International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 10 | Oct 2024 www.irjet.net p-ISSN: 2395-0072

Enhancing LLMs with Indian Multi-Lingual Audio Understanding for AGI Advancement: A Survey

Aniruddha Birage[1], Tanay Thatte[1], Chiranjeev Patil[1], Aarush Balkundi[1], Arati Deshpande[1] Saswati Rabha[2], Chintan Parikh[2]

[1]PICT Pune (of Aff. SPPU), Pune, India
[2]SReverie Language Technologies Limited, Bengaluru, India

Abstract - *The recent trends in speech and language processing include self-supervised learning, multilingual automatic speech recognition, and large-scale language models. New emerging techniques, such as Wav2Vec 2.0 and joint supervised-unsupervised training, can achieve scalability with high performance in low-resource languages, but challenges such as high computational costs may be incurred, and data imbalances will be encountered. Among the newer innovations with few-shot learning, multimodal models, and task generalization, big improvements come on adaptation and efficiency-while similarly so do the challenges of bias and even resource intensity. This paper will explore these breakthroughs and their effects in discipline.*

***Key Words*: Self-supervised learning, Wav2Vec 2.0, Large scale language models, Multimodal models, Task generalization, Low-resource languages, Speech and language processing**

1.INTRODUCTION

This Recent advances in self-supervised learning (SSL), multilingual automatic speech recognition (ASR), and large-scale language models (LLMs) have dramatically impacted the speech and language processing community with so many breakthroughs and establishment of new state-of-the-art benchmarks. This paper discusses over three broad challenges: speaker and cross-lingual representation learning (XLRL) and efficient scaling of LLMs. Another crucial factor is minimizing dependence on labelled data, especially for low-resource languages (LRLs). Therefore, self-supervised models like Wav2Vec 2.0 push the boundaries of scalability and accuracy; they do indeed help us find answers to many challenges that appear to arise from limited labelled datasets but concurrently expose new complexities because of computational costs of fine-tuning very complex models for specific tasks or languages.

XLRL research has advanced significantly, and models can now learn common patterns across LRLs. The property combined with transfer learning techniques makes the flexibility and inclusiveness of speech recognition systems significant. This capability makes these systems generalize effectively across a wide range of languages, thereby making them highly adaptable. Such flexibility has great promise for code-switching and zero-shot learning with such successful applications of the model in languages that it was never explicitly trained on.

The multilingual audio addition from India enhances significantly the performance and the adaptability of the LLMs; bidirectional circuits leverage India's linguistic diversity to provide richer, more nuanced representations, enhancing the processing and understanding of various languages. The challenges with LRLs are addressed, and there is a push forward with cross-lingual transfer learning. Use of multi-lingual audio data enables robust training that captures nuances of various dialects and contexts, culminating in richer, more inclusive, and notably accurate language models.

Scaling large language models to sizes that are like Pathways Language Model (PaLM) and Generative Pre-trained Transformer 3 (GPT-3) has completely revolutionized the nature of natural language processing. These models excel at learning from few examples and adapting to various tasks with minimal task-specific fine-tuning.. These kinds of scale come with enormous computational costs and increase memory requirements and larger environmental impacts. Data imbalance, particularly the lack of low-resource languages in training datasets, is a major issue.

The rapid growth of large model training raises significant concerns for researchers and organizations with limited resources. Therefore, the open-source models like Vicuna have popped up to democratize artificial general intelligence (AGI) so that the organizations can now deploy it for research and development purposes. Other challenges like imbalanced datasets and the search for more efficient and more inclusive models remain.

2. RESEARCH

2.1 Self-Supervised Learning and Speech Recognition

These papers focus on reducing reliance on labeled data and improving speech recognition using self-supervised techniques. kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

- Self-Supervised Cross-Lingual Speech Representation Learning at Scale [1] A scalable self-supervised learning approach for learning speech representations across multiple languages from few or no labeled datasets improves performance on low-resource languages through cross-lingual transfer.

- Wav2Vec 2.0: Self-Supervised Speech Learning [3] It presents Wav2Vec 2.0 which learns speech representations from raw audio using self-supervised learning. Its architecture and results significantly reduce the amount of labelled data needed for training yet allows for state-of-the-art speech recognition performance to be achieved.

- Self-supervised Learning with Random-Projection Quantizer [6] In this approach, more significant gains in self-supervised speech learning are achieved at lower computational costs without affecting the precision and with the use of a random-projection quantizer to make the system efficient.

- w2v-BERT: Combining Contrastive Learning and Masked Language Modelling [8] Combines contrastive learning and masked language modelling to enhance self-supervised speech pretraining for better generalization across speech tasks but requires reduced data for the purpose of being labelled.

- Unsupervised Cross-lingual Representation Learning for Speech Recognition [9] Knowledge transfer of high-resource languages into low-resource languages is focused on in an entirely unsupervised manner to better enhance speech recognition without using labelled data for low-resource languages.

2.2 Multilingual and Multimodal Learning

These papers explore learning across multiple languages or modalities, enhancing system performance in diverse tasks.

- Joint Unsupervised and Supervised Training for Multilingual ASR [2] It combines unsupervised and supervised learning for bettering ASR in multiple languages, especially low-resource languages, by processing labelled and unlabelled data.

- Multimodal Embodied Language Learning [11] It integrates different modalities such as vision, language, and action for tasks like robotics and enables models to talk about and understand the real-world environment more effectively by utilizing several modalities.

- Language-independent Sub word Tokenization [16] A language-independent tokenizer, based on multilingual rules as opposed to specific rules for any one language, is proposed for efficiency in text processing with precision, particularly for rare and unknown words.

2.3 Large Language Models and Few-shot Learning

These papers focus on scaling large models and improving task generalization using minimal task-specific data.

- Language Models are Few-Shot Learners [4] Introduces GPT-3, a large language model that can really carry out multiple tasks with few-shot learning and can generalize robustly across most NLP tasks even when fine-tuned minimally.

- Scaling Vision Transformers to 22 Billion Parameters [10] A new vision transformer with 22 billion parameters significantly improves image recognition and addresses training and scaling challenges, highlighting its potential in diverse computer vision tasks.

- Vicuna: Open-Source Chatbot Development [5] A chatbot that tops 90 percent efficiency of ChatGPT and is capable of quality capability and flexibility in being highly adaptable and customizable.

- PaLM: Scaling Language Modelling with Pathways [7] Scaling large language models. Efficiently with the Pathways system, which only activates parts of the model. Reduces computational overheads while still maintaining strong performance on a wide range of NLP tasks.

2.4 Efficient Adaptation and Transfer Learning

These papers explore techniques for improving the computational efficiency and scalability of models.

- Parameter-efficient Transfer Learning [13] It proposes reducing the number of trainable parameters needed to fine-tune large models without sacrificing performance, reducing computational resources on adaptation to a given task.

- Low-rank Adaptation of Large Models [14] The method applies low-rank matrices to the task of fine-tuning large pre-trained models efficiently in such a way that significant reductions in computational costs are achieved while retaining high performance on new tasks.

2.5 Real-time Speech and Contextual Understanding

These papers deal with enhancing real-time speech recognition and contextual understanding in language models.

- Fast-Slow Encoder-based Speech Transduction [15] This paper proposes a fast-slow encoder framework that improves the accuracy of real-time speech recognition and provides low latency, which makes it suitable for streaming speech-to-text applications.

- Contextual Speech Understanding [12] This framework improves speech comprehension by combining hearing with a reasoning mechanism. Such integrated capabilities improve performance with complex real-time interaction. Virtual assistants are expected to understand their users' commands and get it right even in quite noisy environments.

Table – 1: Comparative Study of Large Language Models (LLMs)

Ref. No.	*Title*	*Method Used*	*Dataset Used*	*Performance*	*Advantages*	*Limitations*
[1]	Self-supervised cross-lingual speech representation learning at scale	• Focuses on language-agnostic speech representation using self-supervised learning. • Learns to predict parts of audio based on context, generalizing speech patterns across languages. • Involves pre-training on large-scale multilingual audio data without labeled transcriptions. • Pre-trained models can be fine-tuned for tasks like speech recognition or translation. • Reduces reliance on labeled data for each language.	• The training corpus includes multilingual speech data, enabling the model to learn robust cross-lingual representations. • Large-scale datasets like Common Voice, BABEL, or proprietary collections are used. • Ensures coverage of a wide range of acoustic and linguistic variations.	• Pretrained models show strong performance in downstream tasks like ASR and speech translation, especially in low-resource languages • Clear improvement in Word Error Rate (WER) compared to models trained from scratch or with labeled data. • Self-supervised pre-training proves highly efficient for cross-lingual speech tasks.	• Reduces reliance on labeled data, especially for low-resource languages. • Cross-lingual knowledge transfer improves performance across languages. • Scalable training with large multilingual datasets. • Enhanced performance in downstream tasks like speech recognition. • Easily adapts to new languages with minimal data. • Cost-efficient by using unannotated data, reducing the need for manual annotations.	• High computational cost for training large models with massive datasets. • Bias towards resource-rich languages due to data imbalance. • Fine-tuning for specific languages or tasks adds complexity. • Performance depends heavily on the quality of unannotated data. • Learned representations are abstract and difficult to interpret.

[2]	Joint Unsupervised and Supervised Training for Multilingual ASR	• The authors propose a joint training framework combining supervised and unsupervised learning. • Unsupervised training uses untranscribed speech data, while supervised training uses labeled data. • The model is trained multilingually, integrating data from various languages to enhance ASR performance across multiple languages.	• The paper utilizes large multilingual datasets of labeled and unlabeled speech data across various languages. • Specific datasets include the MLS dataset and other speech corpora with both transcribed and untranscribed data.	• The joint training scheme shows significant improvements, particularly with low-resource languages. • Authors report a decrease in Word Error Rate (WER) across various languages using this method. • The most drastic WER reduction occurs when labeled data is limited, but there is abundant unlabeled data. • This method outperforms monolingual and multilingual models trained solely through supervised learning.	• Enhances performance across multiple languages. • Works with both labeled and unlabeled data. • Reduces reliance on expensive labeled data. • Improves model robustness and generalization. • Scalable for larger training datasets. • Enhances accuracy in real-world applications.	• Requires complex training configurations. • Higher computational cost due to handling both labeled and unlabeled data. • Quality heavily depends on the quality of the input data. • Limited enhancement for high-resource languages. • Aligning multilingual data can be challenging. • Risk of overfitting the model to specific language families.

[3]	Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations	• Wav2Vec 2.0 comes in with a self-supervised learning framework for robust speech representations learned directly from raw audio waveforms. • The model involves two primary components: feature encoder and context network. • Feature encoder actually transforms the raw audio into a sequence of latent speech representations, while the context network, which is based on a decoder architecture such as Transformers, processes those representations to capture contextual information. • It is trained with a contrastive loss approach so that it learns distinguishing between the true and false representations of the input audio. • Thus, the model may learn rich speech features without requiring labeled data.	• Wav2Vec 2.0 uses a large, diverse dataset composed of large amounts of unlabeled audio from various sources to train its model. • Majorly, the training data comprises 960 hours of reading speech from LibriSpeech corpus. • Further robustness and generalization properties have been added to it by fine-tuning the model with smaller and more labeled datasets involving Common Voice and LibriSpeech datasets. This combination of unsupervised and supervised samples facilitates the model in taking advantages of self-supervised learning while availing supervised fine-tuning for its betterment.	• Wav2Vec 2.0 boasts superior results compared to all the other state-of-the-art models on the Automatic Speech Recognition (ASR) task; it will perform outstandingly with very impressive results on benchmark datasets like LibriSpeech. • Reduction in WER is massive, exceeding previous methods. • For example, for fine-tuning with labeled data on the LibriSpeech test-clean set, it achieves a WER of 1.9%, showing a massive improvement over the existing approaches. • Wav2Vec 2.0 is therefore a versatile framework for speech representation learning, and it can be used effectively in any language or even dialect because of its ability to learn from unlabeled data.	• Self-supervised learning: It does not need pre-training with labeled data but, instead, uses raw audio data for that purpose. • It also outperforms speech recognition models that rely on the classic approaches with fewer numbers of labeled examples. • Transferable features: The method learns robust speech representations which generalize across tasks. • Lower dependency on labels: This makes it suitable for low-resource languages by reducing the requirement for large, labeled datasets.	• Computationally expensive: Large models are pre-trained on massive amounts of raw, unlabeled data. This is very computation-intensive •Implementations complex: Self-supervised learning methods such as Wav2Vec 2.0 also require implementation and fine-tuning.

[4]	Language Models are Few-Shot Learners	• Introducing GPT-3. This large-scale language model was created using a transformer architecture and 175 billion parameters by training on. • It is shown that the paper is capable of applying functions with just a few shots of one-shot or even zero-shot learning; that is, it can produce useful responses with very few examples as input during inference.	• The training data volume was about 570 GB of text in this broad mix of internet text, including Common Crawl, WebText, Wikipedia, and books. • The amount of evaluation tests applied is diverse in NLP benchmarks-for example, question answering (e.g., TriviaQA), close tasks, language translation, and text generation.	• GPT-3 did often outperform smaller models on many NLP tasks and, in some cases, even equaled or surpassed supervised baselines. • It had weaknesses in particular areas such as arithmetic and common-sense reasoning to point out that there were a few capabilities, despite all those few-shot learning abilities.	• Few-shot learning capacity eliminates the requirement of the presence of large amounts of labeled data. • No specialized fine-tuning is necessary on the task for versatile performance on a very wide range of NLP tasks. • Flexible task adaptation from pre-trained knowledge. • Reduced requirement of computational resources for new tasks in comparison to traditional models. • It generalizes effectively to unseen tasks with fewer examples. • It helps in increasing efficiency in carrying out different linguistic tasks with just one model.	• Heavy computational cost in pre-training big models. • Very sensitive to the prompt design, which hurts its performance. • Poor on tasks that are too complex or domain specific. • The model can inherit the biases it learns from the training data to affect its output. • There is limited interpretability concerning how few-shot learning works. • Results may not generalize consistently across different tasks and languages.

[5]	Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality	• Developed by fine-tuning the open-source large language model LLaMA, Vicuna is fine-tuned on human-chat text conversations. • These include high-quality datasets aimed at furthering the capabilities of the chatbot in developing human-like responses. • The model architecture relies on transformer networks with efforts toward balancing its performance efficiency in computation, much like GPT-4.	• The dataset consists of user-chatbot conversation data tailored towards dialogues with ChatGPT. • Fine-tuning utilizes high-quality conversation transcripts in order to enhance natural language understanding and generation.	• At an approximate rating of 90% quality, as developed by benchmarking evaluation metrics, Vicuna stands bright according to GPT-4. • This showed Vicuna to be all the more impressive in any kind of task-dialogue generation to question answering-with bright prospects even if it is open-source and shines in comparison with proprietary models like GPT-4.	• More or less, of quality 90% compared to ChatGPT, yet the same level of performance. • Open source thus can be modified, experimented on, and together developed by users. • In comparison, this will save cost as it is an open-source compared to proprietary models like ChatGPT. • The process of development increases peoples' trust, and they can engage with such a project. • Improvable and fine-tune-able with public data available. • Great access for researchers and developers-a step forward for innovation regarding conversational AI.	• Quality is subjective and relies on the assessment of GPT-4. • It has limited functionalities in certain domains and may fail in certain niches. • Metrics of assessment are not diverse enough to account for the performance of the chatbot. • May fail to handle complex multi-turn dialogues suitably like proprietary models. • Computationally expensive for fine-tuning to encourage better performances. • Lacks wide support that commercial models such as ChatGPT also provide.
[6]	Self-supervised Learning with Random-Projection Quantizer for Speech Recognition	• The paper presents a self-supervised learning paradigm based on a Random-Projection Quantizer (RPQ) to enhance speech recognition tasks.	• The two datasets experimented on include several widely used speech recognition datasets, namely, the LibriSpeech dataset and	• The experimental results clearly show promising improvements in the accuracy of speech recognition by introducing the self-supervised	• Efficiency: The random projection quantizer reduces the computational cost of speech recognition. • Self-	• Complexity: The random-projection approach may add complexity in implementation and tuning. • High computational

		• The method is to learn meaningful representations of speech signals by training a neural network to map high-dimensional features to a low-dimensional space through random projection. • Feature quantization is facilitated by RPQ, which forces a more compact representation while preserving essential information, reducing the complexity of computation, and helping in generalization performance improvement across different speech recognition tasks.	Common Voice, which are large collections of read speech recordings that carry a wide range of speakers, accents, and recording conditions. • This could allow for rich leverage of these datasets, allowing this proposed method to test its effectiveness in learning robust representations of speech and improving recognition accuracy on average.	learning with the RPQ. • This established enhanced robustness to noise and variability in speech with notable margins over the baseline models proposed in this work. • The model also outperformed in terms of efficient computational performance and thus possible for its real-time applications. • Overall, incorporation of random-projection quantization in self-supervised learning helped achieve a state-of-the-art performance on the task of speech recognition.	supervised Learning: It learns from unlabeled data, thus reducing the reliance on large, labeled datasets. • High Performance: Tracked to state-of-the-art performance on speech recognition tasks, many of the time on less labeled data than those training tasks. • Versatile: Works with many different languages as well as low resource settings.	power: Requires significant computational resources for training on large and big datasets.
[7]	Palm: Scaling language modeling with pathways	• The PaLM research of Google studies the efficient scaling of language models using the Pathways architecture. • This model permits a model to scale differently on various tasks and different	• The PaLM model uses a humongous dataset-a combination of high-quality web pages, books, Wikipedia, GitHub, and multilingual corpora-worth around 780 billion tokens. • Diversity in	• By its performance, PaLM achieves state-of-the-art results on a variety of benchmarks concerning both language understanding as well as generation: few-shot learning, reasoning and code generation	• Efficient scalability of models to diverse tasks. • Better utilization of resources through usage of only the relevant portion of the model. • Improved generalization	• High complexity in the implementation and management of the Pathways system. • High computational resources need to be used to train large models. • Potential bias during the

		modalities by allowing parts of the network to specialize and thus only upswing the relevant parts when executing tasks or avoid accomplishing unnecessary computing.	this dataset would ensure good generalization to various domains and languages for the model.	were among the former. • In fact, it was proven to outperform many earlier models, including GPT-3, in many tasks, hence demonstrating better reasoning and comprehension, particularly in few-shot learning scenarios. • It was also shown to perform very well in multilingual environments, displaying versatility on a large range of tasks.	across diverse tasks and domains. • Improved multi-task learning without the need for retraining. • Scalable infrastructure that consumes fewer resources. • Elastic handling of task diversity in a flexible way.	application of large-scale data that was used for training purposes. • High difficulty in fine-tuning for highly specific tasks. • Challenging when interpretable due to the model's size and structure. • Energy consumption issues when scaling up massive models.
[8]	w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training	• Contrasting learning-wav2vec 2.0 and masked language modeling-MLM (as used in BERT - the two approaches are used in the w2v-BERT model). • This is because contrastive learning will enable the model to learn what constitutes similar and dissimilar speech representations. • For its part, MLM predicts masked portions of input speech	• The large-scale speech datasets, which are commonly made up of unlabeled audio data, are used in the training of the model. • Very vital to this pre-training process without manual transcription is the availability of these public corpora, such as LibriSpeech and so many more.	• The obtained results using the w2v-BERT model surpass those by downstream tasks like ASR and speaker identification. • For instance, it achieves substantial improvements in WER and shows better transferability to different speech-related tasks compared to other self-supervised speech models.	• This better represents the speech data by incorporating masked language modeling with contrastive learning. • This also employs self-supervised learning, thereby minimizing the extensive requirement for labeled data, and therefore it is cost-effective as well as scalable. • It obtains	• Complex Implementation: It requires the combination of two learning paradigms, making both design and training complicated. • Extremely Computationally Expensive: It requires a lot of computational resources to train. • Sensitive to Specificity of Training Data: It might degrade performance if the training data is not diversified

		features. • The combination of the two will allow for effective self-supervised learning on raw audio data that will enhance the ability of the model in understanding and generation of speech.			state-of-the-art performances on various speech recognition benchmarks. • This model may be used to support downstream tasks beyond speech recognition. • This enhances improved generalization capability since this technique will learn useful features from diverse data. • This model efficiently uses speech data, leading to faster convergence.	enough. • Sensitive to Hyperparameters Used: The model is sensitive to the choice of hyperparameters. • Limited Explainability: The complexity of this model will make it infeasible to explain the decisions made by the model. • Prone to Overfitting: LIABLE to overfit small samples leading to bad generalization.
[9]	Unsupervised Cross-lingual Representation Learning for Speech Recognition	• This paper proposed an unsupervised cross-lingual representation learning of speech recognition based on self-supervised learning techniques. • The approach was inspired by the shared latent space whereby mapping features of speech in different languages were learned in this space. • A new contrastive learning framework	• The paper works with a variety of speech datasets coming from different languages. • This list includes but is not limited to: LibriSpeech (English), Common Voice (as many languages as possible), VoxForge. • These datasets are characterized by the diversity of speakers, accents, and acoustic environments,	• Thus, the proposed method is tested against some benchmarks for performance evaluation; several improvements in the task of cross-lingual speech recognition are realized. • Results obtained actually show that the unsupervised approach captures cross-lingual representations effectively, and the outcomes	• Cross-lingual learning: It facilitates speech recognition in several languages without having to require labeled samples for every one of those languages. • Data dependency cut down on: Learns good-quality representations from unlabeled multilingual speech data. • Generalization: The low-resource language	• Complex training : Unsupervised cross-lingual training can be computationally expensive. • Requires large datasets: Effective representation learning requires a large amount of unlabeled multilingual speech data.

		helped optimize the model such that it distinguished between similar and dissimilar audio samples with greater accuracy. • Using an enormous amount of speech data without labels, it learns to abstract and generalize phonetics and linguistic information in order to avoid the need for big labeled datasets.	thereby offering a rich spectrum of speech characteristics. • The type of dataset ensures cross-lingual feature learning on the part of the model because it is exposed to wide variations of language and phonetics that would enhance performance in recognizing speech across various languages.	are thus improved for recognition tasks in languages with limited labeled data. • For those models tested on languages unseen during their training, high WER reductions are achieved relative to the traditional methods. • This model generalizes across languages, promising applications in multilingual speech recognition.	benefits from better performance through knowledge transfer from the high-resource languages. • Efficient model: Having fewer resources on languages, it still maintains its high levels of performance..	
[10]	Scaling vision transformers to 22 billion parameters	• Scaling vision transformers up to 22 billion parameters involved a number of very important strategies aimed at keeping efficiency and performance intact. • Sparse mechanisms of attention were engaged in reducing computational costs as well as memory usage that made it possible for the model to enlarge the size of its inputs and	• The model is pre-trained on a vast, diverse dataset built to push its generalization capabilities on many vision tasks. • This dataset combines curated, high-resolution images from public sources, such as ImageNet, with proprietary data holding billions of images. • The aim was to achieve balance in covering as large a domain	• The vision transformer scaled up to 22 billion parameters achieved state-of-the-art performance over various benchmarks. • On ImageNet classification, it has marked new accuracy records with large margins over the previous records. • It showed strong adaptation to downstream tasks such as	• Impressive performance enhancements from previous versions with regard to image recognition and other vision-related tasks. • More advanced capacity for models that capture rich complex visual patterns as well as details. • Improved generalization on a wide and large-scale visual dataset. • Better	• Compute and resource intensiveness: Both training as well as inference are compute-intensive and have high resource requirements. • High memory usage, and hence requires highly advanced hardware. Long training periods due to the size of the model. • Overfitting is more likely to occur when there is limited data and the model has capacity to be

		process them much better. • It also employed gradient checkpointing in order to save memory while in the process of backpropagation. • Model parallelism and pipeline parallelism are used to distribute the work of computation over multiple GPUs. • Custom training loops and memory-efficient techniques made training a model of this size feasible.	as possible, including object categories, visual contexts, and edge cases. • Besides that, pretext tasks such as masked image modeling added to this set proved beneficial for allowing the transformer to learn meaningful visual representations.	object detection, segmentation, and image generation in transfer learning. • Thanks to the optimization of attention mechanisms and computation efficiency, the inference speed was found to be similar to that of much smaller models despite its size. • It marks a key milestone in large-scale vision transformers for scalability and performance balance.	accuracy in fine-grained image classification and detection tasks. • Simplified or enhanced flexibility to handle the wide range of visual applications. • Advanced feature extraction that leads to richer visual representations.	large. • Increased energy consumption and reduced sustainability aspect. • It becomes challenging to manage and deploy the models.

[11]	PaLM-E: An Embodied Multimodal Language Model	• PaLM-E is designed for embodied agents. • It integrates vision with language by processing textual and visual inputs. • It begins at the base of PaLM (Pathways Language Model) but adds the power of visual understanding to it by a pre-trained vision transformer called ViT. • The model may thus enable real-world applications where agents can either interact with objects and environments or make use of perception in combination with language for tasks such as object manipulation and navigation.	• The model was trained with large multimodal data, both containing visual (images and videos) as well as textual (instructions, descriptions) data. • Some such datasets include COCO and the robotic control datasets, used to teach model visual perception and action planning in embodied environments.	• PaLM-E excels in many different multimodal tasks, such as visual question answering and robotic control tasks. • It greatly improves on embodied reasoning and interaction capabilities, performing better than previous models at tasks requiring the understanding of both vision and language. • Generally, it is more generalizable toward new environments and tasks, thereby increasing success rates in robotics and multimodal challenges.	• Multimodal learning: Learn vision, language, and action in parallel, hence effectively interacting with the environment to enhance comprehension. • Embodied understanding: Realistic applications such as robotics and interactive systems are its main focus. • Generalized capabilities: Enables transfer of knowledge across modalities, which enhances its flexibility	• Computationally expensive: Very high computation resources are needed to train large-scale multimodal models. • Adding multiple modalities would add complexity to the architecture of the model.

[12]	Listen, Think, and Understand	• The novelty of the study is in its new approach that involves using audio processing and natural language understanding to enhance tasks of speech recognition. • It follows a multimodal architecture that involves contextual information derived from the text in addition to the audio signals used. • The model enhances input parts' focusing ability by applying a series of attention mechanisms. • It is fully end-to-end trainable architecture, which means that it learns representations that optimize both audio and text inputs towards various tasks of language understanding.	• This paper uses an extremely diverse dataset that actually combines these speech recognition benchmarks with some conversational datasets. • These include recordings from disparate sources, involving some of the public speech corpora, podcasts, and dialogue datasets-thus great linguistics and contextual variability. • The dataset is also supported in terms of transcriptions and contextual tags to support the training and evaluation of the model. • This is a very comprehensive dataset that supports generalization and performance across different audio scenarios and language contexts.	• This is to say, the model executes a number of improvements over baseline approaches in many speech recognition tasks. • It performs better with respect to accuracy rates in transcription tasks and shows robustness in understanding conversational context as against the state-of-art models. The WER and F1 scores' evaluation metrics indeed indicate an improvement quite marked in the comprehension and transcription accuracy, thereby well validating the multimodal approach proposed. • The results show that being combined with contextual understanding, audio performs better in real-world applications.	• Better pronunciation recognition, with a better integration of listening and reasoning processes. • Uses a new framework, hence better enhancing the model to understand context and semantics. • Has considerable improvements in performance for speech understanding benchmarks. • Combines different modalities; hence it offers holistic understanding of spoken language. • Facilitates transfer learning. Thus, it becomes adaptable to various tasks and domains. • The new interaction model for users becomes quite intuitive in real-time applications.	• It is inordinately data-hungry and can be a drawback for being used in resource-poor settings. • The architecture complexity makes it more challenging to deploy and scale. • Computationally intensive with the need of hardware resources on a large scale. • Really prone to overfitting unless regularized appropriately. Since lots of scarce data are available, this can cause a major problem. • Noisy or very diverse speech inputs may cause issues in robustness. • The model is hard to interpret. This means it is not easy to comprehend what the model is deciding

[13]	Parameter-Efficient Transfer Learning for NLP	• It is Parameter-Efficient Transfer Learning (PETL) for NLP fine-tunes pre-trained language models like BERT, GPT, or T5 by adjusting a small subset of their parameters, instead of the whole model. • Some of the popular techniques have been Adapters, Low-Rank Adaptation, and Prompt Tuning. • Adapters add a few layers between existing layers for task-specificity, whereas LoRA injects low-rank matrices to reduce the trainable parameters in the weights. • These methods retain the benefits of large pre-trained models but make fine-tuning computationally more efficient, particularly in the case of multiple downstream tasks.	• Different PETL approaches have been put to the test on numerous general-purpose NLP benchmarks: both GLUE (General Language Understanding Evaluation) and SuperGLUE, SQuAD (Stanford Question Answering Dataset), and task-specific datasets based on the direction of the research in focus. • The corpus collections comprise a variety of tasks that are related to language, including sentence classification, textual entailment, and question answering, among other kinds of tasks. • Testing diverse datasets will be useful to evaluate if PETL methods have generalizability across these different types of NLP tasks.	• These methods of PETL have been proved to be competitive with full fine-tuning: in fact, on some benchmarks they reached the same performance or even surpass the performance of full fine-tuning while having an order of magnitude smaller number of parameters. • For instance, LoRA and Adapters achieved state-of-the-art performance on the GLUE benchmark with more than a 90% reduction in trainable parameters. • Thus, PETL is almost highly efficient for real-world applications where resources of computation and memory are usually limited.	• The efficiency of transfer learning: It highly decreases the trainable parameters of the NLP models. • Flexibility is permitted: Only a small fraction of the whole model can adapt to downstream tasks without fine-tuning. • This reduces the computational cost as it only edits a small subset of parameters. • Scalability-It allows scalable deployment across multiple NLP tasks.	• Poor optimization: The method is unlikely to reach the full exploitation of the achievable advantages in performance by fine-tuning all the parameters. • Task-specific limitations: Quality may be very sensitive to the nature of the downstream tasks.

[14]	LoRA: Low-Rank Adaptation of Large Language Models	• The LoRA method injects low rank trainable matrices into the layers of the pre-trained model to fine-tune it more efficiently. • Instead of updating every model parameter, only a small, low-rank subset can be updated, where the computational cost and the memory requirements are drastically reduced. • Adaptation is faster, yet models perform well under such methods.	• To evaluate LoRA, standard datasets for NLP tasks are used, such as those utilized for the GLUE, SQuAD, and MT datasets for translation purposes. • These are standard benchmarks for text classification tasks and question answering as well as machine translation, thus offering a number of language situations for testing the efficiency of LoRA. .	• That means that LoRA significantly reduces the memory and the computational requirements in comparison to full fine-tuning while being similarly, or even better, performing in comparison to the latter. • In benchmark tasks, it has less overhead than the other two methods and is therefore attractive for adapting large language models like GPT and BERT to specific tasks, considering it is more capable of being used in resource-constrained environments.	• Lower computational cost for fine-tuning large models. • Memory usage is also reduced compared to full updates of the model. • There are fewer parameters to adapt to new tasks. • Fewer time demands during training and deployment. • Strong performance remains with reduced resource requirements. • The large model can be adapted to all kinds of different domains.	• Constraining the rank to be low may depress performance on difficult tasks. • Less expressiveness than fully ranked models. • Also complicates tuning and optimizing low-rank adaptations. • Accuracy as well as generalization could also have trade-offs. •Implementations of low-rank adaptation algorithms are also difficult.

[15]	Improving fast-slow Encoder based Transducer with Streaming Deliberation	• A fast-slow encoder-based transducer architecture is reported along with improvement of a streaming deliberation mechanism. • In this method, two types of encoders are developed: one is a real-time input processing by the model, while the other is a slow encoder for more context and information to enhance the output. • The fast encoder will determine the preliminary prediction, while the slow encoder will refine the prediction based on more context and information. • The deliberation process would allow for better output from both encoders by relying on the strengths of both and ensuring that the result of the final process benefits both speed and accuracy. • This improves transducer performance in speech recognition and language translation.	• The paper executes experiments using several benchmark datasets to show the performance of the presented method. • Among those, LibriSpeech is a typical dataset for speech recognition which was also used with recordings of audiobooks in this work. • Other datasets that might have been used are Common Voice or TIMIT, which are models both for phoneme-recognition and language models. • These datasets provide an all-around performance assessment in terms of linguistic features and acoustic environments in view of the generalizability and robustness of the model in real-world applications.	• The experimental results of the fast-slow encoder-based transducer with deliberation in stream show improvements over baselines. • Word error rates and overall accuracy of the tested dataset set are improved experiments. • Having applied streaming deliberation over it, the model becomes more robust, particularly in processing complex inputs with high precision and reduced errors. • The research could also mention how high efficiency is possible through its real-time applications, providing it could be able to get high throughput without giving up on precise predictions. • The architecture enhances transducer-based speech processing and language understanding.	• Improved accuracy of speech recognition through the refinement of the transducer model. • Streaming deliberation for a real time processing to make the systems more responsive. • Better strategies for deliberation are capable of improving the management of complicated inputs. • Speech-to-text, along with other applications, results in minimizing latency. • More robust noisy or variable input conditions.	• Increased computational complexity due to the additional deliberation process. • Increased latency in processing which might reflect on the real-time performance. • Increased consumption of resources for training and inference. • Implementation and integration issues: Complex and challenging. • There might be a potential danger of degraded performance if the deliberation process is not optimized well.

[16]	SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing	• This paper on SentencePiece introduces a simple sub word tokenization algorithm with regard to the language. • The method is based on unsupervised learning; it can be applied by either unigram language modeling or byte-pair encoding, which is often used as a method of tokenization. • Input text in SentencePiece is assumed to be a sequence of raw bytes, without any assumption of linguistic boundaries, such as in words or phrases, making it effective for languages where word delimiters cannot be defined clearly. • It produces a vocabulary of sub word units, making it easier to address rare words by splitting them into sub words that are shorter and thus probably more frequent, yet still can be combined to make common words.	• Although it does not give particular data sets, this paper notes that SentencePiece has been adapted for various NLP tasks, particularly in NMT. • It has been bench-marked on a multilingual version of several data sets from the WMT and other typical diverse text corpora used in machine translation as well as in language modeling. • Such data sets include different languages and more importantly scripts besides different structures thus making any practical application effective, which is efficient hence represents the flexibility of a SentencePiece.	• In many ways, it integrated an important efficiency and performance improvement into the neural text processing model. • It paved the way to make NMT easier, as well as other NLP tasks, by bringing ease in handling low-resource languages and by reducing many issues related to out-of-vocabulary words. • It also triggered better generalization in downstream tasks and less complexity in preprocessing multilingual data. • It was performing on a par with or better than the traditional tokenizers and was an extremely preferred solution for modern NLP pipelines, including Google's transformer models and multilingual language models like mBERT.	• Language-independent: Can be applied to any language in which no language-dependent rules are used. • Sub word tokenization: Rare and out-of-vocabulary words are processed by breaking them down into semantic sub word units. • Efficiency: Vocabulary size is reduced, which makes their process on neural text processing faster. • Unsupervised Training: No need to train with an already defined token dictionary.	• Potential loss of semantics: The sub word tokenizers tokenize words in such a way that meanings get lost. • Much harder to read like a human: Outputs can be less understandable than full-word tokenizers.

3.15 Faster, more accurate speech recognition

This fast-slow encoder framework provides better accuracy and speed for real-time speech recognition while improving the low latency for ideal applications in streaming transcription.

3.16 Efficient tokenization across languages

It managed to handle sub words in more than one language and was not particularly dependent on particular rules of any language, thus enhancing the performance of the model in processing texts into more than one language and ensuring that the good generalization of the network was across the language barrier.

4. CONCLUSION

This research paper focuses on key advances in speech, language, and multimodal learning with the goals of reducing reliance on labeled data, making models more deployable at scale, and furthering cross-lingual and multimodal capabilities. Techniques like Wav2Vec 2.0 and self-supervised cross-lingual learning have repositioned speech recognition so that such robust representations can be learnt with very little in the way of labeled data. Innovations, such as parameter-efficient transfer learning and low-rank adaptation, have established that efficiency can be achieved in fine-tuning large language models without losing any quality and even saving on computations. Other solutions like the fast-slow encoder also improved real-time speech transduction; thus, they are well suited for live applications.

Contributions like Vicuna opened the access of conversational AI systems to open source while providing equal performance as if the person was accessing the proprietary system. Multimodal learning approaches with vision, language, and action opened the pathway of robotics and real-world interaction. Lastly, cross-lingual processing improves the efficiency of handling rare or unknown words across languages using tokenization systems that are not dependent upon the languages used.

Summary cumulatively, such novelties are pushing progress in speech and language processing towards making such efficient, scalable, and adaptable AI systems that handle diverse real-world tasks.

ACKNOWLEDGEMENT

We would like to place our gratitude on record before others for the opportunity and resources given to us by Reverie Language Technologies Limited, by which we could work out this survey paper. Your support has been invaluable throughout this research.

We would like to thank our guides, Dr. Saswati Rabha and Mr. Chintan Parikh, for the continuous guidance and encouragement in the form of meaningful feedback. At the right moments, expertise and advice helped shape the research direction, so all thanks for the support that we could enjoy.

Further Thanks and Appreciation: We would like to extend our thanks and appreciation to Dr. Geetanjali Kale, Head of the Department at SCTR'S Pune Institute of Computer Technology, for guidance and constant inspiration to create an academic environment of growth and allowed this project to blossom.

Last but not the least, thanks to our project mentor, Dr. Arati Deshpande, for her patience and knowledge-cum-guidance at every stage of this project. Your commitment to our success has been really inspiring, and we thank you.

Thanks to all of you for your contribution toward making this survey paper a rewarding and fulfilling exercise.

REFERENCES

[1] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. “XLS-R: Self-supervised cross-lingual speech representation learning at scale”. In: arXiv preprint arXiv:2111.09296 (2021).

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. “Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020.

[3] Junwen Bai, Bo Li, Yu Zhang, Ankur Bapna, Nikhil Siddhartha, Khe Chai Sim, and Tara N. Sainath. “Joint Unsupervised and Supervised Training for Multilingual ASR”. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. “Language Models are Few-Shot Learners”. In: Advances in Neural Information Processing Systems. 2020.

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. “Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90\%* ChatGPT Quality”. In: (2023).

[6] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. “Self-supervised learning with random projection quantizer for speech recognition”. In:

Proceedings of the 39th International Conference on Machine Learning. 2022.

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. "Palm: Scaling language modeling with pathways". In: arXiv preprint arXiv:2204.02311 (2022).

[8] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. "w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training". In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2021

[9] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. "Unsupervised Cross-lingual Representation Learning for Speech Recognition". In: Interspeech. 2021.

[10] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. "Scaling vision transformers to 22 billion parameters". In: arXiv preprint arXiv:2302.05442 (2023).

[11] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. "Palm-e: An embodied multimodal language model". In: arXiv preprint arXiv:2303.03378 (2023).

[12] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. "Listen, Think, and Understand". In: arXiv preprint arXiv:2305.10790 (2023).

[13] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. "Parameter-Efficient Transfer Learning for NLP". In: Proceedings of the 36th International Conference on Machine Learning. Vol. 97. 2019.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-Rank Adaptation of Large Language Models". In: International Conference on Learning Representations. 2022.

[15] Taku Kudo and John Richardson. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018.

[16] Ke Li, Jay Mahadeokar, Jinxi Guo, Yangyang Shi, Gil Keren, Ozlem Kalinli, Michael L. Seltzer, and Duc Le. "Improving fast-slow Encoder based Transducer with Streaming Deliberation". In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023.