ASR systems using LLMs a Review

Sarah Alhumoud
sohumoud@imamu.edu.sa
College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

Abstract
Large Language Models (LLMs) have significantly advanced Automatic Speech Recognition (ASR) by improving transcription accuracy, handling diverse linguistic contexts, and enabling cross-lingual and low-resource applications. This paper reviews the integration of LLMs into ASR systems, analyzing 19 research papers and 24 datasets. The aim is to examine key methodologies, including fine-tuning, transfer learning, and prompt engineering, and highlight their impact on phoneme recognition and contextual understanding. The datasets reviewed span diverse languages, tasks, and domains, reflecting the growing emphasis on creating inclusive ASR systems. This review also provides an overview of the main ASR architecture with its main 4 modules, to provide a concise synthesis of current advancements, identify existing limitations, and suggest future research directions to enhance the robustness, efficiency, and accessibility of ASR systems powered by LLMs.

Keywords
Spam review detection, CNN-LSTM, CNN-RNN, CNN-GRU, Big data, Deep Learning, Amazon Product Review Dataset

I. Introduction

The Turing test, introduced in 1950, challenged programmers to create computers capable of human-like conversation. Despite numerous attempts over the years, this goal remained elusive until the recent development of chatbots powered by large language models (LLMs), such as ChatGPT. The desire to enable computers to converse with humans extends beyond text-based interactions, encompassing voice communication through the integration of automatic speech recognition (ASR) technology with LLMs. LLMs represented a significant advancement in artificial intelligence, following the developments in machine learning and deep learning. The convergence of LLMs and ASR technology has facilitated substantial progress in the field of artificial intelligence, augmenting capabilities in diverse applications such as voice assistants, automated transcription services, and real-time translation tools.

This review aims to explore the progress in the field of ASR systems based on LLMs, highlighting key algorithms, and datasets, in different tasks. The next section presents the review of the different studies, following it is the ASR archetechterue overview, finally, a review on the datasets used in the studies with a brief description on each.

II. Literature Review

This review investigates 19 study for the automatic speech recognition (ASR) task using LLMs. Those studies focus on different tasks like automatic speech translation (AST), transliteration, Speech emotion recognition (SER), and serval models to enhance ASR and text-to-speech TTS, and Speech-to-text tasks. The studies highlight a notable shift towards end-to-end architectures as opposed to cascading schemes, and the adoption of multimodal datasets, alongside challenges related to computational costs, resource requirements, and scalability. As depicted in Figure 1, research in this domain is predominantly conducted by leading technology companies rather than academic institutions. This inclination can be partially attributed to the extensive in-house datasets and significant computational resources available within these organizations. Table I, present a summarization of the different studies reviewed in this paper.

Yu *et al.* [1] presents a comparative study of three commonly used structures as ASR connectors, fully connected layers, multi-head cross-attention, and Q-Former. Experiments were conducted utilizing the LibriSpeech, Common Voice, and GigaSpeech datasets. LLMs based on Q-Former demonstrated substantial generalization

Manuscript received January 5, 2025
Manuscript revised January 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.1.1

capabilities when applied to out-of-domain datasets, achieving a 12% relative reduction in word error rate (WER) compared to the Whisper baseline ASR model on the Eval2000 test set, all without incorporating any in-domain training data from the Switchboard dataset. Furthermore, a novel segment-level Q-Former has been introduced, allowing LLMs to effectively recognize speech segments that exceed the encoder's duration limitations.

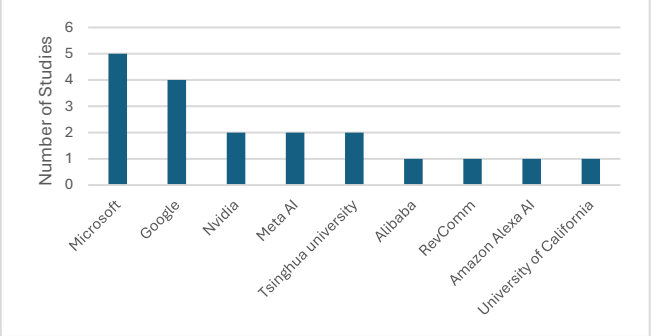

Fig. I. Studies affiliation

This led to a 17% relative improvement in WER over other connector architectures for 90-second speech samples.

Ma et al.[2] investigates the enhancement of speech emotion recognition (SER) through speech pre-trained model (PTM) data2vec, the text generation model GPT-4, and the Azure text-to-speech (TTS) system. They used Azure Emotional TTS for synthesized data and multiple data augmentation strategies, such as random mixing, adversarial training, transfer learning, and curriculum learning, to enhance SER performance using synthetic speech. For evaluation, they used weighted accuracy (WA) and unweighted accuracy (UA), WA corresponds to the overall accuracy while UA corresponds to the average class-wise accuracy. Experiments confirmed the method's effectiveness compared to alternative data augmentation approaches and synthetic datasets, with slight improvements on both.

Park et al. [3] presented the multi-modal decoding process probabilistically and performed joint acoustic and lexical beam searches to incorporate cues from both modalities: audio and text. Experiments demonstrate that infusing lexical knowledge from the LLM into an acoustics-only diarization system improves the overall speaker-attributed word error rate (SA-WER). Showing up to 39.8% relative delta-SA-WER improvement from the baseline system.

Fathullah et al. [4] extend the capabilities of LLMs by prepending a sequence of audio embeddings to the text token embeddings. Where the LLM can be converted to an ASR system and used in the exact same manner as its textual counterpart. Experiments on Multilingual LibriSpeech (MLS) with 8 languages, show that incorporating a conformer encoder into the open-sourced LLaMA-7B allows it to outperform monolingual baselines by 18% relatively in WER and perform multilingual speech recognition, despite LLaMA being trained on English text.

Wu et al. [5] introduced Speech-LLaMA, incorporating acoustic information into text-based large language models. The method leverages Connectionist Temporal Classification and a simple audio encoder to map the compressed acoustic features to the continuous semantic space of the LLM. Experiments done on 13 language speech-to-text translation tasks demonstrated a significant improvement over strong baselines, highlighting the potential advantages of decoder-only models for speech-to-text conversion. It is worth noting that the languages translated include the Arabic language.

Li et al. [6] proposed two zero-shot ASR domain adaptation methods using LLaMA-7B. That is, prompting LLMs for domain adaptation without the need to re-train. Experiments show that, with only one domain prompt, both methods can effectively reduce word error rates on out-of-domain TedLium-2 and SPGISpeech datasets.

Chen et al. [7] studied the ASR of YouTube videos using LLMs. They demonstrate up to 8%

relative reduction in Word Error Rate on US English (en-us) and code-switched Indian English (en-in) long-form ASR test sets and a reduction of up to 30% relative on Salient Term Error Rate (STER) over a strong first-pass baseline that uses a maximum-entropy based language model.

Hu et al. [8] proposed to train a single multilingual language model for shallow fusion in multiple languages. With 84 languages and a generalist language model (GLaM). When the number of experts increases, GLaM dynamically selects only two at each decoding step to keep the inference computation roughly constant. They then apply GLaM to a multilingual shallow fusion task based on a state-of-the-art end-to-end model. Compared to a dense language model of similar computation during inference, GLaM reduces the WER of an English long-tail test set by 4.4% relative. Compared to the baseline model, GLaM achieves an average WER reduction of 5.53% over 43 languages.

Santoso et al. [9] proposed utilizing LLM to annotate emotional speeches, investigating the use of conversation sequence transcription, and incorporating the textual acoustic feature descriptors into the prompt. Experiments using the IEMOCAP dataset show that emotional speech annotation using LLMs can outperform human annotation with possibly lower annotation costs.

Ling et al. [10] explored ASR transcription tasks using 6 different datasets across multiple domains. Experiments demonstrate that the proposed approach can effectively leverage the strengths of pretrained LLMs to produce more readable ASR transcriptions. In specific, GPT2 XL and ZPP achieve slightly better performance in terms of Token Error Rate (TER) on 7 and 9 out of 11 datasets, respectively, improving the average TER from 13.00% to 12.56% and 12.62%.

Tang et al. [11] present the first study that achieves both ASR and Automatic audio captioning (AAC) by connecting an LLM with auditory encoders. Integrating the Whisper encoder for speech and the BEATs encoder for audio events. With 5 datasets used, LibriSpeech, GigaSpeech, WavCaps, AudioCaps, and Clotho. LibriSpeech , and GigaSpeech, achieved a 6% lower WER.

Wang et al. [12] preseted a joint Speech and Language Model (SLM), that combines a pretrained LLM, a pretrained speech encoder, and an adapter. SLM freezes the pretrained foundation models to maximally preserve their capabilities, and only trains a simple adapter with just 1% (156M) of the foundation models' parameters. This adaptation not only leads SLM to achieve strong performance on conventional tasks such as ASR and AST, but also unlocks the novel capability of zero-shot instruction-following for more diverse tasks. This approach proved the superiority of the end-to-end model proposed as opposed to the cascading pipeline, where the speech is fed to an ASR system and the transcripts are sent to LLMs, as BLEU degraded from 38 to 32 due to ASR errors.

Wang et al. [13] applied the SLM [12], presented earlier, to dialog applications. Task-oriented dialogs often contain domain-specific entities, i.e., restaurants, hotels, train stations, and city names, which are difficult to recognize, but, critical for the downstream applications. Inspired by the RAG (retrieval-augmented generation) models, they propose a retrieval-augmented SLM (ReSLM) that overcomes this weakness. They evaluated ReSLM on speech MultiWoz task (DSTC-11 Challenge), and found that the retrieval augmentation boosts model performance, achieving joint goal accuracy (JGA) (38.6% vs 32.7%), slot error rate (SER) (20.6% vs 24.8%) and ASR WER (5.5% vs 6.7%).

Everson at al. [14] introduced a method that utilizes the ASR system's lattice output, aiming to encapsulate speech ambiguities and enhance spoken language understanding (SLU) outcomes. Lattice output is the first-pass ASR system product. They study varying ASR performance conditions and scrutinize the aspects of in-context

learning which prove the most influential. GPT-3.5 achieved the best performance with 30% and 20% enhancement over other LLMs in F1 and exact match (EM) metrics respectively.

Qu et al. [15] presented a personalized voice-based system with Speech emotion recognition (SER) powered by LLMs. It analyzes emotional status from vocal user inputs, with an accuracy exceeding baseline models by 20%.

Lakomkin et al. [16] proposed contextual speech recognition using audio features, along with optional text tokens, to train the system to complete transcriptions in a decoder-only way. As a result, the system implicitly learns how to leverage unstructured contextual information during training. Results prove a 6% WER reduction when additional textual context is provided. Moreover, the proposed method improves WER by 7.5% overall and 17% WER on rare words, compared to a baseline contextualized RNN-T system that has been trained on a speech dataset more than twenty-five times larger.

Wang et al. [17] introduced VALL-E, the first TTS framework with strong in-context learning capabilities as GPT-3. It enables prompt-based approaches for zero-shot TTS, which does not require additional structure engineering, pre-designed acoustic features, and fine-tuning. VALL-E can be used to synthesize high-quality personalized speech with only a 3-second enrolled recording of an unseen speaker as an acoustic prompt. Results show that VALL-E outperforms the state-of-the-art zero-shot TTS systems in terms of speech naturalness and speaker similarity tested on LibriSpeech and VCTK corpus.

Chen et al. [18] present a Speech Augmented Language Model (SALM) with multitasking and in-context learning capabilities. SALM comprises a frozen text LLM, an audio encoder, a modality adapter module, and LoRA layers to accommodate speech input and associated task instructions. SALM achieves performance on par with task-specific conformer baselines for ASR and AST, and also shows zero-shot in-context learning capabilities, demonstrated through keyword-boosting tasks for ASR and AST.

Malkiel et al. [19] proposed SegLLM, for efficient and accurate call segmentation and topic extraction. SegLLM is composed of offline and online phases. The offline phase is applied once to a given list of topics and involves generating a distribution of synthetic sentences for each topic using a GPT3 LLM. The online phase is applied to every call separately and scores the similarity between the transcripted conversation and the topic anchors found in the offline phase. Results show a proven improvement reaching 80% and 44% in Pk score and WindowDiff respectively.

Table I. ASR with LLM

	Affiliation	Training corpus	Task	Encoder, decoder, and LLM	Settings	Results
[1]	Tsinghua university	LibriSpeech [20], Common Voice [21], and GigaSpeech [22]	ASR	Whisper, Vicuna	NVidia A100 80GB GPUs, 90k steps, batch size of 24	17% WER reductions
[2]	Alibaba	IEMOCAP [23]	Speech emotion recognition	GPT-4	Nvidia RTX 3090 GPU with a batch size of 128 for 50 epochs	2% Improvements in WA and UA
[3]	Nvidia	AMI-MH (Mixed Headset) [24] Call Home American English	Speaker diarization (SD)	GPT	NVIDIA TESLA V100 GPU	9.8% relative delta-SA-WER improvement

		Speech (CHAES, LDC97S42) [25]				
[4]	Meta AI	Multilingual LibriSpeech [26]	ASR	Conformer [27] LLaMA-7B	16 NVIDIA A100 40GBs, with a batch size of up to 500 seconds of audio per GPU	18% reductions in WER
[5]	Microsoft	CoVoST 2 [28]	Translation	seq2seq, Whisper, and LLaMA-7B	16 Nvidia V100 GPUs	4.6 absolute BLEU score improvement
[6]	Microsoft	LibriSpeech, TedLium-2 [29], and SPGISpeech [30]	Zero-shot adaptation	HuBERT [31], and LLaMA-7B	8 NVIDIA Tesla V100 32GB GPUs, with a batch size of 1 per GPU	7% and 3% WER reductions on TedLium-2 and SPGISpeech respectively
[7]	Google	Youtube Videos [32], Google's Voice Search traffic [33]	Long-form ASR	T5 [34], PaLM, and HAT factorization [35]		8% WER and 30% STER reductions
[8]	Google	Google Voice Search traffic	Transliterating	Conf-140M, Conf-640M, and GLaM-64E	Google Cloud V4 TPUs	GLaM achieves an average WER reduction of 5.53% over 43 languages
[9]	RevComm	IEMOCAP	Speech emotion recognition	GPT-3.5-turbo		LLM annotation outperform human annotation
[10]	Microsoft	LibriSpeech, Common Voice, GigaSpeech, TedLium-2, SPGISpeech, and VoxPopuli [36]	ASR transcriptions	GPT2-XL, ZPP, and Whisper		GPT2 XL and ZPP achieved slightly better TER
[11]	Tsinghua	LibriSpeech, GigaSpeech, WavCaps[37], AudioCaps[38], and Clotho [37]	Automatic audio captioning (AAC)	Whisper-L-v2 BEATs [39], and Vicuna 13B3 [40]		LibriSpeech and GigaSpeech reduced WER by 6% on average
[12]	Google DeepMind	SpeechStew ASR [41], VoxPopuli ASR, FLEURS ASR [42], CoVoST2[43], and Universal Speech Model (USM) [44]	Translation, and ASR with Contextual Biasing	T5, and MC4 [44] Whisper, mSLAM-CTC, MAESTRO, USM-M, Mu2SLAM, AudioPaLM-2		
[13]	Google DeepMind	Universal Speech Model (USM)	Dialogue state tracking (DST)	T5 XXL		Improvements in JGA 5%, SER 4%, and WER 2%
[14]	Amazon Alexa AI	LibriSpeech, Natural Multi-speaker Spoken Question Answering (NMSQA) [45], and The Airline Travel Information Systems (ATIS) [45]1/24/2025 10:55:00 AM	ASR	BLOOMZ-560M, BLOOMZ-3B, and GPT-3.5-turbo		GPT-3.5 scored 30% and 20% enhancement over other LLMs in terms of F1 and EM metrics respectively
[15]	University of California	Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [46]	Speech emotion recognition (SER)	Whisper		SER-Whisper scored 20% more accuracy compared to baseline models.
[16]	Meta AI	Public inhouse Facebook and Instagram videos	Contextualized ASR	LLaMA-7B	128 A100 GPUs	Speech LLaMA achieved a 5.2% reduction in WER compared to the RNN-T system.

[17]	Microsoft	LibriLight [47], and VCTK [47]	Contextualized TTS	SOTA zero-shot TTS model YourTTS [48]	16 NVIDIA TESLA V100 32GB GPUs, with a batch size of 6k acoustic tokens per GPU for 800k steps.	VALL-E outperforms the zero-shot TTS system in terms of speech naturalness and speaker similarity
[18]	Nvidia	NGC ASR pretrained Fast Conformer-large, and or the Conformer self-supervised learning (SSL) checkpoint from Nvidia	Translation	NeMo, Fast Conformer [49], and GPT-style Megatro [50]	A100 GPUs	SALM performs better on shorter words, compound words, and text normalization than baseline
[19]	Microsoft	Dynamics 365 sales tenants segmented labeled with topics: greetings, closing, pricing, identification, and scheduling	Call segmentation and tagging	GPT-3 (the "davinci-003" model)		SegLLM outperforms other methods by 12% and 8% in Pk and WindowDiff scores

Table II. Datasets used in the studies

	Year	Name	Description
[45]	1990	The Airline Travel Information Systems (ATIS)	41 sessions each of 40 minutes
[25]	1997	Call Home American English Speech	120 unscripted 30-minute telephone conversations between native English speakers
[24]	2005	AMI-MH (Mixed Headset)	100-hour corpus of meetings. Capturing voice, video, electronic pen writings, presentation slides, and white-board contents
[23]	2008	IEMOCAP	12 hours of audiovisual data, including video, speech, motion capture of face, text transcriptions
[29]	2012	TedLium-2	Derived from TED Talks and composed of 216 hours of audio among 698 unique speakers
[32]	2013	YouTube Videos	Two datasets, Ytn08 for YouTube News videos with 11.1 hours and YtiDev11 for YouTube view-count weighted videos with 6.6 hours
[20]	2015	LibriSpeech	Derived from LibriVox audiobooks, and contains 1000 hours of speech sampled at 16 kHz
[46]	2018	Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)	24 professional actors, vocalizing lexically matched statements in a neutral North American accent and a range of emotions, including calm, happiness, sadness, anger, fear, surprise, and disgust, each articulated in two levels of emotional intensity: normal and strong, along with a neutral expression. Resulting in 7,356 recordings and a total size of 24.8 GB.
[47]	2019	VCTK	110 English speakers with various accents. Each speaker reads out about 400 sentences.
[38]	2019	AudioCaps	46K audio clips with human-written text pairs collected via crowdsourcing on the AudioSet dataset
[51]	2019	LibriLight	Derived from open-source audio books from the LibriVox project. It contains over 60K hours of audio.
[33]	2019	Google's Voice Search Traffic	Four datasets, Search 56k hours, Farfield 38k hours, Telephony 4k hours, YouTube 190k hours.
[37]	2020	Clotho	4981 audio samples and each audio sample has five captions (a total of 24 905 captions). Audio samples are of 15 to 30 seconds duration and captions are eight to 20 words long.
[26]	2020	Multilingual LibriSpeech	Derived from LibriVox audiobooks and consists if 8 languages, including about 44.5K hours of English and a total of about 6K hours for other languages
[28]	2020	CoVoST 2	Large-scale multilingual speech-to-text corpus covering translations from 21 languages into English and from English into 15 languages.
[21]	2020	Common Voice	50,000 individuals with 2,500 hours of collected audio in 38 languages
[22]	2021	GigaSpeech	10,000 hours of labeled audio, and 40,000 unlabeled

[30]	2021	SPGISpeech	Derived from company earnings calls by S&P Global, Inc. including 5,000 hours of professionally transcribed earnings calls, with 50,000 speakers
[36]	2021	VoxPopuli	400K hours of unlabeled speech data in 23 languages
[41]	2021	SpeechStew ASR	A collection of 7 datasets: AMI, common voice, English Broadcast News (LDC97S44, LDC97T22, LDC98S71, LDC98T28), LibriSpeech, Switchboard, TED-LIUM v3, Wall Street Journal
[52]	2022	Natural Multi-speaker Spoken Question Answering (NMSQA)	They are the spoken version of the SQuAD v1.1 dataset with 297.18 / 37.61 hours of audio for the train/dev sets
[53]	2023	Universal Speech Model (USM)	pre-trained (unsupervised) on 12 million hours of YouTube audio recordings with BEST-RQ [54]
[55]	2024	WavCaps	400k audio clips with paired captions sourced from audio clips and their raw descriptions from the web and a sound event detection dataset.
[42]	2023	FLEURS ASR	Speech dataset in 102 languages built on top of the machine translation FLoRes-101 benchmark, with approximately 12 hours of speech supervision.

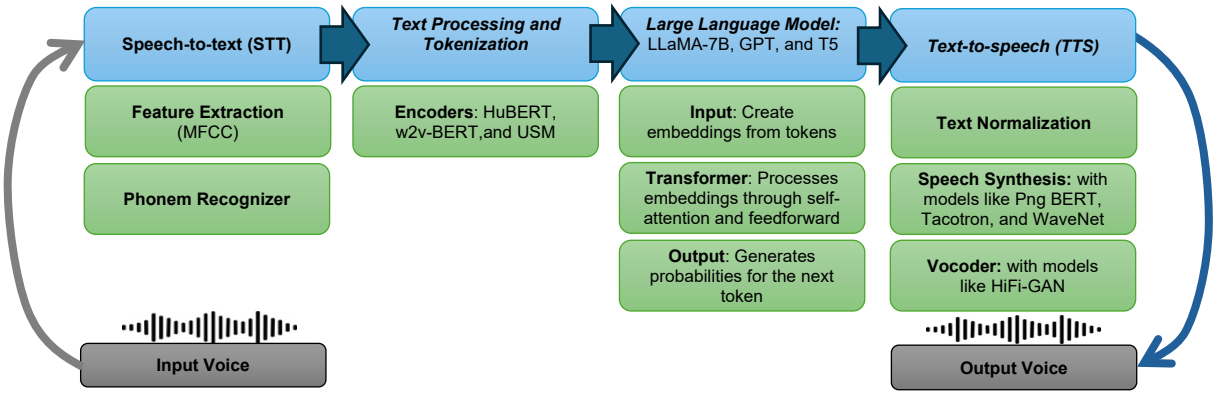

Fig. II: General Architecture of LLM-Based ASR system

III. ARCHITECTURE

The architecture of the speech-based models using LLMs involves several components that work together to handle audio input, process it into textual data, and then use the LLM for natural language understanding and generation. Fig II illustrates the general architecture of LLM-based ASR systems, a reflection of the literature covered. Also, in the following a high-level overview of the architecture's modules:

A. Speech-to-text module

It gets the audio as input and converts it into text for further processing. This module comprises two main components: **Feature Extraction:** Extracts features from raw audio, such as Mel-frequency cepstrum coefficients (MFCCs), using libraries to extract acoustic features such as Librosa [9], [56]. **Phonem recognizer**: Maps audio features to phonemes or graphemes using deep learning models like RNNs, CNNs, or transformers, while previously it was done using classical models like Hidden Markov Models (HMMs) paired with Gaussian Mixture Models (GMMs).

B. Text Processing and Tokenization

Prepares the transcribed text for the LLM by tokenizing it into units (e.g., words, subwords, or

characters) that the model can process. Some of the tokenization encoders include: HuBERT [6], [31], w2v-BERT [57], and Universal Speech Model (USM) [12], [13], [58].

C. Large Language Model

Handles understanding, reasoning, and generating natural language responses based on the text input. According to the literature, examples to thoes LLMs are: LLaMA-7B, GPT, and T5. The architecture of this level is in three parts: **Input Layer**: Encodes tokenized input into embeddings. **Transformer Layers**: Process embeddings through self-attention and feedforward layers to capture contextual relationships. **Output Layer**: Generates probabilities for the next token or sequence of tokens.

D. Text-to-Speech (TTS) Module

Converts the LLM-generated text into natural-sounding speech. It operates using three different essential components: **Text Normalization**: Which handles punctuation and special symbols for smoother speech synthesis. **Speech Synthesis**: Generates audio from text mainly using neural architectures like Png BERT [12], [60], Tacotron [61], [62], WaveNet [63], or FastSpeech [64]. Finally, a **Vocoder**: it converts the synthesized spectrogram into a waveform, with models like Generative adversarial networks for efficient and high-fidelity speech synthesis (HiFi-GAN) [65].

V. Conclusion and Future Work

This review highlights the transformative impact of Large Language Models (LLMs) on Automatic Speech Recognition (ASR) systems, synthesizing insights from 19 research studies conducted in 2023 and 2024 and 24 datasets spanning 1999 to 2024. LLMs have revolutionized ASR by enhancing transcription accuracy,

E. Optimizations for Speech-Based LLMs

End-to-End Models: Some approaches integrate STT and LLM tasks for better performance, like [5], [6], [8], [10], [11], [12], [13], [66].

Contextualized Learning: this refers to the learning enforcement through the injection of external context. This proved to aid the learning and enhance the ASR and TTS in specific, [16] and [17].

IV. Datasets

The datasets covered in this paper leverage a diverse collection of 24 corpora spanning 35 years, from 1990 to 2024. That is said, the studies are conducted in 2023 and 2024 only with 6 and 13 papers in each year respectively. These datasets represent a wide range of linguistic, acoustic, and contextual variations, making them critical for advancing research in ASR systems. The datasets depicted in Table II, vary in size, scope, and purpose, encompassing multilingual corpora, domain-specific recordings, and multimodal resources. This breadth reflects the evolution of data collection techniques and the growing emphasis on inclusivity in language representation. The datasets also highlight the shift towards multimodal approaches, integrating text, audio, and sometimes visual inputs, to support the development of robust and versatile ASR systems. By analyzing these datasets, this paper provides insights into their roles in shaping ASR advancements, their alignment with emerging trends, and their impact on the challenges and opportunities in the field.

enabling cross-lingual capabilities, and addressing challenges in low-resource languages. Key trends identified include a shift toward end-to-end architectures, the growing use of multimodal datasets, and the integration of advanced transfer learning and fine-tuning techniques. However, significant challenges remain, including the high computational costs, scalability issues, and availability of datasets. The dominance of prominent technology companies in the research

in this domain highlights the critical role of computational resources and proprietary datasets, underscoring a disparity in access between industry and academia. Future directions may include breakthroughs in processing and developing energy-efficient models. Additionally, fostering collaboration between industry and academia could bridge gaps and accelerate innovation. This review serves as a foundation for researchers and practitioners to advance the field of ASR by leveraging the capabilities of LLMs while addressing existing limitations.

REFERENCES

[1] W. Yu *et al.*, "Connecting Speech Encoder and Large Language Model for ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12637–12641. doi: 10.1109/ICASSP48485.2024.10445874.

[2] Z. Ma *et al.*, "Leveraging Speech PTM, Text LLM, And Emotional TTS For Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11146–11150. doi: 10.1109/ICASSP48485.2024.10445906.

[3] Tae Jin Park, Kunal Dhawan, Nithin Koluguri, and Jagadeesh Balam, "Enhancing Speaker Diarization with Large Language Models: A Contextual Beam Search Approach," presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Apr. 2024.

[4] Y. Fathullah *et al.*, "Prompting Large Language Models with Speech Recognition Abilities," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 13351–13355. doi: 10.1109/ICASSP48485.2024.10447605.

[5] J. Wu *et al.*, "On Decoder-Only Architecture For Speech-to-Text and Large Language Model Integration," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8. doi: 10.1109/ASRU57964.2023.10389705.

[6] Y. Li, Y. Wu, J. Li, and S. Liu, "Prompting Large Language Models for Zero-Shot Domain Adaptation in Speech Recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8. doi: 10.1109/ASRU57964.2023.10389732.

[7] T. Chen *et al.*, "Large-Scale Language Model Rescoring on Long-Form Data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096429.

[8] K. Hu *et al.*, "Massively Multilingual Shallow Fusion with Large Language Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10094796.

[9] J. Santoso, K. Ishizuka, and T. Hashimoto, "Large Language Model-Based Emotional Speech Annotation Using Context and Acoustic Feature for Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11026–11030. doi: 10.1109/ICASSP48485.2024.10448316.

[10] S. Ling *et al.*, "Adapting Large Language Model with Speech for Fully Formatted End-to-End Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11046–11050. doi: 10.1109/ICASSP48485.2024.10448204.

[11] C. Tang *et al.*, "Extending Large Language Models for Speech and Audio Captioning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11236–11240. doi: 10.1109/ICASSP48485.2024.10446343.

[12] M. Wang *et al.*, "SLM: Bridge the Thin Gap Between Speech and Text Foundation Models," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8. doi: 10.1109/ASRU57964.2023.10389703.

[13] M. Wang *et al.*, "Retrieval Augmented End-to-End Spoken Dialog Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12056–12060. doi: 10.1109/ICASSP48485.2024.10447448.

[14] K. Everson *et al.*, "Towards ASR Robust Spoken Language Understanding Through in-Context Learning with Word Confusion Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12856–12860. doi: 10.1109/ICASSP48485.2024.10447938.

[15] X. Qu, Z. Sun, S. Feng, C. Chen, and T. Tian, "Breaking the Silence: Whisper-Driven Emotion Recognition in AI Mental Support Models," in *IEEE Conference on Artificial Intelligence (CAI)*, 2024, pp. 290–291. doi: 10.1109/CAI59869.2024.00063.

[16] E. Lakomkin, C. Wu, Y. Fathullah, O. Kalinli, M. L. Seltzer, and C. Fuegen, "End-to-End Speech Recognition Contextualization with Large Language Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12406–12410. doi: 10.1109/ICASSP48485.2024.10446898.

[17] C. Wang *et al.*, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers." arXiv, 2023. doi: 10.48550/ARXIV.2301.02111.

[18] Z. Chen *et al.*, "SALM: Speech-Augmented Language Model with in-Context Learning for Speech Recognition and Translation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 13521–13525. doi: 10.1109/ICASSP48485.2024.10447553.

[19] I. Malkiel *et al.*, "SEGLLM: Topic-Oriented Call Segmentation Via LLM-Based Conversation Synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11361–11365. doi: 10.1109/ICASSP48485.2024.10446156.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2015, pp. 5206–5210. doi: 10.1109/icassp.2015.7178964.

[21] R. Ardila *et al.*, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520/

[22] G. Chen *et al.*, "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10, 000 Hours of Transcribed Audio," in *Interspeech 2021*, in interspeech_2021. ISCA, Aug. 2021. doi: 10.21437/interspeech.2021-1965.

[23] C. Busso *et al.*, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Nov. 2008, doi: 10.1007/s10579-008-9076-6.

[24] J. Carletta *et al.*, "The AMI Meeting Corpus," in *Proceedings of Symposium on Annotating and Measuring Meeting Behavior*, 2005.

[25] Canavan, Alexandra, Graff, David, and Zipperlen, George, "CALLHOME American English Speech." Linguistic Data Consortium, 1997. doi: 10.35111/EXQ3-X930.

[26] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech 2020*, in interspeech_2020. ISCA, Oct. 2020. doi: 10.21437/interspeech.2020-2826.

[27] A. Gulati *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition." arXiv, 2020. doi: 10.48550/ARXIV.2005.08100.

[28] C. Wang, A. Wu, and J. Pino, "CoVoST 2 and Massively Multilingual Speech-to-Text Translation." arXiv, 2020. doi: 10.48550/ARXIV.2007.10310.

[29] A. Rousseau, P. Deléglise, and Y. Estève, "TED-LIUM: an Automatic Speech Recognition dedicated corpus," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC`12)*, N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds., Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 125–129. [Online]. Available: https://aclanthology.org/L12-1405/

[30] P. K. O'Neill *et al.*, "SPGISpeech: 5, 000 hours of transcribed financial audio for fully formatted end-to-end speech recognition." arXiv, 2021. doi: 10.48550/ARXIV.2104.02014.

[31] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEEACM Trans Audio Speech Lang Proc*, vol. 29, pp. 3451–3460, Oct. 2021, doi: 10.1109/TASLP.2021.3122291.

[32] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, IEEE, Dec. 2013. doi: 10.1109/asru.2013.6707758.

[33] A. Narayanan, R. Prabhavalkar, C.-C. Chiu, D. Rybach, T. N. Sainath, and T. Strohman, "Recognizing Long-Form Speech Using Streaming End-to-End Models," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, Dec. 2019, pp. 920–927. doi: 10.1109/asru46091.2019.9003913.

[34] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[35] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid Autoregressive Transducer (HAT)," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020. doi: 10.1109/icassp40776.2020.9053600.

[36] C. Wang *et al.*, "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. doi: 10.18653/v1/2021.acl-long.80.

[37] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an Audio Captioning Dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP), 2020, pp. 736–740. doi: 10.1109/ICASSP40776.2020.9052990.

[38] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating Captions for Audios in The Wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 119–132. doi: 10.18653/v1/N19-1011.

[39] S. Chen *et al.*, "BEATs: audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, in ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023.

[40] W.-L. Chiang *et al.*, "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality." Mar. 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[41] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "Speechstew: Simply mix all available speech recognition data to train one large neural network," *ArXiv Prepr. ArXiv210402133*, 2021.

[42] A. Conneau *et al.*, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 798–805.

[43] C. Wang, A. Wu, J. Gu, and J. Pino, "CoVoST 2 and Massively Multilingual Speech Translation.," in *Interspeech*, 2021, pp. 2247–2251.

[44] L. Xue, "mt5: A massively multilingual pre-trained text-to-text transformer," *ArXiv Prepr. ArXiv201011934*, 2020.

[45] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The ATIS spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[46] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS One*, vol. 13, no. 5, p. e0196391, 2018.

[47] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)." University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019. doi: 10.7488/DS/2645.

[48] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*, PMLR, 2022, pp. 2709–2720.

[49] D. Rekesh *et al.*, "Fast conformer with linearly scalable attention for efficient speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–8.

[50] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," *ArXiv Prepr. ArXiv190908053*, 2019.

[51] J. Kahn *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7669–7673.

[52] G.-T. Lin *et al.*, "DUAL: Discrete spoken unit adaptive learning for textless spoken question answering," *ArXiv Prepr. ArXiv220304911*, 2022.

[53] Y. Zhang *et al.*, "Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages." arXiv, 2023. doi: 10.48550/ARXIV.2303.01037.

[54] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised Learning with Random-projection Quantizer for Speech Recognition," in *International Conference on Machine Learning*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246608081

[55] X. Mei *et al.*, "WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 3339–3354, 2024, doi: 10.1109/TASLP.2024.3419446.

[56] B. McFee *et al.*, "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference*, in SciPy. SciPy, 2015, pp. 18–24. doi: 10.25080/majora-7b98e3ed-003.

[57] Y.-A. Chung *et al.*, "w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 244–250. doi: 10.1109/ASRU51503.2021.9688253.

[58] Z. Zhang *et al.*, "Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling." 2023. [Online]. Available: https://arxiv.org/abs/2303.03926

[59] C. Lyu *et al.*, "Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration." 2023. [Online]. Available: https://arxiv.org/abs/2306.09093

[60] Y. Jia, H. Z. (Byungha Chun), J. Shen, Y. Zhang, and Y. Wu, "PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS," in *Interspeech*, 2021. [Online]. Available: https://arxiv.org/abs/2103.15060

[61] A. Dekel, S. Shechtman, R. Fernandez, D. Haws, Z. Kons, and R. Hoory, "Speak While You Think: Streaming Speech Synthesis During Text Generation," in *IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP), 2024, pp. 11931–11935. doi: 10.1109/ICASSP48485.2024.10446214.

[62] J. Shen *et al.*, "Non-Attentive Tacotron: Robust and controllable neural TTS synthesis including unsupervised duration modeling." 2020. [Online]. Available: https://arxiv.org/abs/2010.04301

[63] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," in *Arxiv*, 2016. [Online]. Available: https://arxiv.org/abs/1609.03499

[64] Y. Ren *et al.*, "FastSpeech: fast, robust and controllable text to speech," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019.

[65] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, in NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[66] K. Yamauchi, Y. Ijima, and Y. Saito, "STYLECAP: Automatic Speaking-Style Captioning from Speech Based on Speech and Language Self-Supervised Learning Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11261–11265. doi: 10.1109/ICASSP48485.2024.10445977.