RESEARCH
O&G Forum 2024; 34 – 2s: 317 - 328

A TRANSFORMER APPROACH TO BILINGUAL AUTOMATED SPEECH RECOGNITION USING CODE-SWITCHED SPEECH

Dr. Puspita Dash[1], Sruthi Babu[2], Logeswari Singaravel[3], Devadarshini Balasubramanian[4]
[1]Department of Information Technology,Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry. puspitadashit@smvec.ac.in,
[2]Department of Information Technology,Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry. sruthib274@gmail.com,
[3]Department of Information Technology,Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry. singaravellogeswari@gmail.com,
[4]Department of Information Technology,Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry. devadarshiniit@smvec.ac.in

Abstract
In a bilingual and linguistically diverse country like India, where a significant portion of the population is fluent in multiple languages, the conventional bilingual Transformer neural network architecture faces challenges in accurately translating conversations that seamlessly switch between different languages. This limitation in translation accuracy underscores the need for more sophisticated language models. The proposed solution involves leveraging the Generative Pre- Trained Transformer (GPT) model, a powerful deep learning architecture within the transformer framework. Trained in an unsupervised manner on extensive text data, the GPT model demonstrates enhanced language understanding and generation capabilities. By pre-training on a diverse dataset, GPT learns to capture the intricacies of syntax, semantics, and contextual nuances, enabling it to accurately predict and generate coherent text. We experimented on Tamil-English data and found that the Generative Pre-Trained Transformer model can achieve an 84.37% relative accuracy rate even for short sentences and 73.98% relative accuracy rate for lengthy sentences in bilingual ASR performance. The adaptability of GPT to various downstream tasks, context-aware approach to language processing in linguistically diverse environments.
Index Terms—Generative Artificial Intelligence, Speech Recognition, Generative Pre- Trained Transformer (GPT), Bilingual.

I. INTRODUCTION

The cornerstone for the language understanding of the Generative Pre-Trained Transformer (GPT) model is the wide and varied multilingual content present on the internet, from which it receives its training data. Pre-training and fine-tuning are the two phases of the GPT working model. It gains knowledge of language structure and semantics during the pre-training phase by learning to predict the next word in a sentence. Using a smaller dataset, the fine-tuning phase adjusts the model to certain tasks. GPT is a flexible and potent tool in the field of natural language processing because of its wide range of applications, which include chatbots, content creation, sentiment analysis, language translation, and code development. Automated Speech Recognition System also known as ARS, which analyzes and converts the human speech(audio) into the written text/audio in English or even any other language without any grammatical error and cover all the intra- sentential words like 'the', 'in', 'of' etc. and also covers the intra code-word switching which we can find in Tanglish. So this system effectively converts the Tanglish to English or even any other native language into the common language English without any grammatical error. It is done with the help of Generative AI. In Generative AI we use Generative pre-trained Transformer model which collects its data source from the vast internet by the means of web scraping. Thus makes this system more robust and efficient.

GENERATIVE PRE-TRAINED TRANSFORMER MODEL:

The Generative Pre-Trained Transformer (GPT) model derives its strength from the extensive and diverse corpus of multilingual

content sourced from the vast expanse of the internet. This diverse dataset serves as the bedrock for GPT's remarkable language understanding capabilities. The model operates through a two-phase process: pre-training and fine- tuning. In the pre-training phase, GPT learns to predict the next word in a sentence, enabling it to grasp the intricacies of language structure and semantics. This foundational knowledge is then fine-tuned in thesecond phase using a more specific dataset, tailoring the model to perform particular tasks. GPT's versatility is evident in its proficiency across a spectrum of natural language processing tasks, including but not limited to chatbot interactions, content creation, sentiment analysis, language translation, and even code generation. This adaptability renders GPT a powerful and flexible tool with widespread applications in the rapidly evolving landscape of natural language understanding and generation.

AUTOMATED SPEECH RECOGNITION:

Automatic Speech Recognition (ASR) is atransformative technology designed to convert spoken language into written text through the utilization of sophisticated algorithms and machine learning techniques. This innovation has found wide-ranging applications across diverse fields, such as transcription services, voice assistants, and speech analytics. ASR systems operate by analyzing audio input and transcribing it into a comprehensible text format, enabling seamless communication and interaction with technology. The evolution of ASR technology inrecent years has played a pivotal role in the development of voice-controlled devices and services, enhancing user experience and accessibility. Its ability to accurately and swiftlytranscribe spoken words has made ASR a cornerstone in the advancement of human - machine interaction, marking it as an indispensable tool in our increasingly interconnected and technologically driven world.

BILINGUAL:

Bilingual refers to the ability of a system orindividual to understand and communicate inmultiple languages. In the context of technology,a bilingual system can process and interpret diverse languages, accommodating a global user base. This capability is particularly crucial in applications like machine translation, voice assistants, and natural language processing, where interactions occur in various languages. The development of multilingual technologies involves complex algorithms and linguisticmodels that enable seamless comprehension andgeneration of content in different languages. As our world becomes more interconnected, the demand for bilingual solutions continues to grow, fostering inclusivity and enhancing communication across linguistic barriers.

MECHANISM OF GENERATIVE PRE-TRAINED TRANSFORMER MODEL:

The Generative Pre-Trained Transformer (GPT) model operates on a mechanism known as unsupervised learning to generate human-like text. Initially, the model undergoes pre- training on a vast dataset containing parts of the Internet to learn the patterns, structures, and context of language. The transformer architecture, with attention mechanisms and positional encoding, plays a pivotal role in capturing long- rangedependencies and contextual information. During pre- training, the model learns to predict the next word in a sentence, creating a language model that understands syntax, semantics, and context. Following pre-training, the model is optimized for particular tasks. or domains using supervised learning. This fine-tuning enhances its ability to generate contextually relevant and coherent text for particular applications such as language translation, summarization, or question- answering. The model's success lies inits capacity to generalize from the diverse data encountered during pre-training, enabling it to adapt and perform well on a variety of downstream tasks. The generative nature of GPTallows it to autonomously create coherent and contextually appropriate text by sampling from its learned distribution of language patterns. Themodel can assess the relative relevance of variouswords in a phrase thanks to the self- attention
mechanism, ensuring it captures nuanced relationships and produces contextually rich outputs. Overall, the GPT model showcases the power of unsupervised learning and transformer architectures in capturing the intricacies of natural language, making its versatile tool invarious natural language processing applications. The Figure 1 shows the GPT Architecture.

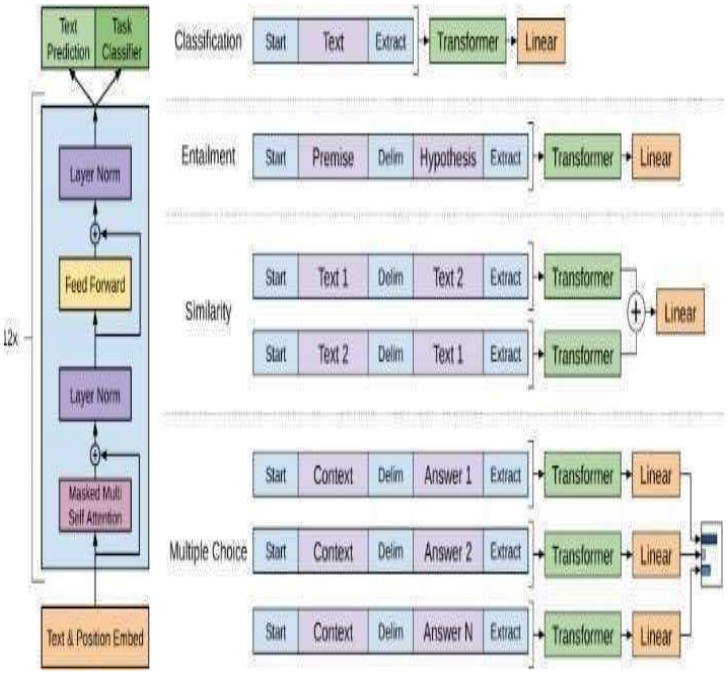

Fig. 1 GPT Architecture

OBJECTIVE OF THE PROJECT

This study's primary objective is to address the challenges that code-switching (CS) presents in multilingual settings. Specifically, it focuses on speech recognition between Tamil and English utilizing. CS in low-resource environments. The project attempts to use Transformer-based frameworks to improve the effectiveness of automated speech recognition (ASR) tasks inlight of the lack of data for Indic CS voice recognition. The study specifically looks at two approaches: using language information as tokens at the targets and using well-trained encoders of Monolingual Transformers as feature extractors for efficient language discrimination. Despite the scarcity of training data, the objectiveis to offer a strong solution for managing CS in speech, with an emphasis on both language discrimination and enhancing overall recognition accuracy.

II. LITERATURE SURVEY

Neural Networks for Bilingual Machine Translation Model

In this paper [1], machine translation can be used in statistical-based, corpus-based, or dataset- based machine translation systems, in addition tolinguistic systems. The goal of this work is to create a high- quality, multilingual English to Arabic translation model that can be easily upgraded to include more language pairs. Additionally, this will build an integratedtranslation environment with computer-assisted tools to improve the caliber of texts generated automatically, boost productivity, and support translators' professional development. Consequently, neural network- based machine translation model will be created. After non- alphanumeric texts are cleaned and removed utilizing linguistic modification tasks for the proposed machine translation model, bilingual dictionaries will be engaged. Thus, for such machine translation, encoder and decoder models are involved.

Bilingual End-to-End ASR with Byte-Level Subwords

Applying Machine Learning and Natural Language Processing Techniques The document[2]. Bilingual End-to End ASR with Byte-Level Sub word examines the impact of an end-to- end neural network's output representation on multilingual automated speech recognition(ASR) [2]. Character-level, byte- level, byte pair encoding (BPE), and byte-level BPE (BBPE) representations are among the various representations that the writers examine. They concentrate on creating a single end- to- end model to facilitate utterance-based bilingual ASR, in which speakers may switch betweenlanguages throughout an utterance rather than alternating between two languages in a single utterance. The authors find that BBPE representation with length and alphabet penalty can, even with fewer outputs and parameters, increase the performance of utterance-based multilingual ASR by 2% to 5% relative. They propose two penalty strategies to modify the bigram statistics that the BPE algorithm employs. The first is the lengthpenalty, which lowers the likelihood of creating erroneous byte sequences, promotes the generation of more single-charactersymbols, and inhibits the BPE algorithm from producing multi-character symbols. Alphabetpenalty is the second scheme of penalties that penalizes alphabetic bigrams to omit any Englishsub words that appear in the Mandarin text. The Mandarin BBPE symbols will get the saved space. They believe. They believe that this approach can be further improved by exploring alternative byte-level representations, as well as the choice of languages.

Bilingual Automatic Speech Recognition A Review Taxonomy and Open Challenges

This paper proposes Bilingual Automatic SpeechRecognition A Review Taxonomy and Open Challenges Bilingual Automatic Speech Recognition (ASR) is a subfield of ASR that focuses on recognizing speech that contains multiple languages [3]. It is a challenging task due to the complexity of dealing with the linguistic and acoustic variations between languages. Bilingual ASR systems can be classified into two main types: code-switching and code-mixing. Code-switching is the phenomenon ofswitching between two or more languages within a single utterance. Code-mixing, on the other hand, isthe phenomenon of using words or phrases from one language withinthe context of another language. The taxonomy of bilingual ASR systems is based on three main factors: language dependency, phone set, and interacting languages. Language dependencyrefers to whether the ASR system is designed to handle single language or multiple languages. Language independence refers to whether the ASR system uses single phone set or multiple phone sets. Interacting languages refer to thelanguages that are being mixed or switched in the speech signal.

Joint Pre-Training with Speech and BilingualText for Direct Speech to Speech Translation

This paper [4] proposes The phenomenon of switching between two or more languages in a single speech is known as an code-switching. conversely code existing is the practice of use terms within the context of another language. The speech-to-units prediction task is used to train the speech encoder and unit encoder to predict the correct units given the input speech. The source-to-target unit's translation task is used to train the unit encoder and unit decoder to translate units from the source language to the target language. The proposed method is evaluated on two S2ST datasets: VoxPopuli and Europarl-ST. The results show that Speech2S outperforms previous S2ST models that are pre-trained on speech data only. The paper presents a promising approach for improving the performance ofS2ST models. The combination of speech and bilingual text pre- training and unit-based HiFi-GAN fine- tuning results in significant improvements over previousmethods.

Code-switching text generation and injection in mandarin-english asr

This work [5] they have discussed code- switching speech recognition which refers to the mixing oftwo or more languages within a single utterance. Code-switching text generationand injection in mandarin-english asr The lack ofdata makes this Using text creation and injectionto boost the Transformer-Transducer (T-T), a popular streaming model performance in Mandarin-English code-switching speech recognition is a difficult issue for Automatic Speech Recognition (ASR). A machine translation model is used to create parallel Mandarin and English phrase pairs. Word alignment is then carried out to align terms that have the same meaning in both languages. They have employed the Text-to-Speech (TTS) method. The generated text is turned into speech using a multilingual text-to-speech system. The paired speech-text training data is supplementedwith the TTS-converted data. Another method iscalled Cross-Modality Learning (CML), in whichthe speech and text latent spaces are connected to introduce text-only data into the T-T model. Mean Squared Error (MSE) loss is one technique used to achieve this.

Data augmentation for end-to-endcode-switching speech recognition

This paper [6] they have discussed raining an end-to-end code-switching model for automated speech recognition (ASR) due to the limited availability of code-switching data The authors propose Three new methods for augmenting code- switching data. words-to-speech (TTS) with word translation, audio splicing, and TTS with word insertion. To create new code- switching data, the audio splicing method consists of splicing together language-dependent fragments from several utterances. Based on alignments between transcripts and auditory features, the segments are chosen. Transcription ofwords using TTS. In this approach, monolingual text is translated into code-switching text using a Mandarin- English dictionary. The translated text is then synthesized into speech using a text-to-speech (TTS) system. Data augmentation techniques aim to increase the diversity of code- switching data on both acoustic and linguistic aspects, thereby improving the performance of code- switching ASR.

Automatic speech recognition of Portuguese phonemes using neuralnetworks ensemble

This paper explains **[7]** titled presents an approach for Portuguese phoneme automated speech recognition (ASR) employing a collection of neural networks [7]. Pre-processing and classification are the two phases that make up thesuggested system. The speech signal ispreprocessed in the pre- processing stage in order to extract pertinent features, such as the Mel- frequency central coefficients (MFCCs). whichrepresent the spectral characteristics of thesignal. In the classification stage, the extracted features are classified using an ensemble of neural networks. The ensemble consists of two sub- ensembles: The Phonetic Expert Model (PEM) and the Class Hierarchy Expert Model (CHEM). The PEM is based on phonetic clusters, while the CHEM is founded on the phonetic classimbalance present in the training set. With an accuracy of 86.525%, the suggested method outperforms the best previously documented system by 7.63%.

Self-Supervised Pre- Training for Attention-Based Encoder-Decoder ASR Model

The paper titled **[8]** proposes a novel for the attention-based encoder-decoder (AED) ASR model (SP- AED), a self-supervised pre-training technique is used [8]. An adaptive combination fine-tuning for the encoder, linguistic pre-trainingfor the decoder, and acoustic pre-training for theencoder comprise the SP-AED approach. The entire system. The unpaired text data is used by the SP-AED approach for the decoder pre- training. By substituting random noise for the actual auditory representations, the decoder is pre- trained as a noise-condition language model. The wav2vec2.0 pre-training approach is used by the SP-AED method for the encoder pre-training. By masking the input acoustic features and pre- training the encoder to reconstruct the input, the wav2vec2.0 approach is a mask-based pre- training technique. Following completion of all pre-training, an adaptive combination fine-tuningmethod is used to combine and fine- tune the encoder and decoder. According to the experimental findings, there can be a relativeimprovement of up to 17% between the random initialized models and the SP-AED pre- trained models. Additionally, we can achieve results on both the English and Chinese corpus that are comparable to other classification-based models with a similar model size or computational cost.

Non- Autoregressive ASR Modeling Using Pre- Trained Language Models for Chinese Speech Recognition

The paper [9] proposes a non-autoregressive method based on pre- trained language models (PLMs) for automated speech recognition (ASR). Traditional ASR models rely on autoregressive decoding, which processes the speech signal one frame at a time, predicting the next word based on the previously predicted words. However, autoregressive decoding can be slow and inefficient, especially for long utterances. Non-autoregressive ASR models, on the other hand, process the entire speech signal at once, using a PLM to predict all the words in the utterance simultaneously. This can be much faster than autoregressive decoding, but it canalso be more challenging to train the PLM to generate accurate predictions without knowing the order of the words. The authors of the paper propose a two-stage training approach for their non- autoregressive ASR model. The first stage involves pre-training a PLM on a large corpus of Chinese text data. The second stage involvesfine-tuning the PLM on a smaller corpus of Chinese speech data, using a non-autoregressiveloss function that encourages the PLM to predict the correct word order. The authors evaluate their non-autoregressive ASR model on a benchmark Chinese speech recognition task and find thatit achieves competitive results with state-of- the-art autoregressive models. This suggests that

non-autoregressive ASR models have the potential to be a viable alternative to traditional autoregressive models.

Learning Deep Generative Clustering viaMutual Information Maximization

Learning Mutual Information [10] Maximizationfor Deep Generative Clustering The title suggests that the paper focuses on a machine learning approach that combines deep generative models with clustering techniques. In this context, "deep generative clustering" likely refers to a method that involves learning representation using deep generative models (generative networks, varitional auto encoders while simultaneously incorporating clustering principles. The term "Mutual Information Maximization" deep generative models (such as variationalauto encoders or generative adversarial networks)while simultaneously incorporating clustering principles. The term "Mutual Information Maximization" indicates that the learning process involves optimizing the mutual information between relevant components of the model. Mutual information measures the dependence between two variables, and maximizing it often serves as an objective to improve the efficiency and effectiveness of learning algorithms method that integrates deep generative models with clustering strategies, emphasizing the role of mutual information maximization to enhance the learning process. The goal is to achieve more effective and meaningful clustering in a deep generativeframework.

In this work, a deep clustering technique combining discriminative and generative models is proposed. The procedure uses a discriminativemodel to classify the data points after they are generated by a generative model. A variational auto encoder (VAE), a kind of neural network that can first decode latent representation back into data and then encode data again, is what thegenerative model is. Convolutional neural networks (CNNs), a particular kind of neural network that works well for image classification,are the basis of the discriminative model. To enhance the performance of clustering, the two models are trained in tandem. It is demonstratedthat the suggested approach beats the most advanced deep clustering techniques on threebenchmark datasets.

A Research On the New Generation Artificial Intelligence Technology Generative Pretraining Transformer ArtificialIntelligence (AI)

This paper [11] facilitates the automation of repetitive tasks performed by humans in the digital age, hence improving living standards. These days, artificial intelligence and its applications permeate every facet of daily life, including energy, transportation, health, education, and tourism as well as agriculture. AI applications are moving quickly toward significant and recent advancements, particularlyin Deep Learning (DL) and Natural Language Processing (NLP). This paper shows [11] the language model known as Generative Pre-Trained Transformer 3 is a specific illustration ofadvancements in these fields. The DL model, which has been effectively implemented in various NLP fields is AI-assisted GPT-3 technology, which can generate lengthy and coherent text that resembles texts created by people with the aid of trained algorithms. By employing the Transformer- based language model, a deep learning technique that relies on attention, the GPT-3 architecture has advanced tothe point where it can outperform humans in numerous domains and generate the best possible answers for a wide range of inputs. The purpose of this essay is to explain to the reader the effectiveness, architecture, and potential of the GPT-3 model with assistance from Transformers, one of the newest NLP technologies.

Design and Construction of a Knowledge Database for Learning Japanese Grammar Using Natural Language Processing and Machine Learning Techniques

Using machine learning and natural language processing techniques, a knowledge database forlearning Japanese grammar is designed and constructed. The title indicates a project that involves creating a knowledge database specifically focused on Japanese grammar using machine learning (ML) and natural language processing (NLP) methods. [12] The subject of "Design and Construction" aspect suggests a comprehensive effort to build a structured and organized database, likely with the intention of facilitating effective learning of Japanese grammar. The incorporation of "Natural Language Processing"suggests the use of computational methods to analyze and understand the complexities of the Japanese language. This could involve techniquessuch as Syntactic parsing, semantic analysis, andpart-of-speech tagging to process and structure information about Japanese grammar. In order to understand Japanese sentence patterns, the study suggests building a knowledge database using machine learning and natural language processing(NLP) methods. Learners of Japanese as a secondlanguage (JSL) can utilize this grammatical knowledge database to prepare for the Japanese Language Proficiency Test (JLPT) by studying Japanese sentence patterns. In order to automatically identify Japanese language patterns from input example sentences, the study suggests a method that combines manual rules with the Conditional Random Fields (CRF) machine learning algorithm. Using NLP and machine learning approaches on the design and building of the Japanese grammatical knowledge database, several tests were carried out, and the experimental findings proved the validity of the proposed methodologies.

An End-to-End Chinese and Japanese Bilingual Speech Recognition Systems withShared Character Decomposition

This paper [13] is entitled It presents a novelbilingual speech recognition system based on a for Chinese and Japanese decomposition strategythat decomposes characters into radicals. The authors propose to decompose Chinese characters into radicals and Japanese Kanji into radicals. This approach not only makes the vocabulary smaller, but it also makes it possible for the two languages to gain from having comparable sub-character level representations during joint training. The paper describes a

multilingual ASR transformer model that is used to train the bilingual systems. The model is based on the attention-based Transformer model. According to the findings, the radical- based bilingual system performs better than both the monolingualsystems and the bilingual system that does not use decomposition. The authors argue that this is because the radical- based system is able to bettercapture the shared phonetic and semantic information between Chinese and Japanese. Theauthors argue that this makes the radical-based system a promising approach for bilingual speech recognition, especially for languages with large vocabularies. In order to automatically identify Japanese language patterns from input example sentences, the study suggests a method that combines manual rules with the Conditional Random Fields (CRF) machine learning algorithm. Using NLP and machine learning approaches on the design and building of the Japanese grammatical knowledge database, several tests were carried out, and the experimental findings proved the validity of the proposed methodologies. This database can beviewed as both an additional useful resource for learning Japanese grammar and an advancement in the efficacy of the already-existing intelligent computer-assisted Japanese language learning system.

Non-Autoregressive Fully Parallel Deep Convolutional Neural speech synthesis

The paper [14] indicates a focused approach to speech synthesis this popular method involves sequence to sequence structure and this method isuses an encoder to understand the language and decode to predict sounds. The traditional methodof speech synthesis involves multiple components and can be slow and prone to errors. The proposed approach eliminates the autoregressive flow and uses a non-autoregressive framework for faster speech synthesis they use another proposed model time varying meta template (TVMT) as decodersinput. The TVMT is represented using a differentconditional distribution and contains sophisticated temporal information. multiple decoders are used in an iterative process which isconverted into spectral features faster and makes them more flexible. This approaches aimsto improve comparison between voice quality andsynthesis speed to traditional models. In this research, a non-autoregressive ASR (NAR-ASR) model for Chinese voice recognition based on pre-trained language models is proposed. An audio encoder and a text generating decoder make up the model. The voice-transformer and LASO-based acoustic encoder takes speech input and extracts high- level acoustic representations from it. The text generation decoder creates the matching text sequence forthe input voice signalbased on a pre-trained language model, like BERT. AISHELL-1 and AISHELL- 2, two freely accessible Chinese speech corpora, are used to train the suggested model. The outcomes ofthe experiments indicate that the suggested model outperforms cutting-edge NAR-ASR models on both datasets, achieving either comparable or better results.

A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning

An Initial Fine-Tuning, Assessment Methodology, and Text-to-Speech Pipeline. Thetitle signifies a comprehensive exploration and development effort in the domain of text-to- speech (TTS) systems. [15] The term "Pipeline" suggests a systematic sequence of processes or components designed for converting text into speech. The inclusion of "Evaluation Methodology" indicates structured approach for assessing and measuring the performance of the TTS system, likely encompassing metrics such as speech quality, naturalness, and intelligibility. The phrase "Initial Fine-Tuning" implies an earlyphase in the optimization process where the TTSmodel is refined or adjusted to enhance its performance. Overall, the title suggests the creation of a complete TTS system, accompanied by a rigorous evaluation strategy and an initial refinement phase to ensure the system's effectiveness and quality. In this research, a training pipeline using kid speech datasets is proposed for optimizing state-of-the-art (SOTA) neural TTS models. The method uses a multi- speaker TTS retuning technique to provide apipeline for transfer learning. The fine-tuningtests were based on a smaller selection of about 19 hours from a publicly available kid speech dataset that had been cleaned. A unique subjective framework for mean opinion score (MOS) evaluations and a pretrained MOSNet for objective evaluations were used for both subjective and objective evaluations. Subjective evaluations yielded MOS values of 3.96 for voice consistency, 3.89 for voice naturalness, and 3.95 for speech intelligibility. Real and synthetic childsounds showed a high association in an objectiveexamination using a pretrained MOSNet. A further method of verifying speaker similarity was to compute the cosine similarity between the speech embeddings. The genuine and synthetic child voices were compared for word error rate (WER) using an automated speech recognition(ASR) model. In as little as five seconds, the finished trained TTS modelwas able to generate speech like that of a child using reference audio samples.

A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning

An Initial Fine-Tuning, Assessment Methodology, and Text-to-Speech Pipeline. Thetitle signifies a comprehensive exploration and development effort in the domain of text-to- speech (TTS) systems. [15] The term "Pipeline" suggests a systematic sequence of processes or components designed for converting text into speech. The inclusion of "Evaluation Methodology" indicates structured approach for assessingand measuring the performance of the TTS system, likely encompassing metrics such as speech quality, naturalness, and intelligibility. The phrase "Initial Fine Tuning" implies an early phase in theoptimization process where the TTS model is refined or adjusted to enhance its performance. Overall, the title suggests the creation of a complete TTS system, accompanied by a rigorous evaluation strategy and an initial

refinement phase to ensure the system's effectiveness and quality. In this research, a training pipeline using kid speech datasets is proposed for optimizing state-of-the-art (SOTA) neural TTS models. Themethod uses a multi-speaker TTS retuningtechnique to provide a pipeline for transfer learning. The fine-tuning tests were based on a smaller selection of about 19 hours from a publicly available kid speech dataset that had been cleaned. A unique subjective framework for mean opinion score (MOS) evaluations and a pretrained MOSNet for objective evaluations were used for both subjective and objective evaluations. Subjective evaluations yielded MOSvalues of 3.96 for voice consistency, 3.89 for voice naturalness, and 3.95 for speech intelligibility. Real and synthetic child sounds showed a high association in an objective examination using a pretrained MOSNet. A further method of verifying speaker similarity was to compute the cosine similarity between the speech embeddings. The genuine and synthetic child voices were compared for word error rate (WER) using an automated speech recognition (ASR) model. In as little as five seconds, the finished trained TTS model was able to generate speech like that of a child using reference audio samples.

Deep transfer learning for automatic speech recognition Towards better generalization

This paper [16] Deep transfer learning (DTL) is apowerful technique gaining traction in automaticspeech recognition (ASR). It addresses the challenges of traditional ASR models that require massive datasets and struggle with generalizationto new domains or tasks.

Deep learning models for **[**15**]** ASR need enormous amounts of labeled speech data for training, making it expensive and time- consuming. Models trained on one domain (e.g., clean studio recordings) often perform poorly on another (e.g, noisy phone calls). ASR systems need to handle diverse accents, dialects, background noise, and unexpected variations in speechpatterns. DTL leverages knowledge from a pre- trained model on a source domain (e.g., clean speech) to improve performance on a target domain (e.g., noisy speech). DTL techniques likefeature adaptation, model adaptation, and adversarial learning help bridge the gap betweensource and target domains, enabling the model to generalize better. Even with limited target domain data, DTL can extract valuable knowledge from the pre-trained model, leading to significant performance gains compared to training fromscratch. DTL models can achieve higher recognition rates onunseen data, especially innoisy or challenging environments. Reduced training time and cost: By leveraging pre- trained knowledge, DTL requires less target domain data, saving time and resources. DTL models canbe fine-tuned for various tasks and domains, making them more versatile for real- world applications. Active research area: Researchers are continuously exploring new DTL architectures, optimization methods, and domain adaptation strategies to push the boundaries of ASR performance. Recent studies have shown significant improvements in accuracy and robustness using DTL techniques for ASR tasks like speech-to-text conversion and speaker identification. Addressing issues like catastrophic forgetting (losing knowledge from the source domain) and bias transfer across domains remains an ongoing researchendeavor.

Joint Multiscale Cross-lingual Speaking StyleTransfer with Bidirectional Attention Mechanism for Automatic Dubbing

Automatic Dubbing via **[**17**]** Joint Multiscale Cross-lingual Speaking Style Transfer and Bidirectional Attention Mechanism The title of this model translates speech to speech directly there is a problem no match data is available because people speak directly from one languageto another to tackle this we have used. The proposed model called Speech2S It is concurrently pre-trained using bilingual text andunpaired speech data The model's objective is tomodel the cross- lingual speech conversion and use paired text data to alleviate the data shortage issue in direct S2ST..pre-training methods, and fine-tuning process. Experimental results, analysis, and subjective evaluation show off the superiority and efficacy of the suggested Speech2S model. Speech2S's performance is assessed using the Vox Populi and Europarl-ST datasets.

A Semi Supervised Complementary Joint Training Approach for Low Resource Speech

This paper proposes a novel [18]. It first generatesspeech- pseudo label (speech-PseL) a semisupervised complementary joint training with limited method of recognition approachestrains an end to end matching pairs using a seed ASR model adjusted on a limited quantity of labeled data and then synthesizes audio for unpaired text using a text-to-speech (TTS) model. The discourse-PseL and synthesizedaudio-text (SynA-text) pairs are then used for complementary joint training of the ASR model.The paper analyzes the complementary propertyof Predictability of speech-PseL and SynA-text pairings probability distribution and feature extraction. It also suggests using parallel layers and label masking techniques to enhance theASR model's performance even more. Results from experiments using the Libri-light dataset demonstrate that the suggested CJT technique can accomplish significant improvements over text-only and speech-only training methods, especially in extreme low- resource cases.Moreover, the suggested technique is expanded to the case of zero paired data by employing an iterative CJT technique. The paper concludes that the suggested CJT technique is a promising approachfor low- resource ASR and can achieve competitive performance with a smaller model size compared to previous pre-training or self- training methods.

Real-Time Machine Translation System between Indian Languages

Automated Translation System**[**19] in Real Time for Indian Languages The title suggests the development and implementation of a systemthat enables instantaneous translation

betweenvarious Indian languages. The term "Real- Time" indicates that the translation process is expected to occur promptly, with minimal latency, making it suitable for dynamic and interactive communication scenarios. The focus on "Machine Translation" implies the utilization of automated translation techniques, likely leveraging state-of the-art algorithms andmodels in the domain of natural language interpretation. Machine translation is the process of translating text or speech from one language toanother using computer techniques. The mention of "Indian Languages" signifies that the system isdesigned to handle translation between multiple languages spoken in India, reflecting the linguistic diversity of the country. In this research, a real-time machine translation systembetween Marathi and Gujarati, two Indianlanguages, is proposed. The system's foundation is an LSTM encoder-decoder architecture driven by deep learning. Because LSTM can forecast future values based on past learning data, sequential data offers higher accuracy. The primary objective of the system is to translatebetween Indian languages, with an accuracy rateof 80–85%. Tensor Flow is the framework used in the implementation of the system. This could include languages such as Hindi, Bengali, Tamil,Telugu, and others. In summary, the topic suggests the development of a real time machine translation system specifically tailored for Indian languages, emphasizing the need for swift andaccurate language translation capabilities withinthe context of the linguistic diversity present inIndia.

Pronunciation Dictionary-Free Multilingual Speech Synthesis Using Learned Phonetic Representations

The research focus on multilingual [20] voice synthesis without the use of pronunciationdictionaries is outlined in the title, The phrase "Pronunciation Dictionary-Free" refers to amethod that provides pronunciation instruction without the need for established dictionaries The term "Multilingual Speech Synthesis" refers to the synthesis of speech in several languages,indicating a wide linguistic range for the study. The most significant innovation is found in "Using Learned Phonetic Representations," which suggests using learned phonetic representations in place of pre-existing language resources. In this framework "the author used Convolutional Neural Network to match and recognize the words with the trained phonetics words in which the author 3 hidden layer, 2x2 pooling layer and 5x5 kernel layer to achieve thispurpose.

Table 1: Literature Reviewed Survey Work

S.NO	TITLE	AUTHOR	METHODS	PROS	CONS
1.	Neural Networks for Bilingual Machine Translation Model	Hassanin M. Al-Barhamtoshy and Ashraf Said Qutb Metwalli(2023)	Statistical-based, corpus-based ordataset-based achine translation systems.	Can be involved in machine translation systems.	Still generating many incorrect translation outputs.
2.	Bilingual End-to-EndASR with Byte- Level Subwords	Liuhui Deng, RogerHsiao, and Arnab Ghoshal(2022)	Deep learning-based LSTM encoder- decoder architecture	Sequential data delivers improved accuracy.	Data scarcity canbe a challenge.
3.	Bilingual Automatic Speech Recognition: A Review, Taxonomy and Open Challenges	Ahmad A. M. Abushariah, Hua-Nong Ting, Mumtaz Begum Peer Mustafa, Anis Salwa Mohd Khairuddin, Mohammad A. M. Abushariah, and Tien-Ping Tan (2022)	Bilingual ASR using deep learning approach.	Can provide acceptable performance.	Many combinations of two languages are still limited.
4.	Joint Pre-Training with Speech and Bilingual Text for Direct Speech to Speech Translation	Kun Wei, Long Zhou, Ziqiang Zhang, Liping Chen, Shujie Liu, Lei He, Jinyu Li, and Furu Wei (2023)	Speech2S model, which is jointly pre-trained with unpaired speech and bilingual text data.	Effective at modeling the cross-lingual speech conversion from source to target language.	Data scarcity is a challenge.
5.	Code-Switching Text Generation and Injection in Mandarin-English ASR	Haibin Yu, Yuxuan Hu, Yao Qian, Ma Jin, Linquan Liu, Shujie Liu,Yu Shi, Yanmin Qian, Edward Lin, and Michael Zeng (2023)	Text generation and injection for improving the performance of an industry commonly- used streaming model, Transformer-Transducer (T-T), inMandarin-English code-switching speech recognition.	Can significantly boost the performance of T-Tmodels.	Can be more effective on the evaluation set which contains more homogeneous datawith the training set.
6.	Data Augmentation for End-to-End Code-Switching Speech Recognition	Chenpeng Du, Hao Li, Yizhou Lu, Lan Wang, Yanmin Qian (2023)	Three novel data augmentation methods for code-switching ASR training to improve both the acoustic and linguistic diversity of code-switching data.	All the three proposed approaches yielded significant improvements on code-switching ASR	Can be combined with recent popular SpecAugment.
7.	Automatic speech recognition of Portuguese phonemes usingneural networks ensemble	Nadia Nedjah , Alejandra D. Bonilla,Luiza de Macedo Mourelle (2023)	Recurrent Neural networks, Expert Network.	The ensemble of neural networks can improve the generalization of the ASR system by reducing overfitting.	The ensemble of neural networks takes longer to train than a singleneural network.

III. PROBLEM DEFINITION

The main issue raised by the problem statement is code-switching (CS) in speech, which is frequent in multilingual nations like India and gives rise to hybrid languages like Hinglish (Hindi-English) and Tanglish (Tamil-English).The problem is that there is not enough information available for Indic CS voice recognition. The research focuses on employing Transformer-based frameworks, which have shown promising results in automatic speech recognition (ASR) tasks, to address this problem. For Tamil- English CS speech recognition, two approaches are investigated: (i) using language information as tokensat the targets and (ii) using well-trained encoders of Monolingual Transformers as feature extractors for language discrimination. The results show that the first method performs well in language discrimination, while the second method handles CS adequately.The goal of this project is to use creative methods inside the Transformer framework to improve the effectiveness of Indic CS speech recognition in extremely low-resource circumstances.

IV. PROBLEM IDENTIFICATION

The issue this study found with speech is code- switching (CS), especially in multilingual nations like India where people frequently switchbetween languages during a single conversation.This linguistic phenomenon presents a problem forautomated speech recognition (ASR) systems; itis seen in urban languages such as Hinglish (Hindi-English) and Tanglish (Tamil-English).The main problem that is brought to light is the lack of data for Indic CS voice recognition, which makes it difficult to create models that work. By investigating strategies to manage situations with extremely limited resources, the research seeks to address this difficulty. It focuses on the use of Transformer-based frameworks for CS speechrecognition in Tamil and English. Twoapproaches are considered: the first uses well- trained Monolingual Transformers encoders as feature extractors for language discrimination, and the second uses language information as tokens at the objectives. The ultimate objective isto improve CS voice recognition performance inconditions of restricted data availability.

Architecture Diagram

The architecture diagram shows the main elements of the suggested system's structure as well as how they interact. The core component of it is the Generative Pre-Trained Transformer (GPT) model, which demonstrates its essential function in managing multilingual dialogues. Modules like Data Preprocessing are highlighted in the figure, with an emphasis on tasks like tokenization and formatting to maximize input data. The Figure 2 shows the architecture of proposed work.

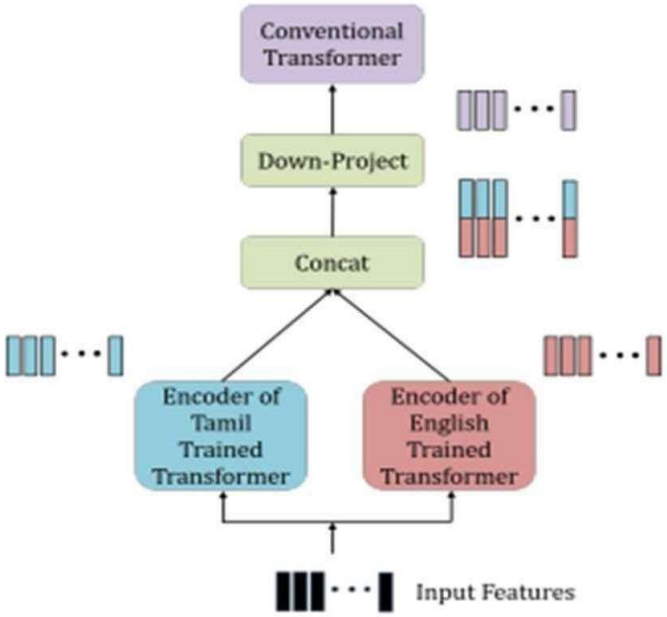

Fig. 2 Architecture of proposedsystem

The part of the GPT model where it autonomously picks up linguistic patterns and contextual knowledge from a variety of datasets is highlighted in the Unsupervised Pre-training module. The next modules focus on particular issues related to multilingual conversation processing, such as token-based language information inclusion and languagediscrimination. The Deployment and Integrationmodule guarantees practical application in real- world circumstances, while the evaluation and testing modules evaluate the system's performance. All things considered, the architecture diagram offers a thorough visualdepiction of the parts of the system and how theywork together, clarifying the methodical progression from data preprocessing to model deployment for effective multilingual conversation processing.

V. RESULT AND DISCUSSION

In the existing work up to till now, only monolingual Tamil and English data have been used in all tests to build the multilingual model. No use was made of CS data. The substantialimbalance and the fact that CS data is less than 5% monolingual prevents us from simply combining CS data with Tamil and English to create a multilingual model. Rather, but in this work employed the subsequent methodology using five hours of English, five hours of Tamil, and three hours of CS data to retrain the original Multilingual.

Table 2: Result% for Tanglish data

Sentence Length	Accuracy% TanglishData
Short	84.37%
Long	73.98%

CONCLUSION

In summary, the suggested system makes use of the Generative Pre-Trained Transformer (GPT) model to tackle the difficulties provided by multilingual discussions, especially when it comes to code- switching. The system incorporates key modules in a methodical manner, such as data preprocessing, unsupervised pre-training, multilingual conversation fine-tuning, and specific language discrimination methods. The system is able Transformer. Compared to a prior multilingual model that was trained just on Tamil and English, this one performs better because it utilized CS data during training. It goes without saying that using CS data for retraining enhances performance much more. In this work postulated that thesystem's decoder would now pick up on the unique character language models of Tamil, English, and CS and contribute to theenhancement of CS outcomes. The provision of linguistic tokens has been found to increase CS accuracy. In this work the same CS data and language tokens to retrain this model, and an improvement in performance show in table 2 to process avariety of language inputs and generatereplies that are both coherent and sensitive to context by utilizing the flexibility and contextualknowledge of the GPT model. The fact that modules for integration, testing, deployment, andassessment are included highlights how practically applicable the suggested approach is in actual situations. When multilingual conversation processing is critical to applicationsor systems, the deployment module guarantees asmooth integration. All things considered, the suggested strategy offers a thorough and reliablemethod for bilingual.

Acknowledgement

Thanks to all the anonymous referees for their helpful guidance that has improved the qualityof this paper. We would also like to our gratitude and sincere thanks to ours guide for the valuablesupport and guidance in the completion of this paper. Thanks to ours research domain & workwas supported by the Research and Development (R & D) experts at our college Sri Manakula VinayagarEngineering College and by our mentor an expert in Deep Learning of our department.

REFERENCES

1. *Hassanin M. Al-Barhamtoshy, Ashraf Said Qutb Metwalli: Neural Networks for Bilingual Machine Translation Model (IEEE Xplore: 16 January 2023)*
2. *Liuhui Deng, Roger Hsiao, and Arnab Ghoshal, " Bilingual End-to-End ASR with Byte-Level Subwords"(IEEE Xplore: 27 April 2022)*
3. *Ahmad A . M .Abushariah, Hua-Nong Ting,Mumtaz Begum Peer Mustafa,Anis Salwa Mohd Khairuddin ,Tien-Ping Tans ," Bilingual Automatic Speech Recognition A Review Taxonomy and Open Challenges" (IEEE 01 November 2022)*
4. *Kun Wei,Long Zhou, Ziqiang Zhang, Liping Chen, Shujie Liu, Lei He, Jinyu Li, Furu Wei, "Joint Pre-Training with Speech and Bilingual Text for Direct Speech to Speech Translation"(IEEE Xplore 05 May 2023)*
5. *Haibin Yu1, Yuxuan Hu, Yao Qian, Ma Jin,Linquan Liu, Shujie Liu,Yu Shi ,Yanmin Qian, Edward Lin, Michael Zeng,"Code-switching text generation and injection in mandarin- english asr"(IEEE Xplore 05 May 2023)*
6. *Chenpeng Du, Hao Li , Yizhou Lu, Lan Wang, Yanmin Qian, " Data augmentation for end-to- end code-switching speech recognition" (IEEE)*
7. *Nadia Nedjah , Alejandra D. Bonilla, Luiza deMacedo Mourelle," Automatic speech recognition of Portuguese phonemes using neural networks ensemble"(Elsevier 1 November 2023)*
8. *Changfeng Gao, Gaofeng Cheng ,Ta Li,Pengyuan Zhang , and Yonghong Yan, "Self-Supervised Pre-Training for Attention-Based Encoder-Decoder ASR Model" (ACM :04 May 2022)*
9. *Fu-Hao Yu, Kuan-Yu Chen, Ke-Han Lu, "Non-Autoregressive ASR Modeling Using Pre- Trained Language Models for Chinese Speech Recognition" (ACM: 11 April 2022)*
10. *Xiaojiang Yang; Junchi Yan; Yu Cheng;Yizhe Zhang "Learning Deep Generative Clustering via Mutual Information Maximization" (IEEE Xplore Issue :4 January, 2022)*
11. *Nazif Aydın: A Research On The New Generation Artificial Intelligence Technology Generative Pretraining Transformer 3 (IEEE Xplore: 29 December 2022)*
12. *Jun Liu; Yuanyu Fang; Zhuohan Yu; Tingkun Wu "Design and Construction of a Knowledge Database for Learning Japanese Grammar Using Natural Language Processing and Machine Learning Techniques" (IEEE Xplore Issue : 19 September, 2022)*
13. *Sheng Li, Jiyi Li, Qianying Liu, Zhuo Gong,"An End-to-End Chinese and Japanese Bilingual Speech Recognition Systems with Shared Character Decomposition " (Springer 14 April 2023)*
14. *Moa Lee; Junmo Lee; Joon-Hyuk Chang "Non-*

Autoregressive Fully Parallel Deep Convolutional Neural Speech Synthesis" (IEEE Xplore Issue : 8 March, 2022)

15. *Rishabh Jain; Mariam Yahayah Yiwere; DanBigioi; Peter Corcoran; Horia Cucu "A Text-to- Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis" (IEEE Xplore Issue : 28 April, 2022)*
16. *Rishabh Jain; Mariam Yahayah Yiwere; DanBigioi; Peter Corcoran; Horia Cucu "A Text-to- Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis" (IEEE Xplore Issue : 28 April, 2022)*
17. *Jingbei Li; Sipan Li; Ping Chen; LuwenZhang; Yi Meng; Zhiyong Wu; Helen Meng "Joint Multiscale Cross-lingual Speaking Style Transfer with Bidirectional Attention Mechanism for Automatic Dubbing" (IEEE Xplore Issue : 10 November, 2023)*
18. *Ye-Qian Du; Jie Zhang; Xin Fang; Ming-Hui Wu; Zhou-Wang Yang "A Semi-Supervised Complementary Joint Training Approach for Low-Resource Speech Recognition" (IEEE Xplore Issue : 11 September, 2023)*
19. *Aarati H. Patil; Snehal S. Patil; Shubham M.Patil; Tatwadarshi P. Nagarhalli "Real Time Machine Translation System between Indian Languages" (IEEE Xplore Issue : 24 May, 2022)*
20. *Chang Liu; Zhen-Hua Ling; Ling-Hui Chen "Pronunciation Dictionary-Free MultilingualSpeech Synthesis Using Learned Phonetic Representations" (IEEE Xplore Issue : 8 September, 2023*