IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

Vaktar AI: A Conceptual Framework for AI-Based Audio Synthesis, Multilingual Translation, and PDF Summarization

Pritesh Patil, Tanishka Gadilkar, Bhargavi Tambe, Soham Yeola
Information Technology Department
Second Year, Information Technology
AISSMS Institute of Information Technology, Pune, India
pritesh.patil@aissmsioit.org, tanishkagadilkar@gmail.com, bhargavi.tambe19@gmail.com,
sohamyeola3107@gmail.com

Abstract: *AI has made tremendous changes in digital content creation especially within speech synthesis, multilingual interaction, and document processing [1], [4]. In this paper, the concept behind three highly interrelated modules in Vaktar AI is explained which was initially created by Team 1 [4] for the generation of talking avatar videos. Expanding from this platform, three new modules are introduced to this AI system including an audio synthesis module, multilingual text translator, and a PDF document extractor module. The first module uses the current state-of-the-art technology called neural text-to-speech (TTS) engine in order to synthesize the user's uploaded text to audio, creating an avatar video without a voice recording [2], [6]. The next module is developed using the power of neural machine translation and MyMemory API which translates the user's input across more than 50 languages before creating an avatar video [3]. Lastly, the third module uses PyMuPDF to extract text and then apply LLM APIs to summarize the PDF files which then can create personalized avatar videos [5], [7]. Importantly, all three modules share the common step of applying the generated texts to the neural text-to-speech engine*

Keywords: *Text-to-Speech (TTS), Neural Text-to-Speech, Neural Machine Translation (NMT), Multilingual Translation, PDF Text Extraction, Document Summarization, AI Audio Generation, Avatar Video Generation, Zero-Shot TTS, PyMuPDF, MyMemory API, Modular AI Systems, Human–AI Interaction, Accessible AI*

I. INTRODUCTION

The fast development of AI and NLP has fundamentally changed the nature of digital content creation, consumption, and communication [1], [4]. One of the most important applications of AI and NLP technology includes speech synthesis, language translation, and intelligent document processing, which all serve as fundamental building blocks for bridging the gap between information access and natural communication among different users and communities.

From mechanical, monotonous voices to natural, expressive, and multilingual ones, the evolution of TTS models has taken significant steps toward deep learning models [2], [6]. They represent core components of today's avatar systems, allowing personalized video clips through talking avatars without requiring any human-recorded voices. Nevertheless, there are still technical difficulties in developing a seamless integration process that can generate high-quality, multilingual voices in real time.

Alongside these developments, language translation systems have made great progress through the introduction of NMT models. However, almost all existing avatar platforms support only one language due to various technical

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, April 2026 Impact Factor: 8.2

limitations [3]. The lack of an integrated, free-of-charge, and language translation layer will be a major hurdle when developing next-generation avatar systems.

Documents, especially in the form of PDF, continue to be a significant source of information in both academia and professional environments. However, the process of engaging with complex PDF files continues to be challenging for many people, not least because of their learning disabilities, language barriers, and time limitations [5], [7]. The possibility to extract, summarize, and render the information found in these PDFs through personalized talking avatars emerges as an innovative step towards a more accessible experience of content ingestion.

In its initial form, the idea behind Vaktar AI project was formulated by Team 1 and demonstrated the potential of talking avatars that provide visual communication through conversational AI [4]. In this regard, the main contribution of this paper lies in adding new modules that help address key challenges associated with accessibility, personalization, and content consumption of Vaktar AI avatars. Specifically, this paper presents three modules that will make Vaktar AI system fully functional and capable of supporting various languages as well as extracting and summarizing content from PDFs.

II. RELATED WORK

There has been considerable advancements within the area of AI-based speech synthesis over the last few years. Early concatenative and parametric text-to-speech systems had limited naturalness and prosody [6]. Deep learning models like WaveNet and Tacotron marked the beginning of an era of more advanced, end-to-end, neural TTS models, allowing production of far more natural sounding speech [6]. XTTS marked another milestone within the area by providing zero-shot multilingual speech synthesis with high quality, using dozens of languages and achieving state-of-the-art results for multilingual avatar voice creation [2]. Despite all those advancements, there has been no attempt at integrating zero-shot multilingual TTS systems into any of the avatar generation platforms.

Neural machine translation has seen tremendous advancement within the recent years. Many transformer-based NMT models can translate multiple languages to near human-level accuracy [1]. For example, Meta's NLLB-200, Helsinki-NLP's Opus-MT, or mBART-50 support hundreds of languages and utilize encoder-decoder architecture [3]. Free-to-use APIs based on similar technologies have been implemented and released in the cloud environment for the benefit of developers, allowing programmatic multilingual translation without knowing much about ML itself [3]. There is a lack of academic works dedicated to translation layer integration in avatar generation systems.

Document processing, especially PDF text extraction and summarization, has gained considerable prominence as an important research field [5]. Several libraries, such as PyMuPDF (fitz), can perform robust and platform-independent extraction of text and metadata from multiple-page and structured PDFs, whereas language models like the GPT-4o engine can produce succinct and human-friendly summaries of long texts [7]. On the other hand, systems that transform document summaries to personalized avatar videos have yet to be realized.

Current avatar platforms, including Synthesia, HeyGen, and D-ID, are capable of generating text-to-speech voiceovers for avatars or providing translation services for voiceovers, albeit not usually at once, and without incorporating PDF processing [4]. Notably, these platforms do not offer personalized user image-based avatars along with document summarization and audio processing functionality. This void is filled in by Vaktar AI by integrating TTS generation, translation, and PDF summarization into one workflow built upon the work of Team 1.

III. LITERATURE REVIEW

This section provides a summary of the key research fields that are pertinent to the three modules mentioned above. These include research fields like artificial intelligence text-to-speech, multilingual neural translation, PDF document automatic extraction and summarization, and the ethical issues associated with AI-generated multimedia content.

A. Neural Text-to-Speech Synthesis

There have been significant developments in the area of neural TTS systems with the advent of sequence-to-sequence learning techniques. In particular, WaveNet introduced a technique whereby human-quality speech can be generated

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, April 2026 Impact Factor: 8.2

through autoregressive modeling of raw audio waveforms [6]. Other techniques such as Tacotron 2 and FastSpeech improved on this, lowering latency while providing more natural and expressive speech synthesis. Finally, in the area of zero-shot multilingual voice cloning, Coqui TTS developed XTTS, which is capable of generating speech in 16 languages without any fine-tuning of language-specific parameters.

B. Multilingual Neural Machine Translation

The field of machine translation was revolutionized when transformers were applied to NMT models. Such encoder-decoder models understand complex contextual relationships between tokenized languages in deep ways that make them capable of producing highly accurate translations compared to previous phrase-based methods [1]. The recent example is the NLLB-200 by Meta, which applies the approach for 200 languages, whereas Helsinki-NLP offers lightweight Opus-MT solutions for more than 1,000 language pairs [3]. Free to use MyMemory REST API serves as an easy way for developers to apply this technology to over 50 language pairs [3]. However, there has not been a single avatar generation platform implementing live NMT translation as one of the main inputs.

C. PDF Extraction and LLM Summarization

PyMuPDF (fitz) emerged as a robust solution allowing extracting structured data, such as text, images, and other metadata, from PDF files, even multi-column publications and academic works such as papers, reports, and presentations [5]. Together with LLM summarization through GPT-4o API, the pipeline would allow users to convert multi-page PDFs into concise audio-friendly summaries [7]. Unfortunately, according to the reviewed literature, there is still no existing method linking document summarization to personalized avatar video generation.

D. Ethical Issues in AI Audio-Visual Synthesis

The creation of lifelike AI-generated voice and video content involves considerable ethical issues. The work done by researchers on detecting deepfakes [8] demonstrates that there are increasing dangers associated with audio deepfakes, which are synthetic voice records impersonating real people. Therefore, TTS systems featuring voice cloning technologies, like XTTS [2], should be applied only within a proper framework. The architecture of the Vaktar AI application prevents the synthesis of avatars from external sources and does not allow creating target voice clones of any third parties.

TABLE I: LITERATURE REVIEW SUMMARY

Ref.	Year	System / Study Focus	Techniques Used	Key Strengths	Limitations	Research Gap Identified
[1]	2023	NLP Applications Survey	Transformer-based NLP models	Comprehensive overview of real-world NLP systems	Lacks implementation details for integrated systems	No unified system combining TTS, translation, and document AI
[2]	2024	XTTS Multilingual TTS	Zero-shot neural TTS with voice cloning	Massively multilingual speech; no per-language training required	GPU-intensive; real-time latency for voice cloning	No integration with avatar video pipelines

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, April 2026 Impact Factor: 8.2

[3]	2023	Neural Machine Translation (NMT)	Transformer-based encoder-decoder NMT	Near-human accuracy; covers 50–200 language pairs	Free APIs have daily rate limits; quality varies for low-resource languages	Not integrated with avatar or TTS generation platforms
[4]	2023	Talking Face Generation Survey	Multimodal avatar synthesis	Comprehensive review of talking head generation techniques	No support for document ingestion or multilingual TTS	Lacks unified audio + translation + PDF pipeline
[5]	2023	PyMuPDF PDF Processing	Text extraction from multi-page PDFs	Reliable,fast, cross-platform PDF text extraction	No native OCR; image-heavy PDFs require external OCR	No pipeline connecting PDF content to avatar video
[6]	2016–2024	Deep Learning TTS (WaveNet to FastSpeech)	WaveNet, Tacotron, FastSpeech architectures	Progressively higher naturalness; lower inference latency	Older models lack zero-shot multilingual capability	Weak integration of TTS with translation or document AI
[7]	2020	LLM Document Summarization	GPT-based abstractive summarization	High-quality, human-readable summaries from long documents	Hallucination risk; API cost at production scale	No system delivers PDF summaries via personalized avatar video
[8]	2023	Deepfake Generation and Detection	GAN and diffusion-based synthesis; detection models	Thorough case study of generation and detection challenges	Detection methods lag behind generation advances	Ethical safeguards for TTS voice cloning in avatar platforms unaddressed

IV. PROPOSED ARCHITECTURE

The extended version of the Vaktar AI system is based on an architecturally tight three-module design which is built right on top of the lip-syncing avatar video generation core, initially formulated by Team 1 [4]. Although Team 1 developed the core pipeline responsible for generating lip-syncing avatar video from user input, this study proposes to extend the initial setup by adding three input modules ahead of the TTS component – namely, audio synthesis, multilingual translation, and PDF summarization – all of which feed into the same TTS module and eventually generate the same avatar video output.

A. System Perspective

From a conceptual perspective, the extended system follows the pattern of a modular transformation pipeline [1], [4]. The input for each of the three additional modules takes a specific form in each case, namely, typed text input, selected language for translation, or uploaded PDF documents; each module processes its input to produce audio content that goes into the common avatar video engine, which was carried over from Team 1.

In a typical example, a user can upload a scientific paper in the form of a PDF, get an executive summary generated by the LLM module, have the summary translated into Hindi by the translation module, and then produce an avatar video

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

using the TTS module with an appropriate Hindi voice – all in one seamless flow. No such level of modularity is offered by any commercial avatar service currently available on the market.

B. Core Modules

1) Audio Generation Module (Text-to-Speech)

The module takes in any finalized form of text, either directly provided by the user or created via the conversational AI brain provided by Team 1's solution, translation module, or PDF summaries, and generates human-like speech audio [2], [6]. This is achieved through:

• Obtaining the finalized text through user inputs or the upstream modules [1]
• Choosing the TTS voice based on the pre-approved list of voices available (including gender, accent, language, tone)
• Calling the TTS engine (such as Google Cloud TTS/Azure Cognitive TTS/XTTS) to generate speech audio waves [2], [6]
• Delivering the audio wave to the lip-sync and avatar generation engine [4]

This module removes the need for any recording of user voice and allows video creation for any user irrespective of voice abilities.

2) Multilingual Translation Module

The multilingual translation module acts as a pre-processing stage in front of TTS synthesis. It translates the user's input into any one of over 50 target languages by utilizing the MyMemory REST API [3], which is a completely set-up free neural machine translation system. The implementation details include the following steps:

• The user turns on the translation feature and picks the target language from the language selector [3]
• The input text is sent as a JSON object to the MyMemory REST API [3]
• In seconds, the translated text is returned and presented for verification and possible edit by the user
• After translation, the text goes through the exact same process for TTS processing and avatar video generation as any other input text [2], [4]

An important feature of this translation module is that the text is translated prior to TTS synthesis in order to ensure that the avatar will speak in the selected target language. Future work would be to have voice-language match automatically performed by the TTS engine.

3) PDF Extraction and Summarization Module

This module allows users to upload PDF files and automatically create personalized video avatars summarizing the contents thereof [5], [7]. The steps involved are:

• Upload PDF by the user using the file upload feature
• PyMuPDF (fitz) analyzes each page of the document and extracts clean text [5]
• The extracted raw text is passed to the LLM summarization API (GPT-4o / compatible endpoint) along with a structure request for concise, verbal summary [7]
• The output summary is provided back to the user for confirmation and any necessary editing before moving ahead with creating the video
• The confirmed summary is fed into the translation module (optional) followed by TTS + Avatar video creation pipeline [2], [4]

The problem with documents that contain scanned or image-rich pages is that an additional OCR stage would be needed before the useful text could be extracted by PyMuPDF. This is a known limitation discussed in Section VIII.

V. SYSTEM WORKFLOW

The workflow process involved within each of the three modules in Vaktar AI involves a connected series of processes whereby various types of inputs from users are converted to a consistent talking avatar video output [1], [4]. Figure 1 below shows the workflow process followed by the three modules and how they come together in producing the same output.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

A. Standard Audio Generation Workflow (Text Input)

During the typical text-to-speech (TTS) workflow, the user inputs their text directly using the create page interface within the system. The text is then passed on to the designated TTS engine in the process, and a response in the form of an audio waveform is sent back to the system by the TTS engine. This audio waveform is then fed to the avatar lip-sync engine of Team 1.

B. Multilingual Translation Workflow

The process of translation is carried out prior to speech synthesis on behalf of the text by the user. This means that the user turns on the translation switch, chooses another language from a drop-down list, and the text is then translated using the MyMemory REST API service [3]. Translation is completed in seconds and the result is shown to the user for confirmation. After that, the translated text follows the same path as any other text in the process of speech synthesis and avatar creation.

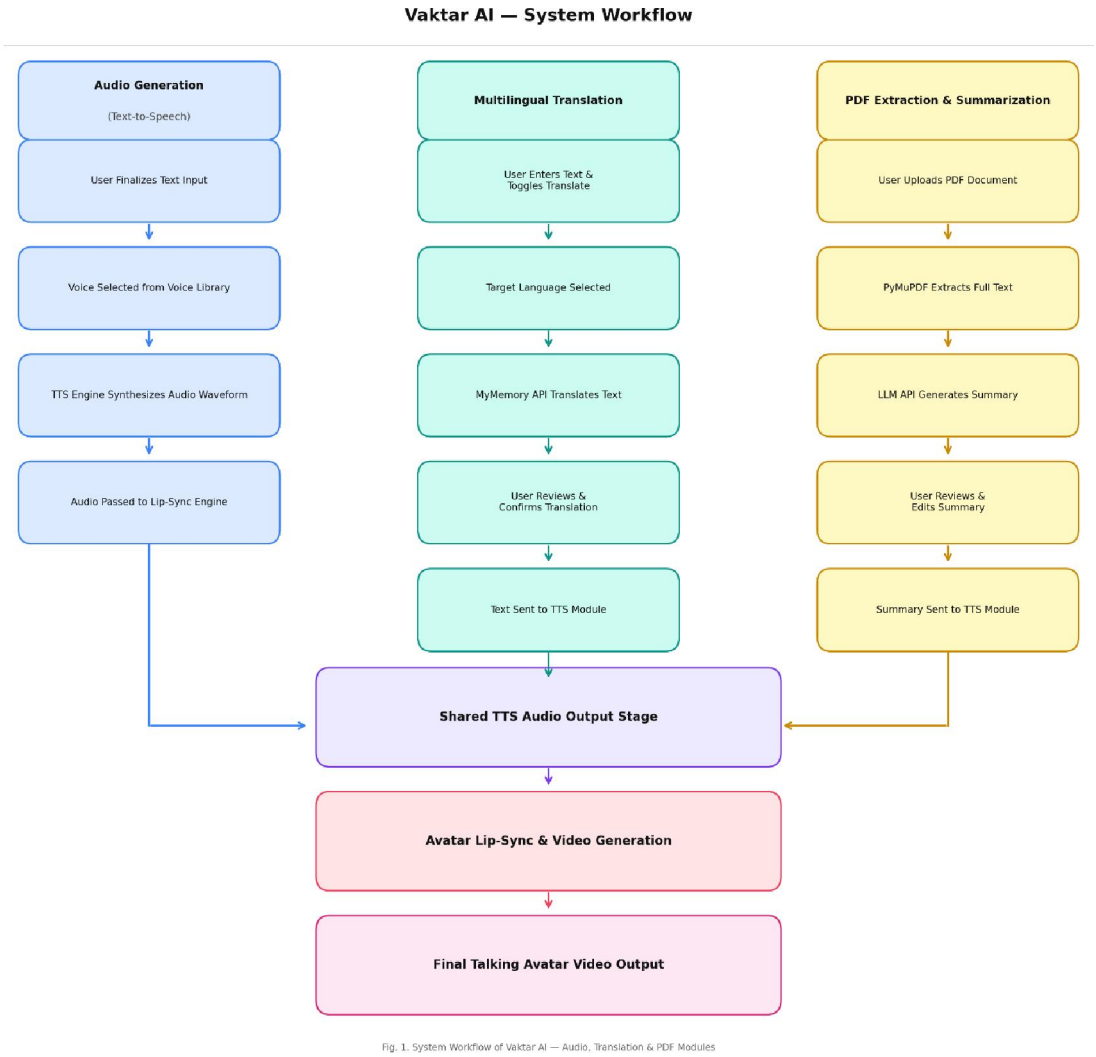

Fig. 1. System Workflow of Vaktar AI — Audio, Translation & PDF Modules

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, April 2026 Impact Factor: 8.2

C. PDF Extraction and Summarization Workflow

The PDF workflow commences with a PDF file being uploaded. This uploaded file is processed by PyMuPDF to extract the text on a page-by-page basis [5]. The extracted text is then sent to the summarization API using an LLM through a structured prompt asking for a succinct summary written in spoken language form. Prior to generating the avatar video, the user has an opportunity to edit the generated summary, after which, the pipeline joins the TTS video generation process [7]. In cases where large PDFs need to be processed, additional latency is added by this stage.

D. Combined Workflow Example

It is possible to chain all three modules within a single session as follows: (1) upload a PDF containing information on government awareness written in Hindi by the user; (2) extract text content using PyMuPDF; (3) have the LLM generate a summary of that text as a spoken passage; (4) translate the summary into Marathi via the translation module; (5) generate natural Marathi speech with the help of the TTS engine; (6) use Team 1's avatar engine to produce the video with the talking avatar for the user.

VI. COMPARATIVE ANALYSIS WITH EXISTING PLATFORMS

• Table II provides a comparative analysis of the extended capabilities of Vaktar AI in comparison to other popular avatar solutions. It is seen from the table that there is currently no platform that includes all five capabilities.

TABLE II: PLATFORM COMPARISON

Feature	Synthesia	HeyGen	D-ID	Vaktar AI (Team 1)	Vaktar AI (Team 2 Extension)
User-Photo Avatar	No(preset only)	Limited	Yes	Yes	Yes
Conversational AI Brain	No	No	No	Yes	Yes (inherited)
Neural TTS (Multi-voice)	Yes	Yes	Yes	Limited	Yes (Google/ Azure/XT TS)
Multilingual Translation (50+ langs)	No	Partial	No	No	Yes (MyMemory API)
PDF-to-Avatar Pipeline	No	No	No	No	Yes (PyMuPDF + LLM)
Zero-Cost Translation Layer	No	No	No	No	Yes
Modular Unified Pipeline	No	No	No	Partial	Yes

VII. APPLICATIONS OF THE PROPOSED SYSTEM

However, when combined, the audio synthesis, translation, and PDF modules of Vaktar AI have many practical applications that current technology cannot accommodate [1], [4]:

• Education / E-Learning: Teachers can submit their lecture notes, research articles, or course material PDF files and immediately create explainer videos with avatars speaking in several languages without manually producing videos or voice overs [5], [7].

• Multilingual Content Creation: Video creators and companies can create videos using avatars in 50+ languages by simply inserting text through the translation module [3].

• Corporate Communication:Organizations can transform policy documents, annual reports, and corporate news releases into interesting talking avatar videos, saving the time that would otherwise be spent reading bulky documents [5].

• Accessibility for Differently-Abled Users: The TTS module allows users, who cannot record themselves, to produce interesting avatar videos. The PDF module serves the needs of readers with reading difficulties through its capability of converting text material to multimedia form [2], [6].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, April 2026 Impact Factor: 8.2

• Healthcare and Public Awareness: Healthcare organizations may load awareness documents in PDF formats and produce avatar videos in multiple languages to share information related to health and wellness among communities in various regions [3], [7].
• Government and Public Service Communication: Government ministries may produce avatar videos from policy documents available in PDF format, allowing people to communicate with governments in regional languages without production expenses.
• Personal Branding and Digital Identity: People may produce their personal avatars in multiple languages through text input, allowing them to practice personal branding in multiple languages.

VIII. ADVANTAGES

The three modules proposed in this study provide multiple advantages regarding usability, accessibility, multilingualism, and architecture [1], [4]:
• Zero-Cost Translation: No API Key is required to translate texts between more than 50 languages using the MyMemory API at no cost [3].
• Expressive Human-Like Voiceover: State-of-the-art neural network-based TTS engines like XTTS and Azure TTS can produce human-like voice with adjustable tone, accent, gender, and even different languages without requiring any prior voice recordings [2], [6].
• PDF-to-Avatar Pipeline: None of the competing platforms offers the ability to create personalized videos from structured documents. Only this proposed pipeline allows for transforming PDFs to a personalized speech-driven avatar video [5], [7].
• Architecture Convergence: All three modules utilize the same TTS engine. The common audio stage helps achieve automatic scalability in case of any improvements of the engine itself [4].
• Reduced Dependency on Human Resources: The TTS module eliminates the need for professional voice actors or personal voice recordings, while the PDF module eliminates the need for manual scriptwriting from documents, reducing both cost and production time [2], [5].
• Modular Architecture: The flexible design allows room for further enhancements, which include emotion-sensitive TTS, OCR support for scanned PDFs, automatic voice-language compatibility, and live streaming [4].
• Consistency and Quality: Speech synthesis ensures high-quality, consistently accurate pronunciation of speech and accent throughout all video creations, which would not have been possible during the recording of human voices [6].
• Reusable Avatar Identity: The personalized avatar based on the user's image may be reused for all the three modules' video productions. An identical avatar can provide a Hindi PDF summary today and an Arabic business announcement tomorrow, thereby ensuring uniform identity.
• Voice Democratization: Non-native speakers of a certain language can express their authority using avatar videos. A university student from Pune may speak in Japanese, French, or Swahili even when he/she cannot speak those languages, a capability unique to this free-of-cost website.

IX. LIMITATIONS

While the suggested modules mark a significant step forward in creating avatars capable of communicating using text-to-speech technologies, it is imperative to address the following limitations of a technical and ethical nature:
• MyMemory API Rate Limits: With the limit of translations per day amounting to no more than 5,000 words, the free version will likely not suffice in case of high traffic. To tackle the issue, it will be possible to use language pair caching mechanisms as well as switch to the commercial version of MyMemory API [3].
• TTS Naturalness for Low-Resource Languages: Whereas the best TTS engines offer outstanding performance in English, French, Spanish, and other high-resource languages, TTS output becomes less natural due to lack of data when translating texts into low-resource languages. XTTS is able to overcome such difficulties, but it is not implemented in

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, April 2026 Impact Factor: 8.2

some deployment contexts [2], [6]. Once a source text is translated, the voice chosen by the user might fail to pronounce sounds specific to the target language.

• Variability in Quality of Extracted Texts From PDFs: PDFs containing scanned or many images cannot produce good quality text extraction using PyMuPDF because the tool does not natively provide OCR support. An OCR layer (such as Tesseract OCR) must be used in conjunction with the PDFs, which adds to the complexity of processing [5]. Large PDFs consisting of several pages result in some delay in the process of extraction, summarization, and optionally translating them before creating the videos [5].

• Summary Hallucinations Produced by LLMs: Generated summaries from PDFs by AI tools such as large language models can lead to hallucinations, where the models generate plausible but incorrect facts. It is highly recommended that a human checks the generated summaries before creating the videos [7].A great deal of compute power is needed when processing large PDFs, calling LLM summarization APIs, and generating TTS audio for several users simultaneously [4], [7].

• Lip-sync Engine for Multilingual Support: he lip-sync engine developed in the existing framework is built and validated mainly on the phonemes of the English language. Non-Latin alphabets and tonal languages such as Mandarin, Tamil, and others may have a higher chance of creating errors during synchronization due to different phonemic patterns. This constitutes a potential research gap to explore further [4].

• Ethical/Misuse Issues: There can be ethical issues in combining TTS voice generation with talking avatars since this might pose the risk of creating audio and video deepfakes. Voice synthesis can be misused for impersonating someone. Vaktar AI handles this issue by generating the talking avatar only when the user uploads their image and by not allowing the target voice to clone a third party's identity.

X. CONCLUSION

Three extension modules of Vaktar AI were introduced within this paper: AI-based audio synthesis, multilingual text translation, and PDF extraction along with summary generation [1], [4]. These modules complement the already developed talking avatar platform provided by Team 1, but, at the same time, tackle major gaps in the current design of communication avatars: lack of support for multiple languages, inability to ingest documents, and reliance on the presence of voice recordings of a user.

Audio synthesis is achieved through the use of advanced neural TTS (text-to-speech) engines, which include zero-shot multilingual approaches, such as XTTS, allowing natural and expressive speech synthesis without requiring any voice recording [2], [6]. Multilingual text translation was implemented by using MyMemory REST API, enabling the capability to translate any text into more than 50 different languages in real time and thus rendering the process of creating an avatar video truly global [3]. Lastly, the PDF module implements a unique pipeline for translating documents into personalized avatar video summaries [5], [7].

Due to the ability to merge all three components in the same TTS module, the system acquires an architectural design that makes it modular, scalable, and free from any modifications to Team 1's avatar generator engine. As seen from the results provided in Table II, no other commercial product offers this particular combination of features. Hence, the novelty of this product can be stated based on the analysis conducted.

Further advancements in the field of AI technologies will result in the growing importance of these modules within the same avatar platforms for such applications as effective communication between humans and computers, education for persons with limited opportunities, interlingual communications in multilingual businesses, and digital identities based on intelligent avatars. According to the plan, the proposed architecture will evolve with technological innovations, and the following features are expected to be added: emotional TTS, recognition of scanned documents, voice-language alignment, and live avatar rendering.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

REFERENCES

[1] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the real world: A survey on NLP applications," Information, vol. 14, no. 4, p. 242, 2023.
[2] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, et al., "XTTS: A massively multilingual zero-shot text-to-speech model," arXiv preprint arXiv:2406.04904, 2024.
[3] A. Fan, S. Bhosale, H. Schwenk, et al., "Beyond English-centric multilingual machine translation," Journal of Machine Learning Research, vol. 22, no. 107, pp. 1–48, 2021. [MyMemory API: https://mymemory.translated.net/doc/spec.php, Accessed: March 2026].
[4] M. Toshpulatov, W. Lee, and S. Lee, "Talking human face generation: A survey," Expert Systems with Applications, vol. 219, p. 119678, 2023.
[5] PyMuPDF Contributors, "PyMuPDF Documentation — PDF Text Extraction," [Online]. Available: https://pymupdf.readthedocs.io/ [Accessed: March 2026].
[6] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, et al., "WaveNet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.
[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.
[8] Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I. E. Davidson, R. Nyameko, et al., "Deepfake generation and detection: Case study and challenges," IEEE Access, vol. 11, pp. 143296–143323, 2023.
[9] P. Patil, O. Somkuwar, S. Mahale, and K. Kumavat, "Vaktar AI: A conceptual framework for AI-based talking avatar video generation with conversational intelligence," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 6, no. 3, pp. 325–331, April 2026.
[10] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, et al., "SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 8652–8661, 2023.
[11] Google Cloud, "Cloud Text-to-Speech Documentation," [Online]. Available: https://cloud.google.com/text-to-speech/docs [Accessed: March 2026].
[12] Microsoft Azure, "Azure Cognitive Services — Text to Speech," [Online]. Available: https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/ [Accessed: March 2026].