International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Intelligent Conversational Agents Based Custom Question Answering System

Nitin Sakhare[1,*], Jyoti Bangare[2], Dr. Deepika Ajalkar[3], Dr. Gajanan Walunjkar[4], Dr. Madhuri Borawake[5], Dr. Anup Ingle[6]

Submitted: 11/02/2023 **Revised:** 12/04/2023 **Accepted:** 09/05/2023

Abstract: Intelligent conversational agents have become increasingly popular in recent years, and they have numerous applications in education, customer service, and entertainment. In this paper, we present an intelligent conversational agent which will act like a historical personality. The goal of this research is to create a system that can provide accurate and engaging information about historical figures in a conversational manner. The digital characters respond to questions by providing audio responses and changing their facial expressions through lip-syncing. The model utilizes the Azure custom answering service to generate question-answer pairs, which are used to train the model to provide accurate answers to questions. The voice of the digital characters is cloned using the Tortoise TTS model of the TortoiseAI team. The audio responses generated by the voice cloning model are then utilized in conjunction with the VOCA and FLAME models and utilize an end-to-end speech-driven facial animation system based on a temporal GAN. The temporal GAN relies on a generator and three discriminators (frame, sequence, and synchronization discriminators) that drive the generation of an auto-lip-sync talking head using only a still 2D image of a person and a voice clip as input. for lip-syncing and facial expressions of the digital characters. The model's subjective listening test evaluated the lip-syncing and facial expressions, demonstrating that the digital characters produced believable and accurate responses. The proposed system allows users to add new characters and is suitable for educational deployment. User study results demonstrate high accuracy and engaging user experience, suggesting our approach is a promising advancement in educational conversational agents.

Keywords*: Tortoise TTS, Custom Question Answering, VOCA, FLAME, GAN.*

1. Introduction

In recent years, advances in artificial intelligence and machine learning have led to the creation of digital characters that can interact with humans in a more natural and engaging way. Digital characters can be used in various applications, such as virtual assistants, customer service chatbots, and video game characters. One important aspect of creating believable digital characters is ensuring that their facial expressions and lip-syncing are accurate and natural. Lip syncing involves synchronizing the movement of a character's mouth with the audio of their speech. In the past, lip-syncing was a time-consuming and difficult task that required a lot of manual work. However, recent advances in deep learning have led to the development of models that can generate accurate lip-syncing automatically [1]. In this research paper, we present a paper that involves the creation of digital characters of historical figures, including Shivaji Maharaj and Albert Einstein, to answer questions visually. The digital characters respond to questions by providing audio responses and changing their facial expressions through lip-syncing. The model utilizes the Azure custom answering service to generate question-answer pairs, which are used to train the model to provide accurate answers to questions. The Azure custom answering service is a cloud-based natural language processing service that allows users to build custom question-answering models. It uses machine learning to extract relevant information from a knowledge base and provide accurate answers to user questions [2]. The voice of the digital characters is cloned using the Tortoise TTS model of the TortoiseAI team. The Tortoise TTS model is a neural text-to-speech model that can generate natural-sounding speech from text[3]. It uses deep learning to synthesize speech that sounds similar to a specific voice. We created a dataset by clipping audio files from movies and YouTube videos to train the Tortoise TTS model to clone the voices of Shivaji Maharaj and Albert Einstein. The audio responses generated by the voice cloning model are further processed using the VOCA and FLAME models for achieving lip-syncing and facial expressions synchronization. These models contribute to creating realistic and expressive audiovisual outputs. By combining the voice cloning model with the VOCA and FLAME models, our system achieves synchronized audio and visual outputs, providing an immersive and engaging experience for users. The model's subjective

[1,*] *nitin.sakhare@viit.ac.in*
[1,6] *BRACT's Vishwakarma Institute of Information Technology, Pune*
[2] *MKSSS's Cummins College of Engineering for Women, Pune*
[3] *G.H. Raisoni College of Engineering and Management, Pune*
[4] *Army Institute of Technology, Pune*
[5] *PDEA's College of Engineering, Pune*

listening test evaluated the lip-syncing and facial expressions of the digital characters, and the results showed that the digital characters produced believable and accurate responses. The research paper concludes with a summary of the work, its objectives, and potential applications.

2. Related work

Image animation methods can be separated into supervised, which require knowledge about the animated object during training, and unsupervised, which does not. Such knowledge typically includes landmarks [4-6], semantic segmentation [7], and parametric 3D models [8-9]. As a result, supervised methods are limited to a small number of object categories for which a lot of labelled data is available, such as faces and human bodies. Early face reenactment work [10] fitted a 3D morphable model to an image, animating and rendering it back using graphical techniques. Further works used neural networks to get higher-quality rendering, sometimes requiring multiple images per identity. A body of works treats animation as an image-to-image or video-to-video translation problem A further group of works re-target animation from a driving video to a source frame. X2Face builds a canonical representation of an input face and generates a warp field conditioned on the driving video. Monkey-Net learns a set of unsupervised key points to generate animations. Follow-up work substantially improves the quality of animation by considering a first-order motion model (FOMM) [11] for each key point, represented by regressing a local, affine transformation. This work utilizes unsupervised improved FOMM [12] which is PCA-based motion estimation.

3. Text Extraction

The Azure Custom Answering Service is a cloud-based natural language processing service that allows users to build custom question-answering models. It is designed to extract relevant information from a knowledge base and provide accurate answers to user questions [13]. In this research, we used the Azure Custom Answering Service to train our model to answer questions about historical figures such as Shivaji Maharaj and Albert Einstein. To train the Azure Custom Answering Service, we provided a carefully curated knowledge base that contained information about the lives and achievements of Shivaji Maharaj and Albert Einstein. The knowledge base was created from various sources, including books, articles, and online resources, and was designed to ensure that it contained accurate and relevant information. The service used our knowledge base to generate question-answer pairs. The service uses machine learning algorithms to analyze the knowledge base and extract relevant information that can be used to generate questions and answers. The question answering system uses a layered ranking approach. The data is stored in Azure search, which also serves as the first ranking layer. The top results from Azure search are then passed through question answering's NLP re-ranking model to produce the final results and confidence score. The algorithms use NLP techniques to ensure that the questions and answers are grammatically correct and understandable [14]. Once the question-answer pairs were generated, they were used to train the model. The Azure Custom Answering Service used a combination of supervised and unsupervised learning techniques to train the model. In supervised learning, the model was trained on labeled data, which means that the correct answers to the questions were provided during training. In unsupervised learning, the model was trained on unlabeled data, which means that the answers to the questions were not provided during training [15]. During training, the model was optimized to provide accurate answers to a wide range of questions about Shivaji Maharaj and Albert Einstein. The model learned to identify the most relevant information in the knowledge base and use it to generate accurate and informative answers. The training process was iterative, and the model was updated based on the feedback received from the subjective listening test. Azure Custom Answering Service is a powerful tool for building custom question-answering models. It uses machine learning and natural language processing techniques to generate question-answer pairs and train the model to provide accurate answers [16]. The carefully curated knowledge base provided by us was crucial in ensuring that the model provided accurate and relevant answers to user questions about Shivaji Maharaj and Albert Einstein.

4. Voice Cloning using Tortoise TTS

4.1 Text to Speech

TTS, or speech synthesis, is a system that takes text as input and generates an audio signal from it. The primary goal of modern TTS is to make a synthesized speech from text sound not only comprehensible but also natural. However, achieving a high level of naturalness is often subjective and evaluated using metrics such as the Mean Opinion Score (MOS). MOS is a subjective rating obtained from human listeners who evaluate synthetic speech samples generated by the TTS system. Listeners rate speech samples on a scale of 1 to 5, where 1 represents poor quality, and 5 represents excellent quality. The scores from multiple listeners are then averaged to determine the overall MOS for the TTS system. The text analysis module converts a text sequence into linguistic features. The acoustic model generates acoustic features from those linguistic features. Finally, the vocoder synthesizes a waveform from those acoustic features.

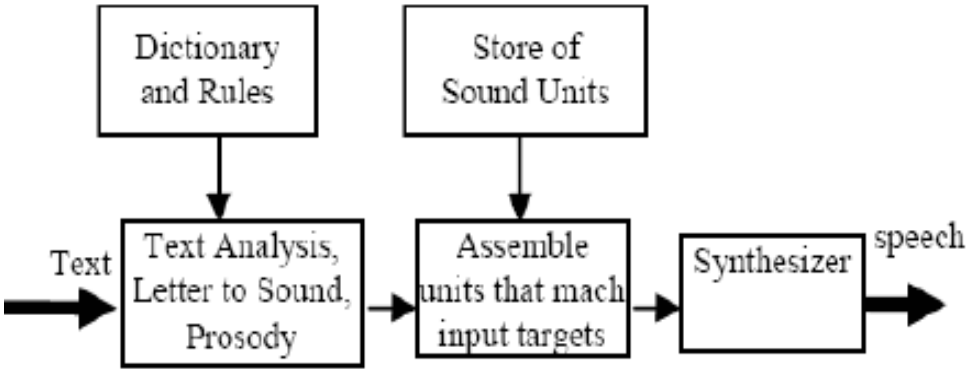

Fig. 1 Basic Components of a TTS model

4.2 Voice Cloning

The field of voice cloning is a challenging subfield of TTS. Its goal is to generate speech that closely matches an audio reference voice. To clone the voice of the digital characters, we used the Tortoise TTS model[14] Tortoise TTS consists of five separately trained neural networks that are pipelined together to produce the final output. Its components draw inspiration from image generation models such as DALL-E[15] and denoising diffusion probabilistic models. The production of the final waveform has the following steps:

Text input and reference clips are fed to the autoregressive decoder that outputs latents and corresponding token codes representing highly-compressed audio data. This step is repeated several times to produce multiple "candidate" latents.

The CLVP (Contrastive Language-Voice Pretraining) and CVVP (contrastive voice-voice pretraining) models select the best candidate. The CLVP model produces a similarity score between the input text and each candidate code sequence, while the CVVP model produces a similarity score between the reference clips and each candidate. These two similarity scores are combined with a weighting provided by the user, and the candidate with the highest total similarity proceeds to the next step.

The diffusion decoder then consumes the autoregressive latents and reference clips to produce a mel-spectrogram representing the speech output. Finally, a UnivNet vocoder is used to transform the mel-spectrogram into actual waveform data.

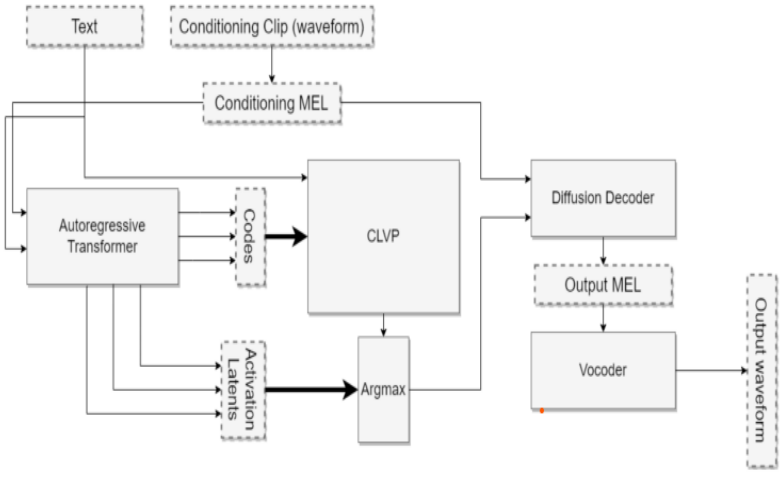

Fig. 2: TorToise architectural design diagram. Inputs of text and a reference audio clip (for speaker cloning) flow through a series of decoding and filtering networks to produce high-quality speech [16]

4.2.1 Conditioning Input

A unique design choice made with TorToise is an additional input which is provided to both the autoregressive generator and the DDPM. The speech conditioning input starts as one or more audio clips of the same speaker as the target. These clips are converted to MEL spectrograms and fed through an encoder consisting of a stack of self-attention layers. The autoregressive generator and the DDPM have their own conditioning encoders, both of which are learned

alongside their respective networks.

The output of these layers is averaged to produce a single vector. The vectors from all of the encoded conditioning clips are then averaged again before being fed as input into the autoregressive or conditioning networks.

The intuition behind the conditioning input is that it provides a way for the models to infer vocal characteristics like tone and prosody such that the search space of possible speech outputs corresponding to a given textual input is greatly reduced.

4.2.2 CLVP

As mentioned earlier, a good strategy for gathering expressive outputs from generative models is using a qualitative discriminator to re-rank several outputs, then choosing only the best. DALL-E uses CLIP for this. The same type of approach used for CLIP can be applied to speech as most TTS datasets are simply pairings of audio clips and text. By training a model on these pairs in a contrastive setting, the model becomes a good discriminator for speech.

For the model we are using, it is trained using Contrastive Language-Voice Pretrained Transformer, or CLVP. It has many of the same properties of CLIP, but notably serves as a scoring model for use in re-ranking TTS outputs from the AR model.

To make this work efficiently in inference, CLVP has been trained to pair discretized speech tokens with text tokens. This way, CLVP can rerank multiple AR outputs without the expensive diffusion model being invoked.

4.2.3 Dataset

To create the dataset used to train the Tortoise TTS model, we collected audio clips of the historical figures from various sources, including movies, speeches, and interviews. We chose clips that were of high quality and were representative of the historical figures' voices. The clips were then carefully curated and organized into a dataset that was used to train the Tortoise TTS model. The Tortoise TTS model accepts audio files in the WAV format, which is a standard digital audio format that is widely used in the industry. The WAV format is capable of storing uncompressed audio data, which makes it ideal for high-quality audio recordings. The audio clips in our dataset were all in the WAV format, which ensured that the Tortoise TTS model was able to clone the voices of the historical figures accurately. The Tortoise TTS model works by analyzing the features of the voice in the audio clips and then generating speech that mimics those features. It does this by using a deep neural network that is trained on the dataset of audio clips. The neural network is trained to learn the relationships between the audio features and the speech generated by the historical figures. Once the model is trained, it can be used to clone the voices of the historical figures and generate speech for the digital characters [117]. Overall, the Tortoise TTS model was a crucial component of the model, as it allowed us to clone the voices of the historical figures and generate speech for the digital characters. The carefully curated dataset of audio clips that we used to train the model ensured that the voices generated by the Tortoise TTS model were very similar to the actual voices of the historical figures. This helped to enhance the realism of the digital characters and made the model more engaging and interactive for users.

5. Video Generation

We have used an unified pipeline that aims to generate talking characters for various applications, including education and healthcare. The pipeline leverages state-of-the-art AI models to produce realistic content featuring humanoid characters with synchronized facial expressions, voice, and motion. Several key models, including VOCA, FLAME, Speech-Driven Facial Animation, and First Order Motion Model, are incorporated into this pipeline to enable the creation of audio and video outputs using diverse data sources.

5.1 VOCA

The VOCA (Voice Operated Character Animation) model is a deep neural network architecture specifically designed for generating realistic and synchronized lip-sync animation for 3D characters based on a provided voice input. It plays a crucial role in the pipeline for generating talking characters with accurate and expressive facial movements.

The architecture of the VOCA model combines several components to achieve its functionality. At a high level, the model consists of an audio encoder, a video decoder, and a facial mesh estimator.

The audio encoder takes the input voice clip and processes it to extract relevant acoustic features. These features capture the temporal and spectral characteristics of the voice, including phoneme information and prosodic cues.

The video decoder, also known as the lip-sync decoder, operates on the extracted acoustic features and produces a sequence of visual features that correspond to the lip movements. These visual features are generated in a synchronized manner with the input voice clip [18]

To estimate the facial mesh, the VOCA model incorporates a facial mesh estimator. This component predicts the 3D shape and pose of the character's face, providing a foundation for generating accurate and realistic lip movements. The facial mesh estimator leverages 4D scans of human faces during training to capture the variability and dynamics of facial

expressions.

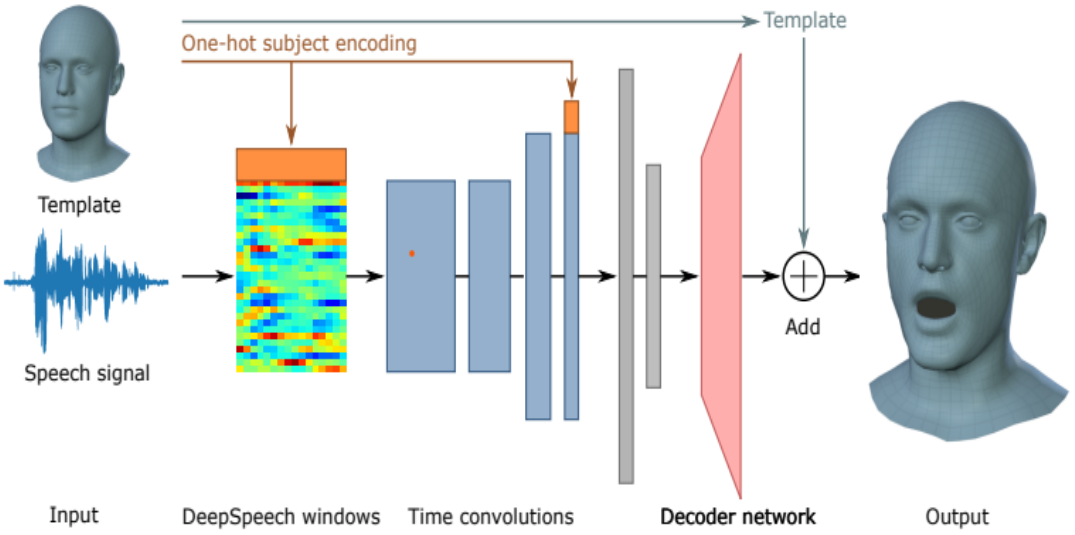

Fig.3 architecture of VOCA [18]

During the training process, the VOCA model is trained on a large dataset of paired audio and video clips. The model learns to map the acoustic features of the voice to the visual features of the lip movements, establishing a correspondence between the two modalities.

One notable aspect of the VOCA model is its ability to generate lip-sync animation for a neutral 3D face. This means that the model can adapt to various characters or faces by aligning the generated lip movements to the specific facial structure and characteristics.

Overall, the architecture of the VOCA model allows for the synthesis of realistic lip movements that are synchronized with the provided voice input. By combining audio and visual information, the model enables the generation of talking characters with accurate and expressive facial animations. The architecture of VOCA is illustrated in Fig.3

5.2 FLAME

The FLAME (Faces Learned with an Articulated Model and Expressions) model is an advanced framework designed to generate realistic and customizable facial expressions for virtual characters. FLAME seamlessly integrates with existing graphics software and provides flexibility in data fitting and expression transfer. At its core, the FLAME model utilizes a combination of shape spaces, articulation structures, and expressive blendshapes to achieve its functionality. The architecture consists of several key components, including a linear shape space, pose-dependent corrective blend shapes, global expression blend shapes, and facial landmarks.

The linear shape space in FLAME is constructed from a dataset of 3,800 3D human head scans. This shape space captures the variations in facial shapes, allowing the model to generate diverse facial structures [18]. FLAME also incorporates articulation structures, such as articulated jaw, neck, and eyeball models. These structures enable more realistic and natural movements of these specific regions in the generated facial expressions.

To handle fine-scale details and correct deformations, FLAME employs pose-dependent corrective blendshapes. These blendshapes are designed to capture the subtle changes in facial expressions that occur due to head pose variations. By incorporating these blendshapes, FLAME can produce more accurate and realistic facial movements across different head orientations. Furthermore, FLAME includes global expression blendshapes that allow for the customization of overall facial expressions. These blendshapes enable users to control the intensity and appearance of specific expressions, such as a smile or a frown. FLAME also leverages facial landmarks to assist in the fitting process. These landmarks provide key points of reference on the face, aiding in aligning the generated model with the target image or video frames.Overall, the FLAME model offers a comprehensive framework for generating expressive and customizable facial animations. Its architecture combines shape spaces, articulation structures, corrective blendshapes, global expression blendshapes, and facial landmarks to enable the creation of realistic and personalized facial expressions for virtual characters.

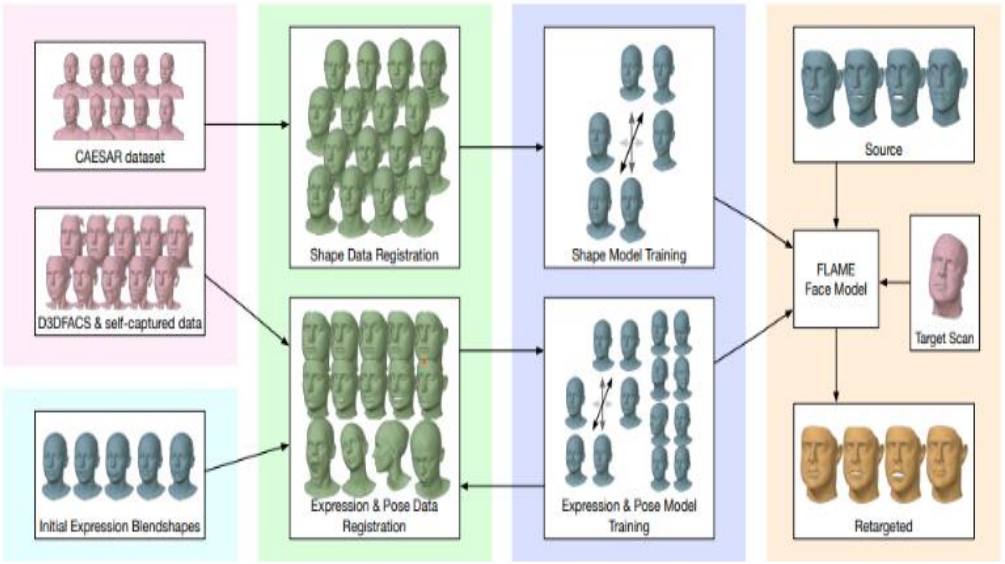

Fig.4 Overview of the face registration, model training, and application to expression transfer [18]

In terms of voice-to-video generation, the pipeline offers two approaches. The first approach involves generating a 3D intermediary talking head that automatically synchronizes with a recorded voice clip. This is achieved by utilizing VOCA to generate lip-sync animation on a neutral 3D face and leveraging FLAME to customize facial expressions and movements as shown in fig.4.

Through the utilization of these models, we have been able to create videos based on our audio files, enhancing the overall quality and realism of the generated content. Voice is a critical element of generated output given that watching a generated video without audio is less likely to be perceived as realistic. To add audio to the generated video, the user provides a recorded voice clip to drive the generated character's speech. As shown in Fig. 5, we explore two pathways for converting the voice clip input into intermediary facial animation before generating the final video and audio output. In fig.5 we have explained overall architecture of our model like the addition of audio file, character image and also driven video which will provide facial expressions to image and also the involvement of VOCA and FLAME model.

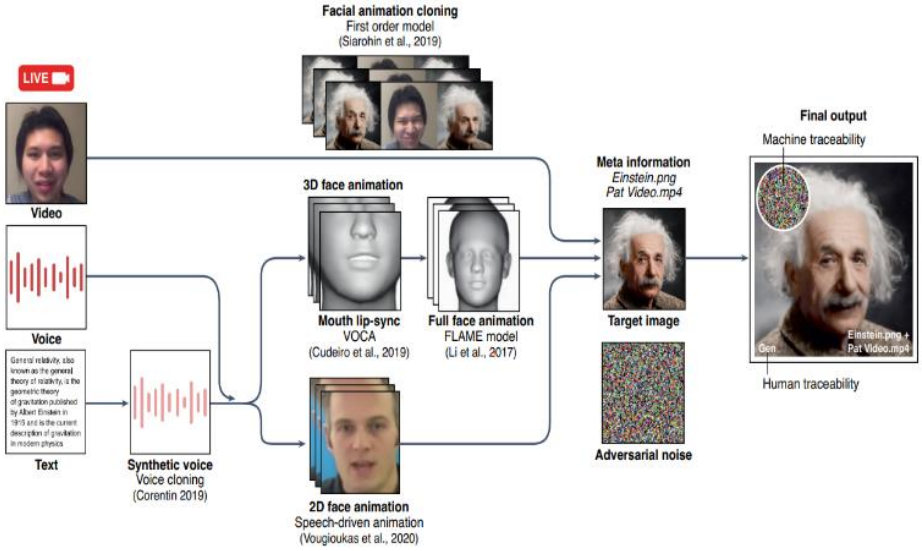

Fig.5 The overall Architecture of the proposed system

We used a driven video to generate facial expression of image according to that video rather add facial expression to the frames of characters according to the frames of driven video. It has been shown in Fig.6.

Fig. 6 Facial Expressions of the Frames

6. Conclusion and Future Work

In this study, we developed a system to create digital characters capable of visually answering questions, using historical figures Shivaji Maharaj and Albert Einstein as examples. Our goal was to generate lifelike characters that not only provide answers but also engage users through facial expressions and lip syncing. To achieve this, we utilized various advanced machine learning models, including Azure Custom Answering Service, Tortoise TTS, VOCA, and FLAME. We utilized the Azure Custom Answering Service to generate question-answer pairs based on a knowledge base of the historical figures. Subsequently, we employed Tortoise TTS to clone the voices of the characters. The video generation models, VOCA and FLAME, played a crucial role in creating videos that matched the lip syncing with the audio. The system we developed introduces natural and captivating methods for user interaction with digital assistants and educational tools. Its potential applications span across virtual assistants, education, and entertainment.

While the lip syncing and facial expressions received positive feedback from participants in our subjective listening test, there are still opportunities for improvement. Enhancements could be made in terms of increasing the variability of facial expressions and improving the smoothness of the lip syncing process. Furthermore, incorporating computer graphics techniques can contribute to making the characters even more realistic. Our study demonstrates the potential of combining machine learning models such as VOCA and FLAME to create visually engaging and lifelike digital characters capable of answering questions. We anticipate that the outcomes of our research can revolutionize the way we interact with virtual assistants, educational tools, and entertainment media, opening up new possibilities for future developments. Our system's contributions lie in the creation of more natural and engaging ways for users to interact with digital assistants or educational tools. It has the potential to be used in various applications such as virtual assistants, education, and entertainment. However, there is still scope for improvement in this system. For instance, while the lip-syncing and facial expressions were rated highly by participants in the subjective listening test, there were still some areas that could be improved upon, such as increasing the variability of facial expressions and improving the smoothness of the lip-syncing. Future work could include expanding the system to include more historical figures or famous personalities, further refining the lip-syncing and facial expression generation, and improving the system's overall naturalness and lifelike quality. Also characters can be made more realistic using computer graphics. Overall, this paper demonstrates the potential of combining several machine learning models to create engaging and lifelike digital characters that can answer questions visually. We believe that our system's potential applications are vast and could transform the way we interact with virtual assistants, educational tools, and even entertainment media in the future.

Acknowledgment

The successful completion of this research paper was made possible through the support and contributions of numerous individuals and organizations, to whom we extend our gratitude.

We would also like to extend our appreciation to Azure Custom Answering Service and TortoiseAI team for providing us with the necessary tools and resources to carry out our research effectively.

Additionally, we would like to thank the National

Geographic and Zee Studios YouTube channels for being useful to get the audio samples of Albert Einstein and Shivaji Maharaj, which were crucial to our research.

References

[1] Khan, M.Z., Jabeen, S., Khan, M., U., 2021. "A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks, in *IEEE Access*, vol. 9, pp. 1250-1260, 2021, doi: 10.1109/ACCESS.2020.3015656.

[2] M. A. Kia, A. Garifullina, M. Kern, J. Chamberlain and S. Jameel, 2022. Adaptable Closed-Domain Question Answering Using Contextualized CNN-Attention Models and Question Expansion in IEEE Access, vol. 10, pp. 45080-45092, 2022, doi: 10.1109/ACCESS.2022.3170466.

[3] Sakhare, N., Shaik, I, Kagad, S., Malekar, H., Dalal, M. 2020. Stock market prediction using sentiment analysis International Journal of Advanced Science and Technology, Vol. 4, issue 3, 2020

[4] Egor, Z., Aliaksandra, S., Egor, B., and Victor L., 2019. Few-shot adversarial learning of realistic neural talking head models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019

[5] Sakhare, N., Shaik, I., 2022. Technical Analysis Based Prediction of Stock Market Trading Strategies Using Deep Learning and Machine Learning Algorithms, International Journal of Intelligent Systems and Applications in Engineering, 2022, 10(3), pp. 411–421.

[6] Sungjoo H., Martin K., Beomsu K., Seokjun S., and Dongyoung K. 2020. Few-shot face reenactment preserving identity of unseen targets. In Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[7] Sakhare, N., Joshi, S., , "Criminal Identification System Based On Data Mining" 3rd ICRTET, ISBN, Issue 978-93, Pages 5107-220, 2015

[8] Zhenglin G., Chen C., and Sergey T., 2019, 3d guided fine-grained face manipulation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[9] Wen L., Zhixin P., Jie M., Wenhan L., Lin M, and Shenghua Gao. 2019. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In Proceedings of the IEEE International Conference on Computer Vision, 2019.

[10] Justus T., Michael Z., Marc S., Christian T., and Matthias Ni., 2019. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[11] Sakhare, N., Joshi, S., 2014. Classification of criminal data using J48-Decision Tree algorithm. IFRSA International Journal of Data Warehousing & Mining, Vol. 4, 2014.

[12] Fan, B., Wang, L., Soong, F., K. and Xie, L. 2015. Photo-real talking head with deep bidirectional LSTM. In International Conference on Acoustics, Speech and Signal Processing, pages 4884–4888, 2015

[13] Chen, C., Qiming, H., and Kun, Z., 2014. Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. 33, 4, Article 43 (July 2014), 10 pages. https://doi.org/10.1145/2601097.2601204

[14] Sakhare, N., Shaik, I.,Saha,S. 2023 Prediction of stock market movement via technical analysis of stock data stored on blockchain using novel History Bits based machine learning algorithm. IET Soft.1–12(2023). https://doi.org/10.1049/sfw2.1209212

[15] Aliaksandr S., Stéphane L., Sergey T., Elisa R., and Nicu S. 2019. First order motion model for image animation. In Proceedings of the Neural Information Processing Systems Conference, 2019

[16] Shengju Q., Kwan-Yee L., Wayne W., Yangxiaokang L., Quan W., Fumin S., Chen Q., and Ran H. 2019. Make a face: Towards arbitrary high fidelity face manipulation. In Proceedings of the IEEE International Conference on Computer Vision, 2019.

[17] Sakhare, N. , Verma, D. , Kolekar, V. , Shelke, A. , Dixit, A. and Meshram, N. 2023. E-commerce Product Price Monitoring and Comparison using Sentiment Analysis. International Journal on Recent and Innovation Trends in Computing and Communication. 11, 5 (May 2023), 404–411. DOI:https://doi.org/10.17762/ijritcc.v11i5.6693.

[18] Yuval N., Yosi K., and Tal H., 2019. Subject agnostic face swapping and reenactment. In Proceedings of the IEEE International Conference on Computer Vision, 2019.