International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 10 | Oct 2019 www.irjet.net p-ISSN: 2395-0072

CARES (Computerized Avatar for Rhetorical & Emotional Supervision)

Ansh Mittal[1], Shilpa Gupta[1], Anusurya[1], Garima Kumar[1]

[1]Computer Science & Engineering, BVCOE (affiliated to GGSIPU, Delhi), New Delhi, India

---

Abstract - *This study proposes a framework for generating a 3D rendered photo-realistic voice-enabled Avatar which can express sentiments and emotions. This virtual agent may act as a personal assistant for people. The application named "CARES (Computerized Avatar for Rhetorical and Emotional Supervision)" will engender an example for using digital technologies to construct an application that can be capable of holding the user's interest through personal communication that further caters to the emotional needs of the user. Using CARES, a completely adaptive human Avatar with aural and visuals that have been cohesively integrated with a chatbot, will be able to provide emotional help in poignant situations by acting as a personal assistant and medical assistant. CARES personal assistant will allow users to interact with their parents' virtual persona to keep themselves motivated and close to their ancestors. After ascending from its inchoate state, the whole framework may be used for multidisciplinary studies.*

***Key Words*: Avatar, Human-Computer interaction, Chatbot, Computerized Human, Phonetic Recognition Algorithm, Eye-blink motion algorithm, Facial Animation Parameters**

1. INTRODUCTION

From the commencement of development of Computerized Avatar, it has faced certain dilemmas and challenges in various domains such as Multi-modal Expressions or behaviours, Speech Synthesis, and Synchronizing of Virtual World Character with their Speech as proclaimed in [1]. The earliest use of Avatars had been initiated by role-playing games such as PLATO in 1979. It had been later used for on-screen player representation in 1985 by Richard Garriott for yet another computer Game-Ultima IV: Quest of the Avatar.

As research progressed, developers faced a problem of speech synchronizing to the computerized avatar face motion movements. This issue had been later elucidated in 1997 by T. Frank et al. in [9] in which major concern for animated characters had been brought up- synchronization of lip movements with the speech signals. He lucidly stated the drawback for animator which comprised of manual processing of speech signals, and need for a multitudinous no. of passes for fine-tuning in a technique called LipSync (usually used for offline synchronization). This encouraged the use of Artificial Neural Network to counter the drawback discussed above in [9]. This wasn't a consummate approach but had been efficacious, in that, it reduced production time of the desired result to less than 30% as opposed to previous Lip-Syncing approach.

Approaches such as Integrated Motion Control techniques [18] and performance of different emotional transitions for Avatars were discussed thoroughly in the late 20th century. Some of these described frameworks such as AGENTlib [22] [18] to inextricably link the smooth transitions from one specific action such as moving his hand to another such as holding a thing in that hand. Analogous to this had been the transitions from one emotion to the other. Example of such a transition can be that of having the initial emotion of slight happiness to the consequent emotion of immense pleasure.

The main idea circumscribes the possible impacts that Virtual World(VW) [4] and Virtual Humans (commonly known as 'Avatar' in the context of this paper) bring about in common populace as had been described by Beth E Kolka [20]. This research cultivates these thoughts to build an integrated framework for Avatar development of real identities for descendants who are in need of their antecedents (who have passed away or are living at a considerable distance from their descendants) to express their emotions which they otherwise, can't express, hence, the name CARES has been designated for this Integrated Framework.

This 3D computer rendered avatar development Integrated Framework has been endorsed with rhetorical and aural capabilities and has been the aspiration of many humans since the genesis of the era of Mixed Reality as has been described in [1]. The framework that has been proposed can only be created for Realists and to some extent for Idealists (as will be seen later in this paper), but it excludes any possibility for Fantasies or Roleplayers; which are the four distinctions proposed by C. Neustaedter and E. Fedorovskaya in VW in [4]. CARES framework has also been concerned with how behavioural history can be used to enhance the avatar's sensitivity towards the user of the computerized Avatar. It emphasizes the mature stages of emotional speech synthesis techniques that are concerned with the discipline of Avatar emotional sensitivity and reaction analysis which can be determined as has been done by MUs (UMUPs (Utterance Motion Unit Parameters) and EMUPs (Expression Motion Unit Parameters)) in [13]; furthermore, this also incorporates a similar way of how Yilmazyildiz et al. in [24] used gibberish speech to train the model to produce its own rhetoric. Furthermore, the talking avatar created from CARES framework will represent countenance of a normal human to some degree (so, used by Realists and Idealists) and be augmented with facial expressions and morphological features advocated with multilingual synthetic voice system which has previously

been depicted by P. Antonius et al in [2]. The only drawback to his system had been the unavailability of Human Voice. For this, the reader can refer to [25] - [33].

Dynamicity and morphology of the face (all components of face mesh including Hair, Jaws, etc.) of the Avatar have been focussed on using the topologically consistent morphological model by C. Cao et al [5]. The proposed model has also tried incorporating advanced 3D rendering features for Avatar as has been discussed in patent presented by J. Kapur et al [3]. But this paper fails to depict the expressions as characteristics of an Avatar for a real identity. Also, it does not enhance the Avatar's persona with the use of actual voices of the people it has been mimicking, which has been stated later in the aforementioned paragraph. Conclusive to the above facts, Human-computer Interaction (HCI) can be defined as having a presence of certain individuals which can be depicted by the use of capture attention, communicate emotional affect, mediate comical cues, etc.

This section dealt with the introductory phases of the model while section 2 deals with related works to our proposed models. Section 3 deals with the research methodology undertaken to form the basis for our proposed model. Section 4 delineated the whole proposed framework for our model. And the last section deals with the conclusion and Future Scope for our proposed model.

2. RELATED WORKS

Although the need for such a model has been immense, such an Integrated framework has not been engendered to encounter the psychological and mental needs of Human subjects who have lost interaction with their antecedents. But, a similar type of Integrated Avatar Development Framework has proliferated in the multi-disciplinary fields of online gaming such as MMO (Massively Multiplayer Online) games such as Maple Story and World of Warcraft [21] as has been discussed. These games usually take into account that players playing them are belonging to a class of Fantasies or Roleplayers [4]. Which results into MMORPGs which are Massively Multiplayer Online Role-Playing Games.

This system has been inextricably related to the works done in [47] in which a model abbreviated for EAVA has been discussed for Emotive Audio Visual Avatar which has been employed for purposes of providing assistance to individuals having oral or aural problems so that they can participate in the quotidian communication. Furthermore, this research has also been partly related to the work done by H. Tang et al [15] in which they disseminate the need of a pipeline model to construct a realistic 3D text driven emotive avatar by utilizing 2D frontal view face image.

At this moment, to make an unbiased approach to this problem, this research takes a look at how other types of Virtual World entities interact with humans and what have been the result on these interactions. This had been done by referring to [19] by James Lester et al. This paper presents how an avatar can use multimodal communicative abilities and Explanatory Lifelike Avatars to engender an animated pedagogical agent. This had been done by generating Task networks to prosecute student-designed tasks. The Task Planner has been used as another feature for computerized Lifelike Avatar. [19] also probed the efficacy of this approach by exegesis of different approaches by a multi-way comparative usability study conducted on different virtual worlds such as Agentless World, an incorporeal narrative in the VW [4], taciturn & life-like avatar, and Full-scale Explanatory Lifelike Avatar.

We revisit the aforementioned pipeline model by H. Tang et al [15] after validating that a Lifelike Avatar has been one of the major Agents in the VW [4]. The steps included in the pipeline model [15] are of crucial significance as they instigate this research towards CARES which can helps make some crucial advancement towards this Integrated Framework. At the inception phase, a fully automated 3D face shape and texture reconstruction framework utilizes its different functionalities in order to inchoately fabricate a 3D face mesh model on the basis of the 2D face image. Next, viseme and expressions are added to the aforementioned face model augmented with visual prosody channel which had been utilized to control head, eye and eyelid movements depending upon different standard MPEG-4 FAPs [56].

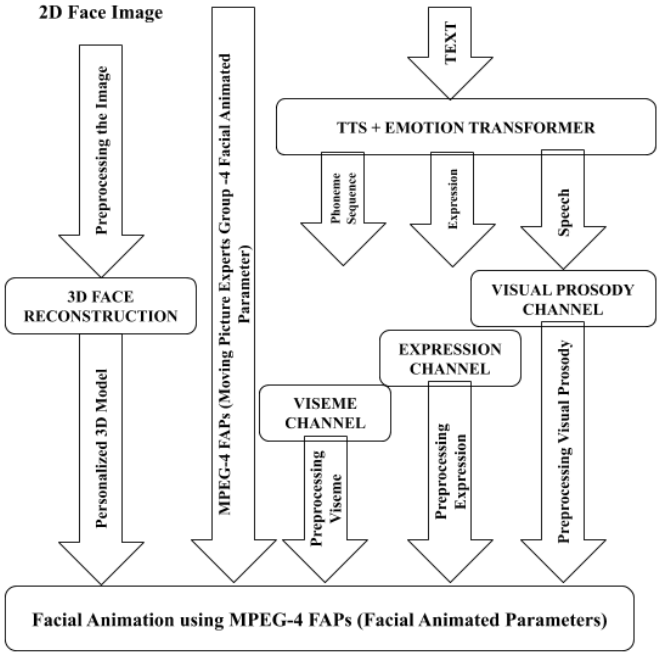

Fig-1: A complete pipeline of frugal and efficacious techniques to build a 3D text-driven EAVA [47] from a 2D face image of the target user [15].

Conclusive to these steps had been the final layer which has been augmented and synchronized with the emotive synthetic rhetoric which can be built by using an augmentation of Festival [57]-MBROLA [56] and emotional transformer.

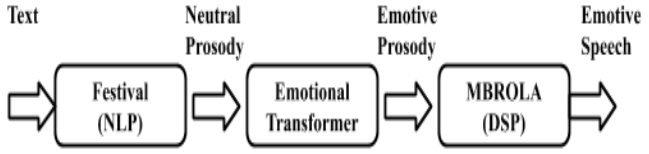

Fig-2: A diagram to represent the final layer the addition and synchronization with emotive synthetic rhetoric in [15].

Although this work has flamboyant and cogent aspects in the disciplines of Avatar creation and Text-to-Speech (TTS), it fails to consider the possibility of advancing semantic networks, to get emotions from Text that has been given as an intake in the model. Furthermore, the whole model has been compromised by the fact that the whole speech had been in a synthetic voice [15]. The solution to this had been ameliorated later in this work.

In 2002, Pengyu Hong et al [13] presented a framework that methodically addressed issues stated-

1. Facial morphology modeling;
2. Automated Face Motion Analysis; and
3. Real-time Speech Driven Face Animation with expression using Artificial Neural Networks.

And [13] also orientates to learn a quantitative visual representation of Facial Deformations which had been coined as Motion Units (MUs). This had been contemporary to the MPEG-4 FAPs which use a set 66 set of points. These are also used for Facial expression synthesis in [12] by S. Zhang et al to generate an effective talking avatar.

The works stated previously give an overview of the frameworks built over the past couple of decades. It also necessitates the need of viewing these virtual assistants using the right perception because a human face and an avatar face may not always conform to the standards specified by each other and this may not be perceivable to humans given their acute senses. And, hence people tend to make fallacies in the determination of emotions (used as class labels) [6] which can lead to viewing positive emotions as completely disgruntled emotions or neutral emotions (in case of Avatars) as has been observed by Sylvie Noël et al [6]. This work also concluded with the fact that humans are more susceptible to the context when they are trying to perceive a visual avatar along with the corresponding text. And this prompts us to create a TTS speech system along with the avatar that has been created. This had also been indicated in [16] which had been used to construct a test-driven talking avatar which employs a statistical parametric approach.

Next, a systematic study of the framework proposed by P. Hong et al [13] had been conducted, that dealt with MUs (Motion Units) which are the quantitative model of facial deformations. These consisted of two types of MUPs (Motion Unit Parameters) which are known as UMUPs (Utterance Motion Unit Parameters) and EMUPs (Emotional Motion Unit Parameters). Both of these use the Mesh model of the face in a fairly different way from that of usual standards of Mesh model for Morphological Analysis of the face. This has been done to get different emotions of the face during vocal and non-vocal gestures as had been clearly stated in [13]. Real-time animation for an avatar i.e. generated from a single image has also been achieved by S. Lucey et al [14].

As has been discussed in [6] by Sylvie Noel et al, the emotional significance of an avatar face had been a reliable characteristic to determine the emotion that the avatar may be depicting. This can further strengthen this research's argument for endorsing the Avatar face with morphological features to depict the real emotion of the Virtual Human that the research aims to create. [13] also depicts how populace would have trouble assigning emotions to the avatar face given the congruent stimuli combinations. This may depict the diverse approach taken by Sylvie Noel et al when compared to Carrol and Russel [58] which limited their stimuli to a few amalgamations of text and face sentiments that could be mistaken for each other. Also, this further depicts the need of a chatbot to understand the context in which the statement had been spoken and what environmental factors encouraged its usage. Before reinstating the need for this model, a clear distinction between other virtual worlds(VWs) can be seen, just like the three virtual realities presented by N. Ducheneaut et al in [21]. This Virtual Reality corresponds to a congruence between Virtual World and Real World subtly giving rise to a Mixed Reality in contrast to Second Life, World of Warcraft, etc. which have been mentioned for their outstanding dynamics or character customization in [21].

We now discuss the works that are inextricably linked to the sub-modules of the integrated framework that have been proposed earlier. As the research delves deeper into this model, the research has been elucidated with an issue of computerized voice and uni-lingual capabilities of the model. To make this computerized human have a multilingual framework, a multilingual system for this avatar [10] had been proposed. This model proposes a "standardized face" format comprising of 3d graphical and morphological features which uniquely comprise the lip-sync as has been used in [9] by T Frank et al, rendering and translation which had been used by M. Luong et al [26] with the comprehensive use of Neural Networks and Zhiyong Wu in [23] by employing Deep Neural Network. This will be studied more comprehensively in the proposed methodology for research.

According to Pengyu Hong et al [13], audio-visual mapping consists of 2 steps which are as described below:

1. Map audio features to UMUPs
    a. Mapping estimated UMUPs calculated to Final UMUPs and EMUPs
2. Evaluation - Calculates normalized MSEs and Pearson Product Moment correlation coefficients between the Ground Truth and Estimated Results

[8] divides the facial motion representation into two Approaches- Image- and Model-based Approach. Some examples of the former approach have been discussed by Cristopher Bregler et al [34] and Tony Ezzat et al [35]; while, the latter approach has been represented in [36] by Demetri Terzopoulos and Keith Waters. Apart from these approaches, there had been yet another approach known as a Performance-Driven approach in which the computerized human learns facial motion from the recorded motion of people. These also use motion-captured data [37] from the compact parameterized facial model and take the Model-based approach which has been discussed earlier.

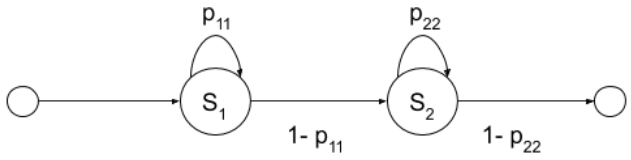

Fig-3: HMM topology for phoneme for CMU set as per [8].

The approach of Visual Speech Synthesis can be said to be defining a key shape for each of the phonemes while smooth interpolating between them as has been discussed by P. Muller et al [38]. Visual Speech synthesis, similar to Acoustic Speech Synthesis falls into 2 categories - HMM-based approach (instances of which are stated in [39] [40]) and Concatenative approach. Former generates facial trajectories based on the parameters generated by the Maximum Likelihood Principle(MLP) while the latter relies on cohesively sticking pre-recorded motion sequences, corresponding to triphones, or phonemes. Figure 3, in particular, corresponds to HMM topology for CMU set where the allowed transitions are between HMM states s1 and s2.

We further look into work done in the fields of voice mimicry, as it has been a crucial part of this framework. Recurrent Sequence Generators have been introduced on input data to perform speech recognition using Attention based models (with minor modification) by Jan Chorowski et al [25], to ameliorate the PER (Phoneme Error Rate) to 17.6%. Meanwhile, the local and global approach for the aforementioned mechanism has individually been studied in [26] by Minh-Thang Luong et al. This can be considered to be a prolongation of work done in [10] as the model described in [26] can be used for WMT translations between languages such as English to German. Until 2015, it had been believed that voice translation reached a bottleneck, but, with the model proposed by Dzmitry Bahdanau et al [27] to automatically soft-search parts of single sentences (phrase-based) to target specific words, a breakthrough had been achieved in the fields of Multilingual Systems.

Conclusive to the works done above had been the generation of the Wavenet model [28] which used raw audio as inputs which trained data on tens of thousands of audio every second. This model had been used as a fully probabilistic model and has been highly autoregressive and a single Wavenet could be used to capture characteristics of individual speakers with equivocal fidelity. It can also be used for phoneme recognition which can provide a good foundation for Viseme channel [15] [47] in the pipeline model shown in Figure 1. A contemporary of the aforementioned model has been the ClariNet [33] which used parallel wave generation instead of the sequential ones generated in a single WaveNet model. Later, in Nov'16, some changes were made to the original Wavenet mechanism (complexity: $O(2^L)$ where L had been the no. of layers in the network), which led a prolongation to Fast Wavenet (complexity: O(L) where information had been cached upon every layer) as discussed by Tom Le Paine [29]. Later, Yuxuan Wang et al [30] proposed an end-to-end TTS model which took phrases or character sequences as inputs in the form of spectrogram and produced an output in the same form.

Conclusively, the research looks into the work of relevance which assists the integrated framework to generate a near-consummate avatar which can help in alleviating mental stress through the application of Human-Computer Interaction(HCI).

3. RESEARCH METHODOLOGY

Avatars provide people with an opportunity to represent their character online in a way that they want [4]. They could look like the users alter-ego since users have an option to tailor-make their characters. These avatars will adopt roles that are completely different from their actual personas of the user. Consequent to this, it can be concluded that talking to a human avatar has been considered different than talking to a human. It may lead to a divergent effect on dialogue or discussion. Avatar will elevate the degree of evenness of speaking to a partner through the materialized form of computerized human. Also, facial expressions will refine the efficacy of avatars to a large extent. Facial expressions find their way to have an effect or a control over people's judgment.

Avatar animations are also known as facial marionettes, as the avatar or the puppet acts like it has been administered by the user's facial movements and expressions. A facial puppetry system had been composed of two main elements:

face tracking and expression transfer. Face tracking records the user's facial distortions. Expression transfer makes the avatar lively so that its facial movements best matches the one which had been recorded from the user.

A personalized avatar captures the dynamics of a user at different poses and the user's actual geometry, appearance and expression are matched with the custom face rig of the user. Amalgamated with real-time face tracking techniques, facial animation induced helps to convey the realism of a person, in contrast to avatars of pre-defined virtual characters, which has been useful and important in many real-world applications, including virtual reality, gaming or teleconferencing.

One major related challenge has been how to easily create compelling user-specific avatars with fine-scale details for non-professionals with commodity hardware at home. Excellent recent introduces techniques that focus on modeling the face and head from image [8]/video [7] recordings of a user. While very impressive results are demonstrated, only the face and head are explicitly modelled; hair, a non-negligible part of a user's appearance, has simply been approximated as a diffuse texture map on the head model. Facial animations are the most expressive way to communicate non-verbal cues. In real time interactions or face-to-face communication, people express themselves not only through speech but through facial expressions and body movements as well. Speaking has not only been moving the lips but also includes eye-contact, blinking, eyebrow-raise, lip movement like smiling, tongue and teeth movements and other facial expressions are also present. All the audio-visual avatars are usually driven through three approaches which are text-driven, speech-driven or performance-driven. A system consisting of dialogues in which the text has been known beforehand, about what needs to be spoken commonly, acts as a text-driven approach. Speech input has been used in applications in which speech has not been known beforehand, similar to multiplayer games and movie characters. For instance, popular mobile game Tomcat consists of a voice repeater which takes a speech as an input and make the cat repeat what you said. You can poke him, pet him and even feed him treats. This all has been done using a conversion algorithm that builds facial parameters from auditory speech. Here the main problem that arises had been that speech should be synchronized with the facial parameters of the audio-visual avatar and this has been known as the lip-sync problem.

Speech synthesis is the unnatural or artificial construction of human speech. The excellent example of this has been the TTS (Text-to-speech) systems. They convert natural language or human language in form of text into speech. Another product of speech synthesis has been a system renderer of symbolic linguistic representations like phonetic transcriptions (how a sound described into a symbol) into speech. Facial animation can also be driven by human performance. The Avatar Kinect 3 makes use of an online video chat which has been capable of driving a whole face of an avatar through dynamic human performance. Audio-visual avatars who are performance driven are solely based on dynamic facial tracking. Their main task has been to map the results from the dynamic facial tracking onto the face model. In contrast, text-driven and speech-driven avatars make use of AV Algorithm or an audio-to-visual (A-V).

Apart from this, the research also tends to use certain features to make the computerized human more interactive to Human-Computer Interaction (HCI) by the use of certain elements which will be discussed in the subsequent sections. The first feature the research focuses on, has been inextricably linked to the language that the Computerized Human/Avatar may tend to use. Using other systems may lead to having a certain type of language which has only been spoken by some speakers while alienating other speakers resulting in a computerized human which may not use the native language of the user. Discussion regarding this topic has been conducted earlier.

For the aforementioned problem, a set of algorithm which can comprehensively be divided into certain subparts of the foregoing algorithm, had been used. The algorithm has been presented in the form of two figures which are stated in the adjacent section (Figure 3 and Figure 4).

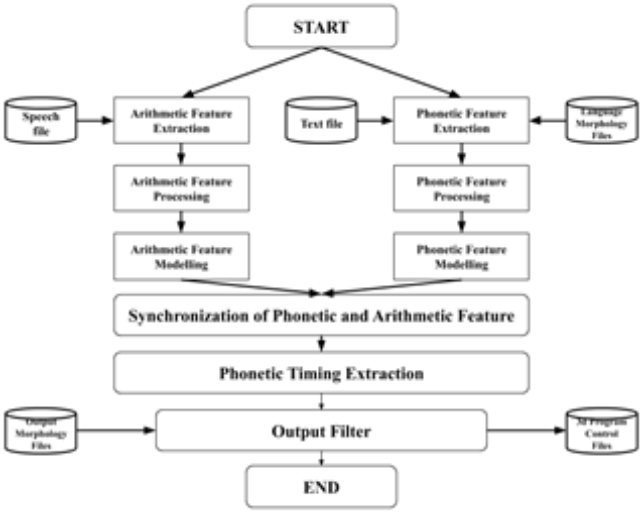

Fig-4: A diagram representing Phonetic Recognition Algorithm as per [10].

This algorithm comprises of 2 parts which are interdependent on each other. These can be resolved as the Phonetic Recognition Algorithm and the Synthesis Algorithm. Although this approach uses Morph Channels that are spread through the face, we, on the other hand, use Mesh model in contrast to these morph channels. The phonetic recognition algorithm has been shown in the aforementioned figure 4.

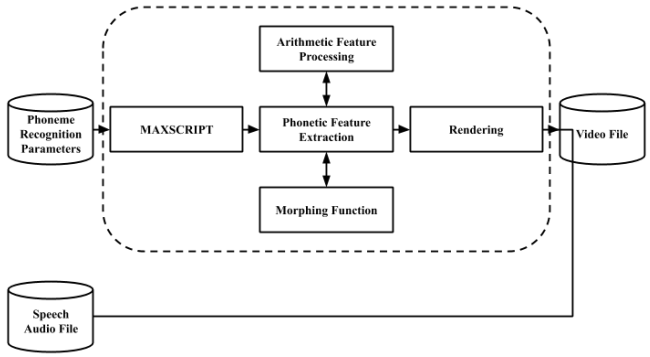

Fig-5: A diagram representing Synthesis Algorithm as per [10].

This algorithm shown in figure 5 has been used as the second phase for Multilingual System algorithm which directly correlates the Phonemic and Visimic exchanges in terms of Glossology as well as Manual Digital Character Lip-Syncing.

Pengyu Hong et al [13] describes a real-time speech-driven synthetic talking avatar, which provides a multimodal communication interface. This has been done through Motion Units which have been described earlier. These Motion Units further correspond to certain parameters which are called MUPs (Motion Unit Parameters), which are in case a linear combination of the aforementioned MUs, describing facial deformation. [13] later presents a comprehensive study of Utterance MUPs and Emotional MUPs which are used in this Integrated Framework. Similar to this work a model for 3 facial changes which represent no expression, angry face and smiling face for calculating the difference between Ground Truth and Estimated Results had been created. While these results may prove to be of immense value but the work done doesn't suggest the modern means for the real-time speech-driven computerized human. Also, the work done has only been limited to 3 emotional changes which limit the usage of UMUPs and EMUPs.

On the other hand, Dmitri Bitouk and Shree K Nayar [8] propose an HMM-based approach which in turn use the meh model implemented to a greater degree of precision. This algorithm takes two parameters into account - Lexical Stress and co-articulation; which helps in producing a realistic animation from both speech or text (which can be beneficial for Human-Computer Interaction).

Facial Motion approaches can be classified into two distinct types based on HMM-model proposed in [8] - Image-Based Approach and Model-Based Approach, which have been discussed in the foregoing section. The synthesis model for facial motion defined in [8] uses a parametric representation of foregoing motions in the face to generate trajectories of facial parameters from time aligned sequence of phonemes.

Another important topic corresponding to Computerized Human has been the Synthesis of blinking and Eye Motion. This research has used a sampling-based Synthesis algorithm as will be mentioned later on. This cohesively interrelates with the Mesh- or Model-based approaches that have been discussed earlier, to result in a more realistic computerized Avatar. This has been discussed in the next section comprehensively. Conclusive to this, the proposed model obtains the complete and final image of eyeballs which have been generated by synthesizing remnant and missing texture inside a) iris; and b) sclera. Eye gaze changes are generated using the previously proposed stochastic model by S. Lee et al [43].

We also face some problems while integrating these modules into this framework and hence are required to mitigate these by providing an alternative. Initially, the HMM-based approach that had been discussed earlier, implicitly assumes synchronicity of visuals and acoustic realizations of phonemes. While the cognitive evidence in [45] by P. Viola and M. Jones suggests a sharp contrast with the aforementioned hypothesis. This can be seen when the facial articulation precedes the sound the entity has been supposed to make. This issue can be mitigated by employing the use of Dynamic Bayesian Networks in addition to the HMM-based approach to formulate an extended HMM approach. Another issue this model may face has been that of an absence of Face Detection Algorithm [46], which directly results in a face marked with points (scarcely separated correspondences between the novel face and prototype face) manually to obtain geometry of a novel face from real-time, while also transforming generic facial motion model. Furthermore, Modern techniques can be used in order to develop a 3D face with characterized information which can have a consummated recover using techniques and algorithms which were developed by Computer Vision Researchers, as has been suggested in [15]. Shape for Shading [48] and Model-based bundle adjustment [49] are some of the techniques which reinforce the aforementioned fact. The high ground for these type of model has been that they directly use the model as a search space concluding in less unknown variables and equations defining above parameters. They also correctly associate tracking point with their correct location as had been stated by Ying Shan et al [49].

Realism can also be compromised if the design of the jaw model of the avatar has been omitted, which can result in less satisfactory results, leading to no significant changes for the target audience. And, hence it had been important to design teeth and jaw model along with the face mesh, eye blink, voice mimicking, and multi-linguistic model. Also, a rapid head movement may enhance the perception of speech-enabled avatars as has been seen in [59] by Lijuan Wang et al. For instance, nodding in case of a reply to a question has been deemed to be of a positive class of replies. Also, there has been a need for integrating a chatbot in the proposed model, in case the speech produced has not been clear to its user. While a text-driven approach for dialog-flow to occur

between generated avatar and human uses the keyboard as a device for input, it will be ineffective as the conversation losses its commonality due to the user typing in the text. So, chatbot has been needed for a one-sided approach (or 2-sided approach if the user has been having any impairment to speech) so that user can understand the virtualized human in case of any impairment to hearing. With Text Input/output the advantage has been that the user can check the validity of input given through speech. However, text input showcases its inefficiency as the time the input needs increases. In addition, the text fails to showcase the emotion behind a sentence.

Figure 1 suggests the fully accomplished pipeline of efficient and low-cost techniques. This can be used for getting a fully Emotive Audio-Visual Avatar which can correspond to the computerized Human that has been in need for emotive HCI. It can create an avatar using just a 2D image, as has been stated earlier. The 3 steps involved in the previously mentioned real time conversion are as stated below:

- Using 3D face shape and texture reconstruction, construct a personalized 3D face model on the basis of a 2D image
- Developing face model animated backed up by viseme and expression channels (which can be used interchangeably with Mesh parameters) and complemented by visual prosody channel by the application of standard MPEG - 4 FAPs [55] (Facial animated Parameters); and
- Synchronize and combine the facial animation in addition to generating emotive synthetic speech, by incorporating an emotional transform into a Festival [57]-MBROLA [56] as has been mentioned earlier.

Moreover, models such as SFS (shape from shading) [48], Model-based bundle adjustment [49], etc. use an effective, frugal, efficient and fully automated framework which has been based upon works done by Yuxiao Hu and Lei Zhang in [50]. The aforementioned work suggests a PIE model for a variant pose, illumination and expression which has immaculately inspired several models (such as [48] [49], etc.).

In addition to the aforementioned, fast and robust face detection had been inspired by the work done by R. Xiao et al [52]. This work compels the need for a 2D face alignment algorithm, similar to the one described by S. Yan et al [53]. This has been used for detecting a face, which then conforms into a rectangular geometry to extract the facial feature parameters. This had been a relatively fast and accurate algorithm for feature extraction from a single photograph in contrast to its antecedent algorithms. An instance and extended mathematical version of this algorithm has been the Bayesian Tangent Shape Model (BSTM) described by Y. Zhou and H. Zhang in [54].

Conclusively, all the work can be either concluded through MPEG-4 FAPs [55] or any other mathematical model. This model has a total of 84 features points that are aligned, 48 of which preferred for 3D face reconstruction. These can be either manually adjusted or can be adjusted with the help of Deep Belief Networks [23]. Conclusively, real-time avatar animations can be achieved through the coarse motion inputs which can be taken from the real world to map into VW [4].

4. PROPOSED MODEL

4.1 Face Modeling

As has been mentioned in the previous section, work done in [52] achieves a high recall rate, enhances detection speed and provide low false alarm rate. The proposed model breaks the procedure into steps as have been discussed in the following:

- Enhance detection performance by removing most non-face like candidates (by employing a linear filtering algorithm)
- Combining a boosting classifier into a “chain” hierarchy (using boosting chain algorithm)
- And finally employing post-filtering algorithm (applications in image pre-processing, SVM filtering and texture filtering)

We now discuss some steps that are involved to a high extent in the face modeling.

4.2 Image Acquisition

It can be stated as the process of taking an image of a user through a provided source to generate a 3D avatar from the capture 2D image (as done in [14]) along with the multiple images which essentially capture the rotational movement of the head of the user [17]. The head poses consist of head rotations over a set of angles with a neutral expression. These can be used to conclusively form an avatar as done in [17].

[50] has been used for reconstruction of a personalized 3D facial model by exploiting a single 2D in contrast to work done by Vetter et al. in [51]. Automatic Face construction module consists of 3 major steps as has been given below:

- 2D face alignment and detection
- 3D face shape reconstruction
- Texture mapping on the virtual face

4.3 Image Pre-processing

This has been a dual step procedure which comprises of segmentation of the captured images into different components, and landmark labelling of all entities which facilitates the subsequent building of geometric prototypes or models. These are described in detail in the following:

- Initially, the model has been required to segment the images into several layers: head, hair (including headwear), eyes, teeth, body and background. As the hair may overlap with other regions in a complex fashion, deep image matting [60] has been done to refine the segmentation.

- We label a few face landmarks (Si) for each facial image (Ii). The eyes and the contour of mouth which form the set of facial features are marked with the 2D positions through these landmarks (as in the case of MPEG-4 FAPs [55] or MUPs [15]). Specifically, the model uses the face tracker to automatically locate these landmarks and then adjust them with a drag-and-drop tool.

Now, the input taken for every part of the face are discussed separately.

4.4 Image-based Hair Modelling

For Techniques in this category, the proposed typically reconstruct a static, 3D geometric model of hair, based on multiple images taken from different angles as well as through different lighting conditions as has been defined by L. Wang et al [17]. Recently, hair modeling methods based on just a single image [15] [47] are proposed to make hair modeling more accessible to non-professional users. One major difference between this work and existing image-based hair modeling work has been that the model does not model the hair geometry to the strand level as it takes up a lot of GPU power. Instead, it takes a hybrid approach that uses a 3D morphable hair model for the coarse motion [11] geometry and images to capture fine details as has been depicted to a minor extent in [8].

4.5 Performance Capture and Tracking

In film and game productions, special equipment such as facial markers, camera arrays and structured lighting, have long been used to capture high-fidelity facial performance. Such pieces of equipment are usually not available for consumer-level applications, where only commodity hardware such as video and RGBD cameras are accessible. So, simple animation [14] has been reliable for this model.

Face tracking which has been done through video input has been exhaustively researched in both fields' computer vision as well as the graphics. The latest techniques demonstrate robust real-time tracking and animation from an ordinary web camera as has been shown in [5] [11] [14]. Facial details with high-fidelity like wrinkles, using shading cues [48] [49] in real time can also be reconstructed.

The head motion and facial expression coefficients tracked by these methods can be used to drive the facial animation of the image-based avatars in real-time as seen in [11]. The foregoing figure depicts the whole encapsulated flow of the CARES integrated framework. This section deals with Avatar Generation phase as per this figure.

4.6 Eyeball Movement and Eyeblink motion

For Eye Blink motion the model uses points marked around eyes, and extract image regions around the eye to get a lucid picture of how the eyeball shall move for corresponding different facial motions. Following which, to segment eye region into the sclera and iris, a generalized Hough Transform [42] has been used for positioning and shaping of the iris that has been generated as a result of the previous step.

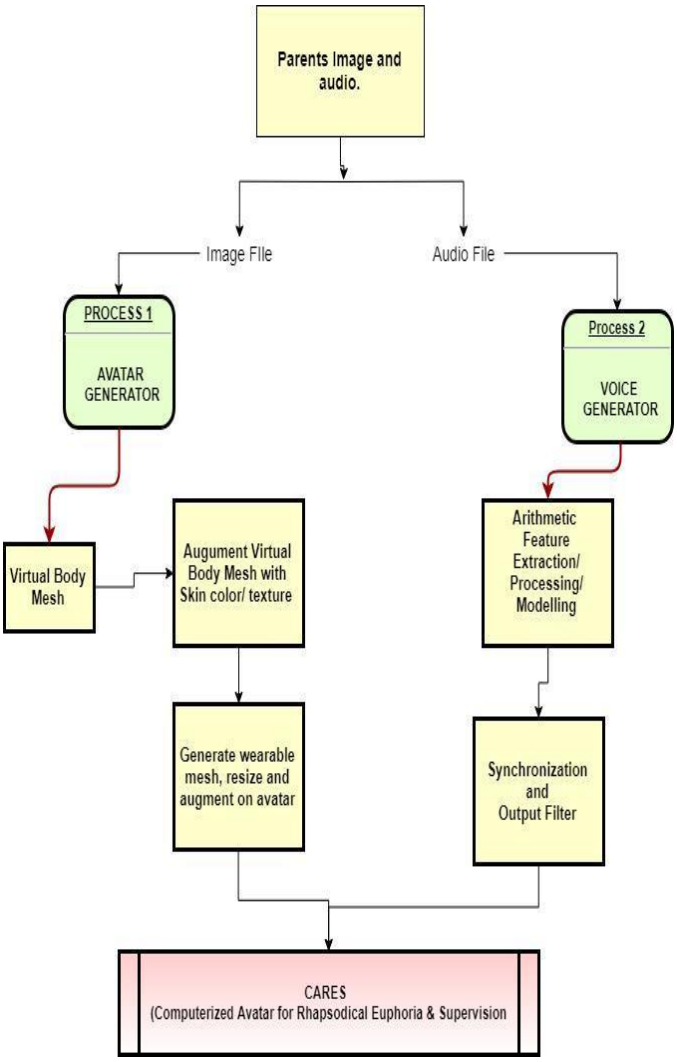

Fig-6: A complete flowchart for the Proposed Framework taken into consideration.

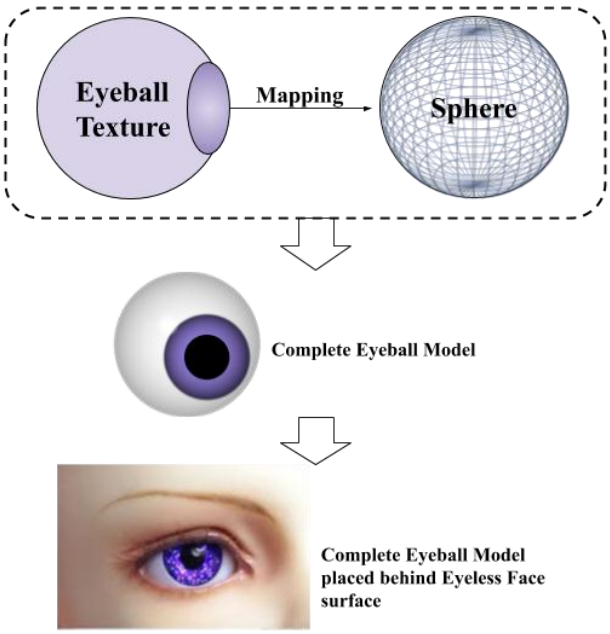

Fig-7: A complete module for eyeball rotation, reflection, translation as suggested in the proposed framework.

This work (as discussed in the earlier section) also encompasses the use of a blend shape approach described in [44] for synthesizing eye blinks in addition to the stochastic model [43]. Meanwhile, eye blink motion in the prototype face model generated by the framework can be generated by a linear interpolation between 2 eye key shapes which correspond to:

- Eyelids in open position; and
- Eyelids in the closed position

We chose the duration for eye blinks to be around 200ms and the interval chosen between successive blinks had been randomly chosen with a mean of 4 seconds. Conclusively, to map eye blinks to novel facial motion, key shapes deformed using the approach in [8].

Chatbot - There are provided certain platforms which are web-based in nature. These are called chatbots and are a user-friendly and conversational agent accessible through any web browser. Backing a facial model having MUs (which comprises UMUPs and EMUPs [15]) with a corresponding chatbot may provide a good design as has been discussed in [2] by Antonius et al. These allow the developers to create conversational experiences for users by giving them the tools to shoulder two kinds of tasks: understand and generate natural language statements and manage the flow of dialogue.

1. Agent - Agents are reported as NLU (Natural Language Understanding) components. They can be easily integrated into an application, service or a product and map the requests from the user into data which has been actionable. The application that you have created has the Agent as its name. The name of the agent had been very important. An agent can be called on by saying: "Okay Chatbot, talk to <app name>"

2. Intent - Any time the user asks a question from the agent, it will match the question to the corresponding Intent. Intent has a very crucial role. An intent incorporates elements that interprets information from the user and answer their requests. To comprehend the question in the best possible way the developers have to give as much as data they can.

3. Entity - The Integrated Framework has to know that the information that has been present with him, which of them has been relevant and meaningful and can be used to answer the requests made by the user. These pieces of data are called entities. They are dynamic in nature.

4. Context - Context plays an important role in the Integrated Framework. How? Context helps the assistant to talk more like a human by maintaining the context and replying in the context to end users.

5. Platform Integration - This integrated Framework allows the chatbot to be connected to other application to engender new research in the related discipline.

Traditionally, a chatbot can be described as an agent i.e. able to communicate with users pertaining to a topic by employing natural language through a text-driven approach. For reasons mentioned in the previous section, a chatbot with avatar and voice interaction aids to make the conversation more alive and interesting for the user. One solution to improve chatbot efficiency has been to use an interface other than text, for instance, voice-text interface. Voice-text interface has been developed with a technology incorporating both speech recognition technology and text-to-speech. The ability to converse with chatbot would make the avatar more explanatory and lifelike as in [19]. This can be achieved through certain mesh models which have been discussed in the earlier sections of this work.

Speech Replication and Voice Mimicking - Speech recognition had been also called Automatic Speech Recognition (ASR) or computer speech recognition. Speech Recognition may be defined as the process of conversion of a speech signal to a text sequence with the help of an algorithm that has been administered through a computer program. Speech recognition technology makes it easy for the computer to make sense of human voice commands and interpret the meaning behind the human languages. The main objective of speech recognition has been to develop

techniques and systems for converting speech input into a machine.

Speech recognition engine also known as speech recognizer takes an audio stream as an input and turns it into a text transcription. Speech recognition process can be composed as having a front end and back end. The front end processes the speech given as the input that has been isolated by the segments of sound and characterizes the various vocal sounds present in the signal to a series of numeric values. The back end has been a unique search engine which takes the output produced by the front end and searches across three databases: an acoustic model, a lexicon, and a language model.

- The acoustic model represents the phonemic sounds in the language and has been trained to recognize the characteristics of a particular user's speech patterns.
- The lexicon lists a large number of words in the language and provides information on the pronunciation of each word.
- The language model represents the ways in which the words of a language are combined.

Several models have been used as speech recognition algorithms. HMMs [40] are popularly used in many systems as they can be trained automatically. Dynamic Time Wrapping (DTW)-based speech recognition, neural networks, and Deep Neural Networks are some other models that are also used worldwide across various systems.

Speech synthesis is the artificial production of human speech and one of the example text-to-speech (TTS) systems. Speech Synthesizer works in contrast to speech recognizer in a way that the task of speech recognizer had been to convert speech into text and the task of the speech has been to convert text into speech. TTS system thus converts the natural language or human language in form of text into speech. Another module of speech synthesis has been a system renderer of symbolic linguistic representations like phonetic transcriptions (how a sound described into a symbol) into speech.

TTS system had been composed of two parts: a front-end and a back-end. The front-end has two major tasks. First, has been to normalize text, tokenization or text pre-processing. Text normalization has been an analysis of the sequence of words or a string of characters to determine where the words are. The front-end then allocates phonetic transcriptions to each word, and splits up and marks the text into prosodic units, like phrases, clauses, and sentences. This process has been known as text-to-phoneme or grapheme-to-phoneme conversion which had been basically the allotment of phonetic transcriptions to words. Phonetic transcriptions and prosody information together form the symbolic linguistic representation which has been the output of the front-end.

The back-end will be commonly called as the synthesizer then converts the symbolic linguistic representation which had been the output from the front end into sound. In some of the systems, this synthesis module also includes the computation of the target prosody like frequency, pitch contour, phoneme durations etc. which will then be imposed on the output speech.

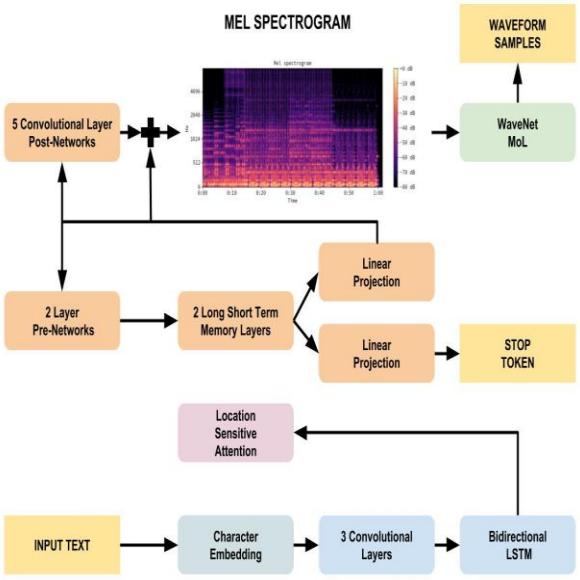

Fig-8: A consummate model of Tacotron 2 as has been discussed in [31].

As has been discussed about models such as WaveNets previously, after the addition of Wavenet as the Vocoder, the Tacotron had been extended using the recurrent sequence-to-sequence network which mapped embedded characters to a mel-spectrograms, resulting in Tacotron 2 as proposed by Jonathan Shen et al [31]. Following which Deep Voice 3 [32] had been introduced which used millions of queries for voice recognition from the training of more than eight hundred hours. Another important result had been that of the model described in [33] which conditioned the hidden representation of this end-to-end model to successfully distil parallel waveform synthesizer. In the revised work of [31], a consummate model has been described in which the vocoder used had been further modified to get a better accuracy.

DISCUSSIONS

We observe that the model that has been created, in order to generate a computerized avatar of a target's affectionate person may still be suffering from some defects which can be

explained due to reduced Degree of Freedom (DOF) which in contrast are numerous in case of work done in [2] [18]. Also, the model's power may be compromised due to the use of certain discrete functions which increases the accumulated overhead. Furthermore, Figure 10 suggests that the avatar has not yet been trained on multitudinous inputs and hence lacks the appeal for getting effectuated to the current target audience. This model can be programmed in a way such that the aforementioned extricable functions used to create it reach a minimum threshold value and reduce the corresponding overhead while increasing the cohesiveness of the amalgamated framework. Conclusively, a successful attempt had been made to integrate the various motions, actions and emotive speech and gestures of computerized avatar while minimizing the causes of errors which were subtly cloistered in the previous works that have been mentioned in this text

5. RESULTS AND CONCLUSION

We have effectuated a model which amalgamates the efficacy of all the modules defined in the integrated framework in a consummate approach which can help alleviate the problems that the target audience may be feeling and hence, ameliorate those issues. Besides this, a framework [18] for the integrated model has also been designed, that can help to better understand how the proposed model will be working as it has been effectuated to other users'. This framework has 3 categories of users which comprise of the end user (i.e. Target Audience), Content Provider (which may be users' loved ones or the user themselves) and the Avatar creator (which will be the proposed framework itself). This framework also has certain requirements which have been stated in the following:

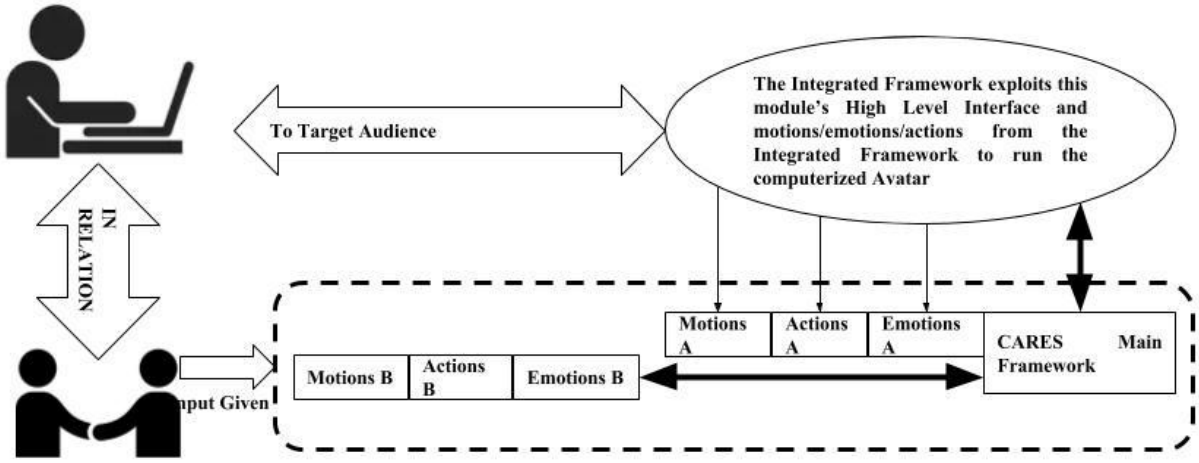

Fig-9: Framework for AGENT.

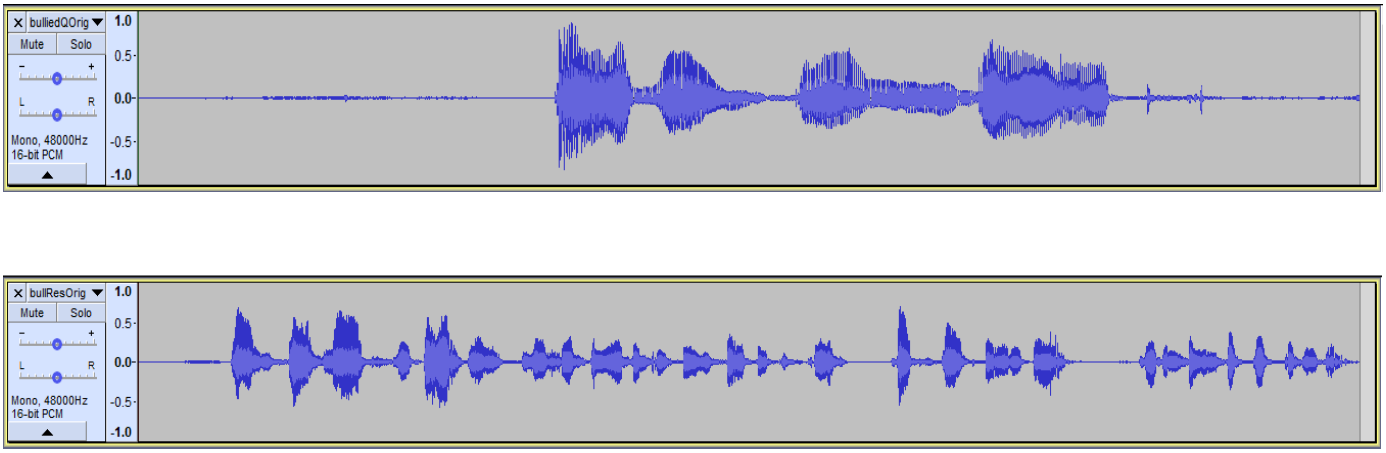

Fig-10: (at top) The input graph for the sentence "I was being bullied" and (at the bottom) the corresponding output that the Computerized Avatar gives to the user i.e. "I'm so glad you told me. You should be able to feel safe at school; that's not fair at all'. This can be dealt with together.

- At framework's side, the model provides an HL interface allowing a combination of several motions/emotions/actions by freely activating and deactivating the motions/emotions/actions performed at anytime
- Internally, it shall ensure a simultaneous and correct execution of successive motions/emotions/actions
- At End User's side, this framework should elucidate the synchronicity and continuity of resulting motions/actions/emotions

The Model in Figure 9 showcases the basic CARES module which will help encapsulate the whole integrated framework into a single entity so, each of the modules can cohesively and inextricably work together to provide the user with a virtualized human in VW [4], so the user can be alleviated of their longingness for their loved ones.

In figure 10, the top graph corresponds to input graph for the sentence "I was being bullied" and the bottom graph corresponds to output that the Computerized Avatar presents to the user i.e. "I'm so glad you told me. You should be able to feel safe at school; that's not fair at all'. We will deal with this together."

REFERENCES

[1] L. Xie, J. Jia, H. Meng, Z. Deng and L. Wang, "Expressive talking avatar synthesis and animation," Multimedia Tools and Applications, vol. 74, no. 22, pp. 9845-9848, 2015. 1, 1993.

[2] Angga, P. Antonius, W. Fachri, A. Elevanita and R. Agushinta, "Design of chatbot with 3D avatar, voice interface, and facial expression.," In Science in Information Technology(ICSITech), 2015 International Conference, 2015.

[3] J. Kapur, S. Jones and K. Tsunoda "Avatar-based virtual dressing room." U.S Patent 9,646,340, 9 May 2017

[4] C. Neustaedter and E. Fedorovskaya, "Presenting identity in a virtual world through avatar appearances," in Proceedings of graphics interface, 2009.

[5] C. Cao, H. Wu, Y. Weng, T. Shao and K. Zhou, "Real-time facial animation with image-based dynamic avatars.," ACM Transactions on Graphics, vol. 4, no. 35, 2016.

[6] S.Noel, S. Dumoulin and G. Lindgaard, "Interpreting human and avatar facial expressions," in IFIP Conference on Human-Computer Interaction, 2009.

[7] A. Ichim, S. Bouaziz and M. Pauly, "Dynamic 3D avatar creation from hand-held video input," ACM Transactions on Graphics (ToG), vol. 34, no. 4, 2015.

[8] D. Bitouk and S. Nayar, "Creating a speech-enabled avatar from a single photograph," in Virtual Reality Conference, 2008, 2008.

[9] T. Frank, M. Hoch and G. Trogemann, "Automated lip-sync for 3d-character animation," in 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics, 1997.

[10] K. Avdelidis, C. Dimoulas, G. Kalliris, C. Bliatsiou, T. Passias, J. Stoitsis and G. Papanikolaou, "Multilingual automated digital talking character," in Proceeding of the International Broadcasting Convention, Amsterdam, 2002.

[11] K. Pietroszek, P. Pham, S. Rose, L. Tahai, I. Humer and C. Eckhardt, "Real-time avatar animation synthesis from coarse motion input," in Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology, 2017.

[12] S. Zhang, W. Zhiyong, M. Helen and C. Lianhong, "Facial expression synthesis based on emotion dimensions for affective talking avatar," Modeling machine emotions for realizing intelligence, pp. 109-132, 2010.

[13] Hong, Pengyu, W. Zhen and T. Huang, "Real-time speech-driven face animation with expressions using neural networks.," IEEE Transactions on neural networks, vol. 13, no. 4, pp. 916-927, 2002.

[14] Saragih, J. M, S. Lucey and J. F. Cohn, "Real-time avatar animation from a single image" in IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, 2011.

[15] H. Tang, H. Yuxiao, F. Yun, H.-J. Mark and S. H. Thomas, "Real-time conversion from a single 2D face image to a 3D text-driven emotive audio-visual avatar.," in Multimedia and Expo IEEE International Conference on Multimedia and Expo, 2008.

[16] L. Xie, N. Sun and B. Fan, "A statistical parametric approach to video-realistic text-driven talking avatar," Multimedia Tools and Applications, vol. 73, no. 1, pp. 377-396, 2014.

[17] L. Wang, H. Wei, S. Frank and H. Qiang, "Text driven 3D photo-realistic talking head," in Twelfth Annual Conference of the International Speech Communication Association., 2011.

[18] Boulic, Ronan, B. Pascal, L. Emering and D. Thalmann, "Integration of motion control techniques for virtual human and avatar real-time animation.," in Proceedings of the ACM symposium on Virtual reality software and technology, 1997.

[19] J. Lester, L. C, S. Zettlemoyer, J. P. Grégoire and W. H. Bares, "Explanatory lifelike avatars: performing user-centered tasks in 3D learning environments.," in Proceedings of the third annual conference on Autonomous Agents, 1999.

[20] B. E. Kolko, "Representing bodies in virtual space: The rhetoric of avatar design.," The Information Society, vol. 15, no. 3, pp. 177-186, 1999.

[21] N. Ducheneaut, W. Ming-Hu, N. Yee and G. Wadley, "Body and mind: a study of avatar personalization in three virtual worlds.," in Proceedings of the SIGCHI conference on human factors in computing systems, 2009.

[22] R. Boulic, Z. Huang and D. Thalmann, "A Comparison of Design Strategies for 3D Human Motions," in Human Comfort and Security of Information Systems, Berlin, 1997.

[23] Z. Wu, K. Zhao, X. Wu, X. Lan and H. Meng, "Acoustic to articulatory mapping with deep neural network," Multimedia Tools and Applications, vol. 74, no. 22, pp. 9889-9907, 2015.

[24] S. Yilmazyildiz, W. Verhelst and H. Sahli, "Gibberish speech as a tool for the study of affective expressiveness for robotic agents," Multimedia Tools and Applications, vol. 74, no. 22, pp. 9959-9982, 2015.

[25] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho and Y. Bengio, "Attention-based models for speech recognition," Advances in neural information processing systems, pp. 577-585, 2015.

[26] M. Luong, H. Pham and C. Manning, "Effective approaches to attention-based neural machine translation," arXiv preprint arXiv:1508.04025, 2015.

[27] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473., 2014.

[28] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499., 2016.

[29] T. Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M. Hasegawa-Johnson and T. Huang, "Fast Wavenet Generation Algorithm," arXiv preprint arXiv:1611.09482, 2016.

[30] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio and Q. Le, "Tacotron: Towards end-to-end speech synthesis," arXiv preprint arXiv:1703.10135, 2017.

[31] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan and R. Saurous, "Natural TTS synthesis by conditioning Wavenet on Mel spectrogram predictions," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779-4783, 2018.

[32] W. Ping, K. Peng, A. Gibiansky, S. Arik, A. Kannan, S. Narang, J. Raiman and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," 2018.

[33] W. Ping, K. Peng and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," arXiv preprint arXiv:1807.07281., 2018.

[34] C. Bregler, M. Covell and M. Slaney, "Video rewrite: Driving visual speech with audio," in Proceedings of the 24th annual conference on Computer graphics and interactive techniques, 1997.

[35] T. Ezzat, G. Geiger and T. Poggio, "Trainable video realistic speech animation", ACM Transactions on Graphics, 2002, vol. 21, pp. 388-398.

[36] D. Terzopoulos and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 6, pp. 569-579, 1993.

[37] D. Bitouk. "Head-pose and Illumination Invariant 3-D Audio-Visual Speech Recognition", Ph.D. thesis, The Johns Hopkins University, 2006.

[38] P. Müller, G. Kalberer, M. Proesmans and L. Van Gool, "Realistic speech animation based on observed 3-D face dynamics," in IEE Proceedings-Vision, Image and Signal Processing, 2005.

[39] M. Brand, "Voice Puppetry," in Proceedings of the 26th annual conference on Computer graphics and interactive techniques, 1999.

[40] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi and K. Tokuda, "Text-to-visual speech synthesis based on parameter generation from HMM," Icassp, vol. 6, pp. 3745-3748, 1998

[41] A. Efros and T. Leung, "Texture synthesis by non-parametric sampling," iccv, p. 1033, 1999.

[42] C. Kimme, D. Ballard and J. Sklansky, "Finding circles by an array of accumulators," Communications of the ACM, vol. 18, no. 2, pp. 120-122, 1975.

[43] S. Lee, J. Badler and N. Badler, "Eyes alive," ACM Transactions on Graphics (TOG), vol. 21, no. 3, pp. 637-644, 2002.

[44] F. Parke, "Computer-generated animation of faces," in Proceedings of the ACM annual conference, 1972.

[45] K. Grant, V. Van Wassenhove and D. Poeppel, "Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony," Speech Communication, vol. 44, no. 1-4, pp. 43-53, 2004.

[46] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proceedings of the 2001 IEEE Computer Society Conference, 2001.

[47] H. Tang, Y. Fu, J. Tu, T. Huang and M. Hasegawa-Johnson, "EAVA: a 3D emotive audio-visual avatar," Applications of Computer Vision, pp. 1-6, 2008.

[48] R. Zhang, P. Tsai, J. Cryer and M. Shah, "Shape-from-shading: a survey," IEEE transactions on pattern analysis and machine intelligence, vol. 21, no. 8, pp. 690-706, 1999.

[49] Y. Shan, Z. Liu and Z. Zhang, "Model-based bundle adjustment with application to face modeling," in Computer Vision, 2001. ICCV 2001. Proceedings, 2001.

[50] Y. Hu, D. Jiang, S. Yan and L. Zhang, "Automatic 3D reconstruction for face recognition," in Proceedings. Sixth IEEE International Conference, 2004.

[51] S. Romdhani, V. Blanz and T. Vetter, "Face identification by fitting a 3d morphable model using linear shape and texture error functions," in European Conference on Computer Vision, 2002.

[52] R. Xiao, M. Li and H. Zhang, "Robust multi-pose face detection in images," in IEEE Transactions on Circuits and Systems for Video Technology, 2004.

[53] S. Yan, C. Liu, S. Li, H. Zhang, H. Shum and Q. Cheng, "Face alignment using texture-constrained active shape models," Image and Vision Computing, vol. 21, no. 1, pp. 69-75, 2003.

[54] Y. Zhou and H. Zhang, "Bayesian tangent shape model: Estimating shape and pose via Bayesian inference," in IEEE Conf. on CVPR, 2003.

[55] Igor S. Pandzic, Robert Forchheimer (Eds), "MPEG-4 Facial Animation: The Standard, Implementation and Applications", John Wiley & Sons, Inc., 2002.

[56] The MBROLA Project, http://mambo.ucsc.edu/psl/mbrola/.

[57] The Festival Project, http://www.cstr.ed.ac.uk/projects/festival/.

[58] Carroll, J.M., Russell, J.A.: "Do facial expressions signal specific emotions? Judging emotion from the face in context.", J. Person. Soc. Psych. 70(2), 205–218 (1996)

[59] Lijuan Wang, Wei Han, Frank K. Soong, Qiang Huo, "Text-Driven 3D Photo-Realistic Talking Head", Conference: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication

[60] Ning Xu, Brian Price, Scott Cohen, Thomas Huang, "Deep Image Matting", arXiv preprint arXiv:1703.03872, 2017.

BIOGRAPHIES

Ansh Mittal received his B. Tech (Computer Science & Engineering) from Bharati Vidyapeeth's College of Engineering, New Delhi which is affiliated to Guru Gobind Singh Indraprastha University. His area of interests includes Machine Learning and Deep Learning. He is a member of IAASSE. He has 3 research papers in reputed International conferences and journals.

Ms. Shilpa Gupta is working as an Assistant Professor in Bharati Vidyapeeth's College of Engineering, New Delhi which is affiliated to Guru Gobind Singh Indraprastha University since 2008. She is Gold Medalist in M. Tech(CSE) from IFTM University. Published 8 papers in various reputed journal and conferences.

Anusurya received her B. Tech (Computer Science & Engineering) from Bharati Vidyapeeth's College of Engineering, Delhi which is affiliated to Guru Gobind Singh Indraprastha University. Her area of interests includes Machine Learning and Data Science. She has a research paper in a reputed conference.

Garima Kumar received her B. Tech (Computer Science & Engineering) from Bharati Vidyapeeth's College of Engineering, Delhi which is affiliated to Guru Gobind Singh Indraprastha University.