalance is a separation between language-resource development and sociolinguistic representation. Low-resource ASR research often organizes speech around named languages, benchmarks, or model adaptation tasks, because such categories are easier to document, license, and evaluate. This structure explains why Javanese, Sundanese, and Balinese may appear as dataset labels, while Javanese-influenced Indonesian, Sundanese-influenced Indonesian, or Makassarese-influenced Indonesian remain harder to classify. The issue is not caused by a single model or institution; it emerges from dataset taxonomies, benchmark conventions, and standard-language assumptions. This finding supports Markl's [8] argument that algorithmic bias can arise from how language variation is operationalized before recognition occurs. It also resonates with Wassink et al. [9], who show that dialect-related disparities are not merely technical failures but consequences of how legitimate speech varieties enter evaluation systems.

The second finding indicates that error-evaluation readiness is uneven across corpus items. Its implication is methodological: Indonesian ASR fairness cannot be assessed responsibly unless speech data, gold transcripts, model outputs, and error categories are jointly available. Literature-based evidence provides strong guidance on possible error domains, including code-switching, lexicon mismatch, multilingual modeling, and medical dictation. Yet such evidence cannot substitute for utterance-level comparison on the same corpus. This distinction matters because ASR bias may be hidden when evaluation relies on published performance summaries rather than direct transcript alignment. Feng et al. [6] and Patel et al. [10] argue that inclusive ASR requires subgroup-sensitive metrics and transparent error interpretation. This study's corpus confirms that such evaluation is possible but not evenly ready. Its contribution lies in identifying readiness itself as a precondition of fair machine listening.

The reason for this uneven readiness lies in the technical and documentary demands of ASR auditing. A dataset may be publicly accessible but still require transcript verification, alignment, segmentation, or domain adjustment before it can support accent-sensitive error analysis. Scripted speech, TTS datasets, medical dictation, and code-switching materials are not interchangeable, because each speech condition shapes what kinds of errors can be meaningfully interpreted. This partly supports work by Abidin et al. [1], Tazakka et al. [5], and Jarin et al. [15], which demonstrates that lexicon design, domain specificity, and code-switching require different modeling assumptions. At the same time, this finding complicates purely performance-oriented approaches: a higher-performing model on a controlled dataset may still tell us little about spontaneous regional Indonesian speech. Evaluation readiness therefore depends not only on data availability, but also on analytic compatibility.

The third finding shows that sociotechnical visibility is fragmented across direct speech evidence, preprocessing-dependent resources, literature-mediated accounts, and model-opaque documentation. Its implication is that Indonesian voices may be present in the broader ASR ecosystem without being equally testable as speech. This is a crucial distinction. Model documentation can identify candidate systems such as Whisper or MMS, but it cannot demonstrate how those systems handle regional Indonesian accents without direct testing. Likewise, scholarly articles can describe Indonesian ASR progress without making accent audibility the central object of inquiry. This study therefore reframes visibility as a layered condition: a voice becomes sociotechnically visible only when it can be accessed, transcribed, compared, coded, and interpreted within a transparent evaluation protocol. This position converges with Prinos et al. [11], who warn that accent misconceptions can persist even inside technical ASR research. Fawaid et al. [42] compare human intelligence and algorithmic precision in Indonesian translation pedagogy, a reminder that algorithmic mediation of Indonesian deserves scrutiny in its own right.

This fragmented visibility is structured by opacity at several levels: training data opacity, benchmark selection, accent labeling, transcript availability, and the institutional preference for standardized or clean speech. Large multilingual systems may appear inclusive because they cover many languages, but coverage does not guarantee equitable recognition across local speech forms. Radford et al. [13] show the power of large-scale weak supervision, while Graham and Roll [7] and Jahan et al. [37] caution that strong ASR systems can still produce disparities across accents and speaker traits. This study's findings sit between those positions: it does not reject foundation models, but it questions whether their documented multilingual capacity is sufficient evidence of archipelagic fairness. The deeper structure is therefore not technological failure alone, but a mismatch between Indonesia's sociolinguistic plurality and evaluation regimes built around standard, countable, and easily benchmarked speech. Fawaid and Handayani [43] uncover gender bias in Indonesian Islamic textbooks, another case in which unequal representation becomes visible only when materials are examined systematically.

6. CONCLUSION

This study shows that unequal ASR recognition in Indonesia begins before transcription occurs: it is already structured by how speech resources classify, separate, and make voices available for evaluation. The most important lesson is that Indonesian voice technology cannot be judged only by national-language coverage or aggregate recognition performance. Regional-language datasets, dialectal Indonesian materials, ASR evaluation studies, and model documentation form an uneven evidentiary ecology in which some voices are directly testable while others remain mediated or opaque. This study contributes by shifting ASR inquiry from technical accuracy alone toward archipelagic audibility, combining corpus representation audit, accent-sensitive error readiness, and sociotechnical visibility assessment.

This study is limited by its reliance on public and verifiable corpus items rather than direct large-scale inference across newly collected speaker recordings. Several resources require transcript alignment, utterance segmentation, and metadata enrichment before robust subgroup comparison can be performed. The analysis therefore identifies representational and evaluative conditions rather than claiming causal effects of accent, region, or speaker identity on ASR performance. Future research should build a controlled, ethically

documented Indonesian speech corpus that includes spontaneous speech, regional Indonesian accents, local-language influence, code-switching, gender and age metadata where appropriate, and comparable ASR outputs across multiple systems. Such work would enable stronger fairness-sensitive evaluation.

ACKNOWLEDGMENTS

The author thanks the reviewers and editorial team for their constructive feedback on this manuscript.

FUNDING INFORMATION

Author states no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Imam Masrur: conceptualization, methodology, formal analysis, writing – original draft, writing – review and editing.

CONFLICT OF INTEREST STATEMENT

Author states no conflict of interest.

AI USE DECLARATION

The author used artificial intelligence tools to assist with the generation of illustrative figures and with the formatting and layout of this manuscript. AI was not used to generate the substantive content, conceptualization, data, analysis, or conclusions of the research, all of which remain the sole responsibility of the author.

INFORMED CONSENT

Not applicable – this study did not involve direct human participants. Data consisted solely of publicly accessible official, institutional, and scholarly corpus materials.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible corpus, institutional, and scholarly materials, and therefore no institutional review board approval was required. Where research does involve human subjects, the author affirms that such research would be conducted in full compliance with all relevant national regulations and institutional policies, in accordance with the tenets of the Declaration of Helsinki, and would be approved by the author's institutional review board or equivalent committee.

DATA AVAILABILITY

The corpus of publicly accessible official, institutional, and scholarly materials analyzed in this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] T. Abidin, A. Misbullah, R. Ferdhiana, L. Farsiah, M. Z. Aksana, and H. Riza, "Acoustic model with multiple lexicon types for Indonesian speech recognition," *Applied Computational Intelligence and Soft Computing*, vol. 2022, Art. 3227828, 2022, doi: 10.1155/2022/3227828.

[2] P. Arisaputra and A. Zahra, "Indonesian automatic speech recognition with XLSR-53," *Ingénierie des Systèmes d'Information*, vol. 27, no. 6, 2022, doi: 10.18280/isi.270614.

[3] P. Arisaputra, A. T. Handoyo, and A. Zahra, "XLS-R deep learning model for multilingual ASR on low-resource languages: Indonesian, Javanese, and Sundanese," arXiv, 2024, doi: 10.48550/arXiv.2401.06832.

[4] 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," in *Proc. 2024 27th Conf. Oriental COCOSDA*, 2024, pp. 1-6, doi: 10.1109/O-COCOSDA64382.2024.10800336.

[5] R. V. M. Tazakka, D. Lestari, A. Purwarianti, D. Tanaya, K. Azizah, and S. Sakti, "Indonesian-English code-switching speech recognition using machine speech chain based semi-supervised learning," in *Proc. 3rd Annual Meeting SIGUL @ LREC-COLING 2024*, 2024, doi: 10.63317/3hriruv4bj9u.

[6] S. Feng, B. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, Art. 101567, 2024, doi: 10.1016/j.csl.2023.101567.

[7] C. Graham and N. Roll, "Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits," *JASA Express Letters*, vol. 4, no. 2, 2024, doi: 10.1121/10.0024876.

[8] N. Markl, "Language variation and algorithmic bias: Understanding algorithmic bias in British English automatic speech recognition," in *Proc. 2022 ACM Conf. Fairness, Accountability, and Transparency*, 2022, doi: 10.1145/3531146.3533117.

[9] A. Wassink, C. Gansen, and I. Bartholomew, "Uneven success: Automatic speech recognition and ethnicity-related dialects," *Speech Communication*, vol. 140, pp. 50-70, 2022, doi: 10.1016/j.specom.2022.03.009.

[10] T. Patel, W. Hutiri, A. Y. Ding, and O. Scharenborg, "How to evaluate automatic speech recognition: Comparing different performance and bias measures," arXiv, 2025, doi: 10.48550/arXiv.2507.05885.

[11] K. Prinos, N. Patwari, and C. A. Power, "Speaking of accent: A content analysis of accent misconceptions in ASR research," in *Proc. 2024 ACM Conf. Fairness, Accountability, and Transparency*, 2024, doi: 10.1145/3630106.3658969.

[12] K. Koźmińska, "Reimagining accents and speech recognition with sociolinguistic perception studies and research on listening subjects," *Journal of Sociolinguistics*, 2025, doi: 10.1111/josl.12720.
[13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Machine Learning*, 2022, pp. 28492-28518.
[14] H. Kheddar, M. Hemis, and Y. Himeur, "Automatic speech recognition using advanced deep learning approaches: A survey," *Information Fusion*, 2024, doi: 10.1016/j.inffus.2024.102422.
[15] A. Jarin, A. Santosa, M. T. Uliniansyah, L. R. Aini, E. Nurfadhilah, and G. Gunarso, "Automatic speech recognition for Indonesian medical dictation in cloud environment," *IAES International Journal of Artificial Intelligence*, 2024, doi: 10.11591/ijai.v13.i2.pp1762-1772.
[16] H. Pardede, P. Adhi, V. Zilvan, A. Ramdan, and D. Krisnandi, "Deep convolutional neural networks-based features for Indonesian large vocabulary speech recognition," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 610-617, 2023, doi: 10.11591/ijai.v12.i2.pp610-617.
[17] D. Dwi, P. D. Pamungkasari, A. P. Rifai, I. D. Sholihati, and G. Sakarkar, "Comparison of feature extraction techniques for long short-term memory models in Indonesian automatic speech recognition," *Green Intelligent Systems and Applications*, 2025, doi: 10.53623/gisa.v5i1.605.
[18] H. Azis, M. F. F. M, H. Darwis, P., T. Hasanuddin, and S., "Development of a low-resource automatic speech recognition system for the Makassar dialect," in *Proc. 2026 20th Int. Conf. Ubiquitous Information Management and Communication (IMCOM)*, 2026, pp. 1-8, doi: 10.1109/IMCOM69009.2026.11360893.
[19] K. Tomanek, V. Zayats, D. Padfield, K. Vaillancourt, and F. Biadsy, "Residual adapters for parameter-efficient ASR adaptation to atypical and accented speech," in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing*, 2021, doi: 10.18653/v1/2021.emnlp-main.541.
[20] 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.
[21] A. Hadziqi, "Small models, many languages: Parameter-efficient adaptation of multilingual language models for low-resource Indonesian languages," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.
[22] G. Brown and S. Hellmuth, "Computational modelling of segmental and prosodic levels of analysis for capturing variation across Arabic dialects," *Speech Communication*, vol. 141, pp. 80-92, 2022, doi: 10.1016/j.specom.2022.05.003.
[23] M. Del Rio, C. Miller, J. Profant, J. Drexler-Fox, Q. McNamara, N. Bhandari, N. Delworth, I. Pirkin, M. Jette, S. Chandra, P. Ha, and R. Westerman, "Accents in speech recognition through the lens of a World Englishes evaluation set," *Research in Language*, 2023, doi: 10.18778/1731-7533.21.3.02.
[24] L.-F. Lai, J. G. Van Hell, and J. Lipski, "Dialect bias in automatic speech recognition: Analysis of Appalachian English," *American Speech: A Quarterly of Linguistic Usage*, 2024, doi: 10.1215/00031283-11466494.
[25] D. Serditova and K. Tang, "Automatic speech recognition biases in Newcastle English: An error analysis," in *Proc. Interspeech 2025*, 2025, doi: 10.21437/interspeech.2025-1973.
[26] N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. H. Chau, "Best of both worlds: Robust accented speech recognition with adversarial transfer learning," in *Proc. Interspeech 2021*, 2021, pp. 1314-1318, doi: 10.21437/Interspeech.2021-1888.
[27] X. Wang, Y. Long, and D. Xu, "Universal and accent-discriminative encoders for conformer-based accent-invariant speech recognition," *International Journal of Speech Technology*, vol. 25, pp. 987-995, 2022, doi: 10.1007/s10772-022-10010-z.
[28] D. Prabhu, P. Jyothi, S. Ganapathy, and V. Unni, "Accented speech recognition with accent-specific codebooks," arXiv, 2023, doi: 10.48550/arXiv.2310.15970.
[29] T. Banerjee and V. Ramasubramanian, "Accent-robust speech recognition for English in low-resource settings using manifold mixup," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, 2025, doi: 10.1186/s13636-025-00435-0.
[30] W. Lee, H. Kim, and G. Lee, "Mixture-of-experts with intermediate CTC supervision for accented speech recognition," arXiv, 2026, doi: 10.48550/arXiv.2602.01967.
[31] A. Bera and A. Agarwal, "Enhancing non-native accent recognition through a combination of speaker embeddings, prosodic and vocal speech features," *Signal & Image Processing: An International Journal*, 2023, doi: 10.5121/sipij.2023.14302.
[32] S. Ghorbani and J. H. L. Hansen, "Advanced accent/dialect identification and accentedness assessment with multi-embedding models and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 155, no. 6, pp. 3848-3860, 2023, doi: 10.1121/10.0026235.
[33] M. A. Dar and J. Pushparaj, "Machine learning and deep learning approaches for accent recognition: A review," *IEEE Access*, vol. 13, pp. 51527-51550, 2025, doi: 10.1109/ACCESS.2025.3552935.
[34] S. Feng, O. Kudina, B. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," arXiv, 2021.
[35] C. Li, T. Cohen, and S. V. S. Pakhomov, "Reexamining racial disparities in automatic speech recognition performance: The role of confounding by provenance," arXiv, 2024, doi: 10.48550/arXiv.2407.13982.
[36] A. DiChristofano, H. Shuster, S. Chandra, and N. Patwari, "Performance disparities between accents in automatic speech recognition," arXiv, 2022, doi: 10.48550/arXiv.2208.01157.
[37] M. Jahan, P. Mazumdar, T. Thebaud, M. Hasegawa-Johnson, J. Villalba, N. Dehak, and L. Moro-Velázquez, "Unveiling performance bias in ASR systems: A study on gender, age, accent, and more," in *Proc. ICASSP 2025*, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10887767.
[38] A. Rai, S. D. Jaiswal, S. Prakash, B. P. Sree, and A. Mukherjee, "DENOASR: Debiasing ASRs through selective denoising," in *Proc. 2024 IEEE Int. Conf. Knowledge Graph (ICKG)*, 2024, pp. 283-290, doi: 10.1109/ICKG63256.2024.00043.
[39] A. Martin, R. L. MacDonald, P. Jiang, M. A. Ladewig, J. Cattiau, R. Heywood, R. Cave, J. Tobin, P. C. Nelson, and K. Tomanek, "Project Euphonia: Advancing inclusive speech recognition through expanded data collection and evaluation," *Frontiers in Language Sciences*, 2025, doi: 10.3389/flang.2025.1569448.
[40] J.-B. Sun, Y. Xiao, S. Chung, Q. Hu, G. Huang, E.-J. Holden, and T. Dang, "Activation steering for accent adaptation in speech foundation models," 2026.
[41] M. A. Sumadi, "Who gets represented by Indonesian AI? Measuring regional, gender, and sociolinguistic bias in large language models," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.
[42] A. Fawaid, R. Assyabani, I. Abdullah, C. Muali, M. S. Itqan, and S. Islam, "Human intelligence and algorithmic precision: An experimental study of Indonesian translation pedagogy in higher education," *Asian Journal of University Education*, vol. 21, no. 3, pp. 779-792, 2025, doi: 10.24191/ajue.v21i3.53.
[43] A. Fawaid and P. Handayani, "Visualizing inequality: Uncovering gender bias in Islamic textbooks of Indonesia," *Cogent Education*, vol. 12, no. 1, 2025, doi: 10.1080/2331186X.2025.2593342.

BIOGRAPHY OF AUTHOR



Imam Masrur is affiliated with Universitas Islam Negeri Syekh Wasil Kediri, Indonesia, where his work engages corpus-based speech-technology auditing, sociolinguistic discourse analysis, and the study of accent variation and linguistic justice in Indonesian automatic speech recognition. His current research examines how Indonesian ASR resources, evaluation practices, and model documentation make regional and accent-varied voices differentially visible, testable, and interpretable. He can be contacted at email: imammasrur@uinkediri.ac.id