iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34048

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Virtual Self: A Text-driven Facial Animator

Waheeda Shaikh[1], Batul Petiwala[2], Moiz Sitabkhan[3], Mohammed Suratwala[4]
[1]*Assistant Professor,* [2, 3, 4]*Student, Department of Computer Engineering, M. H. Saboo Siddik College of Engineering, Mumbai, India*

Abstract: ***In this paper we present a system that generates a video of a person from only one image that is given to it, with complete facial animation and speech generated from the text message. The purpose of this project is to create a talking head that can (a) replicate the facial movements of the person whose image is given and (b) sync those movements with the speech that will be generated from the text input that is written to deliver the message the user wants their digital human to convey. Only two inputs, viz. An image of the user and text message that needs to be delivered are required by the system. This paper presents a method that uses only a single base model to create a personalized video, eliminating the hassle of training it every time. In a nutshell, the output of the system will be a two-dimensional clone of an individual made entirely of a still image of a person, that conveys the given text message to the target audience through replicated facial expressions.***

Keywords: Expression Synthesis; Facial Animation; Lip-Synching; Text-Driven Animation; Speech-Driven Animation.

I. *INTRODUCTION*

The quest for replicating human actions and behaviours in the past few decades has been fuelled by the advancements that are coming about in the fields of Machine Learning, Artificial Intelligence and Deep Learning. Human-like robots and digital assistants have already gained popularity in the recent past, due to the great inventions of voice assistants like Siri, Google Home and Alexa that are available in almost every smart device and in almost every household.

Even tools like a Text-to-Speech converter have evolved to include more "life-like" voices that read whatever is given to them in text format. The visual representation of such tools would improve speech retention drastically. It would be even more personal if this voice could be a two-dimensional representation of ourselves, this would make it possible for us to be present in places without being there physically. A visual clone or digital human can be made that looks and expresses just like the person, with only one photograph of them. This digital human can be made to deliver the speech that was given as text input to it.

There are a number of ways that this project can be used:

1) In distance learning, where video lectures could be delivered by digital clones in the absence of the teacher.
2) A chatbot, audio bot or robotic video bot could be replaced with a life-like representation of ourselves.
3) A presentation can be given without having to actually worry about what needs to be said.
4) In public announcement systems where instructions need to be given to a mass crowd, it would be more interactive and appealing to the audience to watch someone than only listening.

This automation would reduce the effort we have to take to give our message. We only need to provide a text input and an image to see a duplicate of us say and express things that we need it to. All this can be done with minimum effort from our side.

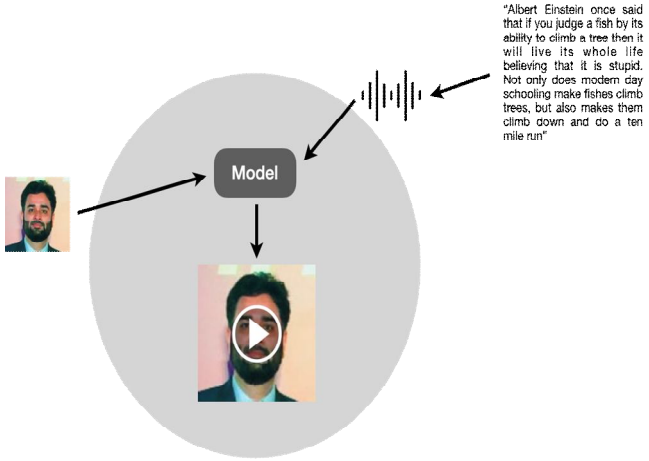


Fig. 1. An overview of the system

II. RELATED WORK

Various techniques have been proposed to animate a face model powered by text or voice, which is generally referred to in the literature as a "talking head." Text or speech-driven video synthesis is not a recent concept in computer vision; in reality, it has piqued researchers' interest for decades.

The relationship between acoustics, vocal-tract, and the facial motion was first researched by Yehia et al. (1998) [1], who discovered a close relationship between video and auditory features, as well as a weak pairing between head motion and the speech signal's fundamental frequency (Yehia et al. 2002) [2].

A text-driven approach employs text-to-speech and a text-to-face form synthesizing unit to produce facial animation [10, 11]. A similar approach is used by many speech-driven techniques, which is to directly map an acoustic feature sequence into a visual feature sequence. The underlying face model can also be used to categorize the above approaches, into model-based [3, 11, 12, 13, 14, 15] and image-based [6, 10, 16, 17, 18, 19]. The most popular methods used in each approach are presented below.

A. Using Text to Generate Photorealistic Facial Animations (Combination of Animation Unit with DNN and HMM)

It is a technique that uses text to produce a photorealistic facial animation, which is based on a combination of Deep Neural Network (DNN) and Hidden Markov Model (HMM) for an immersive agent implementation with facial features. The use of Animation Unit (AU) has been made in the method put forward.

AU is used for the expression of the condition for every region on the face and can be gathered using Kinect. HMM is used to create facial features, which are then translated into pixel intensities using DNN. We look into the best conditions for HMM and DNN training. The proposed methodology is then compared to a traditional technique based on principal component analysis (PCA) in an objective assessment. [20]

B. Deep Neural Networks-Based Synthesis

Many current methods use neural networks as a result of recent developments in deep learning. Taylor et al.(2017) have made the use of a deep neural network (DNN) to translate a chain of phonemes into a variety of forms for the section of the face, nose below [7]. Since phonemes are used instead of unprocessed audio, the process is subject-independent. Convolution Neural Networks (CNNs) are used by Karras and others (2017) construct 3D meshes of a particular human by remodelling audio information using Convolution Neural Networks (CNNs). The aforementioned system is broken down into smaller networks that are responsible for capturing articulation dynamics and estimating the mesh's 3D points.

C. Hidden Markov Models-Based Synthesis

Hidden Markov Models (HMMs) were used in the primitive methods to capture the minute distinctions of speech and video progressions for facial animation. Simons and Cox (1990) used vector quantization to construct a concrete depiction of audio / video components, which thereafter were used in their completely connected Markov Model as states [4]. Yamamoto et al. (1998) use a similar approach to measure the lip variables series[5]. The Video Rewrite methodology (Bregler et al. 1997) produces an array of triphones that are used to scan a repository for mouth photos by employing the same standards [6]. Ultimately, the observed solution is built by real-time synchronizing the voice with the frames, after which, spatially binding the jaw pieces to the background face is done.

D. Blending and Selecting Visual Features

Cao et al. (2005) created an anime graph, which is a graph of visual representations that correspond to audio features [3]. Under some co-articulation and smoothness constraints, the graph is searched for a sequence that best reflects a given utterance. This method learns to recognize the emotion in a speaker's voice and adjusts the animes to create gestures around the complete face. The end goal is met by taking the anime sequence and time-warping it to sync with the rhythm of the spoken statement and smoothen it by blending, the final result is achieved.

III.SYSTEM DESCRIPTION

A. Input Interface

A Node.js web application is used to provide an interface to the user. To be able to create and generate a personalized video, 3 inputs will be required by the web app, viz., a text input that has the message needed to be conveyed by the talking head, an image of the user that will be used for the personalized animation of that talking head and gender selection of the person so as to generate a voice in the same gender as that of the user to power the 2D clone.

B. Proposed System

The following are the 3 important inter-dependent modules of the project:

1) *Speech Synthesis:* Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech computer or speech synthesizer, and can be implemented in software or hardware products [23]. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

The breakdown of this module is as follows:

a) *Text Normalization:* Text normalization is the process of transforming text into a single canonical form that it might not have had before. It converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words.

b) *Text-to-phoneme:* In this process, phonetic transcriptions are assigned to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation [25].

c) *Synthesizer:* The synthesizer then converts the symbolic linguistic representation into sound.

2) Facial Animation System: Based on an incoming audio stream, a face image is animated with full lip synchronization and expression. An animation sequence using optical flow between visemes is constructed, given an incoming audio stream and still pictures of a face speaking different visemes. Rules are formulated based on coarticulation and the duration of a viseme to control the continuity in terms of shape and extent of lip opening. In addition to this new viseme expression combinations are synthesized to be able to generate animations with new facial expressions.

a) *Frame Discriminator Module:* The Frame Discriminator evaluates individual frames taken from synthetic/real sequences. This drives the Generator to produce frames that are detailed.

b) *Sequence Discriminator Module:* The Sequence discriminator evaluates sequence -audio pairs to determine if they are real or synthetic. This drives the audio and video to be in sync and encourages the generation of facial expressions (e.g., blinks)

c) *First Order Motion Model:* It basically copies the features of generated animation on the image provided by the user. Motion Model consists of generating a video sequence so that an object in a source image is animated according to the motion of a driving video (The Facial Animation created from Audio). The framework addresses this problem without using any annotation or prior information about the specific object to animate. Once trained on a set of videos depicting objects of the same category (e.g., faces, human bodies), method can be applied to any object of this class. Steps 1, 2 and 3 are generalized to create a mesh like animation for every audio and text input that is given to it. This mesh makes the process of customization easier, so that the system does not have to animate a new user's image every time. Instead, it can just overlay the image of the user onto the mesh that is already animated.

3) *Audio and Video Synchronization using FFmpeg:* FFmpeg is a free and open-source software project consisting of a large suite of libraries and programs for handling video, audio, and other multimedia files and streams. At its core is the FFmpeg program itself, designed for command-line-based processing of video and audio files. FFmpeg will sync the audio generated and the video created and produce the final output.

4) *Web Application*

a) *The web interface will obtain the following 3 inputs from the user:*

- Image of the user
- Text Input that needs to be digitalized
- Gender of user for voice selection

b) *The final output would be displayed after the processing is done and user can download the video from the website. Node.js will be used to create a web application framework wherein all the python scripts will be called from the backend and will be used to display output once complete.*

The entire process can be seen in the following flowchart of how the expected system will be:

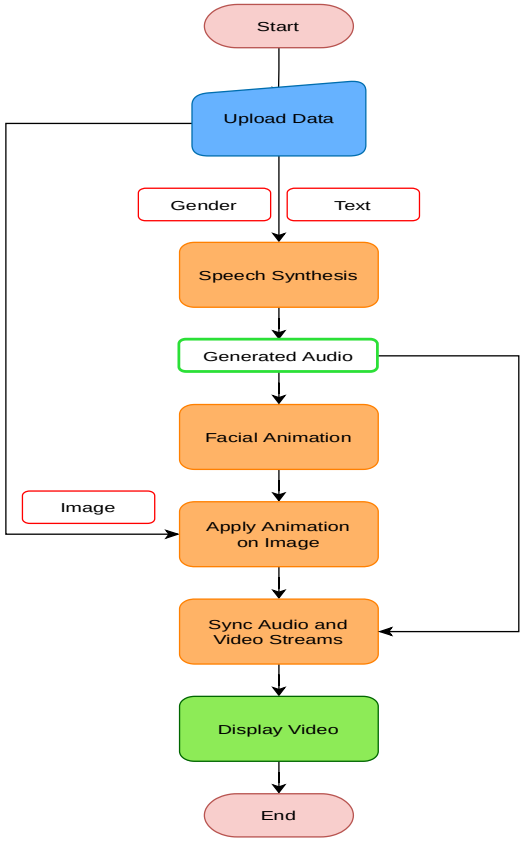


Fig. 2. Flowchart

C. Integration and Rendering of the final Output

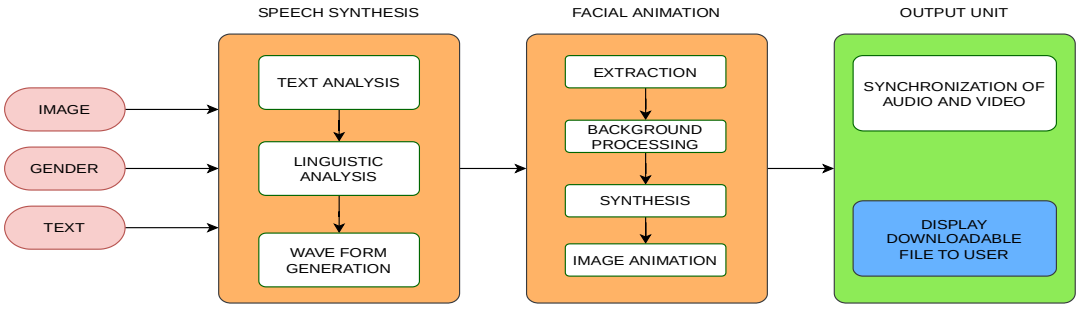


Fig. 3. Data Flow Diagram

The above dataflow diagram is better explained in the following steps. The complete project with all the integrated modules will work in 3 steps:

1) Speech Synthesis

a) Text Analysis: The input sentence is segmented into token. After tokenization, each word is determined as part of speech (POS) tagging. Part-of-speech is a process assigning correct POS tag to each word in a sentence from a given set of tags [27].

b) Linguistic Analysi

- *Phasing:* Converting each word into smaller predefined parts which have individual phonetic sound
- *Intonation:* The objective of this module is to determine the melodic pattern of an utterance. Intonation is primarily a matter of variation in the pitch level of the voice.
- *Duration:* This is the part where the length of the audio will be determined based on the size of the text input

c) Wave Form Generator: Waveform generators like WaveNet which are deep neural network for generating raw audio from the linguistic data achieved from the analysis. It is able to create very realistic-sounding human-like voices

2) Facial Animation

a) The collection of facial animation specifications adequately defines the dynamic alterations in the manufactured video sequence. Every single parameter describes a basic movement that impacts a particular region of the face. By combining the individual action units, the ultimate facial movements are produced.

b) Estimating distortion and facial movement from 2D images is the most difficult aspect of facial expression study.

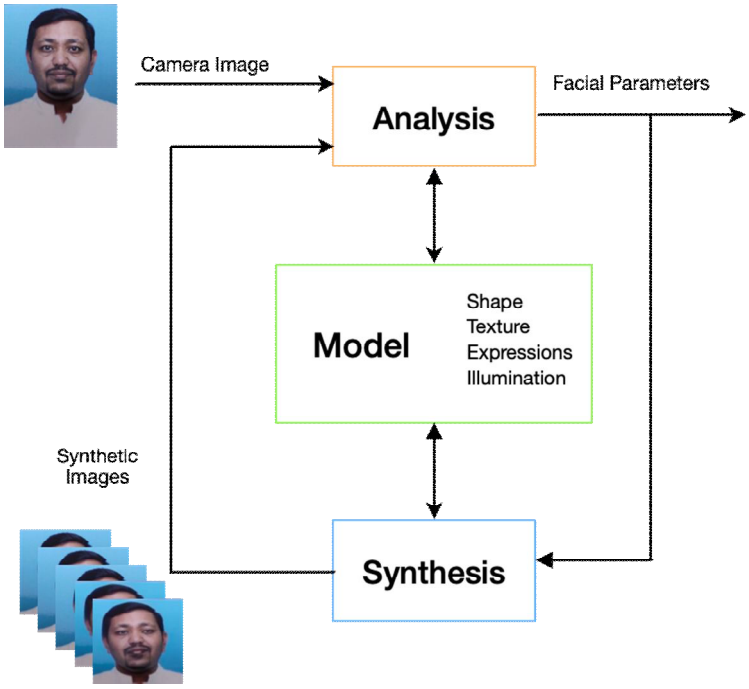


Fig. 4. Synthesis loop of facial animation model

3) Audio and Video Synchronization: Audio and Video encoding: The rendered video frames are encoded, interleaved with the speech data, and finally packed to the target file container (AVI container with MPEG-4 encoders). Finally, the video is available for download through the web application.

IV. RESULTS

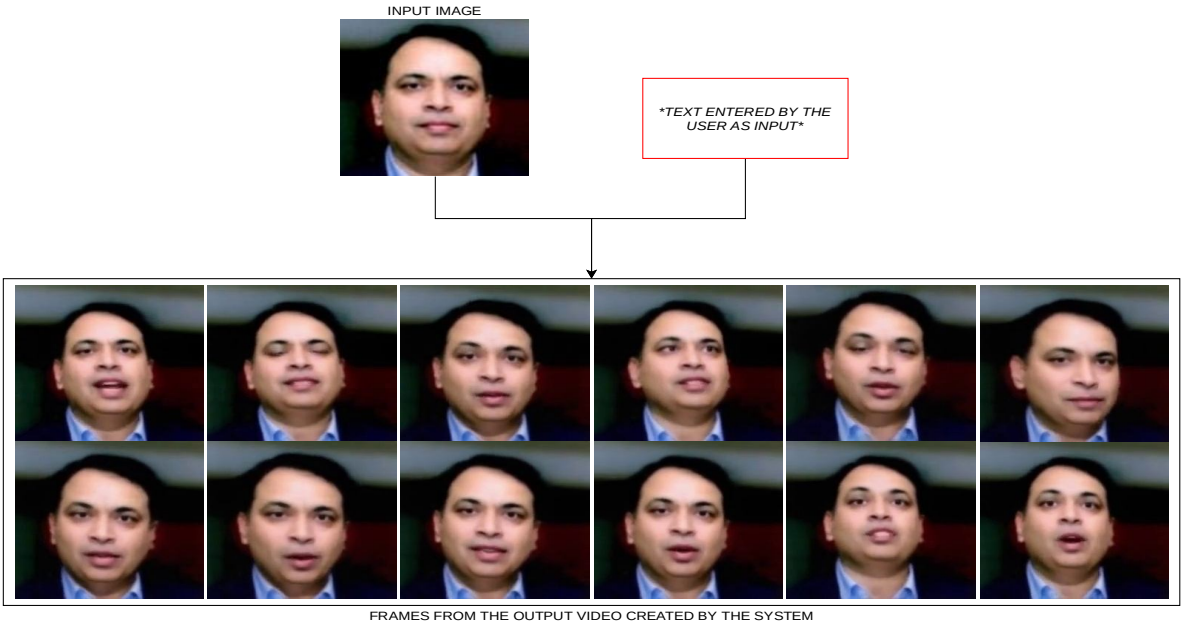


Fig. 5. Illustration of results frame by frame

V. CONCLUSION

This paper presents a different project with varied uses in a wide range of industries. The previous work done in the individual modules of this project are combined and altered to create a new project that can prove to be efficient, visually interactive and more personalized as compared to the existing digital assistants. Video creation using facial animation and speech synthesis, can tend to the target audience, be it clients or students or even public announcement systems without putting in too much effort, especially with the world being forced to move online in this pandemic. From the education system, using distance learning to businesses providing virtual assistance and work-from-home policies, the presence of visual information in addition to audio could improve speech understanding and better interpretation of the message. The proposed project can be summarized as a medium to remove the dependency of a person's physical presence for a virtual job, and instead create a video of that person that would deliver the exact same message, only automatically.

VI. ACKNOWLEDGMENT

We wish to express our sincere thanks to our director Dr. Mohiuddin Ahmed and our principal Dr. Ganesh Kame, M.H. Saboo Siddik College of Engineering for providing us all the facilities, support, and wonderful environment to meet our project requirements. We would also take the opportunity to express our humble gratitude to our Head of the Department of Computer Engineering Dr. Zainab Pirani for supporting us in all aspects and for encouraging us with her valuable suggestions to make our project a success.

We are highly thankful to our internal project guide Er. Waheeda Dhokley whose valuable guidance helped us understand the project better, her constant guidance and willingness to share her vast knowledge made us understand this project and its manifestations in great depth and helped us to complete the project successfully. We would also like to acknowledge with much appreciation the role of the staff of the Computer Department, especially the Laboratory staff, who gave us the permission to use the labs when needed and the necessary material to complete the project. We would like to express our gratitude and appreciate the guidance given by other supervisors and project guides, their comments and tips helped us in improving our presentation skills.

Although there may be many who remain unacknowledged in this humble note of appreciation, there are none who remain unappreciated.

REFERENCES

[1] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. 1998. Quantitative association of vocal-tract and facial behavior. Speech Commun. 26, 1–2 (Oct. 1998), 23–43. DOI:https://doi.org/10.1016/S0167-6393(98)00048-X

[2] Yehia, Hani C., Takaaki Kuratate, and Eric Vatikiotis-Bateson. "Linking facial animation, head motion and speech acoustics." Journal of Phonetics 30, no. 3 (2002): 555-568.

[3] Yong Cao, Wen C. Tien, Petros Faloutsos, and Frédéric Pighin. 2005. Expressive speech-driven facial animation. ACM Trans. Graph. 24, 4 (October 2005), 1283–1302. DOI:https://doi.org/10.1145/1095878.1095881

[4] Simons AD. Generation of mouthshape for a synthetic talking head. Proc. of the Institute of Acoustics. 1990.

[5] E. Yamamoto, S. Nakamura and K. Shikano, "Lip movement synthesis from speech based on hidden Markov models," Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 154-159, doi: 10.1109/AFGR.1998.670941.

[6] Christoph Bregler, Michele Covell, and Malcolm Slaney. 1997. Video Rewrite: driving visual speech with audio. In Proceedings of the 24th annual conference on Computer graphics and interactive techniques (SIGGRAPH '97). ACM Press/Addison-Wesley Publishing Co., USA, 353–360. DOI:https://doi.org/10.1145/258734.258880

[7] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. ACM Trans. Graph. 36, 4, Article 93 (July 2017), 11 pages. DOI:https://doi.org/10.1145/3072959.3073699

[8] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Trans. Graph. 36, 4, Article 94 (July 2017), 12 pages. DOI:https://doi.org/10.1145/3072959.3073658

[9] Alice Wang, Michael Emmi, and Petros Faloutsos. 2007. Assembling an expressive facial animation system. In Proceedings of the 2007 ACM SIGGRAPH symposium on Video games (Sandbox '07). Association for Computing Machinery, New York, NY, USA, 21–26. DOI:https://doi.org/10.1145/1274940.1274947

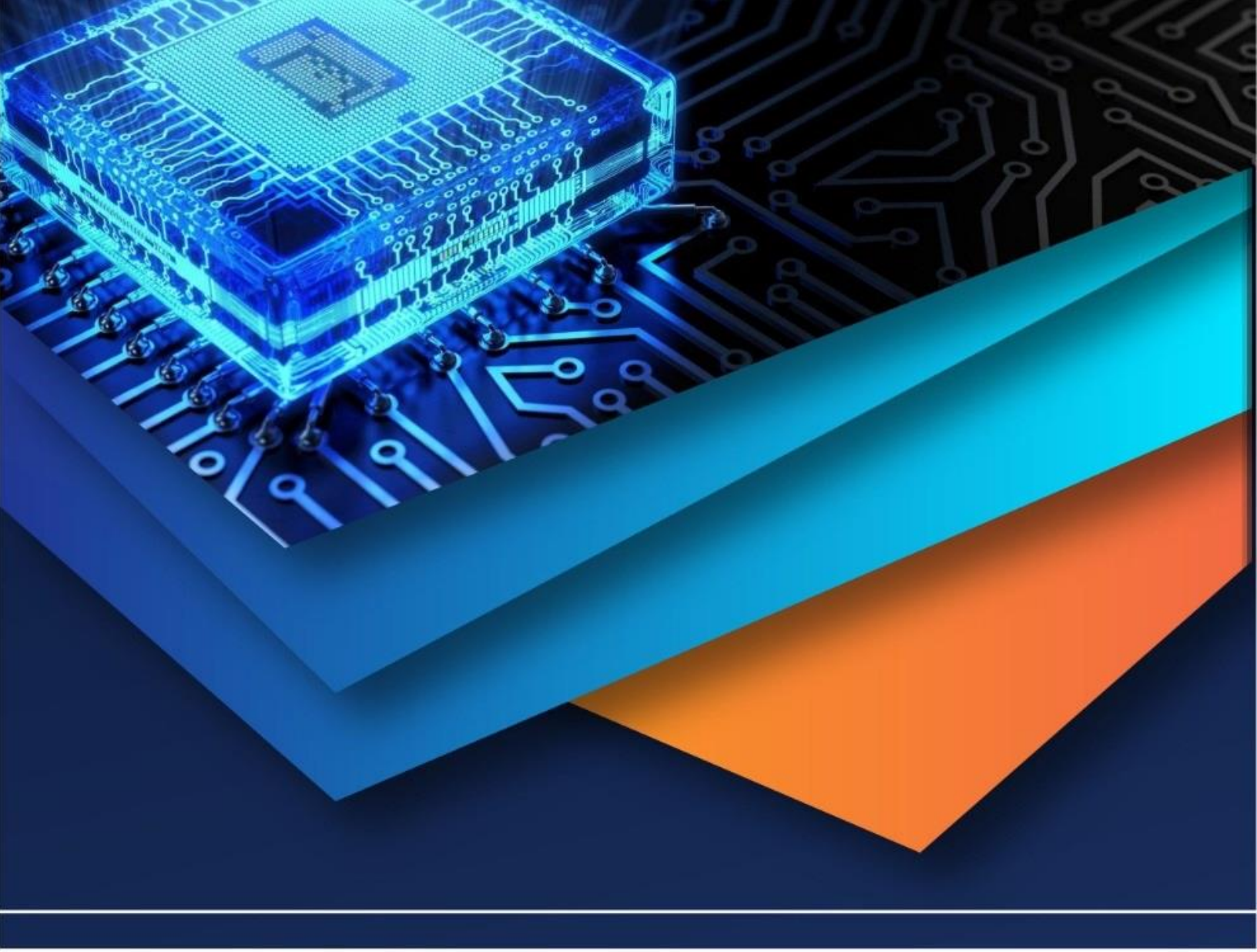
[10] Cosatto, Eric, Jorn Ostermann, Hans Peter Graf, and Juergen Schroeter. "Lifelike talking faces for interactive services." Proceedings of the IEEE 91, no. 9 (2003): 1406-1429.

[11] Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH '99). ACM Press/Addison-Wesley Publishing Co., USA, 187–194. DOI:https://doi.org/10.1145/311535.311556

[12] Blanz, Volker, Curzio Basso, Tomaso Poggio, and Thomas Vetter. "Reanimating faces in images and video." In *Computer graphics forum*, vol. 22, no. 3, pp. 641-650. Oxford, UK: Blackwell Publishing, Inc, 2003.

[13] Salvi, Giampiero, Jonas Beskow, Samer Al Moubayed, and Björn Granström. "SynFace—speech-driven facial animation for virtual speech-reading support." EURASIP journal on audio, speech, and music processing 2009: 1-10.

[14] Wu, Zhiyong, Shen Zhang, Lianhong Cai, and Helen M. Meng. "Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar." In Ninth International Conference on Spoken Language Processing. 2006.
[15] Ding, Chuang, Lei Xie, and Pengcheng Zhu. "Head motion synthesis from speech using deep neural networks." Multimedia Tools and Applications 74, no. 22 (2015): 9871-9888.
[16] Tony Ezzat, Gadi Geiger, and Tomaso Poggio. 2002. Trainable videorealistic speech animation. ACM Trans. Graph. 21, 3 (July 2002), 388–398. DOI:https://doi.org/10.1145/566654.566594
[17] Wang, Lijuan, Xiaojun Qian, Wei Han, and Frank K. Soong. "Synthesizing photo-real talking head via trajectory-guided sample selection." In Eleventh Annual Conference of the International Speech Communication Association. 2010.
[18] Xie, Lei, and Zhi-Qiang Liu. "Realistic mouth-synching for speech-driven talking face using articulatory modelling." IEEE Transactions on Multimedia 9, no. 3 (2007): 500-510.
[19] Bo Fan, Lei Xie, Shan Yang, Lijuan Wang, and Frank K. Soong. 2016. A deep bidirectional LSTM approach for video-realistic talking head. Multimedia Tools Appl. 75, 9 (May 2016), 5287–5309. DOI:https://doi.org/10.1007/s11042-015-2944-3
[20] K. Sato, T. Nose and A. Ito, “Synthesis of Photo-Realistic Facial Animation from Text Based on HMM and DNN with Animation Unit,” in *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, Cham, Springer International Publishing, 2017, pp. 29--36.
[21] Siarohin, Aliaksandr & Lathuilière, Stéphane & Tulyakov, Sergey & Ricci, Elisa & Sebe, Nicu. (2020). First Order Motion Model for Image Animation.
[22] Jia, Ye & Zhang, Yu & Weiss, Ron & Wang, Quan & Shen, Jonathan & Ren, Fei & Chen, Zhifeng & Nguyen, Patrick & Pang, Ruoming & Moreno, Ignacio & Wu, Yonghui. (2018). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis.
[23] Jonathan Allen, M. Sharon Hunnicutt, Dennis H. Klatt, Robert C. Armstrong, and David B. Pisoni. 1987. From text to speech: the MITalk system. Cambridge University Press, USA.
[24] Rubin, Philip, Thomas Baer, and Paul Mermelstein. "An articulatory synthesizer for perceptual research." *The Journal of the Acoustical Society of America* 70, no. 2 (1981): 321-328.
[25] Van Santen JP. Assignment of segmental duration in text-to-speech synthesis. Computer Speech & Language. 1994 Apr 1;8(2):95-128.
[26] Vougioukas, Konstantinos, Stavros Petridis, and Maja Pantic. "End-to-end speech-driven facial animation with temporal gans." *arXiv preprint arXiv:1805.09313* (2018).
[27] Htun, Hay Mar, Theingi Zin, and Hla Myo Tun. "Text to speech conversion using different speech synthesis." *International Journal of Scientific & Technology Research* 4, no. 7 (2015): 104-108.
[28] Dhokley, W., Petiwala, B., Sitabkhan, M. & Suratwala. M. (2021). “Video Creation Using Facial Animation and Speech Synthesis”. Unpublished paper, SSRN - Elsevier’s online digital publication under ICAST-2021 conference proceeding



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY