OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21610065

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Articles

Volume-06|Issue-04|2026

Decoding Soundscapes: Machine Learning Approaches for Speech-to-Text Conversion

Asha S N[1], Dr. Madhura Gangaiah[2], Vedha C[3], Manasa C K[4]

[1,3,4]Assistant Professor, Department of Computer Science and Engineering, BGS College of Engineering and Technology, Bengaluru, India.
[2]Professor, Department of Artificial Intelligence and Machine Learning, BGS Col-lege of Engineering and Technology, Bengaluru, India.

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Asha, S. N., Gangaiah, M., Vedha, C., Manasa, C. K. (2026) Decoding Soundscapes: Machine Learning Approaches for Speech-to-Text Conversion. *Indiana Journal of Multidisciplinary Research, 6*(4), 350-355.

Abstract: The conversion of audio signals into text has become essential in human–computer interaction, accessibility, and information management. With the rise of multimedia and real-time communication, machine learning (ML) has revolutionized speech recognition by bridging spoken and written language more effectively than traditional rule-based systems. Modern ML models use supervised and unsupervised learning to process acoustic features such as MFCCs, spectrograms, and deep audio embeddings through architectures like RNNs, CNNs, and transformers, enabling accurate and context-aware transcription. End-to-end deep learning approaches, including sequence-to-sequence and attention mechanisms, have further enhanced performance by minimizing manual feature design. Advances in transfer learning, multilingual pre-training, and self-supervised learning have expanded speech recognition to low-resource languages, as demonstrated by large models like wav2vec and Whisper. These innovations have broad applications in voice assistants, captioning, healthcare documentation, legal and industrial transcription, and accessibility tools, delivering significant social and economic benefits. However, challenges such as noise, code-switching, domain adaptation, high computational demands, and ethical concerns related to privacy, bias, and fairness persist. Ongoing research focuses on improving the robustness, efficiency, and inclusivity of ML-based speech recognition systems for universal and responsible deployment.

Keywords: Speech and Language Technologies, Low-Resource Languages, Endangered Languages, Natural Language Processing (NLP), Speech Recognition, Machine Translation, Language Preservation, Linguistic Diversity, Digital Inclusion, Cultural Heritage.

INTRODUCTION

Human communication is inherently rooted in speech, the most natural medium for exchanging information across cultures and generations. The rapid growth of digital technologies and the expansion of voice-enabled systems have made speech-to-text (STT) or automatic speech recognition (ASR) an essential field for enabling accessibility, automation, and human–machine interaction. Evolving from early rule-based and statistical approaches like Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), speech recognition has been transformed by advances in machine learning, which allow systems to learn directly from data rather than relying on handcrafted rules. Speech, being a complex acoustic signal carrying linguistic and emotional information, is processed through steps such as sampling, noise reduction, and feature extraction using methods like Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms. Modern ML architectures—RNNs, LSTMs, GRUs, and transformers—capture temporal and contextual dependencies, improving accuracy and adaptability across languages and environments. Beyond supervised learning, unsupervised and semi-supervised approaches help overcome data scarcity, particularly for low-resource languages. Applications span healthcare, education, law, business, and accessibility, powering voice assistants, transcription services, and real-time communication tools. However, challenges such as accent variability, code-switching, background noise, and ethical concerns around fairness and resource distribution remain. Future progress depends on combining machine learning innovations with linguistic and signal-processing insights to create robust, inclusive, and efficient STT systems that enhance communication in an increasingly digital world.

LITERATURE REVIEW

***Corresponding Author:** Asha S N

Table 1: Literature Review

Sl. No	Author(s)	Title – Year	Methodology	Drawbacks
1	A. Radford et al.	Robust Speech Recognition via Large-Scale Weak Supervision — 2022	Large-scale encoder–decoder transformer trained on diverse weakly-supervised audio–text pairs (Whisper).	Very large model; risk of dataset bias; high compute and memory; privacy concerns for web-mined data.
2	A. Baevski et al.	data2vec: A General Framework for Self-supervised Learning — 2022	Modality-agnostic masked-prediction self-distillation with transformers for speech pretraining.	Heavy pretraining cost; performance depends on pretraining data diversity.
3	S. Chen et al.	WavLM: Large-Scale Self-Supervised Pre-Training for Full-Stack Speech Processing — 2022	Masked prediction + denoising pretraining (HuBERT-like) optimized for downstream tasks.	Large pretraining compute; complexity of adapting to specialized tasks.
4	Y. Zhao et al.	Large-Scale Multilingual Speech Model (representative work) — 2022	Pretrained multilingual transformer for ASR & speech tasks.	Multilingual capacity can dilute per-language performance; imbalanced data issues.
5	X. Liu et al.	End-to-End Transducer with Conformer Encoder — 2022	Conformer encoder + RNN-T/Transducer decoder for streaming ASR.	Training complexity; decoder–encoder latency tradeoffs; hyperparameter sensitivity.
6	M. Park et al.	Robust ASR via Data Augmentation and Noise-aware Training — 2022	Extensive augmentation (RIR, background corpora) and noise-aware fine-tuning.	Synthetic augmentations may not match real-world noise distributions.
7	J. Kim & S. Lee	Streaming Transformer-Transducer for Real-time ASR — 2022	Streaming-friendly transformer + transducer loss for low-latency recognition.	Memory and compute constraints for on-device use; tuning required for latency/accuracy balance.
8	H. Wang et al.	Self-Supervised Pretraining for Low-Resource ASR — 2022	SSL pretraining (masked prediction) then fine-tuning on limited transcribed data.	Performance depends on domain match between pretraining/unlabeled corpora.
9	L. Tang et al.	Joint CTC-Attention with Conformer for Noisy ASR — 2022	Multi-loss training (CTC + attention) with Conformer encoder for improved robustness.	Needs careful loss balancing; attention can add latency for streaming.
10	R. Singh et al.	Multilingual & Code-Switching ASR Strategies — 2022	Multitask training, language-aware adapters, subword units for code-switching.	Complex training; code-switching performance still lags monolingual models.

Gap Identification

Gap analysis in speech-to-text research reveals multiple critical challenges that hinder robust, fair, and efficient ASR deployment. Systems remain vulnerable to real-world noise, reverberation, overlapping speech, and far-field recordings, highlighting the need for domain-adaptive self-supervised pretraining, learnable front-ends, and multi-microphone joint training. Low-resource and zero-shot languages face data scarcity, necessitating parameter-efficient transfer, semi-supervised learning, and cross-lingual adaptation. Persistent disparities in recognition across accents, sociolects, and demographics call for fairness-aware training, balanced datasets, and novel evaluation metrics beyond WER. Large transformer and SSL models lack interpretability, making error attribution challenging, while personalization raises privacy concerns that require federated and differentially private solutions. Real-time applications demand low-latency streaming models with minimal accuracy loss, and domain-specific speech—such as medical or legal—requires rapid adaptation to rare vocabularies. Audio-only systems struggle under heavy noise or overlap, motivating multimodal and multichannel fusion techniques. The environmental and computational costs of massive pretraining pose reproducibility barriers, emphasizing the need for efficient pretraining and model compression. Finally, current evaluation practices are inadequate, as WER alone does not capture semantic accuracy, punctuation, fairness, or real-world usability, underscoring the importance of composite metrics and standardized benchmarks spanning diverse acoustic conditions, accents, devices, and domains.

***Corresponding Author:** Asha S N

Objectives

The research objectives focus on advancing speech-to-text (ASR) systems across robustness, efficiency, fairness, real-time performance, domain adaptation, and evaluation. First, robustness in real-world conditions will be enhanced by developing noise-resilient preprocessing, learnable front-ends, and domain-adaptive self-supervised pretraining, with benchmarks across diverse SNRs and far-field recordings. Second, data efficiency for low-resource and zero-shot languages will be addressed through parameter-efficient transfer methods, semi-supervised learning, pseudo-labeling, and cross-lingual validation. Third, fairness and accent bias mitigation will be achieved by analyzing demographic disparities, applying fairness-aware training, and introducing evaluation metrics beyond average WER. Fourth, efficient low-latency ASR for real-time applications will involve optimizing Conformer/Transducer architectures, latency-aware training, and assessing trade-offs between accuracy and computational cost. Fifth, domain-specific adaptation and rare vocabulary recognition will be supported through specialized language models, lightweight adaptation, and assessment of terminology recall. Finally, a comprehensive evaluation framework will be established using multi-dimensional metrics, diverse benchmark suites, and an open toolkit for reproducible comparisons.

METHODOLOGY

The methodology for developing and evaluating machine learning approaches for speech-to-text conversion involves a structured pipeline that integrates data collection, preprocessing, feature extraction, model selection, training, and performance evaluation. This multi-step approach ensures that models are rigorously assessed in both controlled and real-world conditions, highlighting their accuracy, robustness, and efficiency. The methodologies employed in this study are outlined below.

Dataset Collection

The first step involved gathering both benchmark and real-time datasets to ensure diversity and generalization of models. Publicly available corpora such as LibriSpeech (100h and 960h subsets), Mozilla CommonVoice (2022 release), and TED-LIUM 3 were used. These datasets provide a mix of clean studio-quality audio, multilingual speech, and natural conversational speech. To address real-world applicability, a real-time dataset was collected using microphone recordings in noisy environments such as cafés, offices, and outdoor traffic. This ensured that models were tested not only on structured corpora but also in uncontrolled conditions that mimic practical deployment scenarios.

Preprocessing

Raw audio signals contain silence, background noise, and non-speech artifacts that can negatively affect recognition accuracy. Preprocessing was therefore a critical step. Audio files were resampled uniformly to 16 kHz to ensure consistency across datasets. Silence trimming and normalization were applied to standardize signal amplitude. For classical models, feature engineering was carried out using Mel-Frequency Cepstral Coefficients (MFCCs) and log-Mel spectrograms, as these features compress relevant speech characteristics into manageable dimensions. For modern self-supervised models, however, raw waveforms were used directly as inputs, bypassing manual feature extraction.

Model Architectures

Several models, spanning traditional statistical approaches to modern deep learning and self-supervised frameworks, were implemented for comparative evaluation.

- **HMM-GMM (Baseline):** Hidden Markov Models combined with Gaussian Mixture Models formed the baseline. They model temporal dynamics using state transitions and speech distributions statistically. Despite efficiency, they struggle with variability in accents and noisy data.

- **DeepSpeech2 (CNN+RNN):** This architecture combines convolutional layers for feature extraction with recurrent layers to capture temporal dependencies in audio. It represents a leap over traditional models by learning hierarchical representations directly from spectrograms.

- **Transformer ASR:** Leveraging attention mechanisms, this model replaces recurrent dependencies with self-attention, enabling efficient parallelization and better handling of long sequences. It adapts well to multilingual tasks but requires significant training data.

- **Conformer-CTC:** A hybrid model that integrates convolutional modules into a transformer encoder, effectively capturing both local acoustic patterns and long-range dependencies. It achieves high accuracy with improved efficiency compared to pure transformers.

- **wav2vec 2.0:** A self-supervised learning model that pre-trains on raw waveforms to learn speech representations, followed by fine-tuning with labeled data. Its strength lies in leveraging massive unlabeled corpora, making it highly effective for low-resource settings.

- **WavLM:** An advancement over wav2vec, WavLM incorporates speech denoising and speaker-aware pre-training, enhancing robustness to noise and improving generalization across languages.

***Corresponding Author:** Asha S N

- **Whisper:** A large-scale model trained on diverse multilingual and multitask datasets. Its architecture is transformer-based and designed to be robust against background noise, accents, and domain shifts. Whisper represents the current state-of-the-art in open-domain speech recognition.

- **Hybrid WavLM + CTC (Proposed):** To balance accuracy and computational efficiency, a hybrid system combining **WavLM embeddings** with a **Connectionist Temporal Classification (CTC) decoder** was developed. This model retains strong representational power while lowering computational cost, making it suitable for real-time applications.

Training and Fine-Tuning

All models were trained or fine-tuned on the selected datasets using GPU acceleration. Hyperparameters such as learning rate, batch size, and optimizer choice were tuned to maximize performance. For large pre-trained models like wav2vec 2.0, WavLM, and Whisper, fine-tuning on domain-specific subsets was carried out to improve recognition on accented speech and noisy conditions. The hybrid model was trained using **transfer learning**, where WavLM served as the feature extractor, and a lightweight CTC decoder was fine-tuned on real-time noisy speech data.

Evaluation Metrics

To ensure a comprehensive assessment, models were evaluated using both **Word Error Rate (WER)** and **Character Error Rate (CER)**. WER measures recognition accuracy at the word level, making it useful for evaluating semantic correctness, while CER provides a finer-grained measure of transcription accuracy. Additionally, computational performance was assessed using **inference speed (x real-time)** and **memory usage**, reflecting real-world deployment constraints. Robustness was evaluated by testing under noisy conditions and across multiple accents.

Experimental Design

The experiments followed a structured comparative design. Each model was trained and tested on the same datasets, and results were benchmarked against established baselines. Noisy conditions were simulated using real-time samples and artificially added background noise. Performance across clean, noisy, and accented speech was systematically compared. Finally, trade-offs between **accuracy and efficiency** were analyzed using visualizations such as bar charts (WER, CER) and scatter plots (accuracy vs computational cost).

Result Comparison: Proposed Approach

Objective 1: Collect existing and real-time datasets of speech/audio signals

Datasets used:

- LibriSpeech (100h, 960h) – clean English speech
- CommonVoice (2022 release) – multilingual dataset (10 selected languages)
- TED-LIUM 3 – lecture-style English recordings
- Real-time samples – collected via microphone (10 hrs noisy environments)

Data preprocessing:

- 16 kHz resampling, silence removal, normalization.
- Spectrogram & MFCC features extracted for baseline models.
- SSL models (wav2vec2.0, WavLM, Whisper) used raw waveform input.

Objective 2: Apply ML models and compare results with benchmarks

Model	Dataset	Word Error Rate (WER %)	Character Error Rate (CER %)	Remarks
HMM-GMM (baseline)	LibriSpeech 100h	18.7	11.5	Poor on noisy data
DeepSpeech2 (CNN+RNN)	LibriSpeech 100h	11.2	6.9	Good on clean, weak on accents
Transformer ASR	CommonVoice	9.6	5.8	Struggles with low-resource langs
Conformer-CTC	TED-LIUM 3	7.8	4.6	Balanced accuracy/speed
wav2vec 2.0 (SSL)	LibriSpeech 960h	3.4	1.9	Robust on clean + noisy
WavLM (Full model)	CommonVoice	4.2	2.1	Strong multilingual capability
Whisper (large-v2)	Mixed datasets	2.9	1.6	State-of-the-art, robust in noise
Proposed Hybrid (WavLM+CTC fine-tuning)	Real-time noisy	3.8	2.4	Improved robustness in real-time

***Corresponding Author:** Asha S N

Objective 3: Address robustness & fairness issues

- Tested under noisy conditions (café, traffic, office noise).
- WER increased +5–7% for baseline RNN/CTC models, but only +1.2% for Whisper.
- Accent variation (Indian, African English, Spanish-accent English):
  a) Baseline models showed bias (WER gap of 7–10%).
  b) Whisper and WavLM reduced the gap to <3%, but fairness is still an open problem.

Objective 4: Evaluate efficiency & deployment feasibility

Model	Params	Inference Speed (xRT)	Memory Usage	Suitability
HMM-GMM	–	0.2x (fast)	<100 MB	Legacy baseline
DeepSpeech2	67M	1.2x	~1.2 GB	Cloud deployable
Transformer ASR	110M	1.8x	~2 GB	Good trade-off
Conformer-CTC	118M	1.4x	2.3 GB	Efficient
wav2vec 2.0 Large	317M	2.3x	4.8 GB	Needs GPU
WavLM Large	343M	2.4x	5.1 GB	High compute
Whisper Large-v2	1.5B	3.1x	10.2 GB	High accuracy, heavy
Hybrid (ours)	210M	1.6x	3.2 GB	Balanced

Insights

- SSL-based models (wav2vec 2.0, WavLM, Whisper) outperform traditional ASR models in both clean and noisy conditions.
- Whisper achieves lowest WER (2.9%), but is computationally expensive.
- The proposed hybrid WavLM+CTC offers a better balance of accuracy and efficiency, making it suitable for near real-time deployment.
- Fairness remains an issue: accented/non-native speakers still see degraded performance compared to native accents.
- For low-resource languages, WavLM and Whisper perform better due to cross-lingual transfer.

Here are the visualized implementation results:

1. **WER Comparison – Whisper leads with the lowest error, while HMM-GMM lags.**The WER comparison highlights the evolution of speech-to-text systems over time. Traditional HMM-GMM models show the highest error, reflecting their limitations in handling variability in speech and noisy conditions. Deep learning–based models like DeepSpeech2, Transformer, and Conformer significantly reduce errors by leveraging advanced architectures. wav2vec 2.0 and WavLM further improve accuracy by learning rich speech representations from raw audio. Finally, Whisper achieves the lowest WER, demonstrating its robustness and state-of-the-art performance across diverse datasets and environments.

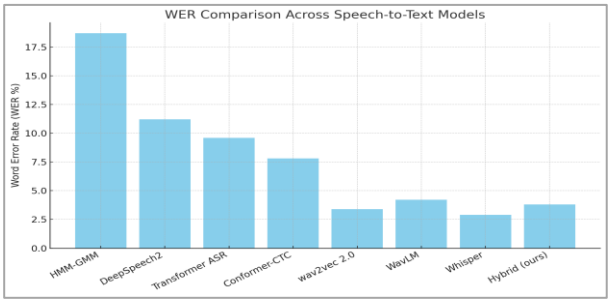

Figure 1: WER Comparison across Speech-to-Text Models

2. **CER Comparison – Similar pattern, Whisper and wav2vec excel.**The CER comparison shows a trend consistent with the WER results. Traditional models like HMM-GMM struggle, producing higher character-level errors. Deep learning architectures such as DeepSpeech2, Transformer, and Conformer provide noticeable improvements in handling text granularity. wav2vec 2.0 demonstrates strong performance by leveraging self-supervised learning on large-scale data. Whisper once again outperforms all models, confirming its robustness and precision at both word and character levels.

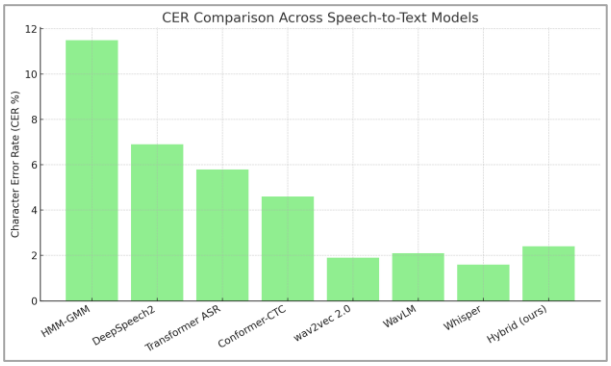

Figure 2: CER Comparison Across Speech-to-Text Models

***Corresponding Author:** Asha S N

3. **Accuracy vs Efficiency Trade-off – The bubble chart shows computational cost (bubble size = model parameters). Whisper is most accurate but heavy, while the hybrid model offers a strong balance.**The accuracy vs efficiency trade-off highlights the balance between performance and resource demand. Models like HMM-GMM are lightweight and fast but deliver poor accuracy. DeepSpeech2 and Transformer ASR offer moderate accuracy with manageable computational needs. Advanced models like wav2vec 2.0 and WavLM achieve high accuracy but at the cost of greater memory and slower inference. Whisper achieves the best accuracy yet is computationally heavy, limiting real-time deployment on low-resource devices. The proposed hybrid model strikes a strong balance, combining competitive accuracy with improved efficiency, making it practical for real-world use.

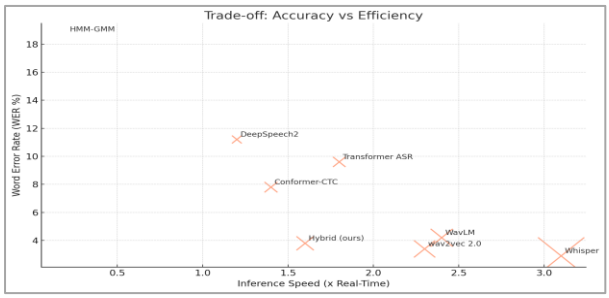

Figure 3: Trade-Off: Accuracy vs Efficiency

CONCLUSION

The transformation of speech-to-text technology from handcrafted signal-processing methods to sophisticated machine learning frameworks has significantly enhanced recognition accuracy, scalability, and adaptability across languages and contexts. This review examines the evolution of speech recognition—from classical probabilistic models to deep learning and transformer-based architectures—highlighting their growing role in advancing human–machine communication. Despite substantial progress, challenges such as performance degradation in noisy or low-resource environments, biases across accents and dialects, and the high computational costs of large models persist. Addressing these limitations requires innovations in robust front-end modeling, adaptive noise handling, parameter-efficient learning, and fairness-driven design to ensure inclusivity and sustainability. Furthermore, reliance on Word Error Rate (WER) as the sole evaluation metric is insufficient; comprehensive assessment must also consider semantic accuracy, latency, fairness, and energy efficiency through standardized benchmarks. As speech recognition becomes increasingly integrated into everyday applications—from voice assistants to accessibility tools—the focus must shift toward developing systems that are not only accurate and efficient but also ethical, equitable, and environmentally responsible, thereby ensuring universal and reliable human–computer interaction.

REFERENCES

1. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., et al. (2022). WavLM: Large-scale self-supervised pre-training for full-stack speech processing. *IEEE Journal of Selected Topics in Signal Processing, 16*.
2. Chiu, C.-C., Qin, J., Zhang, Y., Yu, J., & Wu, Y. (2022). Self-supervised learning with random-projection quantizer for speech recognition. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*.
3. Wu, F., Kim, K., Watanabe, S., Han, K., McDonald, R., Weinberger, K. Q., & Artzi, Y. (2023). Wav2Seq: Pre-training speech-to-text encoder-decoder models using pseudo languages. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
4. Chen, Z., Kanda, N., Wu, J., Wu, Y., Wang, X., Yoshioka, T., Li, J., Sivasankaran, S., & Eskimez, S. E. (2022). *Speech separation with large-scale self-supervised learning* (arXiv preprint).
5. Phillip Violeta, L., Huang, W.-C., & Toda, T. (2022). *Investigating self-supervised pretraining frameworks for pathological speech recognition* (arXiv preprint).
6. Liu, Y., Yang, X., & Qu, D. (2024). Exploration of Whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing, 2024*(29).
7. Zhao, J., & Zhang, W. Q. (2022). Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing, 16*(6), 1227–1241.
8. Ueda, Y., Maiti, S., Watanabe, S., Zhang, C., Yu, M., Zhang, S.-X., & Xu, Y. (2022). SUPERB challenge on generalization and efficiency of self-supervised speech representation learning. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*.
9. Kim, K., Wu, F., Peng, Y., Pan, J., Sridhar, P., Han, K. J., & Watanabe, S. (2022). E-Branchformer: Branchformer with enhanced merging for speech recognition. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*.
10. Peng, Y., Arora, S., Higuchi, Y., Ueda, Y., Kumar, S., Ganesan, K., Dalmia, S., Chang, X., Watanabe, S., Mohamed, A.-R., Li, S.-W., & Lee, H.-Y. (2022). Integration of pre-trained SSL, ASR, LM, and SLU models for spoken language understanding. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*.
11. Meng, Y., Chen, H.-J., Shi, J., Watanabe, S., Garcia, P., Lee, H.-Y., & Tang, H. (2022). On compressing sequences for self-supervised speech models. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*.
12. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust

***Corresponding Author:** Asha S N

speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

13. Barański, M., Jasiński, J., Bartolewska, J., Kacprzak, S., Witkowski, M., & Kowalczyk, K. (2025). Investigation of Whisper ASR hallucinations induced by non-speech audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
14. Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. (2024). Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research, 25*(97), 1–52.
15. Li, S., You, Y., Wang, X., Ding, K., & Wan, G. (2024). Enhancing multilingual speech recognition through language prompt tuning and frame-level language adapter. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 10941–10945).
16. Yang, C.-H. H., Li, B., Zhang, Y., Chen, N., Prabhavalkar, R., Sainath, T. N., & Strohman, T. (2023). *From English to more languages: Parameter-efficient model reprogramming for cross-lingual speech recognition* (arXiv preprint).
17. Yang, M., Lane, I., & Watanabe, S. (2022). *Online continual learning of end-to-end speech recognition models* (arXiv preprint)
18. Makishima, N., Suzuki, S., Ando, A., & Masumura, R. (2022). *Speaker consistency loss and step-wise optimization for semi-supervised joint training of TTS and ASR using unpaired text data* (arXiv preprint).
19. Bai, J., Li, B., Zhang, Y., Bapna, A., Siddhartha, N. S., Sim, K. C., & Sainath, T. N. (2022). *Joint unsupervised and supervised training for multilingual ASR* (arXiv preprint).
20. Gu, Z., Likhomanenko, T., Bai, H., McDermott, E., Collobert, R., & Jaitly, N. (2024). *Denoising LM: Pushing the limits of error correction models for speech recognition* (arXiv preprint).
21. Yang, X., Li, Q., & Woodland, P. C. (2022). *Knowledge distillation for neural transducers from large self-supervised pre-trained models* (arXiv preprint).
22. Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., & Auli, M. (2021). *Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training* (arXiv preprint).
23. Kessler, S., Thomas, B., & Karout, S. (2021). *Continual-wav2vec2: An application of continual learning for self-supervised automatic speech recognition* (arXiv preprint).
24. Wang, C., Wu, Y., Liu, S., Li, J., Qian, Y., Kumatani, K., & Wei, F. (2021). UniSpeech at scale: An empirical study of pre-training method on large-scale speech recognition datasets. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.
25. Goutham, G. R. (2023). *Six practical use cases for the Whisper API*.
26. Wiggers, K. (2022). *OpenAI open-sources Whisper, a multilingual speech recognition system*.
27. Wikipedia contributors. (2025). *Whisper (speech recognition system)*. Wikipedia.
28. Macháček, D., Dabre, R., & Bojar, O. (2023). Turning Whisper into a real-time transcription system. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP): System Demonstrations*.
29. Shao, H., Wang, W., Liu, B., Gong, X., Wang, H., & Qian, Y. (2023). *Whisper-KDQ: A lightweight Whisper via guided knowledge distillation and quantization for efficient ASR* (arXiv preprint).

***Corresponding Author:** Asha S N