A comparative study on End-to-End and traditional Automatic Speech Recognition

Nancy Agarwal* and Sandeep Gupta†

*Galgotias University, Greater Noida, India

†Centre for Secure Information Technologies (CSIT), Queen's University Belfast, UK

***Abstract*—Automatic Speech Recognition (ASR) is playing an increasingly significant role in daily life, facilitating features such as voice-based search, virtual assistants, and other applications across diverse contexts. Recent advancements in deep learning have led to a paradigm shift in the design of ASR systems. End-to-end models, powered by integrated deep learning frameworks, are increasingly replacing the traditionally used Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) for speech-to-text conversion. This paper presents a study on E2E and traditional ASR, focusing on the strengths and weaknesses of E2E approaches. Our investigation covers key aspects of E2E ASR, including its integrated architecture, simplified training and inference, coherence optimization, reduced reliance on feature and statistical engineering, handling of long-distance dependencies, integration with generative AI, need for interpretability and non sequential alignment.**

***Index Terms*—Automatic speech recognition, HMM, GMM, E2E**

I. Introduction

Automatic speech recognition (ASR) is transforming human-computer interaction by enabling a wide range of functionalities, including command and control, dictation, transcription, audio search, and interactive spoken dialogues [1]. ASR technology has been adopted across various domains, such as automobiles, education, and healthcare. Figure 1 highlights applications such as call center agent assistance, automatic captioning, speaker diarization, and voice command systems that exploit ASR technology. Undeniably, ASR is becoming an integral part of our daily life, powering features like voice search and virtual assistants in diverse environments.

Advances in deep learning have led to a paradigm shift in ASR system design. End-to-end models based on deep learning frameworks are now the dominant approach, superseding the previously ubiquitous Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) methods for speech-to-text conversions [2]. Although HMM-GMM models have been proven valuable for modeling complex patterns in large-vocabulary speech recognition, E2E architectures offer a compelling alternative. By directly mapping raw audio to text, E2E models [3], [4], [5] eliminate the need for separate acoustic, pronunciation, and language models, resulting in a more unified and streamlined training process that aligns with sequence-to-sequence deep learning principles [6].

In this paper, we present a study on E2E and traditional ASR discussing the strengths and weaknesses of E2E ASR. E2E approaches can be beneficial in minimizing the expected word error rate, which can be specified as the primary objective of ASR systems [2]. Additionally, they can help reduce the time and memory complexity of the resulting decoder. In particular, the ability of transformer and transducer architectures to effectively model long-range dependencies in speech significantly enhances the accuracy of automatic speech recognition [4], [5]. In contrast to the surveys listed in Table I, our study aims to provide insights into aspects such as integrated architecture, simplified training and inference processes, coherence optimization, feature and statistical engineering, handling of long-distance dependencies, and the use of generative AI in E2E ASR.

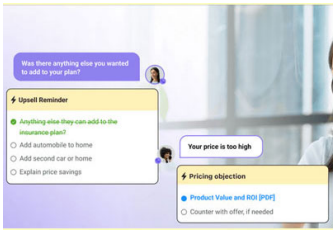
TABLE I
A COMPARISON WITH PREVIOUS SURVEYS PAPERS

Ref/Year	Focus Area
Present Review	Comparison of traditional (HMM-GMM) and E2E approaches for ASR.
Prabhavalkar et al. [2], 2023	Taxonomy of E2E ASR models discussing of their properties and relationship to classical HMM-based ASR architectures.
Bhardwaj et al. [7], 2022	Review children speech recognition systems in the wider field of ASR.
Fendji et al. [8], 2022	Classification of the acoustic-phonetic-, pattern recognition-, and artificial intelligence approaches.
Malik et al. [9], 2021	Examine the components of an ASR system, spanning from feature extraction to language modeling.
Wang et al. [10], 2019	Advantages and disadvantages of HMM-based model and end-to-end models.
Deng et al. [11], 2013	Machine learning techniques, including supervised, semi-supervised, unsupervised, and active learning, have been used in building recognition systems.

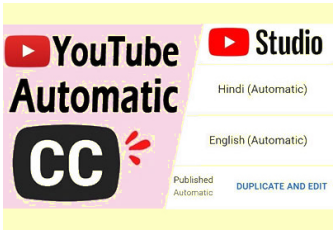
The rest of the paper is organized as follows: Section II establishes the theoretical background of speech recognition. Section III and IV detail the architectures of traditional HMM-GMM and modern end-to-end deep learning-based systems, respectively. Section V provides a comparative study of these approaches. Finally, Section VI concludes the paper.

II. Background

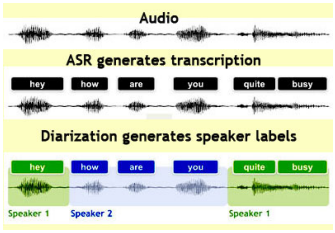
This section provides background on speech recognition theory and popular speech recognition models.



(a) Call center agent assist for live call guidance



(b) Automatic captioning to automatically create captions for videos/broadcasting



(c) Diarization generating speaker-wise transcription.

(d) virtual assistant for smart devices.

Fig. 1. Illustration of ASR applications across a wide range of industries. Credit: Google Images

A. Speech recognition

Speech recognition involves understanding and interpreting human speech to convert it into machine-readable text [12]. It can also be specified as automatic speech recognition, computer speech recognition or speech-to-text. It is essential to note the technical distinctions between speech recognition and voice recognition systems. As illustrated in Figure 2, speech recognition focuses on converting spoken language into text, whereas voice recognition is concerned with authenticate an individual by analyzing the unique characteristics of the voice such as pitch, accent and frequency [13].

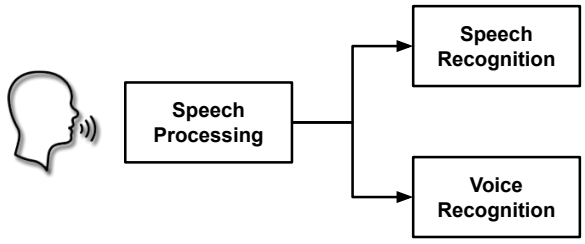


Fig. 2. Illustration of speech processing for speech recognition and voice recognition

B. Speech recognition theory

A speech recognition system converts audio signals into text by dividing the audio into fixed-length segments (L). The ASR system aims to identify the most probable word sequence, $W = w_1, ..., w_n$, corresponding to an acoustic input sequence, $O = o_1, ..., o_T$, where T is the number of frames in the utterance [14]. For a given set of observed acoustic features (O), the speech recognition problem can be expressed as the word sequence (W) that is most likely caused by acoustic features (O) as shown in Equation (1).

$$\widehat{W} = argmax\ P(W|O) \quad for \quad W \in L \tag{1}$$

Further, given a sequence of words (W) and the corresponding acoustic signal frames, where $P(W)$ represents the prior probability of the word sequence, the probability $P(W|O)$ can be evaluated using Bayes' Theorem, as shown in Equation (2).

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \tag{2}$$

Considering $P(O)$ is the same for each candidate sequence (W), Equation (2) can be rewritten as Equation (3). $P(O|W)$ represents the likelihood of observing acoustic sequence O given that the speaker uttered word sequence W, primarily determined by an HMM.

$$W = argmax \frac{P(O|W)P(W)}{P(O)} = P(O|W)P(W) \quad \forall \quad W \in L \tag{3}$$

C. Popular speech recognition models

- Hidden Markov Models (HMMs) are built on a Markov process with hidden states [15]. A Markov model is characterized by the memory-less property, i.e., the transition from one state to another depends only on the current state. HMMs are used as sequence models in speech recognition, labeling each unit (words, syllables, or sentences) in a sequence [16]. This mapping between labels and input allows the system to identify the most likely sequence of labels.
- Gaussian Mixture Model (GMM) is a statistical method used to estimate spectral parameters from the speech spectrum. GMMs are widely used for acoustic modeling in ASR, particularly when applied to Fourier spectrum-based speech features [17]. They can also be adapted using techniques like Maximum A Posteriori (MAP), Maximum Likelihood Linear Regression (MLLR), and Feature Space MLLR (fMLLR) [18].
- N-grams are probabilistic language models that estimate the probability of a word based on the preceding N-1 words [19]. These models can effectively identify words in noisy, ambiguous, and varied speech input.
- Dynamic Time Warping (DTW) calculates the optimal alignment between two sequences, such as an audio signal and a reference pattern, even when they have varying time scales [20]. It is commonly used to align speech signals with phoneme sequences.
- Recurrent Neural Networks (RNNs) consist of convolutional layers for feature extraction, multiple recurrent layers to capture sequential data, and a fully connected dense layer for predicting the target output [21]. Their ability to model temporal dependencies between speech frames makes them well-suited for speech recognition tasks.

- Long Short-Term Memory (LSTM) networks are a specialized type of RNNs that have demonstrated state-of-the-art performance in sequence learning tasks [22]. LSTMs address the vanishing gradient problem through the use of multiplicative gating mechanisms. These networks utilize memory cells, which are regulated by three key gates: forget, input, and output gates. Gated Recurrent Units (GRUs) are a simplified and computationally more efficient variant of LSTM networks [23].
- Transformers are neural networks specifically designed for sequence-to-sequence modeling, making them highly effective for a wide range of natural language processing tasks, including translation, summarization, and question answering [24]. Unlike RNNs, which process information sequentially, Transformers process information globally across the entire input sequence. This eliminates the vanishing information problem inherent in RNNs and enables the use of parallel computation to accelerate training. Moreover, Transformers leverage an attention mechanism to identify and focus on relevant information within the input sequences.

III. Traditional Automatic Speech Recognition System

In this section, we discuss traditional HMM-GMM-based speech recognition models. They typically involves four key components: acoustic feature extraction, acoustic modeling, language modeling, and search based on Bayes' decision rule [2]. These systems have historically used phonetic lexicons, e.g., CMU dictionary, to map words to phoneme sequences, serving as a crucial bridge between acoustic signals and linguistic units [25].

A. Traditional ASR architecture

As shown in Figure 3, traditional ASR comprises a feature extractor, statistical models, and a decoder. The decoder utilizes acoustic and language models to perform statistical inference and produce a sequence of words.

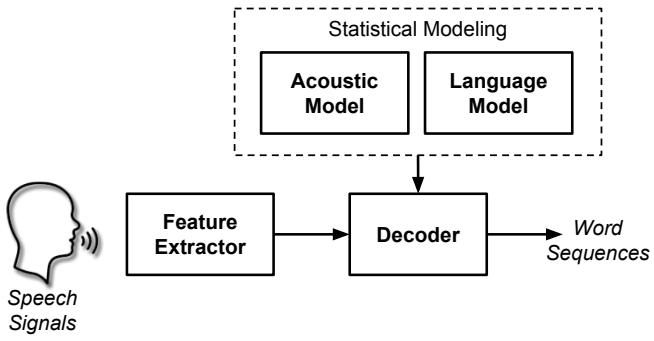


Fig. 3. Illustration of a traditional ASR architecture

1) Feature extractor: The feature extractor transforms the input audio waveform into a sequence of fixed-size audio vectors using techniques such as Principal Component Analysis (PCA), Filter-Bank Analysis, Linear Discriminant Analysis (LDA), or Kernel-Based Feature Extraction [14]. Studies have highlighted that the Mel-Frequency Cepstral Coefficient (MFCC) is the most commonly used feature representation technique [26].

Figure 4 illustrates the MFCCs computation process [27]. The frequency axis is scaled to the non-linear Mel scale (using triangular overlapping windows) after applying the Fourier transform to a window of the voice signal. Next, a Discrete Cosine Transform (DCT) is performed on the log of the power spectrum of each Mel band. The MFCCs are the amplitudes of the resulting spectrum, which forms a $2-D$ vector of size $13 \times variable\ length$ (the length of the vector depends on the duration of the voice signal).

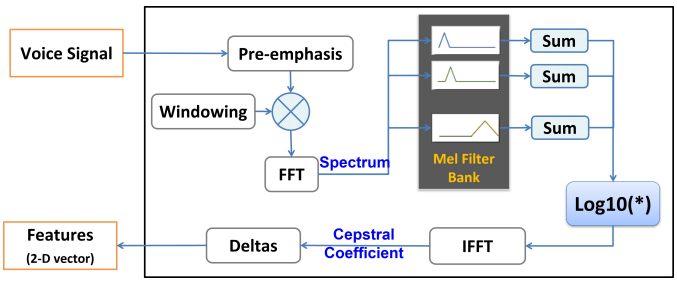


Fig. 4. Illustration of MFCC computation process.

2) Statistical Modeling: Statistical models for speech recognition combine language and acoustic models [14]. A language model, denoted as $P(W)$, estimates the prior probability of that word sequence W, reflecting its linguistic plausibility. An acoustic model, denoted as $P(O|W)$, calculates the likelihood of observing a sequence of acoustic features (O) given a hypothesized word sequence (W).

- *Language model*: A language model is a collection of constraints applied to determine the acceptability of sequences of words. For example, the sentence, *I am best* is grammatically acceptable, while *I are best* is not. These models are typically trained using N-gram probabilities, which compute the likelihood of grouping words in a sequence [19].
- *Acoustic model*: HMM is one of the most commonly used approaches to build acoustic models [16]. A database is used to store the statistical representations of phonemes that make up words. Each phoneme is associated with its own HMM based on three probabilities (initial state probabilities, transition probabilities, and emission probabilities). These three probability-based parameters are further trained using baum-welch algorithm. The acoustic model evaluates the probability of a specific acoustic feature given a base phoneme. The traditional ASR system utilizes the HMM-GMM framework to represent the sequential structure of audio signals [28]. In this framework, a GMM models the distribution of acoustic features associated with a phoneme, which is a distinct speech sound. Words are composed of sequences of phonemes. For example, the English word "spin" consists of four phonemes: [s], [p], [i], and [n]. HMM is used to build a statistical model for evaluating the likelihood of transitions between phonemes and their corresponding observed acoustic signals. However, in HMM-GMM settings,

HMM utilizes GMM to model the spectral representation of the sound wave.

3) Decoder: The task of the decoder component is to processes acoustic feature vectors derived from the input audio by the feature extractor and determine the optimal word sequence based on acoustic and language models. The Viterbi algorithm is mostly used to yield the best sequence of words, for the given observation sequence.

B. Traditional ASR systems

Zhao et al. [29] propose a regularization approach for parameter estimation in GMMs by enabling simultaneous clustering and subspace identification for each cluster. The approach assumes that the components of the mixture (or clusters) reside in low-dimensional subspaces. This assumption is motivated partly by empirical evidence from various applications that support the use of both global and local dimensionality reduction techniques, and partly by studies analyzing the structure of high-dimensional data. Lu et al. [30] investigate cross-lingual acoustic modeling using SGMMs and demontrate that the global parameters of these models are transferable across languages, especially when trained on multilingual data. SGMMs factorize acoustic model parameters into global set, i.e., common to all the states of a HMM, and state-specific sets, i.e., unique to individual HMM states. Cross-lingual SGMM acoustic models can significantly improve accuracy by leveraging a large proportion of parameters estimated from source language data in contrast to conventional speech recognizers.

Povey et al. [28] propose an SGMM for acoustic modeling. The HMM states share a common structure, but the means and mixture weights are allowed to vary within a subspace of the full parameter space, which is controlled by a global mapping from a vector space to the space of GMM parameters. Sun and Chol [18] propose an adaptive continuous space language modeling approach that combines the long-term context of recurrent neural network (RNN) with the adaptability of SGMM. The authors obtain word clusters using top-down or bottom-up clustering approaches that exhibit similarities to traditional word classes employed in class-based language models. These models are extensively studied to address the data sparsity limitation. Reducing the vocabulary to a smaller set of word classes significantly reduces the number of parameters to estimate. This approach mitigates the data sparsity problem and leads to the development of more compact language models.

IV. End-to-end automatic speech recognition system

This section discusses end-to-end automatic speech recognition (E2E-based ASR), which leverages deep learning frameworks to address the inherent challenges of sequential structure and latent alignment in speech recognition [2]. In contrast to traditional ASR systems, which rely on separately optimized acoustic and language models, E2E architectures offer a unified, streamlined approach, directly mapping raw audio input to text output. This approach aligns with the principles of sequence-to-sequence deep learning simplifying the training process. Figure 5 illustrates the transformer and transducer architectures, which are prevalent in E2E methods [4], [5]. These architectures are particularly effective at capturing long-range dependencies in speech, leading to more accurate transcriptions.

A. Transformer-based E2E ASR architecture

As shown in Figure 5a, transformer-based ASR consists of blocks of encoders and decoders [31]. The encoder and decoder models correspond to the acoustic model and language model, respectively, in a conventional system. The feature vectors derived from the audio signals, $X = (x_1, ..., x_T)$, are fed as input to the encoder, which generates a sequence of intermediate representations. The decoder takes these dense representations as input and outputs the text (W) as a sequence of words, $W = w_m = (w_1, ..., w_M)$.

The two components use several layers of self-attention, which are interconnected to identify dependencies and relevant information within a sequence [5]. The attention mechanism evaluates how much focus each token or audio frame should give to other subsequential parts of the sequence. Furthermore, it enables the processing of different parts of the input sequence in parallel, making transformers a highly efficient methodology.

B. Transducer-based E2E ASR architecture

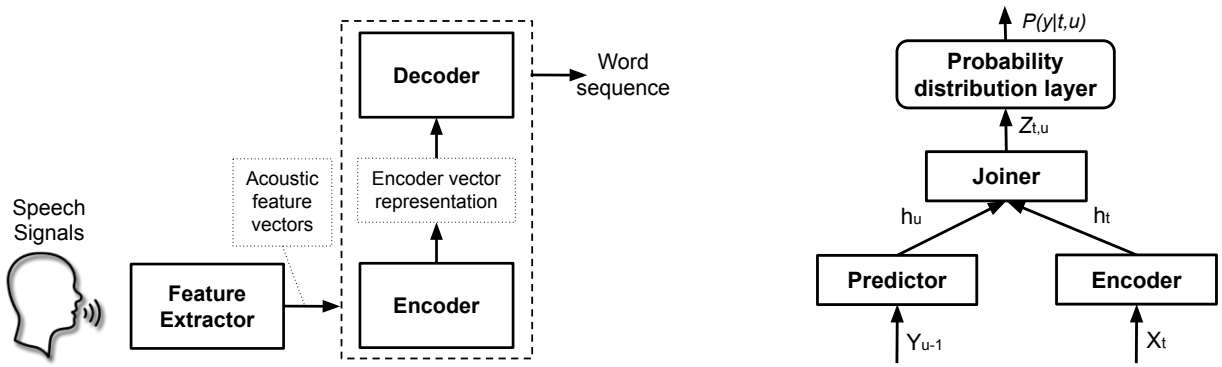
The transducer is commonly used for developing streaming encoder-decoder frameworks as it overcomes the limitations of the transformer approach by enabling the decoder to generate output as soon as the first input is encoded, without waiting for the entire sequence to be available [32]. The recurrent neural network transducer (RNN-T) is widely regarded as the de facto strategy for designing streaming recognition architectures [5]

Given an input sequence of acoustic features of length T, $X = (x_1, ..., x_T)$, the RNN-T model aims to predict the text sequence, $y = (y_1, ..., y_U)$, of length U. It comprises three components: encoder, predictor, and joiner, as illustrated in Figure 5b.

- The encoder component receives speech segments as input and generates high-level representations as output. x_t denotes the acoustic feature at time step t and h_t represents its high-level representation, which can computed using $h_t = f^{encoder}(x_t)$
- The predictor component is an autoregressive model that receives the previous output, $y_{(u-1)}$ emitted at time step $(t-1)$ and generates the feature vector to be used for identifying the next output computed using $h_u = f^{predictor}(y_{(u-1)})$
- Lastly, the joiner component is a type of feed-forward network that fuses the predictor and encoder output to generate the text sequence which is monotonically aligned with speech sequence as $z_{(t,u)} = f^{joiner}(h_t, h_u)$

C. E2E deep learning approaches

Connectionist temporal classification (CTC), the RNN transducer (RNN-T), the recurrent neural aligner (RNA), and the hybrid auto-regressive transducer (HAT) are some of the



(a) Transformer-based ASR architecture

(b) Transducer-based ASR architecture

Fig. 5. Illustration of prominent E2E methods for ASR

early E2E modeling approaches. Li et al. [32] propose a framework for combining hybrid systems and attention-based encoder–decoder (AED) models using N-best and lattice re-scoring. The authors describe that hybrid systems offer advantages in handling streaming data and integrating structured knowledge like lexicons, whereas AED models can be modified to enable streaming ASR. Wang et al. [10] describe the end-to-end model as one that directly maps audio to characters or words using a single model, replacing the traditional engineering process with a learning-based approach. This eliminates the need for domain expertise, making the end-to-end model simpler to construct and train and suitable for large vocabulary continuous speech recognition (LVCSR).

Tomashenko and Esteve [33] investigate speaker adaptation techniques for bidirectional long short-term memory (BLSTM) RNN-based AMs trained using the CTC objective function. BLSTM-CTC AMs are fundamental components of E2E ASR systems, enabling accurate speech-to-text transcription. The paper describes three different feature-space adaptation approaches for CTC AMs, i.e., feature-space maximum linear regression, i-vector based adaptation, and maximum a posteriori adaptation using GMM-derived features. Li and Vu [34] propose a method that combines CycleGAN and inter-domain losses for semi-supervised E2E ASR. The combination of inter-domain loss and CycleGAN enables a more effective shared latent space for unpaired speech and text data, thereby enhancing the accuracy of speech-to-text mapping. Wang et al. [31] propose transformer-based acoustic models (AMs) for hybrid ASR that use self-attention to replace the RNNs in the acoustic encoder.

Borgstrom et al. [35] propose a multiscale autoencoder (MSAE) for mask-based E2E neural network speech enhancement framework. The framework provides the encoder and decoder mappings required by mask-based approaches. The autoencoder is a crucial component of mask-based end-to-end (E2E) enhancement networks, mapping input signals to an embedding space that effectively separates speech and noise components, while the decoder synthesizes the output waveform. The MSAE decomposes an input waveform into separate band-limited branches, each operating at a distinct rate and scale, to extract a sequence of multiscale embeddings. Smit et al. [36] propose subword modeling for weighted finite-state transducer (WFST)-based hybrid DNN-HMM speech recognition, using graphemes instead of phonemes. The subword modeling is independent from the acoustic model. The acoustic model, trainable on phoneme or grapheme sequences, can be integrated with diverse language models, including character-, subword-, or word-based systems.

Hadian et al. [37] propose end-to-end HMM-based ASR system that eliminates the need for traditional techniques like GMMs and decision trees. The proposed acoustic model is fully neural and is trained in a flat-start manner in a single stage using lattice-free Maximum Mutual Information (LF-MMI). Notably, the approach eliminates the need for initial alignments, pre-training, prior estimation, or transition training. Tanaka et al. [38] propose a joint end-to-end and DNN-HMM hybrid ASR system that shares the network to transfer knowledge between the systems. The continuous vectors extracted from the DNN acoustic model are utilized as auxiliary features to enhance the end-to-end ASR system. These vectors, which encapsulate knowledge of phoneme estimation, contribute to improved ASR accuracy.

V. A COMPARISON BETWEEN TRADITIONAL AND E2E ASRS

This section provides a detailed analysis of the strengths and weaknesses of E2E methodologies compared to HMM-GMM frameworks. Table II presents a comparison of key attributes between traditional and E2E ASRs.

A. Strength of E2E

1) Integrated Architecture: A conventional ASR consists of various models including language, lexicon and acoustic models designed using probabilistic theory. Compared to traditional approach, E2E ASR simplifies the above factorized structure into a single integrated network architecture using deep learning framework. E2E approaches directly map an acoustic feature sequence to a sequence of characters or even words. In this framework, Encoder and predictor/decoder components

TABLE II
COMPARISON OF TRADITIONAL AND E2E ASRS

Attributes	Traditional ASR	E2E ASR
Architecture	Each component (HMM, GMM, Language Model, Acoustic Model) requires independent training.	A single encoder-decoder model is trained end-to-end.
Dataset	Performs well on small datasets but relies on manually created phoneme dictionaries.	Suitable for large-scale labeled datasets (e.g., LibriSpeech).
Noise tolerance	Leveraging phoneme-level modeling and statistical smoothing, the system achieves strong robustness.	Although sensitive to noise, this can be effectively addressed through data augmentation.
Interpretability	Enhanced interpretability is a key benefit. This architecture allows for granular analysis of individual model components, enabling precise error tracing to specific modules like the Acoustic Model or Language Model, facilitating targeted improvements.	The black-box nature of E2E models results in a lack of transparency and hinders error resolution.
Computational requirements	Higher efficiency for low-resource devices due to the use of statistical modeling.	End-to-end ASR models demand significant computational resources, typically requiring GPUs or TPUs.
Context Awareness	Capturing long-range dependencies can be challenging.	Transformers can overcome the challenge of capturing long-range dependencies.

are analogus to acoustic and language model of traditional ASR [39].

2) Simplified training and inferencing: E2E being an integrated framework is free of the requirements of building lexicon, text normalization, finite state transducers, or any conditional independence assumptions. Training such models and inferencing the output sequence are much simplified than conventional systems [40]. The learning and decoding time is also reduced in this integrated architecture. Furthermore, it is observed that despite such a simpler implementation, they also outperform the traditional ones.

3) Coherence Optimization: The underlying architecture of E2E allows joint optimization of encoder and decoder components. In contrast, the different components in conventional ASR (i.e., acoustic, lexicon and language component) are optimized individually with separate objectives. It may lead to in incoherence in optimization, where one model is not trained according to the context of another model [R15].

4) No Feature Engineering: An HMM model requires a set of parameters, $[T, O, \pi]$. T matrix holds the transition probabilities, O matrix holds the observation probabilities and π corresponds to initial state probability. Designing effective matrices may require domain expertise. Also, to integrate acoustic and language models, a lexicon model is required which is usually a handcrafted pronunciation dictionary to map word to sequence of phonemes. E2E speech recognition models follow deep learning methodologies where feature engineering is not required.

5) No statistical Assumptions: The role of GMM in conventional speech recognition system is to model the probability distribution of phonemes in order to identify the most likely phoneme given an input signal. However, GMM relies on underlying assumption that data is generated from a mixture of normal distributions. If the input features are from non-gaussian distribution, which is usually in the case of speech recognition, its performance may fail. Similarly, HMM leans on conditional independence assumptions, i.e., Markov assumption where each observation only depends on the immediate hidden state. In real scenario, sequence of words do have longer dependencies rather than just one previous state. Using deep learning, no assumptions about the statistical distributions are made and therefore, E2E is a suitable candidate for modeling complex distributions.

6) Leverage LLM potential: Since deep learning scaffolds the foundation of E2E, these systems can incorporate large language models (LLM) for better understanding the natural language [41]. The role of LLMs in speech-to-text conversion lies in their potential to learn complex structures, context and semantics of a language, thereby enhancing the overall performance of speech recognition.

7) Capture long-distance dependencies: E2E models include attention and RNN-based mechanisms which are known for capturing long-distance relationships. It allows the systems to not only understand the individual words, but also learn the context of entire sentences or even a complete paragraph. On the other hand, HMM can capture the limited dependencies as it has a fixed-size context window. In ordinary HMM, the probability of the current word depends only on the previous adjacent word. However, high-order HMMs allow the consideration of dependencies over a window of previous words. Nonetheless, the size of this window is very limited because increasing the window size significantly increases the complexity of the model.Therefore, HMM fails to capture the context and dependency that is outside the window of words.

B. Weaknesses of E2E

1) Interpretability: The conventional HMM-GMM framework is more interpretable than E2E. The parameters of GMM (i.e., mean, covariances and weights) have a clear interpretation which highly assists in understanding the underlying structure and distribution of data. HMMs also provide a transparent probabilistic structure of HMMs also make it easier to interpret the predictions and behavior of the system. On the other hand, E2E systems are based on deep learning models which are

known for their black-box nature because of their complex non-linear hierarchical layers. The lack of interpretability of E2E models raises the trust issue in the predicted output. Interpretable models offer the features to know the reasons of their decisions and improve the performance [42], [43].

2) Non Sequential Alignment: The transformer-based E2E supports non sequential alignment. In this alignment, the mapping between the output and input tokens is not strictly or linear ordered. This alignment is beneficial for the applications such as translator where the grammatical structure of the two languages is different. But in the case of speech recognition, there is alignment between input audio signals and the sequence of words. Therefore, monotonic alignment becomes a necessary task to ensure that translated words appear in same order as they are spoken in speech signal. The conventional HMM-GMM supports sequential alignment framework and thereby, is still adopted in designing speech recognition. However, there are studies which attempt to address the alignment problem in E2E [44].

VI. Conclusions

A significant paradigm shift in ASR system design is evident, as the widely adopted HMM and GMM approaches for speech-to-text conversion have been replaced by E2E models. This study delved into a comparative analysis of E2E and traditional ASR, focusing on the advantages offered by E2E. In contrast to traditional ASR, E2E ASR systems streamline the process by directly mapping audio to text, eliminating the need for separate acoustic and language model optimization. Notably, E2E models enable direct optimization for the primary objective of ASR, minimizing word error rate, while also significantly reducing the time and memory footprint of the decoding process. Researchers and industry professionals can leverage our findings to optimize the performance and applicability of E2E models across various industrial sectors. In the future, we plan to evaluate the security aspects of E2E ASR systems.

References

[1] M. Gales, S. Young, *et al.*, "The application of hidden markov models in speech recognition," *Foundations and Trends® in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[2] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[3] X. Chang, S. Watanabe, M. Delcroix, T. Ochiai, W. Zhang, and Y. Qian, "Module-based end-to-end distant speech processing: A case study of far-field automatic speech recognition [special issue on model-based and data-driven audio signal processing]," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 39–50, 2025.

[4] X. Gong, Y. Wu, J. Li, S. Liu, R. Zhao, X. Chen, and Y. Qian, "Advanced long-content speech recognition with factorized neural transducer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[5] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7829–7833, IEEE, 2020.

[6] S. Gupta and B. Crispo, "Towards autonomous device protection using behavioural profiling and generative artificial intelligence," *IET Cyber-Physical Systems: Theory & Applications*, 2024.

[7] V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam, "Automatic speech recognition (asr) systems for children: A systematic literature review," *Applied Sciences*, vol. 12, no. 9, p. 4419, 2022.

[8] J. L. K. E. Fendji, D. C. Tala, B. O. Yenke, and M. Atemkeng, "Automatic speech recognition using limited vocabulary: A survey," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2095039, 2022.

[9] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, pp. 9411–9457, 2021.

[10] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, 2019.

[11] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060–1089, 2013.

[12] IBM, "What is speech recognition?." https://www.ibm.com/topics/speech-recognition, *Accessed on 10-12-2024*. online web resource.

[13] S. Gupta, C. Maple, B. Crispo, K. Raja, A. Yautsiukhin, and F. Martinelli, "A survey of human-computer interaction (hci) & natural habits-based behavioural biometric modalities for user recognition schemes," *Pattern Recognition*, vol. 139, p. 109453, 2023.

[14] S. Karpagavalli and E. Chandra, "A review on automatic speech recognition architecture and approaches," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 4, pp. 393–404, 2016.

[15] D. Jurafsky and J. H. Martin, "Hidden markov models," *Speech and Language Processing*, pp. 1–20, 2017.

[16] B. Mor, S. Garhwal, and A. Kumar, "A systematic review of hidden markov models and their applications," *Archives of computational methods in engineering*, vol. 28, pp. 1429–1448, 2021.

[17] D. Yu, L. Deng, D. Yu, and L. Deng, *Gaussian mixture models*, pp. 13–21. Springer, 2015.

[18] R. H. Sun and R. J. Chol, "Subspace gaussian mixture based language modeling for large vocabulary continuous speech recognition," *Speech Communication*, vol. 117, pp. 21–27, 2020.

[19] S. Avasthi, R. Chauhan, and D. P. Acharjya, "Processing large text corpus using n-gram language modeling and smoothing," in *Proceedings of the Second International Conference on Information Management and Machine Intelligence*, pp. 21–32, Springer, 2021.

[20] Y. Permanasari, E. H. Harahap, and E. P. Ali, "Speech recognition using dynamic time warping (dtw)," in *Proceedings of the Journal of physics: Conference series*, vol. 1366, pp. 1–5, IOP Publishing, 2019.

[21] P. B. Atosha, E. Özbilge, and Y. Kırsal, "Comparative analysis of deep recurrent neural networks for speech recognition," in *Proceedings of the 32nd Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2024.

[22] S. Maleki, S. Maleki, and N. R. Jennings, "Unsupervised anomaly detection with lstm autoencoders using statistical data-filtering," *Applied Soft Computing*, vol. 108, p. 107443, 2021.

[23] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.

[24] A. Pölz, A. P. Blaschke, J. Komma, A. H. Farnleitner, and J. Derx, "Transformer versus lstm: A comparison of deep learning models for karst spring discharge forecasting," *Water Resources Research*, vol. 60, no. 4, p. e2022WR032602, 2024.

[25] D. Le, X. Zhang, W. Zheng, C. Fügen, G. Zweig, and M. L. Seltzer, "From senones to chenones: Tied context-dependent graphemes for hybrid speech recognition," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 457–464, IEEE, 2019.

[26] M. Baelde, C. Biernacki, and R. Greff, "Real-time monophonic and polyphonic audio classification from power spectra," *Pattern Recognition*, vol. 92, pp. 82–92, 2019.

[27] S. Gupta and B. Crispo, "Usable identity and access management schemes for smart cities," in *Proceedings of the Collaborative Approaches for Cyber Security in Cyber-Physical Systems*, pp. 47–61, Springer, 2023.

[28] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, *et al.*, "The subspace gaussian mixture model—a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[29] Y. Zhao, A. K. Shrivastava, and K. L. Tsui, "Regularized gaussian mixture model for high-dimensional clustering," *IEEE transactions on cybernetics*, vol. 49, no. 10, pp. 3677–3688, 2018.

[30] L. Lu, A. Ghoshal, and S. Renals, "Cross-lingual subspace gaussian mixture models for low-resource speech recognition," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 1, pp. 17–27, 2013.

[31] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang, *et al.*, "Transformer-based acoustic modeling for hybrid speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6874–6878, IEEE, 2020.

[32] Q. Li, C. Zhang, and P. C. Woodland, "Combining hybrid dnn-hmm asr systems with attention-based models using lattice rescoring," *Speech Communication*, vol. 147, pp. 12–21, 2023.

[33] N. Tomashenko and Y. Estève, "Evaluation of feature-space speaker adaptation for end-to-end acoustic models," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1–8, 2018.

[34] C.-Y. Li and N. T. Vu, "Improving semi-supervised end-to-end automatic speech recognition using cyclegan and inter-domain losses," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pp. 822–829, IEEE, 2023.

[35] B. J. Borgström and M. S. Brandstein, "A multiscale autoencoder (msae) framework for end-to-end neural network speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[36] P. Smit, S. Virpioja, and M. Kurimo, "Advances in subword-based hmm-dnn speech recognition across languages," *Computer Speech & Language*, vol. 66, p. 101158, 2021.

[37] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi.," in *Proceedings of the Interspeech*, pp. 12–16, 2018.

[38] T. Tanaka, R. Masumura, T. Moriya, T. Oba, and Y. Aono, "A joint end-to-end and dnn-hmm hybrid automatic speech recognition system with transferring sharable knowledge.," in *Proceedings of the INTERSPEECH*, pp. 2210–2214, 2019.

[39] G. Saon, Z. Tüske, D. Bolanos, and B. Kingsbury, "Advancing rnn transducer technology for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5654–5658, IEEE, 2021.

[40] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4774–4778, IEEE, 2018.

[41] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli, *et al.*, "Prompting large language models with speech recognition abilities," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13351–13355, IEEE, 2024.

[42] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised automatic speech recognition: A review," *Speech Communication*, vol. 139, pp. 76–91, 2022.

[43] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond," *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3197–3234, 2022.

[44] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.