Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/23364
DOI URL: http://dx.doi.org/10.21474/IJAR01/23364

CONFERENCE PAPER

A COMPREHENSIVE REVIEW OF AUTOMATIC SPEECH RECOGNITION FOR ODIA LANGUAGE: CHALLENGES, TECHNIQUES, AND FUTURE DIRECTIONS

Surabika Hota[1], Hammad Ahmad[2] and Md Samir Ansari[2]

1. Asst. Prof.(MCA) College of Engineering,Bhubaneswar.
2. College of Engineering Bhubaneswar.

Manuscript Info

Manuscript History
Received: 8 February 2026
Final Accepted: 10 March 2026
Published: April 2026

Key words:-
Automatic Speech Recognition, Odia Language, Low-Resource Languages, Deep Learning, Speech Processing

Abstract

This paper highlights an extensive review of ASR for the low-resource language Odia by focusing on methodologies, datasets, metrics, applications, and future scopes. In terms of evolution, the current paper examines the progression of technology from conventional approaches like Hidden Markov Models to recent innovations, such as CNN, RNN, and attention-based systems. The challenges that have been discussed extensively within this review paper include insufficient data, dialectic issues, lack of benchmarking resources, and sensitivity to noises. The ongoing research has been primarily focused on isolated speech, although continuous speech has not received sufficient attention. Transfer learning and multilingual techniques may offer some potential in addressing existing problems.

Introduction:-

Automatic Speech Recognition (ASR) converts spoken language into text and has achieved significant progress with deep learning and large-scale datasets; however, these advancements are mainly limited to high-resource languages such as English [1], [2]. In contrast, Odia, an Indo-Aryan language spoken by millions in India, remains underrepresented due to limited annotated data and linguistic resources, posing challenges for effective ASR development [2], [3].

Early ASR systems included statistical models like Hidden Markov Models and Gaussian Mixture Models, together with handmade features like MFCC [3]. While excellent for single word identification, most methods struggled with continuous speech. Subsequent machine learning algorithms improved performance, although they relied on feature engineering.

Recent improvements have centered on deep learning models such as CNNs, RNNs, and attention-based architectures, which automatically learn feature representations and achieve greater accuracy [4]. However, these methodologies necessitate huge datasets, which are scarce in Odia.

Corresponding Author:- Surabika Hota
Address:-Asst. Prof.(MCA) College of Engineering,Bhubaneswar.

To overcome this, multilingual and transfer learning approaches have been investigated as ways to harness knowledge from high-resource languages [1]. Despite these attempts, difficulties such as data scarcity and resource constraints remain.

Fundamentals of Automatic Speech Recognition:-

Overview of ASR:-

Automatic Speech Recognition (ASR) refers to the process of transforming spoken language into a textual format by modeling the relationship between acoustic signals and linguistic units.

$$W^* = \arg\max_W P(W|X) = \arg\max_W P(X|W) \cdot P(W)$$

WhereP(X|W) represents the acoustic model and P(W) represents the language model. A standard ASR system comprises several components, including feature extraction, acoustic modeling, language modeling, and decoding, which collectively function to produce the most probable transcription of a given speech signal [5], [6].

As shown in Fig. 1, an ASR system consists of multiple stages including feature extraction, acoustic modeling, language modeling, and decoding.

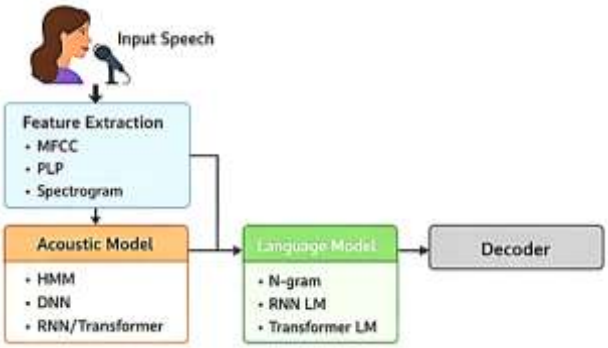

Fig. 1. ASR Pipeline

Feature Extraction:-

Feature extraction constitutes the initial phase of an Automatic Speech Recognition (ASR) system, wherein raw speech signals are converted into compact and informative representations. The primary objective is to capture pertinent phonetic information while minimizing noise and variability. Techniques such as Mel-Frequency Cepstral Coefficients (MFCC) and log-Mel spectrograms are extensively employed, as they effectively model the characteristics of human auditory perception [7].

Acoustic Modeling:-

Acoustic modeling establishes the relationship between extracted speech features and phonetic units. Traditional ASR systems depend on Hidden Markov Models (HMM) combined with Gaussian Mixture Models (GMM) to model temporal and statistical variations in speech. Conversely, contemporary systems employ deep learning methodologies such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), which can autonomously learn complex feature representations and enhance recognition accuracy [8].

Language Modeling:-

The language model offers contextual information by estimating the probability of word sequences, thereby ensuring that the recognized output is both grammatically and semantically coherent. Initial methodologies employed statistical n-gram models, whereas recent advancements utilize neural language models and Transformer-based architectures, which are adept at capturing long-range dependencies in speech [5].

Decoding:-
The decoding process integrates acoustic and language models to produce the most probable transcription of the input speech. Efficient search algorithms, such as Viterbi decoding and beam search, are employed to explore potential word sequences and select the optimal output [6].

Traditional vs End-to-End ASR:-
Automatic Speech Recognition (ASR) systems have transitioned from traditional hybrid architectures to end-to-end models that directly map speech signals to text. Traditional systems consist of distinct components for feature extraction, acoustic modeling, and language modeling. In contrast, end-to-end approaches, which are based on Connectionist Temporal Classification (CTC), attention mechanisms, and Transformer architectures, streamline the pipeline and achieve enhanced performance [8].

Literature Review:-

Research in Odia Automatic Speech Recognition (ASR) has progressed from traditional statistical methods to modern deep learning methodologies. However, the construction of robust systems is still constrained due to the language's low resource requirements. This section examines previous work by categorizing methodologies as traditional, machine learning, deep learning, multilingual techniques, and dataset contributions, followed by a critical examination and identification of research gaps. A taxonomy of Odia speech recognition approaches is illustrated in Fig. 2.

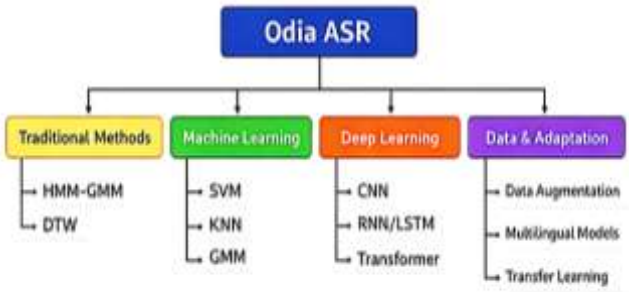

Fig. 2. Taxonomy of approaches used in Odia speech recognition

Traditional Approaches:-
Early Odia ASR systems primarily used HMM–GMM models, which were effective for small-vocabulary and isolated word recognition tasks. For example, Karan et al. [9] employed MFCC features with a bi-gram language model, achieving moderate performance under controlled conditions. However, these methods rely on handcrafted features, limiting scalability and robustness, and perform poorly in noisy environments and continuous speech scenarios.

Machine Learning Approaches:-
The introduction of machine learning marked an intermediate stage in Odia ASR development, with methods such as Support Vector Machines (SVM), k-Nearest Neighbours (KNN), and ensemble classifiers improving recognition performance. Sahoo et al. [10] proposed an ensemble-based approach for voice command recognition, achieving better results than individual models. However, these methods still relied on handcrafted features and were limited to specific tasks, with inadequate capability to model sequential dependencies in continuous speech.

Deep Learning Approaches:-
Recent advances in deep learning have dramatically altered the landscape of Odia voice recognition. Deep neural networks enable the automatic learning of hierarchical representations from raw or slightly processed voice input, decreasing the reliance on handmade features. Convolutional Neural Networks (CNN) have been effectively deployed to tasks such as isolated digit identification, with Bag [11] outperforming previous techniques. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks improved the modeling of temporal dependencies in voice sequences. Recently, attention-based encoder-decoder designs have gained popularity. Majhi

and Saha [12] suggested an attention-based ASR system for Odia that uses data augmentation approaches to improve recognition accuracy despite a short training dataset. Although deep learning approaches surpass classical and machine learning methods, their usefulness is limited by the availability of big, annotated datasets. The computational complexity and data requirements of these models present substantial hurdles in low-resource environments like Odia.

Multilingual and Transfer Learning Approaches:-

To address data scarcity, recent studies have explored multilingual and transfer learning approaches that leverage knowledge from high-resource languages to improve Odia ASR performance. Majhi and Saha [13] proposed the OdiSR-TL framework, utilizing pre-trained models and transfer learning to enhance recognition accuracy. Multilingual ASR systems exploit shared phonetic and linguistic features across languages, particularly within the Indic family, improving generalization and reducing data requirements. However, challenges remain in adapting models to language-specific phonetics and handling dialectal variations in Odia.

Dataset Contributions:-

The creation of speech datasets is crucial for developing ASR research. The majority of public datasets in Odia are of limited size and scope. Previous research has generally used custom-built datasets concentrating on single words, numerals, or command-based speech tasks. For example, Bag [11] and Sahoo et al. [10] evaluated their models using task-specific datasets, whereas Majhi and Saha [12] used relatively small corpora enriched using data augmentation approaches. Despite these efforts, the dearth of large-scale, publicly accessible, and standardized datasets remains a significant impediment. The low diversity of speakers, accents, and recording settings further limits the generalizability of ASR systems built for Odia.

Critical Analysis of Existing Work and Research Gaps:-

A comprehensive study of the current literature indicates numerous significant tendencies and limits. There is a distinct shift from classic HMM-based systems to deep learning and attention-based architectures, which reflects larger advances in speech recognition research. However, most studies focus on isolated speech recognition tasks, with little examination of continuous and conversational speech contexts.

Data scarcity is seen as the most critical difficulty across all techniques. The lack of large-scale annotated datasets hinders the performance of deep learning models and their reproducibility. Furthermore, the lack of established benchmarking frameworks makes it impossible to compare findings from different investigations. Another noticeable difference is the low adoption of advanced architectures such as transformer-based and self-supervised learning models, which have shown cutting-edge performance in high-resource languages. Furthermore, present algorithms fail to account for dialectal variances and linguistic diversity within Odia, resulting in lower robustness in real-world applications.

While multilingual and transfer learning approaches show promise, they are still underexplored and require additional research to properly capitalize on cross-lingual knowledge transfer. Addressing these issues is critical for creating scalable, accurate, and real-world-relevant ASR systems for the Odia language. Table 1 summarizes key studies in Odia speech recognition, including methodologies, datasets, and limitations.

Table 1. Summary of Existing Work on Odia Speech Recognition

Ref	Author & Year	Approach	Dataset	Task Type	Key Contribution	Limitation
[1]	Karan et al. (2015)	HMM-GMM	Custom dataset	Isolated words	Early Odia ASR system using statistical modeling	Limited vocabulary, low scalability
[2]	Sahoo et al. (2025)	Ensemble ML	Command dataset	Command recognition	Improved accuracy using ensemble classifiers	Domain-specific dataset
[3]	Bag (2023)	CNN	Digit dataset	Isolated digits	Deep learning improves recognition accuracy	Limited to digit recognition
[4]	Majhi & Saha	Attention-	Custom	Continuous	Data augmentation + attention improves	Small dataset size

	(2024)	based DL	dataset	speech	performance	
[5]	Mishra et al. (2023)	Review	Multiple datasets	Survey	Comprehensive overview of Odia ASR research	No experimental validation
[6]	Mohanty & Nayak (2022)	CNN	Keyword dataset	Keyword spotting	Efficient keyword detection system	Limited vocabulary scope
[7]	Basu et al. (2021)	Multilingual ASR	Indic corpus	Continuous speech	Supports cross-lingual learning	Limited Odia-specific data
[8]	Sethiya et al. (2025)	Transformer	Indic-ST dataset	Speech-to-text	Large-scale multilingual dataset	Odia subset limited
[9]	Panigrahi (2022)	Dataset creation	21+ hours corpus	General ASR	Public Odia speech dataset	Still relatively small
[10]	Singh et al. (2020)	Survey	Indian languages	Survey	Overview of ASR in Indian languages	Limited Odia focus

Datasets For Odia Asr:-

Robust Automatic voice Recognition (ASR) systems rely substantially on the availability of huge, high-quality annotated voice datasets. However, Odia is regarded as a low-resource language, and dataset availability remains a significant difficulty. The majority of extant datasets are tiny, domain-specific, and frequently not publically available, limiting the performance and generalizability of ASR models. Panigrahi [14] made a significant addition with his public-domain Odia speech corpus, which contains over 55,000 utterances and nearly 21 hours of speech data. This dataset contains recordings from numerous speakers and is a significant resource for ASR research. Nonetheless, it is still tiny when compared to datasets for high-resource languages. Several research use custom-built datasets tailored to certain goals. For example, Majhi and Saha [15] employed continuous speech data for attention-based ASR, whereas Bag [16] and Sahoo et al. [20] created datasets for digit and command recognition. These datasets are often small in size and focus on specific application domains. Another useful resource is the ILCI corpus [17], which contains multilingual data that can help with Odia ASR in cross-lingual contexts. However, it is not primarily intended for large-scale speech recognition. Despite these attempts, substantial limits remain. Most datasets have fewer than 10-20 hours of speech and lack variation in speakers, accents, and recording settings, resulting in poor real-world performance. Furthermore, the lack of defined benchmark datasets makes it difficult to compare models between studies. Recent research emphasizes the importance of large-scale data collecting, crowdsourcing, and data augmentation, as well as multilingual and transfer learning methodologies, in addressing data scarcity. However, the production of comprehensive and publicly available datasets is critical for enhancing Odia ASR.

Challenges In Odia Asr:-

The development of Odia Automatic Speech Recognition (ASR) systems is hindered by several challenges arising from its low-resource nature, including limited data, linguistic variability, and technological constraints. The scarcity of large-scale annotated datasets—often restricted to only a few hours—negatively impacts model accuracy and generalization [18]. Additionally, regional dialectal variations introduce differences in pronunciation and vocabulary, complicating the design of a unified system [20]. The absence of standardized benchmark datasets and evaluation protocols further limits performance comparison across studies, while ASR systems trained on clean data often perform poorly in noisy, real-world conditions [21]. Moreover, the lack of linguistic resources such as pronunciation lexicons and language models, along with code-switching involving Hindi and English, increases system complexity [18]. Finally, the high computational and data requirements of modern deep learning models pose further challenges for effective Odia ASR development [22].

Evaluation Metrics:-

Automatic Speech Recognition (ASR) systems are commonly evaluated using quantitative metrics that compare recognized output with reference transcriptions, providing insights into accuracy and robustness, especially in low-resource settings like Odia. Error-based metrics such as Word Error Rate (WER) and Character Error Rate (CER) are the most widely used [23]. WER measures transcription accuracy based on substitutions, deletions, and insertions but has limitations in morphologically rich languages like Odia [23]. CER, operating at the character level, offers finer-grained evaluation and is more suitable for low-resource and multilingual contexts [24]. Additionally, Phoneme Error Rate (PER) evaluates phonetic accuracy, while metrics such as Match Error Rate (MER) and Word Information Lost (WIL), along with recent semantic-based measures, provide further insights into alignment and meaning preservation [24], [25]. Despite these alternatives, WER remains the standard metric; however, combining it with CER enables more comprehensive evaluation for low-resource languages.

Future Directions:-

The development of Automatic Speech Recognition (ASR) for low-resource languages such as Odia opens up numerous avenues for future research. Despite recent advances, numerous paths remain critical for developing strong and scalable systems. Adoption of self-supervised learning models, such as wav2vec 2.0, which use huge volumes of unlabeled data to increase performance in low-resource contexts, is critical [26]. Similarly, Transformer-based and end-to-end architectures, such as Conformer and attention-based models, provide better contextual modeling but require additional investigation for Odia because to restricted data availability [27]. Another potential option is the employment of multilingual and cross-lingual learning models such as XLSR and Whisper, which allow knowledge transfer across related languages [28]. Furthermore, the development of large-scale, diversified, and publicly accessible datasets is crucial for boosting model generalization and real-world application [27]. Future study should also focus on code-switching, noise robustness, and domain adaptability, as real-world speech frequently involves many languages and diverse acoustic conditions. Furthermore, the creation of lightweight and edge-deployable models is critical for allowing real-time applications on resource-limited devices. Overall, improving self-supervised learning, multilingual modeling, dataset generation, and system robustness will be critical in closing the gap between low- and high-resource ASR systems.

Conclusion:-

This paper provides a complete assessment of Automatic Speech Recognition (ASR) for the Odia language, including methodology, datasets, evaluation metrics, applications, and future developments. The paper focuses on the transition from classic statistical models to newer deep learning methodologies. Emerging approaches, such as multilingual, transfer learning, and self-supervised models, provide intriguing alternatives. Furthermore, Odia ASR offers great potential in applications like voice assistants, e-governance, and assistive technology. Overall, improving dataset development, robust modeling, and consistent evaluation will be critical to developing scalable and viable Odia ASR systems.

References:-

1. Diwan, R. Vaideeswaran, S. Shah, and A. Singh, “Multilingual and code-switching ASR challenges for low resource Indian languages,” arXiv preprint arXiv:2104.00235, 2021. https://arxiv.org/abs/2104.00235
2. N. Mishra, S. R. Dash, S. Parida, and R. S. Prasad, “A Systematic Review on Automatic Speech Recognition for Odia Language,” in Proc. Springer Conf., 2023. https://link.springer.com/chapter/10.1007/978-981-99-6706-3_9
3. B. Karan, J. Sahoo, and P. K. Sahu, “Automatic speech recognition based Odia system,” in Proc. Int. Conf., 2015. https://ieeexplore.ieee.org/document/7489765
4. M. K. Majhi and S. K. Saha, “An automatic speech recognition system in Odia language using attention mechanism and data augmentation,” International Journal of Speech Technology, 2024. https://link.springer.com/article/10.1007/s10772-024-10132-6
5. J. Li, L. Deng, and Y. Gong, “Machine learning paradigms for speech recognition: An overview,” IEEE, 2014.https://ieeexplore.ieee.org/document/6732927
6. S. Karpagavalli and E. Chandra, “A review on automatic speech recognition architecture and approaches,” 2016. https://www.researchgate.net/publication/302915903
7. E. Koffi, “A tutorial on acoustic phonetic feature extraction for ASR,” 2020. https://www.academia.edu/download/76099380/viewcontent.pdf

8. C. Yu et al., "Acoustic modeling based on deep learning for low-resource speech recognition," IEEE Access, 2020. https://ieeexplore.ieee.org/document/9180290
9. G. Sahoo, P. Mohanty, C. K. Patra, and A. K. Nayak, "Enhancing Voice Command Recognition in the Low-Resource Odia Language Using Ensemble Techniques," 2025. https://journal.uob.edu.bh/items/5507e04a-3d4a-4043-a81c-47bafd53ee01
10. B. R. Bag, "Speech recognition of isolated Odia digits using convolutional neural network," Proc. IEEE Conf., 2023. https://ieeexplore.ieee.org/document/10263195
11. S. Panigrahi, "Building a public domain voice database for Odia," Proc. ACM Web Conf., 2022. https://dl.acm.org/doi/10.1145/3487553.3524931
12. T. Dalai, T. K. Mishra, and P. K. Sa, "Part-of-speech tagging of Odia language using statistical and deep learning-based approaches," ACM Trans., 2023. https://dl.acm.org/doi/10.1145/3588900
13. Diwan, R. Vaideeswaran, S. Shah, and A. Singh, "Multilingual and code-switching ASR challenges for low resource Indian languages," arXiv preprint, 2021. https://arxiv.org/abs/2104.00235
14. N. Mishra, S. R. Dash, S. Parida, and R. S. Prasad, "A Systematic Review on Automatic Speech Recognition for Odia Language," Springer Conf., 2023. https://link.springer.com/chapter/10.1007/978-981-99-6706-3_9
15. B. Bhukta, M. K. Jena, and P. Patnaik, "Analyzing Dialectal Variations in Odia: Computational Approaches," 2025.https://ijaas.in/wp-content/uploads/2025/11/Paper-5-1.pdf
16. Singh et al., "ASRoIL: A comprehensive survey for automatic speech recognition of Indian languages," Artificial Intelligence Review, 2020.https://link.springer.com/article/10.1007/s10462-019-09775-8
17. D. N. Raja and K. J. Giri, "AUDIRE: A comprehensive review of speech recognition technologies—methods, uses, and challenges," Int. J. Speech Technology, 2025. https://link.springer.com/article/10.1007/s10772-025-10213-0
18. H. S. Chadha et al., "Vakyansh: ASR toolkit for low resource Indic languages," arXiv, 2022. https://arxiv.org/abs/2203.16512
19. Dhaka and N. R. Mahapatra, "Practical Considerations for Selecting Metrics to Evaluate Automatic Speech Recognition Systems," Springer Conf., 2024. https://link.springer.com/chapter/10.1007/978-3-031-94937-1_9
20. P. Shah et al., "Is Word Error Rate a good evaluation metric for Speech Recognition in Indic Languages?" arXiv, 2022. https://www.academia.edu/download/112787541/2203.16601v1.pdf
21. J. James and D. P. Gopinath, "Advocating Character Error Rate for Multilingual ASR Evaluation," arXiv, 2024. https://arxiv.org/abs/2410.07400
22. R. P. Magalhaes and D. J. R. Vasconcelos, "Evaluation of automatic speech recognition approaches," J. Information Systems, 2022. https://journals-sol.sbc.org.br/index.php/jidm/article/view/2514
23. S. Roy, "Semantic-WER: A unified metric for the evaluation of ASR transcript usability," arXiv, 2021. https://arxiv.org/abs/2106.02016
24. M. A. Dar and J. Pushparaj, "Wav2Vec2 model-based ASR for low-resource languages," Int. J. Speech Technology, 2026. https://link.springer.com/article/10.1007/s10772-025-10228-7
25. Y. O. Sharrab et al., "Advancements in speech recognition: A systematic review of deep learning transformer models, trends, innovations, and future directions," IEEE, 2025.https://ieeexplore.ieee.org/document/10924161
26. J. Li, "Recent advances in end-to-end automatic speech recognition," arXiv preprint, 2021. https://arxiv.org/abs/2111.01690
27. S. H. Imam et al., "Automatic speech recognition for low-resource languages: Challenges and future directions," ACL Workshop, 2025. https://aclanthology.org/2025.africanlp-1.13/