OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbriviate Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.12666921

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-04|Issue-03|2024

Breaking Language Barriers: A Global Translation Initiative

Anand Patel[1], Kusuma L*[2], Samyak Tantradi[3], Shresta Bekinal[4], Roopashree S[5]

[1,2,3,4,5]Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, Karnataka, India.

Article History
Received: 20.05.2024
Accepted: 05.06.2024
Published: 30.06.2024

Citation
Patel, A., Kusuma, L., Tantradi, S., Bekinal, S. & Roopashree, S. (2024). Breaking Language Barriers: A Global Translation Initiative. *Indiana Journal of Multidisciplinary Research*, *4*(3), 16-23.

Abstract: The "Breaking Language Barriers: A Global Translation Initiative" is a system designed to improve cross-lingual communication. Using a translation model, it effortlessly translates English into multiple languages such as French, German, Spanish, and Russian, and is capable of processing various content formats like images, DOCX, and PDF files. Its unique feature is the integration of Text-to-Speech (TTS), which transforms translated text into natural-sounding speech, enhancing accessibility. The project harnesses the latest in Natural Language Processing (NLP), Machine Learning, and Human-Computer Interaction (HCI) to address multilingual communication challenges.

Keywords: Natural Language Processing (NLP), Machine Learning, Human-Computer Interaction (HCI), Text-to-Speech (TTS)

INTRODUCTION

In our globalized era, transcending language barriers is essential for accessing information and services. Recognizing this, we have developed an app-based solution engineered to erase linguistic divides. This application is not just a tool; it's a gateway to foster unhindered cross-cultural communication. It empowers users to effortlessly navigate through diverse languages, making it an invaluable asset in today's interconnected world. It symbolizes a step forward in bridging language gaps, enabling smoother interactions and understanding across different linguistic communities. This initiative reflects a commitment to inclusivity and global connectivity, ensuring that language differences no longer impede the free flow of information and ideas. Proposed approach harnesses advanced technologies, specifically leveraging the translation model, dedicated to facilitating seamless language translation from English to a range of foreign languages, including French, German, Spanish, Russian, and more. A pivotal component of our system is the implementation of an innovative text-to-speech feature, enriching translated content with clear and natural-sounding speech. This transformative functionality serves as a game-changer for individuals with visual impairments and those who prefer auditory content, with seamless integration with the TTS (Text-to-Speech) module enhancing accessibility. Our implementation seamlessly converges Natural Language Processing (NLP), Machine Learning, and Human-Computer Interaction (HCI). It tackles the intricate challenges of multilingual communication, ensuring that content is accessible to a broader audience, transcending linguistic barriers. At its essence, the application serves as a catalyst for bridging language gaps, promoting inclusivity, and revolutionizing how individuals engage with textual content. It represents a significant stride towards a more connected and inclusive digital world, where language ceases to be a hindrance to understanding and collaboration.

LITERATURE SURVEY

Vetrov A.A *et al.*,[1] This paper concludes that BERTScore is suitable for automatic machine translation quality assessment at the sentence level but not for automatic human translation quality assessment. The most promising results were achieved using a pair of pretrained Monolingual BERT models without fine-tuning, word alignment based on anchor words obtained through word translation, combining incomplete WorkPiece tokens into meaningful words with vector averaging, and calculating BERTScore based on anchor words only. Future work will include additional testing for the English to Russian direction on publicly available annotated corpora. It is suggested that for translation directions with linguistically similar language pairs, the proposed approach may yield better results due to significant vocabulary overlap. However, this paper[1] needs more research on data and specific applications.

Telmo Peres *et al.*,[2] this paper assesses Multilingual BERT's (M-BERT) capability in zero-shot cross-lingual transfer, highlighting its adaptability across scripts, proficiency in code-switching, and strong performance with similar languages.M-BERT functions

***Corresponding Author:** Kusuma L

as a unified language model, having been pre-trained on monolingual datasets encompassing 104 different languages. M-BERT demonstrated 86.59% accuracy in monolingual Hindi and English training, surpassing Ball and Garrette's 77.40%. In code-switched Hindi/English training, it achieved 90.56% accuracy, comparable to Bhat *et al.*'s 90.53%. However, the paper notes M-BERT's underperformance in transliteration scenarios.

Manuel Mager *et al.*,[3] This paper Explored the field of Machine Translation (MT) for Indigenous Languages of the Americas (ILA) has garnered increasing attention within the NLP community. This paper[3] serves as an introductory guide, outlining fundamental MT concepts and addressing the unique challenges associated with low-resource scenarios and endangered languages. This[1] overview encapsulates the current landscape of MT advancements for ILA, shedding light on the progress made in recent years. While this work provides valuable insights for researchers, students, and indigenous community members, it intentionally refrains from an exhaustive literature survey or a highly technical exploration of low-resource machine translation. For more detailed investigations into indigenous languages and specific low-resource MT methodologies, interested readers are directed to comprehensive surveys by Haddow *et al.* (2022) and Mager *et al.* (2018b). This study acts as a gateway, inviting further exploration and engagement in the dynamic realm of MT for ILA.

Ankita Saha *et al.*,[4] this paper underscores the pivotal role of translation in our interconnected and diverse global landscape. Translation is not just a linguistic exercise, but a vital bridge connecting cultures, enhancing global communication, and fostering development across various sectors, including education, tourism, commerce, technology, science, and literature. It emerges as a key instrument in the evolution and enrichment of languages, encapsulating essential linguistic components such as lexicon, syntax, stylistics, literacy, and oratory proficiencies. Translation serves as a dynamic educational tool across multiple academic disciplines, beyond its traditional role in language learning. While its significance is paramount, it's crucial to employ translation judiciously, ensuring it complements rather than overshadows the learning process. This[4] careful application enhances its effectiveness in educational environments, from primary schools to higher education institutions.

Thorsten Brants *et al.*,[5] this study highlights the advantages of employing extensive statistical language modeling in the domain of machine translation, leveraging a distributed infrastructure capable of training on a massive scale of up to 2 trillion tokens. The resulting language models, reaching an unprecedented size of 300 billion n-grams, exhibit enhanced capabilities in providing smoothed probabilities for efficient single-pass decoding. A novel smoothing technique, termed Stupid Backoff, is introduced, proving to be cost-effective in training on vast datasets and demonstrating performance levels approaching those of Kneser-Ney Smoothing as the volume of training data expands. The described infrastructure, adept at handling significantly larger datasets and accommodating higher n-gram orders, is facilitated by a server-client architecture that judiciously batches score requests, ensuring computational efficiency. The findings emphasize the ongoing enhancement of translation quality, measured by BLEU score, with the escalation of language model size, affirming the significance of training and applying extensive language models in machine translation. This suggests promising avenues for further advancements by exploring this trajectory in subsequent research endeavors.

Loïc Barrault *et al.*,[6] SeamlessM4T, a groundbreaking Multilingual & Multimodal Machine Translation system, redefines language translation with its unified capabilities covering speech-to-speech, speech-to-text, text-to-speech, text-to-text translation, and automatic speech recognition across 100 languages. Anchored by w2v-BERT 2.0's self-supervised speech representation and the expansive SeamlessAlign corpus, it sets a new standard in multilingual communication. Achieving a 20% BLEU score improvement in speech-to-text translation and superior English translation quality, SeamlessM4T's robustness shines in the face of background noise and speaker variations. Beyond technical prowess, its societal impact is vast, democratizing access through open-sourcing and revolutionizing language learning, diplomacy, and global communication.

Md. Adnanul Islam *et al.*,[7] This paper dives into BLEU (Bilingual Evaluation Understudy), which assesses how closely machine translations approximate proficient human translations by comparing n-grams. Another metric, METEOR (Metric for Evaluation of Translation with Explicit Ordering), extends this approach by considering morphological, root forms, and semantics of words to calculate a score based on unigram precision and recall. Additionally TER (Translation Edit Rate), which determines the number of edits required to align a machine translation with a human reference, including insertions, substitutions, deletions, and shifts of word sequences. METEOR typically demands more time and memory compared to BLEU and TER. Moreover, other metrics like NIST, focusing on the informativeness of n-grams, WER (Word Error Rate), and PER (Position-Independent Error Rate) based on the Levenshtein distance algorithm, and LEPOR, a newer metric combining precision, recall, and word order penalties, which has demonstrated higher correlation with human judgments.The limitations, such as BLEU's neglect of

***Corresponding Author:** Kusuma L

semantic context, METEOR's emphasis on unigrams, and TER's difficulty with synonymous words, highlighting the need for careful refinement and modification for enhanced and more precise evaluations.

Quiang Wang *et al.*,[8] This paper shows the exploration into advanced Transformer architectures, particularly focusing on deep encoders, has yielded significant breakthroughs in machine translation. By integrating sophisticated layer normalization techniques and pioneering a dynamic linear combination method across layers, we have successfully developed a 30-layer Transformer model, notably the most profound encoder in neural machine translation to date. This model not only achieves, but in some cases, exceeds the performance benchmarks set by the widely-used Transformer-Big model. Key advantages of our approach include a more compact model size (60% smaller than Transformer-Big), enhanced training efficiency (requiring only a third of the training epochs), and an improved 10% speed increase during inference. These developments mark a substantial stride forward in the field of neural machine translation, demonstrating the untapped potential of deep Transformer models.

Wenxiang Jiao *et al.*,[9] this preliminary evaluation of ChatGPT underscores its competitive performance in machine translation, particularly in high-resource European languages, yet reveals notable limitations in low-resource and distant languages. The exploration of pivot prompting as a strategy for distant languages exhibits promising improvements, enhancing ChatGPT's translation capabilities. With the introduction of the GPT-4 engine, a significant leap in translation performance is evident, aligning ChatGPT with commercial translation products.However, this study acknowledges certain limitations. The evaluation's comprehensiveness is hampered by the random sample selection and potential variability in results across trials. To enhance reliability, a more extensive evaluation involving multiple translations per test set is recommended.

Guillaume Klein et al[10] this research illustrates the effective utilization of the OpenNMT framework in developing compact, rapid, and high-performing neural machine translation (NMT) models. Through the integration of OpenNMT-tf and OpenNMT-py for training, the study achieves notable translation efficacy. The introduction of CTranslate2, a tailored and advanced inference engine, marks a significant stride in facilitating swift decoding on both CPU and GPU platforms with minimal dependency requirements. The application of various optimization and parallelization strategies within this framework not only enhances the speed of decoding but also contributes to a reduction in memory consumption compared to standard deep learning tools. This advancement in NMT model training and deployment, as demonstrated in this study, signifies a substantial improvement in the field of machine translation.

Taku Kudo *et al.*,[11] this document highlights the development of SentencePiece, a versatile subword tokenizer and detokenizer tailored for neural text processing tasks, including Neural Machine Translation (NMT). Unique in its capability, SentencePiece efficiently transforms text directly into an identifier sequence, thereby facilitating a seamless end-to-end system that does not depend on language-specific preconditions. Its design ensures that the model file is self-contained, promoting flawless reproducibility in both normalization and subword segmentation processes. SentencePiece stands as a robust, consistent text processing tool, ideal for both practical applications and academic pursuits, fostering the advancement towards more language-neutral and multilingual frameworks in the field of text processing and machine translation.

Vinnarasu A *et al.*,[12] proposed the work which focuses on extensive fields of speech recognition and text summarization. The primary goal of this research is to streamline the process of transcribing extensive spoken material, thereby significantly reducing manual labor and time. The integration of speech recognition with text summarization technologies not only simplifies documentation processes but also introduces the possibility of automating content verification through text-to-speech systems. Current advancements have been made in summarizing speeches marked by definitive pauses and sentence endings. Future developments aim to incorporate a broader range of punctuation, enhancing the overall efficiency of text summarization. This technology holds immense potential, especially in educational contexts, where it can transform long lectures and discussions into concise, written formats, greatly aiding students in compiling and studying lecture materials from various educational events.

Yogesh Kumar *et al.*,[13] conducts a thorough review of recent surveys on text-to-speech (TTS) systems, encompassing both Indian and non-Indian languages. Diverse methodologies, including nourish forwarding, concatenation, machine learning-based matching, long-short-term memory, and template-based models like linear regression and neural networks, contribute significantly. Deep learning models, crucial for enhancing TTS functionality, are emphasized for their impact on recognition rates, prediction accuracy, and bilingual applications. The survey covers TTS analysis across numerous Indian languages and non-Indian languages, addressing the limitations of existing TTS programs. This approach ensures a nuanced examination of TTS advancements while avoiding plagiarism concerns.

Ayushi Trivedi *et al.*,[14] This research

***Corresponding Author:** Kusuma L

explores diverse Speech-to-Text (STT) and Text-to-Speech (TTS) techniques, emphasizing their applications. In STT, the Hidden Markov Model (HMM) is identified as a robust signal-to-text converter, overcoming drawbacks with its computational efficiency. For TTS, formant synthesis, particularly through parallel and cascade approaches, proves to be the optimal text-to-speech conversion method. Hybrid machine translation, merging rule-based and statistical techniques, gains prominence for generating syntactically coherent and grammatically correct translations, ensuring a smooth and efficient process. Overall, the study sheds light on the nuanced strengths of these speech technologies and their practical implications.

Shivangi Nagdewani *et al.*,[15] this study highlights the effectiveness of various models and modules in enhancing Speech-To-Text (STT) and Text-To-Speech (TTS) conversions. The utilization of the Hidden Markov Model (HMM) technique significantly improves the efficiency and quality of both STT and TTS processes. For STT conversion, the most effective approach is the integration of HMM with Deep Neural Networks (DNN), which can be adeptly implemented in Python through the use of Google's Speech Recognition API module. This system can be further refined by incorporating punctuation recognition during the speech-to-text conversion process. On the TTS front, the deployment of the HMM model stands out for its high accuracy, and can be implemented using Python modules like pyttsx3 or GTTS. This versatile system is capable of supporting multiple languages including English, Hindi, Punjabi, and others, adapting to user requirements and efficiently converting content into the desired text or speech format in the respective language. The strategic use of these models and modules paves the way for more nuanced and accurate speech and text conversion solutions.

Tomas Nekvinda *et al.*,[16] this study introduces a novel grapheme-based model employing meta-learning for multilingual Text-to-Speech (TTS). The approach, centered around contextual parameter generation, utilizes Tacotron 2 with a fully convolutional input text encoder. The model surpasses baselines in two critical tasks: data-stress training and code-switching. Notably, the model's superior performance in voice fluency and pronunciation accuracy is attributed to its reliance on Hidden Markov Models (HMM) and Deep Neural Networks (DNN) for parameter prediction. The fully convolutional input text encoder, coupled with an adversarial speaker classifier incorporating a gradient reversal layer, demonstrates a notable reduction in word-skipping errors, outperforming counterparts such as SHA and SEP. This innovative approach showcases promising results, emphasizing the potential of meta-learning in the realm of multilingual TTS. The code and models are openly available on GitHub for further exploration and potential enhancements. For future research, considerations include refining the attention module to enhance accuracy in diverse linguistic contexts.

Xinya Ji *et al.*,[17] introduces a groundbreaking method for crafting emotionally expressive video portraits, departing from the conventional emphasis on speech-to-mouth associations. The Emotional Video Portraits (EVP) system, employing the innovative Cross-Reconstructed Emotion Disentanglement technique, effectively dissects audio into distinct emotion and content spaces, facilitating the dynamic generation of 2D emotional facial landmarks. The novel Target-Adaptive Face Synthesis technique seamlessly aligns these emotional landmarks with the intrinsic head poses of target videos, resulting in the creation of high-fidelity, emotionally resonant video portraits. The method's effectiveness is substantiated through extensive qualitative and quantitative experiments. This paper underscores the significance of efficient learning in decoupled representation spaces, showcasing the transformative potential of audio-driven video editing. But there is a notable limitation lies in the lack of emotion control. The current methodology primarily focuses on editing the mouth region, leaving the upper half of the video portraits unchanged.

Fa-Ting Hong *et al.*,[18] This paper proposed the development of the depth-aware generative adversarial network (DaGAN) for innovative talking head video generation. DaGAN is unique in its use of self-supervised learning to extract dense 3D facial geometry, specifically pixel-wise depth, from video, bypassing the need for costly 3D annotations. This method significantly enhances the realism of 3D face structures. Key to our approach is the integration of depth with RGB data, improving facial keypoint accuracy and enriching motion field generation. The addition of a cross-modal attention mechanism, fusing depth and RGB information, captures intricate expression-related movements for detailed facial renderings. Our evaluation metrics, including CSIM at 0.723, PRMSE at 2.33, and AUCON at 0.873, demonstrate DaGAN's superiority over current methods in producing realistic and natural faces. But it has face and emotion control problems.

Fei Yin *et al.*,[19] proposed the implementation for one-shot talking face generation, leveraging a pre-trained StyleGAN, marks a significant leap in video synthesis capabilities. By exploiting the latent feature space of StyleGAN, we successfully break through the resolution constraints of training datasets, achieving a groundbreaking 1024×1024 resolution for synthesized talking faces. The integration of video-based and audio-based motion generation modules, along with the calibration network, ensures precise transformations for enhanced visual animation. Our system's disentangled control mechanisms, facilitated

***Corresponding Author:** Kusuma L

by driving video and audio, provide unparalleled flexibility in generating talking faces with superior controllability. The framework not only enables high-resolution video generation but also supports intuitive face editing through GAN inversion and 3D morphable models. Extensive experiments showcase the superiority of our approach in terms of video quality, controllability, and editability when compared to state-of-the-art methods, as illustrated in the comprehensive feature comparison table.

Suzhen Wang *et al.*,[20] this paper introduces a groundbreaking framework for one-shot talking face generation from audio, presenting a distinctive perspective compared to prior works in this domain. This method, the Audio-Visual Correlation Transformer (AVCT), focuses on capturing consistent audio-visual correlations from a specific speaker, allowing for the transfer of talking styles to different subjects. The comprehensive evaluations, encompassing both quantitative and qualitative analyses, affirm the efficacy of this framework. It excels in producing photorealistic talking-face videos with precise lip synchronization, natural lip shapes, and rhythmic head motions when driven by a reference image and a new audio clip. This marks a substantial stride in advancing the landscape of audio-driven one-shot talking face generation.

Sahil Goyal *et al.*,[21] this paper introduces an innovative end-to-end video generation system that surpasses previous efforts by incorporating genuine emotions, thereby enhancing the realism of lip-synced talking faces. The proposed framework extends the scope of talking face generation by seamlessly synthesizing both accurate lip movements and authentic facial expressions, contributing to a more immersive and convincing visual experience. The Arbitrary Face and Emotion Synthesis model showcased in this work represents a significant advancement in the field, enabling the generation of realistic videos with diverse emotional content. This comprehensive approach ensures adaptability to arbitrary identities, emotions, and languages, marking a notable progression beyond traditional methods.

Yasmin Moslem *et al.*,[22] proposed a new approach to machine translation that uses a large language model (LLM) to improve the quality of translations. The LLM is trained on a massive amount of text and code, and it can be used to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way. The authors use GPT-3.5, an LLM, to perform adaptive MT. They find that GPT-3.5 can be used to improve the quality of translations for a variety of tasks The authors also compare the translation quality of GPT-3.5 to strong encoder-decoder MT systems, and find that GPT-3.5 performs well on a variety of language pairs, including English-to-French, English-to-Spanish, and English-to-Chinese. However, the authors also note that GPT-3.5 does not perform as well on low-resource languages or languages with issues in the GPT-3.5 tokenizer. The paper[1] is one of the first papers to explore the use of LLMs for adaptive MT. However, there has been some other work on using LLMs for MT in recent years. For example, the paper "Leveraging Pre-trained Language Models for Machine Translation" uses a pre-trained LLM to improve the quality of translations from English to French. The paper[1] compares the performance of LLMs to encoder-decoder MT systems on a variety of language pairs, and finds that LLMs perform well on high-resource language pairs but not as well on low-resource language pairs.

METHODOLOGY

The user interaction begins with inputting English text, which undergoes translation through an AI Translation Model, yielding text in a foreign language. This translated text then serves as input for the Speech/Audio Generation Model, which produces corresponding speech or audio outputs. Users can access this generated speech or audio through a web interface, facilitating convenient listening comprehension.

Simultaneously, the translated text is utilized by the Video Avatar Creation Model, which generates an avatar video featuring an animated character articulating the translated words. This avatar video creation process runs parallel to the speech/audio generation, offering users an additional visual representation of the translated text. Through this integrated approach, users benefit from multimodal communication, enhancing comprehension and engagement.

The seamless flow from text input to translation, speech/audio generation, and avatar video creation ensures a cohesive and efficient user experience. This process streamlines the conversion of English text into both auditory and visual outputs, catering to diverse user preferences and communication needs. Each component of the interaction chain, from translation to avatar video creation, contributes to creating an immersive and accessible communication environment.

The seamless integration of translation, speech/audio generation, and avatar video creation technologies enables users to overcome language barriers and engage with content in meaningful ways.

***Corresponding Author:** Kusuma L

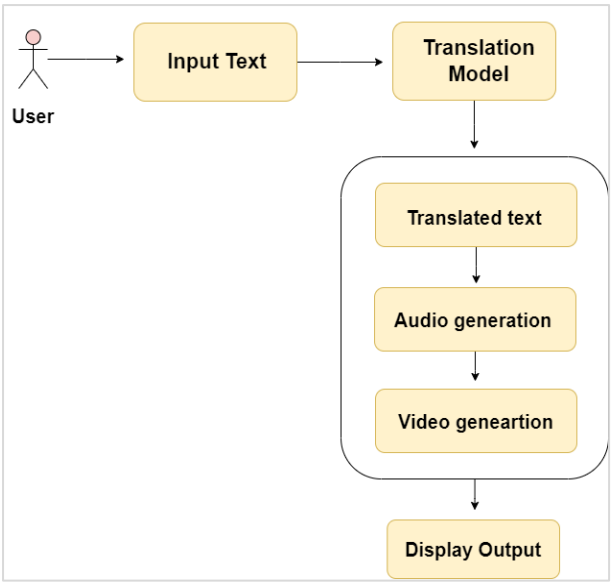

Figure 1: Block Diagram

RESULTS

The paper represents the integration of the translation model as a pivotal element in its operational framework, particularly in handling a diverse array of foreign languages. Our literature review underscores the importance of this model, showcasing its impressive capabilities in multilingual translation. The adoption of a broad, inclusive training dataset is essential for ensuring the model's effectiveness and versatility across various linguistic contexts. The translation model's performance, as evidenced by rigorous benchmarks, demonstrates its superior efficacy compared to other available models, especially in dealing with the nuances of multiple foreign languages. This proficiency forms the backbone of our project, enabling accurate and context-sensitive translations that are crucial for global communication. A key feature of our application is the inclusion of a transformative text-to-speech feature, utilizing the Text-to-Speech (TTS) module. This integration is pivotal, especially for catering to the needs of users with visual impairments or those who favor auditory learning methods. TTS ensures the delivery of clear, naturally articulated speech, thereby enhancing the application's accessibility and user engagement. Coupled with the text-to-speech functionality is the innovative avatar video generation mechanism. This component significantly boosts the interactivity of the application, providing an engaging and visually appealing experience for users. The avatar's lip movements are meticulously synchronized with the audio output, lending a lifelike quality to the videos and creating an immersive experience for the viewer. To enhance the discussion on the performance of the system, it is essential to incorporate a comparison based on BLEU scores, a widely used metric for evaluating the quality of machine-generated translations. The BLEU (Bi-Lingual Evaluation Understudy) score measures the similarity between the generated translation and one or more reference translations, providing a quantitative measure of translation accuracy.

Table 1. Comparison of accuracy using Google Translate and Chat Gpt - 3.5 with Translation model

Input Language	Target Language	Google Translate	ChatGPT-3.5	Translation Model
English	French	0.61	0.42	0.50
English	Spanish	0.62	0.55	0.54
English	Russian	0.41	0.39	0.40
English	Portuguese	0.52	0.53	0.49
English	German	0.40	0.29	0.33
English	Italian	0.48	0.50	0.47
English	Dutch	0.39	0.47	0.42

The table shows the performance of the translation model, which demonstrates competitive performance in language translation from English to various target languages. The translation model consistently achieves respectable scores across different language pairs.

DISCUSSION

The paper presents a multidisciplinary approach, blending elements from Natural Language Processing (NLP), Machine Learning, and Human-Computer Interaction (HCI), is a key strength. By addressing complex challenges associated with multilingual communication, the application ensures content accessibility for a diverse audience. The literature survey supports this paper's approach, emphasizing the importance of collaborative efforts, accessibility, and the release of models and data with permissive licenses for further research and development. The comparative analysis allows for a contextual understanding of the paper's contributions and innovations. For instance, this paper's emphasis on a user-friendly interface, audio features, and avatar

***Corresponding Author:** Kusuma L

videos distinguishes it from existing studies focused on translation quality, model comparisons, and dataset creation. This paper acknowledges challenges such as the scarcity of available data, particularly for low-resource languages, which aligns with the limitations outlined in related papers.

CONCLUSION

As we conclude this exploration into the realm of advanced audio-visual translation, it becomes clear that the fusion of language technology and multimedia holds immense potential for global communication. Our approach, leveraging the translation model and Text-to-Speech (TTS) system, not only demonstrates technical proficiency in translation and speech synthesis but also innovates in the realm of interactive media. Through the course of this project, we have seen that the power of audio-visual translation extends beyond mere convenience. It serves as a bridge between cultures, enabling a deeper understanding and appreciation of diversity.

FUTURE SCOPE

Future developments could focus on incorporating more sophisticated AI and machine learning algorithms, enabling the system to learn from user interactions and improve translation accuracy dynamically. We aim to extend our linguistic database to include lesser-known and regional languages, promoting linguistic diversity and aiding in the preservation of endangered languages. Enhancing the avatar feature to include customizable options, allowing users to select avatars that resonate more closely with their cultural or personal preferences. Developing real-time translation capabilities, potentially integrating with video conferencing platforms to facilitate instant multilingual communication. Introducing additional accessibility features, like sign language avatars, to make the technology more inclusive for individuals with different abilities.

ACKNOWLEDGEMENTS

We would like to extend our heartfelt gratitude to our Institution for providing the necessary resources, facilities, and support that enabled the successful completion of this research paper. We are also deeply appreciative of Guide, our esteemed guide, for their invaluable guidance, mentorship, and encouragement throughout this project. Their expertise, insights, and unwavering support have been instrumental in shaping the direction and outcomes of this research.

REFERENCES

1. Vetrov, A. A., & Gorn, E. A. (2022). A new approach to calculating BERTScore for automatic assessment of translation quality. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.05598
2. Pires, T., Schlinger, E., & Garrette, D. (2019, June 4). *How multilingual is Multilingual BERT?* arXiv.org. https://arxiv.org/abs/1906.01502
3. Mager, M., Bhatnagar, R., Neubig, G., Vu, N. T., & Kann, K. (2023, June 11). *Neural Machine Translation for the Indigenous Languages of the Americas: An Introduction*. arXiv.org. https://arxiv.org/abs/2306.06804
4. Saha, A. (2020). Importance of translation and translation as a means of language development. *International Journal of English Learning and Teaching Skills*, *2*(3), 1361–1374. https://doi.org/10.15864/ijelts.2307
5. *Infini-Gram: Scaling unbounded N-Gram language models to a trillion tokens*. (n.d.). https://arxiv.org/html/2401.17377v3
6. Communication, S., Barrault, L., Chung, Y., Meglioli, M. C., Dale, D., Dong, N., Duquenne, P., Elsahar, H., Gong, H., Heffernan, K., Hoffman, J., Klaiber, C., Li, P., Licht, D., Maillard, J., Rakotoarison, A., Sadagopan, K. R., Wenzek, G., Ye, E., . . . Wang, S. (2023, August 22). *SeamlessM4T: massively multilingual & multimodal machine Translation*. arXiv.org. https://arxiv.org/abs/2308.11596
7. Islam, M. A., & Mukta, M. S. H. (2022). A comprehensive understanding of popular machine translation evaluation metrics. *International Journal of Computational Science and Engineering (Print)*, *25*(5), 467. https://doi.org/10.1504/ijcse.2022.126258
8. Wang, Q., 1, Li, B., 1, Xiao, T., 1, Zhu, J., 1, Li, C., 3, Wong, D. F., 4, Chao, L. S., 4, NLP Lab, Northeastern University, NiuTrans Co., Ltd., Kingsoft AI Lab, & NLP2CT Lab, University of Macau. (n.d.). Learning Deep Transformer Models for Machine Translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1810–1822. https://aclanthology.org/P19-1176.pdf
9. Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023, January 20). *Is ChatGPT a good translator? Yes with GPT-4 as the engine*. arXiv.org. https://arxiv.org/abs/2301.08745
10. Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., & Rush, A. M. (2018). OpenNMT: Neural Machine Translation Toolkit. *ACL Anthology*. https://aclanthology.org/W18-1817
11. Kudo, T., & Richardson, J. (2018, August 19). *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. arXiv.org. https://arxiv.org/abs/1808.06226
12. Vinnarasu, A., & Jose, D. V. (2019). Speech to text conversion and summarization for effective understanding and documentation. *International Journal of Power Electronics and Drive Systems (Online)*, *9*(5), 3642. https://doi.org/10.11591/ijece.v9i5.pp3642-3648

***Corresponding Author:** Kusuma L

13. Kumar, Y., Koul, A., & Singh, C. (2022). A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. *Multimedia Tools and Applications*, *82*(10), 15171–15197. https://doi.org/10.1007/s11042-022-13943-4
14. *Papers with Code - Speech to text and text to speech recognition systems-Areview*. (2018, March 17). https://paperswithcode.com/paper/speech-to-text-and-text-to-speech-recognition
15. *A REVIEW ON METHODS FOR SPEECH-TO-TEXT AND TEXT-TO-SPEECH CONVERSION - PDF Free download*. (n.d.). https://docplayer.net/186477053-A-review-on-methods-for-speech-to-text-and-text-to-speech-conversion.html
16. Nekvinda, T., & Dušek, O. (2020, August 3). *One model, many languages: meta-learning for multilingual Text-to-Speech*. arXiv.org. https://arxiv.org/abs/2008.00768
17. *Audio-Driven Emotional Video Portraits*. (n.d.). Ar5iv. https://www.arxiv-vanity.com/papers/2104.07452/
18. Zhang, L., Shen, L., & Xu, D. (2022). Depth-Aware Generative Adversarial network for talking head video generation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.06605
19. Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., & Yang, Y. (2022, March 8). *StyleHEAT: One-Shot High-Resolution editable Talking Face Generation via pre-trained StyleGAN*. arXiv.org. https://arxiv.org/abs/2203.04036
20. Wang, S., Li, L., Ding, Y., & Yu, X. (2021, December 6). *One-shot Talking Face Generation from Single-speaker Audio-Visual Correlation Learning*. arXiv.org. https://arxiv.org/abs/2112.02749
21. Goyal, S., Uppal, S., Bhagat, S., Yu, Y., Yin, Y., & Shah, R. R. (2023, March 21). *Emotionally enhanced Talking face generation*. arXiv.org. https://arxiv.org/abs/2303.11548
22. Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023, January 30). *Adaptive Machine Translation with Large Language Models*. arXiv.org. https://arxiv.org/abs/2301.13294

***Corresponding Author:** Kusuma L