10.48047/jocaaa.2026.35.03.13

Understanding Modern Neural Speech Recognition with Transformer Architectures

Uma Shankar Koushik Kethamakka
Independent Researcher, USA

Abstract

Neural speech recognition has been transformed by new transformer architectures, self-supervised learning and end-to-end training models. This article reviews the current state of the art systems from a representation learning perspective, and how their performance is achieved with effective representation learning from large volumes of unlabelled audio data. These techniques have evolved from early HMM and GMM systems, to state-of-the-art DNN-based acoustic models, and now to E2E transformer-based systems, with over 70 years of research. Self-supervised methods such as wav2vec 2.0, HuBERT, data2vec, and WavLM enable building systems to learn representations of speech that greatly reduce the need for expensive human transcript annotations. Some transformer-based architectures (including the Conformer architecture) integrate multi-head self-attention and convolutional layers, to obtain both short-term acoustic and longer-term context. E2E learning with CTC and transducer architectures enables direct predictions of text transcriptions from audio signals, bypassing the need for the construction of complex systems. This article surveys streaming architectures, model compression techniques, and practical challenges such as the benchmark to production gap. The use of streaming machine learning systems in voice assistants, health documentation and accessibility tools highlights the impact they have on billions of users around the world. Mentioned elsewhere are the implications for equitable access across language and acoustic conditions, environmental concerns, and the fast pace of improvements to the base models.

Keywords: Transformer Architectures, Self-Supervised Learning, End-to-End Speech Recognition, Attention Mechanisms, Conformer, Streaming ASR, Multilingual Acoustic Modeling, Model Compression

1. Introduction

Speech recognition is used as an integrated component of modern computing infrastructure and its applications, namely voice assistants, transcription software, speech-to-speech translation systems, accessibility technology, and multilingual applications. Speech recognition is used to convert spoken language into a textual format. It eliminates language barriers, supports natural human-computer interaction, and provides accessibility to individuals with hearing or motor disabilities. Systems using large-scale weak supervision to train strong speech recognition models have demonstrated that multilingual-multitask models can achieve human-level robustness and accuracy across a wide distribution of acoustic environments and speaking styles [1].

The global speech & voice recognition market was valued at around 20.25 billion US dollars in 2023. It is forecast to reach over 53.67 billion US dollars by 2030, with a CAGR of over 19 percent [2]. Factors driving the market include its adoption in healthcare documentation, customer service, automotive voice interfaces, smart home devices, and enterprise productivity applications. Beyond the specialized acoustic labs where it was first developed, speech recognition is expected to operate in 24/7/365 production environments under noisy conditions, across languages, in real time, and on resource-constrained devices.

The improvements in performance of neural speech recognition systems over time are important: for instance, the word error rate on the LibriSpeech benchmark dataset has dropped from around 10.3% for the previous hybrid deep neural network systems to near-human performance for recent transformer-based models [8]. In languages with little data, self-supervised pre-training is even more impactful: for instance, a model pre-trained on self-supervised data achieves a word error rate of less than 10.2% when trained on one hour of labeled data; without the unsupervised pre-training, the same amount of labeled data yields greater than 23.5% error. This article highlights the most important technology breakthroughs

10.48047/jocaaa.2026.35.03.13

that have made modern speech recognition systems practical, and discusses best practice for deploying them.

2. Historical Evolution of Speech Recognition

Knowledge-based models dominated early automatic speech recognition (ASR) technology. The first two decades of speech recognition in the 1950s and 1960s involved rules created by expert engineers, limited to isolated word recognition and small vocabularies. In the 1970s and 1980s, hidden Markov models provided a principled statistical basis for continuous word recognition over vocabularies of tens of thousands of words, an approach that predominated for three decades. In this framework, Gaussian mixture models (GMMs) were used as an acoustic model to represent the acoustic feature probability density functions (PDFs) given the phonetic state.

The deep learning revolution for speech recognition began around 2010 by replacing Gaussian mixture models (GMMs) with deep neural networks (DNNs) for the acoustic model while retaining the use of hidden Markov models for alignment. This hybrid approach achieved a 30 percent relative improvement in word error rates on benchmark databases. In the mid-2010s, RNNs (particularly LSTMs) became the standard for ASR. Seq2seq models could learn to align speech audio with text. Where this alignment is unknown, a connectionist temporal classification provides a loss function for seq2seq models [7].

The period from around 2020 has been characterized by self-supervised learning objectives and deep transformer networks for speech. Two such systems, Wav2Vec 2.0 and HuBERT, showed that unsupervised training on raw audio, followed by a second semi-supervised training stage with a small amount of read transcripts, can reach performance on par with supervised systems with orders of magnitude more speech data [3][4]. Whisper and other models trained on web-scale weakly-supervised data were found to generalize well across a variety of acoustic conditions and languages [1]. The Conformer architecture, which combines convolutions and transformers, was also found to improve performance on speech recognition tasks [6].

3. Self-Supervised Pre-training: Learning from Unlabeled Audio

3.1 Contrastive and Predictive Learning

Another important speech recognition problem is self-supervised representation learning on unlabeled audio. Contrastive predictive coding tasks a model with the learning of representations that can predict future audio frames given the past audio context, while using a contrastive loss which distinguishes between the true future audio and negative audio from elsewhere [14]. The model can be trained by this model to capture slowly varying attributes that determine the content of speech, while ignoring fast varying attributes such as channel noise, and the learned representations can be easily transfer-learned to downstream tasks with limited labeled data for fine-tuning.

The wav2vec 2.0 project is an extension of the original framework, which uses quantization of audio and masked prediction tasks to train the model to learn discrete units representing speech sounds [3]. The encoder is designed as a convolutional feature encoder, which learns latent feature representations from audio waveforms, which are then quantized into discrete targets. This, in turn, was able to learn acoustic and linguistic information without human transcription, and later fine-tuned on a much smaller amount of labeled data, achieving results comparable to other systems trained on orders of magnitude more labeled data.

3.2 Masked Prediction and Iterative Refinement

HuBERT extends this approach by using cluster assignments as target labels to perform masked prediction with offline clustering to create a self-consistent loop where cluster assignments improve as training progresses. The masked prediction task predicts cluster assignments on masked time steps, given the unmasked steps' context. Learning to make these predictions can be thought of as learning contextualized embeddings similar to masked language modeling in NLP, resulting in embeddings clustering to phonetic and linguistic attributes.

10.48047/jocaaa.2026.35.03.13

Multi-task self-supervised objectives combine pretext tasks to learn more powerful representations. Data2Vec unifies self-supervised learning across modalities by predicting the contextualized latent representations of masked portions of the input which lead to state-of-the-art results on speech, vision, and language tasks [11]. WavLM uses masked speech prediction and denoising objectives during pre-training to learn noise and reverb-strong representations [10]. These objectives are useful for training on languages with limited resources or no transcribed data, it democratizes speech technology, and enables state-of-the-art models for languages where supervised models cannot be trained due to lack of transcribed data.

Framework	Core Mechanism	Training Strategy	Key Advantage
Contrastive Predictive Coding	Predicting the future from the past context	Contrastive loss distinguishing true vs negative samples	Captures slowly-varying factors, ignores channel noise
wav2vec 2.0	Quantized representations with contrastive learning	Masked prediction on discrete speech units	Efficient learning with minimal labeled data requirements
HuBERT	Iterative clustering with masked prediction	Offline cluster assignments as targets	Progressive refinement of phonetic/linguistic structure
data2vec	Cross-modal self-supervised learning	Predicting contextualized latent representations	Unified framework across speech, vision, and language
WavLM	Masked prediction + denoising	Multi-task objectives for robustness	Noise and reverberation robustness

Table 1: Self-Supervised Pre-training Methodologies [3, 4]

4. Transformer Architectures: Attention-Based Processing

4.1 Adapting Transformers for Speech

A meaningful change in neural speech recognizers has been the replacement of recurrence with transformers, which have access to longer-range context. Speech is sampled at 16,000+ samples/second, thus speech sequences are several orders of magnitude longer than text sequences. Because self-attention is quadratic in the sequence length, speech transformers now typically preprocess the audio using a convolutional front-end, which downsamples the input (by a factor of 10 to 40) while preserving important acoustic information. Relative positional encodings are applicable to variable-length input and have been shown to generalize to longer sequences than trained on [5].

The Speech-Transformer (ST) extends the vanilla transformer to speech recognition by using time-based positional embeddings and time-based attention functions on high-dimensional acoustic inputs. The ST model uses an encoder-decoder architecture to process the input acoustic features through stacks of N multi-head self-attention and feed-forward layers. The encoder extracts acoustic-phonetic features and the decoder generates the target text sequences. The multi-head self-attention enables the model to simultaneously observe many aspects of the input, building features corresponding to formants, temporal dependencies, and context at multiple time scales.

4.2 The Conformer Architecture

The current state-of-the-art, the Conformer, combines convolutions and self-attention, taking advantage of the former's ability to capture local acoustic phenomena, such as formants and phonetic transitions, and the latter's ability to capture long-range effects, such as prosody and cross-word coarticulation [6]. The

10.48047/jocaaa.2026.35.03.13

model uses depthwise separable convolutions with gating to control which acoustic variations to attend to or ignore, and adds macaron-style feed-forward neural network layers applied between attention and convolution modules to increase representational power.

The Conformer achieved state-of-the-art results on several language tasks. It achieves this by combining local feature extraction and global modeling, resulting in representations of speech that are both local spectral features and long-range temporal dependencies. The success of Conformer has led to improvements in streaming, multilingual training and low-power implementations. Layer normalization and residual connections allow the training of deeper networks, and for example, stable optimization of networks with hundreds of millions of parameters.

Architecture	Design Principle	Component Integration	Performance Characteristic
Speech Transformer	Pure attention mechanism	Multi-head self-attention with positional encoding	Parallel processing for long-range dependencies
Conformer	Hybrid local-global modeling	Convolution modules within transformer blocks	Balanced fine-grained and temporal patterns
Emformer	Streaming with memory	Cached past computations with chunked attention	Low-latency real-time recognition
Macaron Feedforward	Enhanced capacity	Sandwich architecture with dual feed-forward	Increased non-linear transformation capability

Table 2: Transformer Architecture Innovations [5, 6]

5. End-to-End Learning: Simplifying the Recognition Pipeline

5.1 Connectionist Temporal Classification

Recent state-of-the-art ASR systems have moved towards directly modeling the audio-waveform-to-text mapping using a sequence-to-sequence model without intermediate representation. This is possible with the Connectionist Temporal Classification (CTC) framework, which allows for training such a model without pre-segmenting or aligning it to input sequences. Instead, it uses the forward-backward algorithm to sum over all possible alignments [7]. We can define the loss function by marginalizing over all possible alignments between the input and output sequences, while controlling for repeated symbols and variable-length outputs using blank symbols. This can be computed using efficient dynamic programming instead of considering a number of possible alignments that is exponential in the sequence length.

5.2 Attention-Based and Transducer Architectures

Attention-based encoder-decoder architectures (listen-attend-spell) train a decoder to predict output tokens from the full encoder output at each prediction step, with previous tokens as additional context [15]. Cross-attention allows the model to focus on specific parts of the encoder to predict tokens, capturing complex output dependencies and outperforming CTC alone. Hybrid objectives combine CTC and attention losses during training to enable alignment learning while also modeling dependencies.

Transducer architectures can be used to model the stream as well as the output. The recurrent neural network transducer and transformer transducer are types of neural net transducer which model joint distributions over alignments and outputs [7]. Because training with joint probability is difficult, models reach near-optimal accuracy under streaming conditions. Two-pass models first decode using low-latency streaming models, then rescore using more accurate models that have access to the full audio and a more complex language model [8]. Transducer architectures are also commonly used for production systems as they allow easy accuracy-latency trade-off.

10.48047/jocaaa.2026.35.03.13

Framework	Alignment Approach	Operational Mode	Primary Benefit
CTC	Implicit marginalization over alignments	Single-pass direct mapping	No pre-segmented data or explicit alignment needed
Listen-Attend-Spell	Cross-attention alignment learning	Encoder-decoder sequence-to-sequence	Complex output dependency modeling
RNN-T / Transformer-T	Joint alignment and output distribution	Streaming with output dependencies	Near-optimal streaming accuracy
Two-pass Architecture	Streaming first-pass + full-context rescoring	Dual-stage processing	Low latency with high accuracy through deliberation

Table 3: End-to-End Learning Frameworks [7, 17]

6. Streaming Architectures and Production Deployment

6.1 Real-Time Processing Techniques

Streaming architectures modify the transformer to operate in real time rather than waiting until the end of the utterance. These can do limited past attention, use chunked embedding inputs with fixed-length attention, or cache previous computations with Emformer architectures [12]. This sacrifice in accuracy has enabled applications such as live captioning or voice assistants where low latency is essential, and with recent architectural advances the gap between streaming and non-streaming models has narrowed, with some two-stage models nearing the accuracy of non-streaming models on some benchmarks.

6.2 Model Compression and Edge Deployment

On-device processing is important due to privacy concerns and low latency. Techniques such as quantization, pruning, and knowledge distillation can be applied to allow state-of-the-art speech recognizers to run on smartphones and wearables that do not have access to external processing resources. Quantization decreases model weights, reducing memory requirements and increasing inference speed on integer-based hardware, converting 32-bit floating point values to lower precision, such as 8-bit or 4-bit integers. Pruning removes weights or entire attention heads whose contributions were minimal. In knowledge distillation, smaller student models are trained to imitate larger teacher models, retaining comparable levels of accuracy.

New hardware accelerators for transformer inference are becoming available in consumer hardware. Specialized neural processing units tuned for matrix multiplication and attention can run transformers efficiently on-device for speech recognition. This shift enables new use cases in offline applications and with privacy concerns over sending voice data to cloud servers. The economics of speech recognition depend on inference efficiency. Cost varies with traffic volume for cloud-based applications and hardware capabilities when running local applications.

7. Evaluation Frameworks and Expert Insights

7.1 Systematic Evaluation Methodology

Nonetheless, there remains a considerable lack of correspondence between benchmark performance and realistic conditions due to background noise, overlapping speech, domain-specific vocabulary and channel variability compared to clean evaluation sets. Practitioners should evaluate models on data representative of what they will see in production, rather than solely on published benchmarks. This requires considering tradeoffs around privacy, consent, and sampling, and generating data representative of diversity in production traffic. Word error rate remains the most important measure, but more detailed comparisons may also include latency, computational cost, and performance across demographics.

10.48047/jocaaa.2026.35.03.13

7.2 Production System Considerations

Conversely, pre-trained models provide high zero-shot performance, but may have lower accuracy on out-of-domain data or out-of-vocabulary words, such as certain keywords, or less common accents. Fine-tuning on in-domain data addresses this issue but may overfit small or unrepresentative datasets. Analyzing error patterns offers insights into where attention is needed: additional data or fine-tuning. Using methods such as shallow fusion with domain language models on top of fine-tuning can improve model performance beyond either method alone.

Other system components also affect the quality as perceived by the user, such as endpointing (determining whether the user has stopped speaking), punctuation, capitalization, inverse text normalization (conversational to written form), and system latency. To achieve high accuracy, all of these accessories require separate corpora, models, and optimization, often at a similar level of engineering effort as the recognition itself. Production speech recognition is a systems problem comprising the pipeline, not just the accuracy of the acoustic model.

8. Applications: Transforming Human-Computer Interaction

The current generation of voice assistants, driven by transformer-based recognition engines, serve billions of users on smart speakers, smartphones, cars and wearables. They work well in multiple acoustic environments and can generalize to new speakers through fine-grained tuning, such as accent adaptation. Research has shown user satisfaction and continuing use of the interface becomes negatively affected when recognition accuracy falls below 95 percent. Models trained on multiple languages have shown advantages for code-switching and multilingual utterance accuracy as there is greater acoustic-phonetic similarity amongst languages compared with most accent variation. Zero-shot and few-shot capabilities from scaling up pre-training benefit underrepresented languages [9].

Healthcare is one of the major application domains as physicians spend a large percentage of their time documenting care. Current systems can be integrated into the clinical documentation process. In fine-tuned medical models, the word error rate on clinical dialogue is below 5 percent, in contrast with 15-20 percent for prior systems. Medical models can be used in applications for which classification errors have safety-critical consequences for patients. Media and entertainment applications, such as automated captioning, subtitling, and large-scale indexing of media, are the most demanding applications of speech recognition inference, driving further work on compression and hardware efficiency.

Application Domain	Technology Integration	Functional Capability	Impact Area
Real-time Transcription	Low-latency streaming models	Live captioning with minimal delay	Educational accessibility and meeting documentation
Multilingual Interfaces	Cross-lingual transfer learning	Zero-shot and few-shot language support	Global device interaction and linguistic inclusivity
Accessibility Tools	Robust noise handling	Voice control and automated captioning	Independence for users with hearing or motor impairments
Multimodal Systems	Audio-visual fusion	Context-aware understanding with complementary signals	Enhanced robustness in challenging acoustic environments

Table 4: Application Domains and Capabilities [9, 10]

9. Broader Implications and Future Directions

9.1 Environmental and Economic Considerations

Large self-supervised speech models come with meaningful environmental impact due to the carbon emissions from the hardware and energy used for training. The state-of-the-art self-supervised speech

10.48047/jocaaa.2026.35.03.13

models require thousands of GPU hours to train. However, fine-tuning these models on downstream tasks, such as on a new language or domain, has a much lower carbon footprint compared to training from scratch. Other research focuses on efficient architectures and training, and distillation methods that reduce the environment cost, while retaining accuracy.

9.2 Rapid Base Model Improvements

These advancements in the base model raise important questions for the deployment of base models, as every base model generation seems to provide major improvements to capability. In particular, this could result in a fine-tuned model on an old base model becoming worse than a new base model with simple prompts or a few fine-tuning steps, making it wasteful to invest in fine-tuning. Organizations can weigh the costs of fine-tuning against the benefits of migrating to newer base models. Other mechanisms to reduce this risk include operationalization of the fine-tuning process, including automatic generation and annotation of training instances, reproducible pipelines for training and hyperparameter selection that quickly adapt to new base models, and evaluation pipelines for rapid iteration.

9.3 Social Impact and Accessibility

When transcription quality is high, speech recognition can help remove barriers and provide accessibility for the hearing impaired in conversations, meetings, and media. Higher accuracy can also be helpful for motor disabled users, for people speaking dialects or languages that have not been as well supported for training, and for those who have strong accents or speech impediments. There is a need for explicit, active auditing of accuracy across demographic groups, and organizations using these systems should evaluate their models across demographic groups as well as across acoustic conditions.

10. Conclusion

Transformer architectures, self-supervised learning, and end-to-end training represent a paradigm shift in speech recognition, moving from phoneme-based systems requiring extensive linguistic expertise to data-driven systems learning directly from examples. Self-supervised pre-training has reduced reliance on annotated data and democratized access to high-quality models for low-resource languages and domains. Attention-based transformers have replaced recurrent networks because they naturally capture long-term dependencies and train in parallel. Hybrid architectures like the Conformer combine local and global information extraction, matching the two-fold representation of speech as local spectral features and temporal dependencies.

End-to-end learning through CTC, attention-based decoders, and transducer architectures allows joint optimization toward transcription with reduced engineering requirements. Billions of users apply these methods in virtual assistants, transcription, and accessibility applications, demonstrating strong performance across acoustic conditions and languages. Streaming architectures enable real-time processing while model compression techniques support edge deployment. The gap between benchmark and production performance requires systematic evaluation and careful attention to auxiliary components, including endpoint detection, punctuation, and inverse text normalization.

Looking forward, multimodal models jointly processing speech and text, audio language models for generation and recognition, and integration with large language models will expand applications. The rapid pace of base model improvements necessitates robust infrastructure for continuous adaptation. Voice as a user interface will continue expanding as modeling advances, language coverage grows, context understanding deepens, and technology becomes universal across languages and acoustic environments. Through collective effort across researchers, practitioners, and policymakers, the transformative potential of accurate, accessible speech recognition can be realized while managing associated risks.

References

[1] Alec Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv>eess>arXiv:2212.04356, 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

10.48047/jocaaa.2026.35.03.13

[2] Grand View Research, "Voice And Speech Recognition Market (2024 - 2030)". Available: https://www.grandviewresearch.com/industry-analysis/voice-recognition-market
[3] Alexei Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," arXiv:2006.11477, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477
[4] Wei-Ning Hsu et al., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9585401
[5] Ashish Vaswani et al., "Attention is All You Need," Computation and Language, 2023. [Online]. Available: https://arxiv.org/abs/1706.03762
[6] Anmol Gulati et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," arXiv>cs>arXiv:2005.08100, 2020. [Online]. Available: https://arxiv.org/abs/2005.08100
[7] Alex Graves, "Sequence Transduction with Recurrent Neural Networks," Neural and Evolutionary Computing, 2012. [Online]. Available: https://arxiv.org/abs/1211.3711
[8] Jordan Hosier et al., "Lightweight domain adaptation: A filtering pipeline to improve accuracy of an Automatic Speech Recognition (ASR) engine," ACAI'21: 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence, 2021. [Online]. Available: https://dl.acm.org/doi/fullHtml/10.1145/3508546.3508641
[9] Arun Babu et al., "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," Interspeech 2022, 2022. [Online]. Available: https://www.isca-archive.org/interspeech_2022/babu22_interspeech.pdf
[10] Sanyuan Chen et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,"IEEE Journal of Selected Topics in Signal Processing, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9814838
[11] Alexei Baevski et al., "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language," 2022. [Online]. Available: https://arxiv.org/abs/2202.03555
[12] Jianbo Ma et al., "Low latency transformers for speech processing", 2023. [Online]. Available: https://www.researchgate.net/publication/368842358_Low_latency_transformers_for_speech_processing
[13] Shinji Watanabe et al., "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," IEEE Journal of Selected Topics in Signal Processing, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8068205
[14] Aaron van den Oord et al., "Representation Learning with Contrastive Predictive Coding," arXiv preprint arXiv:1807.03748, 2019. [Online]. Available: https://arxiv.org/abs/1807.03748
[15] William Chan et al., "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016. [Online]. Available: https://www.researchgate.net/publication/304372721_Listen_attend_and_spell_A_neural_network_for_large_vocabulary_conversational_speech_recognition