# COMPARATIVE ANALYSIS OF END-TO-END NEURAL SPEECH RECOGNITION FRAMEWORKS

**Hasan Gyulyustan, Pavel Kyurkchiev**

***Abstract.*** *This paper presents a comprehensive comparative analysis of modern end-to-end automatic speech recognition (ASR) frameworks, emphasizing architectural diversity, performance trade-offs, and real-world deployment considerations. The study examines representative toolkits such as Wav2Letter++, WeNet, ESPnet, Fast Conformer, and PyTorch-Kaldi across multiple evaluation metrics, including word and character error rates, training and inference efficiency, and hardware requirements. Beyond controlled benchmark results, the paper investigates robustness under noisy and demographically diverse conditions, highlighting persistent biases in commercial ASR systems. Furthermore, it explores the integration of neural language models and multi-task learning architectures, which enhance alignment, adaptability, and contextual understanding. Emerging domains such as cross-lingual ASR, speech emotion recognition, and silent speech interfaces are analyzed to reveal limitations of traditional evaluation frameworks when applied to multi-domain and non-auditory data. The findings underscore the necessity of holistic performance metrics that incorporate fairness, robustness, and resource efficiency for next-generation ASR system development.*

**Key words:** End-to-End Speech Recognition; Neural Architectures; WeNet; ESPnet; Fast Conformer; Wav2Letter++; PyTorch-Kaldi; Automatic Speech Recognition (ASR); Multi-Task Learning; Language Model Integration; Noise Robustness; Demographic Bias; Cross-Domain Generalization; Speech Emotion Recognition; Silent Speech Recognition.

## Introduction

Modern end-to-end speech recognition frameworks have converged on transformer-based architectures while pursuing distinct optimization strategies for production deployment. WeNet introduced the unified streaming and non-streaming U2 approach, combining CTC and attention mechanisms with dynamic chunk-based processing to achieve 5.03% relative character error rate reduction on AISHELL-1 while maintaining reasonable RTF and latency [1]. This dual-mode capability addresses a critical deployment challenge where sys-

tems must balance real-time processing requirements with transcription accuracy. Building on established accuracy measurement methodologies, comparative analysis reveals that architectural choices profoundly impact real-world deployment metrics beyond simple error rates [2]. The evolution from pipeline-based hybrid systems to end-to-end models represents a paradigm shift in ASR development, yet systematic evaluation of hardware requirements versus performance trade-offs remains limited across framework implementations [3].

Recent toolkit developments prioritize production readiness through training efficiency and inference speed optimizations. Wav2Letter++ demonstrated 2× faster training than competing frameworks while maintaining competitive accuracy, achieving linear scaling to 64 GPUs for models with 100 million parameters [4]. Fast Conformer further advanced this trajectory, delivering 2.8× speedup over standard Conformer while supporting scaling to billion-parameter models without architectural modifications [5]. These improvements translate directly to reduced computational costs and faster iteration cycles during model development. Framework selection for practitioners increasingly depends on the interplay between training efficiency, inference speed, model size, and deployment complexity [6]. PyTorch-Kaldi bridges the efficiency of Kaldi with PyTorch's flexibility, enabling rapid prototyping while maintaining production-grade performance [6].

*Table 1. Summary of ASR Frameworks: Architecture, Efficiency, and Accuracy Metrics*

| **Framework** | **Architecture Type** | **Training Time (relative)** | **Inference Speed** | **Model Size** | **Hardware Requirements** | **Accuracy (LibriSpeech test-clean WER %)** |
|---|---|---|---|---|---|---|
| WeNet | Transformer/ Conformer U2 | Baseline | RTF < 0.1 | 50-100M params | GPU: 8GB+ VRAM | 2.0-3.0 |
| Wav2Letter++ | Fully Convolutional | 0.5x (2x faster) | RTF < 0.05 | 30-80M params | GPU: 4GB+ VRAM, Linear scaling to 64 GPUs | 3.5-4.5 |
| ESPnet | Transformer/ Conformer | 1.2x | RTF 0.1-0.15 | 80-150M params | GPU: 16GB+ VRAM | 2.0-2.5 |
| Fast Conformer | Conformer + Downsampling | 0.36x (2.8x faster) | RTF < 0.08 | 100M-1B params | GPU: 16GB+ VRAM | 1.9-2.3 |
| PyTorch-Kaldi | Hybrid TDNN-HMM | 0.8x | RTF 0.15-0.2 | 20-60M params | GPU: 8GB+ VRAM, CPU compatible | 3.0-4.0 |

*Table 2. ASR Framework Performance Across Standard Benchmarks and Computational Requirements*

| Framework | LibriSpeech (960h) WER% | AISHELL-1 (170h) CER% | Switchboard (300h) WER% | Scaling with Data | Memory Footprint |
|---|---|---|---|---|---|
| WeNet | 2.0 / 4.6 (clean/other) | 4.6 | 8-10 | Linear improvement | 6-12GB |
| Wav2Letter++ | 3.8 / 10.5 | 6.5 | 10-12 | Strong scaling, linear GPU scaling | 4-8GB |
| ESPnet | 2.2 / 5.2 | 4.5 | 7-9 | Excellent with large data | 12-24GB |
| Fast Conformer | 1.9 / 4.1 | 4.2 | 7-8 | Excellent, handles up to 1B params | 16-32GB |
| PyTorch-Kaldi | 3.5 / 8.9 | 5.8 | 9-11 | Good with moderate data | 4-10GB |

## Robustness and Adaptability of ASR Systems Under Real-World Conditions

While controlled accuracy metrics provide useful baselines, real-world ASR systems face substantial degradation from environmental noise and demographic variation. Commercial ASR services show pronounced demographic disparities, with WERs of 0.35 for African American speakers versus 0.19 for white speakers, indicating systematic inequity rather than simple accuracy loss [7]. Similar biases appear for ESL speakers, people of color, and gender groups, though these transcription biases do not necessarily propagate into downstream model scores [8]. Such disparities can reduce trust and increase failure rates for marginalized users.

The AURORA framework established early robustness benchmarks by introducing eight real-world noise types across multiple SNR levels, supporting systematic comparison of feature extraction techniques [9]. Modern deep learning systems build on this foundation: CNN-LSTM models reach 97.96% clean accuracy and 90.72% under noise on Aurora-2, outperforming simpler architectures [10]. Some systems further show 25.7% relative WER reduction through noise-disentanglement modules [11]. The AURORA-2J corpus extends this by incorporating speaker-level analyses that reveal individual-level performance differences [12].

Personalized ASR demonstrates that targeted adaptation can accommodate atypical speech patterns. For deaf and hard-of-hearing speakers, as little as 1000 utterances (1–2 hours) substantially improves performance via speaker-

specific fine-tuning [13]. For dysarthric speech, speaker-dependent systems consistently outperform commercial speaker-independent ASR for both words and sentences [14]. These results highlight viable pathways toward more inclusive ASR, though initial data collection requirements remain a practical barrier.

## Integration of Language Models and Multi-Task Learning in Modern ASR Architectures

Modern ASR frameworks rely increasingly on advanced language-model integration to improve accuracy and handle domain-specific vocabulary. Neural language models using letter-based features and importance sampling (e.g., Kaldi-RNNLM) achieve competitive results for large vocabularies while remaining computationally efficient [15]. This mitigates the central trade-off between vocabulary coverage and model complexity. WeNet 2.0 illustrates practical LM integration by combining n-gram LMs, WFST decoding, and contextual biasing, yielding up to 10% relative improvement and enabling rapid user-specific adaptation [16].

Multi-task learning has become a key strategy for addressing alignment and robustness limitations in end-to-end ASR. Joint CTC-attention architectures achieve 5.4–14.6% relative CER improvements by leveraging both CTC and attention mechanisms, particularly benefiting noisy or long-sequence conditions [17]. Hybrid acoustic-to-word and character-level CTC architectures match or exceed DNN-HMM accuracy while enabling 25× faster decoding [18]. Additional work on encoder-decoder models with specialized focus mechanisms shows improved sequence labeling robustness [19]. Adversarial training approaches using criticizing language models further extend multi-task objectives, allowing models to exploit large unpaired text corpora while preserving compatibility with existing decoders [20].

Modern LM-integration strategies can be summarized into four categories:

- N-gram WFST Models – Efficient deterministic decoding via pre-composed search graphs but with large memory demands (>1 GB) and limited contextual flexibility.
- Neural LM Rescoring – Second-pass RNNLM reranking that yields major WER gains (e.g., WSJ 11.1% → 4.5%) when combined with Transformer-CTC pipelines [21].
- Multi-Task LM Estimation – Joint CTC/attention optimization that learns internal LM priors and supports fusion techniques for end-to-end systems.
- Hybrid Approaches with Contextual Biasing – Incorporating user-specific

vocabulary or domain terms through WFST or attention-based biasing, enabling rapid adaptation without full retraining [16].

## Emerging Applications and Cross-Domain Transfer for Speech Recognition Systems

Large-scale multi-domain corpora have reshaped ASR evaluation by enabling systematic study of cross-domain generalization. As an example we can take GigaSpeech that offers 10,000 hours of transcribed audio from audiobooks, podcasts, and YouTube, covering diverse topics and providing professionally transcribed evaluation sets [22]. WenetSpeech extends this paradigm to Mandarin with 22,400 hours of mixed-condition audio and introduces OCR-based segmentation and automated label error detection to ensure annotation quality [23]. These resources allow researchers to test whether evaluation methods validated on single-domain datasets remain reliable for systems trained on heterogeneous data, illuminating open questions about transfer learning and domain adaptation.

The emerging applications such as speech emotion recognition and silent speech recognition challenge traditional reliance on word error rate. Emotion recognition systems require both transcription accuracy and robust affect classification across cultures, with noise and cultural variability posing major evaluation difficulties [24]. Silent speech recognition relies on non-auditory signals – e.g., graphene strain sensors detecting throat muscle movement – achieving 55–85% word accuracy using machine learning decoding [25]. These domains demand expanded evaluation frameworks that incorporate task-specific metrics such as emotion classification accuracy, cultural robustness, and sensor signal quality.

Cross-lingual and multilingual ASR systems introduce language-specific accuracy measurement challenges that conventional frameworks often overlook. Limited vocabulary approaches targeting under-resourced languages focus on small word sets to enable human-to-human and human-to-machine systems for illiterate populations, requiring evaluation metrics sensitive to vocabulary coverage and adaptation efficiency rather than large-vocabulary performance alone [26]. Arabic ASR exemplifies language-specific complications, where diacritized recognition systems must balance transcription accuracy with diacritic placement correctness, with end-to-end deep learning approaches using ESPnet and Espresso frameworks achieving substantial error rate reductions on Modern Standard Arabic but revealing the scarcity of benchmark datasets limiting systematic evaluation [27, 28]. Extending accuracy measurement methodologies to accommodate morphologically rich languages, tonal languages, and di-

alectal variations would enable more equitable international ASR development by providing evaluation frameworks that capture language-specific recognition challenges rather than imposing metrics optimized for well-resourced languages.

*Table 3. Comparison of ASR Performance Across Application*

| Application Domain | Primary Input Modality | Standard WER/CER | Task-Specific Metrics | Domain Transfer Challenges | Evaluation Corpus Size |
|---|---|---|---|---|---|
| Multi-Domain Conversational (GigaSpeech) | Audio (read & spontaneous) | 4% cap for XL subset | Domain-specific WER across topics | Style variation, topic diversity | 10,000h transcribed |
| Multi-Domain Conversational (WenetSpeech) | Audio (podcast & video) | Baseline on Dev/Test sets | Cross-domain generalization | Noisy conditions, spontaneous speech | 10,000h labeled, 22,400h total |
| Speech Emotion Recognition | Audio with affective content | Standard WER | Emotion classification accuracy, cultural robustness | Noisy data, cultural variation in expression | Variable, limited benchmarks |
| Silent Speech Recognition | Throat muscle strain signals | 55-85% word accuracy | Sensor signal quality, movement detection accuracy | Non-auditory signal processing, user adaptation | 15 words × 20 repetitions (limited) |
| Limited Vocabulary ASR | Audio (under-resourced languages) | WER on small lexicons | Vocabulary coverage, adaptation efficiency | Resource scarcity, illiterate user support | Typically <100 hours |
| Arabic ASR with Diacritics | Audio (Modern Standard Arabic) | 5.24% WER (CNN-LSTM), 2.62% WER (attention) | Diacritic error rate, dialect-specific accuracy | Diacritization, dialectal variation | SASSC: 7 hours (limited) |

## References


[1] Z. Yao et al., WeNet: Production Oriented Streaming and Non-Streaming End-to-End Speech Recognition Toolkit, *Interspeech*, Aug. 2021, doi:10.21437/interspeech.2021-1983

[2] B. Zhang et al., WeNet: Production First and Production Ready End-to-End Speech Recognition Toolkit, arXiv / Cornell University, Feb. 2021

[3] A. Georgescu, A. Pappalardo, H. Cucu, M. Blott, Performance vs. hardware requirements in state-of-the-art automatic speech recognition, *Springer Nature*, Jul. 2021, doi:10.1186/s13636-021-00217-4

[4] V. Pratap et al., Wav2Letter++: A Fast Open-source Speech Recognition System, *ICASSP*, Apr. 2019, doi:10.1109/icassp.2019.8683535

[5] D. Rekesh et al., Fast Conformer With Linearly Scalable Attention For Efficient Speech Recognition, *ASRU*, May 2023, doi:10.1109/ASRU57964.2023.10389701

[6] M. Ravanelli, T. Parcollet, Y. Bengio, The Pytorch-kaldi Speech Recognition Toolkit, *ICASSP*, Apr. 2019, doi:10.1109/icassp.2019.8683713

[7] A. Koenecke et al., Racial disparities in automated speech recognition, *PNAS*, Mar. 2020, doi:10.1073/pnas.1915768117

[8] L. Hickman, M. Langer, R. Saef, L. Tay, Automated speech recognition bias in personnel selection: The case of automatically scored job interviews, *Journal of Applied Psychology*, Oct. 2024, doi:10.1037/apl0001247

[9] D. Pearce, H. Hirsch, The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, *ICSLP*, Oct. 2000, oi:10.21437/icslp.2000-743

[10] N. Djeffal, H. Kheddar, D. Addou, S. Selouani, Combined CNN-LSTM For Enhancing Clean And Noisy Speech Recognition, *Conference Publication*, Dec. 2024, doi:10.38169/0661-030-002-001

[11] S. Chen, S. Wei, D. Xu, Y. Long, Noisy Disentanglement with Tri-stage Training for Noise-Robust Speech Recognition, *arXiv*, Sep. 2025, doi:10.48550/arXiv.2509.01087

[12] S. Nakamura et al., AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition, *IEICE*, Mar. 2005, doi:10.1093/ietisy/e88-d.3.535

[13] L. Violeta, T. Toda, An Analysis of Personalized Speech Recognition System Development for the Deaf and Hard-of-Hearing, *APSIPA ASC*, Jun. 2023, doi:10.1109/APSIPAASC58517.2023.10317318

[14] I. Calvo et al., Evaluation of an Automatic Speech Recognition Platform for Dysarthric Speech, *Folia Phoniatrica et Logopaedica*, Nov. 2020, doi:10.1159/000511042

[15] H. Xu et al., Neural Network Language Modeling with Letter-Based Features and Importance Sampling, *ICASSP*, Apr. 2018, doi:10.1109/icassp.2018.8461704

[16] B. Zhang et al., WeNet 2.0: More Productive End-to-End Speech Recognition Toolkit, Interspeech, Sep. 2022, doi:10.21437/interspeech.2022-483

[17] S. Kim, T. Hori, S. Watanabe, Joint CTC-attention based end-to-end speech recognition using multi-task learning, *ICASSP*, Mar. 2017, doi:10.1109/icassp.2017.7953075

[18] S. Ueno, H. Inaguma, M. Mimura, T. Kawahara, Acoustic-to-Word Attention-Based Model Complemented with Character-Level CTC-Based Model, *ICASSP*, Apr. 2018, doi:10.1109/icassp.2018.8462576

[19] S. Zhu, K. Yu, Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding, *ICASSP*, Mar. 2017, doi:10.1109/icassp.2017.7953243

[20] A. Liu, H. Lee, L. Lee, Adversarial Training of End-to-end Speech Recognition Using a Criticizing Language Model, *ICASSP*, Apr. 2019, doi:10.1109/icassp.2019.8683602

[21] S. Karita et al., Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration, *Interspeech*, Sep. 2019, doi:10.21437/interspeech.2019-1938

[22] G. Chen et al., GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio, *Interspeech*, Aug. 2021, doi:10.21437/interspeech.2021-1965

[23] B. Zhang et al., WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition, *ICASSP*, Apr. 2022, doi:10.1109/icassp43922.2022.9746682

[24] Y. Dixit, S. Chauhan, S. Yadav, S. Singh, Surabhi, Speech Emotion Recognition, *International Journal for Multidisciplinary Research*, May 2024, doi:10.36948/ijfmr.2024.v06i03.21105

[25] D. Ravenscroft et al., Machine Learning Methods for Automatic Silent Speech Recognition Using a Wearable Graphene Strain Gauge Sensor, *MDPI Sensors*, Dec. 2021, doi:10.3390/s22010299

[26] J. Fendji, D. Tala, Y. Omer, M. Atemkeng, Automatic Speech Recognition Using Limited Vocabulary: A Survey, *Applied Artificial Intelligence*, Jul. 2022, doi:10.1080/08839514.2022.2095039

[27] A. Dhouib et al., Arabic Automatic Speech Recognition: A Systematic Literature Review, *Applied Sciences (MDPI)*, Sep. 2022, doi:10.3390/app12178898

[28] H. Alsayadi, A. Abdelhamid, I. Hegazy, Z. Fayed, Arabic speech recognition using end-to-end deep learning, *IET Signal Processing*, Jun. 2021, doi:10.1049/sil2.12057

Hasan Gyulyustan[1], Pavel Kyurkchiev[1]
[1] Paisii Hilendarski University of Plovdiv,
Faculty of Mathematics and Informatics,
236 Bulgaria Blvd., 4027 Plovdiv, Bulgaria
Corresponding author: hasan@uni-plovdiv.bg