Lip Syncing Method for Realistic Expressive 3D Face Model

Itimad Raheem Ali[1], Hoshang Kolivand[2], Mohammed Hazim Alkawaz[3]

[1] Collage of Business Informatics, University of Information Technology and Communication, Baghdad, Iraq

[2] Department of Computer Science, Liverpool John Moores University, Liverpool,UK, L3 3AF

[3]Faculty of Information Sciences and Engineering, Management and Science University, Shah Alam,Selangor, Malaysia

Abstract

Lip synchronization of 3D face model is now being used in a multitude of important fields. It brings a more human, social and dramatic reality to computer games, films and interactive multimedia, and is growing in use and importance. High level of realism can be used in demanding applications such as computer games and cinema. Authoring lip syncing with complex and subtle expressions is still difficult and fraught with problems in terms of realism. This research proposed a lip syncing method of realistic expressive 3D face model. Animated lips requires a 3D face model capable of representing the myriad shapes the human face experiences during speech and a method to produce the correct lip shape at the correct time. The paper presented a 3D face model designed to support lip syncing that align with input audio file. It deforms using Raised Cosine Deformation (RCD) function that is grafted onto the input facial geometry. The face model was based on MPEG-4 Facial Animation (FA) Standard. This paper proposed a method to animate the 3D face model over time to create animated lip syncing using a canonical set of visemes for all pairwise combinations of a reduced phoneme set called ProPhone. The proposed research integrated emotions by the consideration of Ekman model and Plutchik's wheel with emotive eye movements by implementing Emotional Eye Movements Markup Language (EEMML) to produce realistic 3D face model.

Keywords: Lip Syncing, Facial expression, MPEG-4, PCA, Plutchik's wheel.

1. Introduction

Face is definitely the fundamental element for expressing human emotions and personality. Many different types of vital information are detectable on faces. Lately, miscellaneous applications of computer facial animation in the foundation of pragmatic virtual humans with diverse facial expressions are used in entertainments and other fields. An attractive application can advance the interaction between users and devices via interactive virtual speech, and thereby attract users by providing a pleasant interface. With the advent of computer-aided technologies, animated virtual characters are widely used in movies, games and embodied

conversational agents (ECAs) to provide an effective and realistic human computer interaction [1]. Expressive visual speech being the emotional manifestation of a personality requires efficient and pragmatic communication. Advanced research on facial animation mainly focussed to understanding clearly the animated avatar that explores the expressiveness, communication and interactivity aspects of ECA development [2]. Human face being the mirror of internal sensation is undoubtedly the most significant art object and pivotal part that plays a meaningful role in this interaction process. It is needless to mention that the proportions and the facial appearances are highly important to recognize the origin, emotional tendencies, health qualities, sentiments, and often fundamental to human social interaction.

Creation of a realistic facial expression with expressive visual speech is a challenging task due to complex structure of the face. Truly, the combination between eye movement, lip syncing, gesture, emotional facial expression and body orientation provide information about flow of ideas, sequences of thoughts in decision making, and depth of understanding and knowing. Eye movements blended with the gaze convey significant non-verbal information and emotional intentions when a person speaks. Many efforts are dedicated to realize the behaviour of embodied expression including the Behavior Expression Animation Toolkit (BEAT). This allows avatar to speak from input text written by an animator through nonverbal expressive behaviors and synthesized speech [3]. It is essential to create virtual characters that can produce realistic utterances with correct synchronized facial animations involving close-to-nature lip movements and face appearances. Thus, creation of interesting talking and gazing scenes are noteworthy domains in animation research. This remains an utter challenge because nonverbal behaviours are inherently complex. In this view, this paper attempts to develop an interactive model capable in synthesizing expressive visual talking of avatar supported through MPEG-4. The proposed system possesses several notable advantages including: Cheap production and storage costs, only video that is viewed need be created, updates are immediately effective, multiple characters can be used, multiple languages can be used, and delivery of the information can easily be modified using faster or slower speech.

2. Previous Studies

Over the last few years, Virtual Worlds (VWs) especially the 3D graphic contents have been greatly developed and implemented. Rapid progress in this area has enabled almost every user to have access to different tools and applications for the VWs. Building an avatar needs the sharing of many skills to create the precise integration of eye gaze with lip syncing, facial expression, and eye behavior. The major problem in the previous work has been synthesized of realistic visual speech animations corresponding to text or prerecorded acoustic speech input. Fig. 1 displays several important facial features that exemplifies much verbal and non-verbal information require for such combination [4].

Fig. 1. Synthesized facial expressions based on emotion of talking Avatar [4].

Sonke Frantz et al. (1998) investigated the localization performance of the two-step procedure that generalized from the multi-step procedures for subpixel localization of 2D point landmarks to an analytical model of a Gaussian blurred L-corner. Frantz et al. (2000) introduced a new approach to 3D landmark localization based on deformable models, they used quadric surfaces combined with global deformations to model the surface at a landmark.

According to Balc et al. (2007) it is easy to extend research on virtual human characters by using Xface Open Source Project and SMIL-Agent Scripting Language. Queiroz et al. (2009) have introduced new techniques to develop a usable, extendable, and robust facial animation platform capable of easily animating MPEG-4 parameterized face. They used high-level description of facial actions and behaviors by providing interaction between the user and a

virtual character. Bailly et al. (2010) acknowledged the process of creating interesting virtual human character by examining audio and visual face-to-face interaction between human-human and a human-virtual conversational agent. A. Čereković and Igor S. Pandžić (2010) applied preprocessed sets of realized behaviors to virtual human character modules using Back-propagation Neural Network (BNN). Gillies et al. (2010) presented a real time multimodal interaction to the avatar animation system in virtual reality setting [7]. This work was further improved by A. Čereković and Igor S. Pandžić (2011) by innovating multi-platform Real Actor Animation System (RAAS) for realizing real time behavior of ECAs. This work relied on a solution for gestures and speech synchronization based on neural networks.

Lip-syncing is a process of speech assimilation with the lip motions of a virtual character. A virtual talking character is a challenging task because it should provide control on all articulatory movements and must be synchronized with the speech signal (Queiroz et al., 2009; Lee et al., 2011; Serra et al., 2012; Taylor et al., 2012; Leuski and Richmond, 2014; Wei and Deng, 2015). Lee et al. (2010) designed an expressive avatar of real human process for the Lifelike Responsive Avatar Framework (LRAF), which was implemented to analyze the efficiency of expressive avatars. Shapiro (2011) achieved a high level of realism and control by describing a system for the movement of virtual characters by incorporating a set of important aspects of simulated character models and games. It is realized that the lip movements and sound track should be synchronized to provide realistic lip syncing animation. To obtain the capabilities used in spoken dialog, higher-level syncing between two modules is necessary. Most systems used a set of visemes that are activated by a text-to-speech engine (TTS). The TTS engine translates an utterance in text format into a series of phonemes. This technique is used to generate a realistic speech animation without having to manually set the positions for a set of visemes [11]. Serra et al. (2012) presented a visual speech animation method in speeding up and assessing the quality of phonemes to viseme mappings device for the English language (Brett Kessler and Rebecca Treiman, 2002; Serra et al, 2012). Enrico Vezzetti et al. (2013) described the improvements made of the previous algorithm developed by using a new parameterization of the mesh, new geometrical descriptors, and new conditions. Vezzetti, Enrico, and Federica Marcolin (2014) proposed a structured methodology for the soft-tissues landmarks formalisation in order to provide a methodology for their automatic identification.

As aforementioned, facial expressions being the most convincing essence of human's environmental emotional transfer it provides and clarifies the speeding up of intentions and the interactions tuning with others. Majority of the available datasets of facial expression encloses only the postured effective displays of six basic emotions such as natural, happiness, sadness, surprise, disgust, and anger. This paper aims to provide a new framework of realistic facial animation with emotional facial expressions, eye behaviour, and lip syncing. The main aspiration is to enhance the realism of the virtual human character and achieve a visually expressive talking avatar. Several important questions being the core of the research topic need to be answered to achieve these perspectives.

A pragmatic 3D facial model with enhanced interaction and convincing avatar is proposed. Face Gen software (Singular Inversions Facegen software, 2006) is used to achieve the complete facial geometric features. Previous efforts offered a series of eye behaviours, which are combined with diverse emotions. Issues concerning the creation of 3D facial animation model by blending emotional expression with the eye blinks are resolved and a lifelike face is synthesized. Zhao *et al.*, 2012). Previous issues concerned on how to increase the realism of 3D face model by augmenting their expression representation. . Facial expression and speech synthesis seems to be potential to solve these issues and problems. However, the authors have moved further and seen the potential of avatar as an important feature other than facial expression and speech. An in-depth investigation is performed on how to enhance the realism of 3D face model. The integration of facial expression, speech and eye movements are found to be the greatest solution. A benchmarking to the proposed system is performed to evaluate and compare the results with the existing system.

Furthermore, the MPEG-4 generated expressive visual speech is able to build convincingly the dialogued virtual human character via accurate lip syncing and phonemes. In short, a more realistic, synchronized, and effective dialoguing face animation is achieved with MPEG-4. Open source 3D facial animation model Xface is implemented depending on the viability of MPEG-4 approach (Balci et al., 2007). In Xface a set of vertices are selected for all feature points (zones) and a Raised Cosine Function (RCF) is used to deform the region and transfer the vertices in the neighbourhood when a feature point is moved. This is being a distance transform achieves satisfactory results. Another features based real-world example of MPEG-4

FA is Greta (Pasquariello et al., 2001), which is useful for creating wrinkles [15]. The LUCIA works on standard Facial Animation Parameters and speaks with the Italian version of FESTIVAL TTS [16]. Fig. 2 shows the existing talking head system.

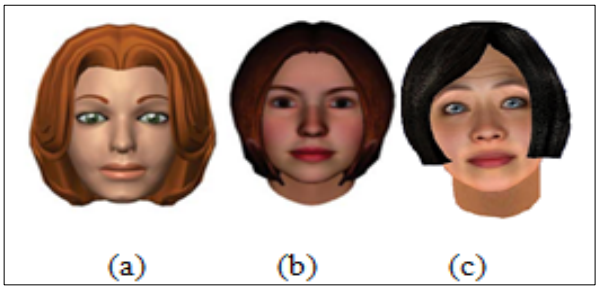


Fig. 2. Talking Head Systems (a) Greta [12], (b) Xface [14][17][17][17], and (c) Lucia[13].

3. Methods

This section explains the step-by-step process to achieving the pragmatic virtual human with visually expressive speech, this method proved by MAGIC-X committee in University of Technology Malaysia. Fig. 3 depicts the framework of the proposed system.

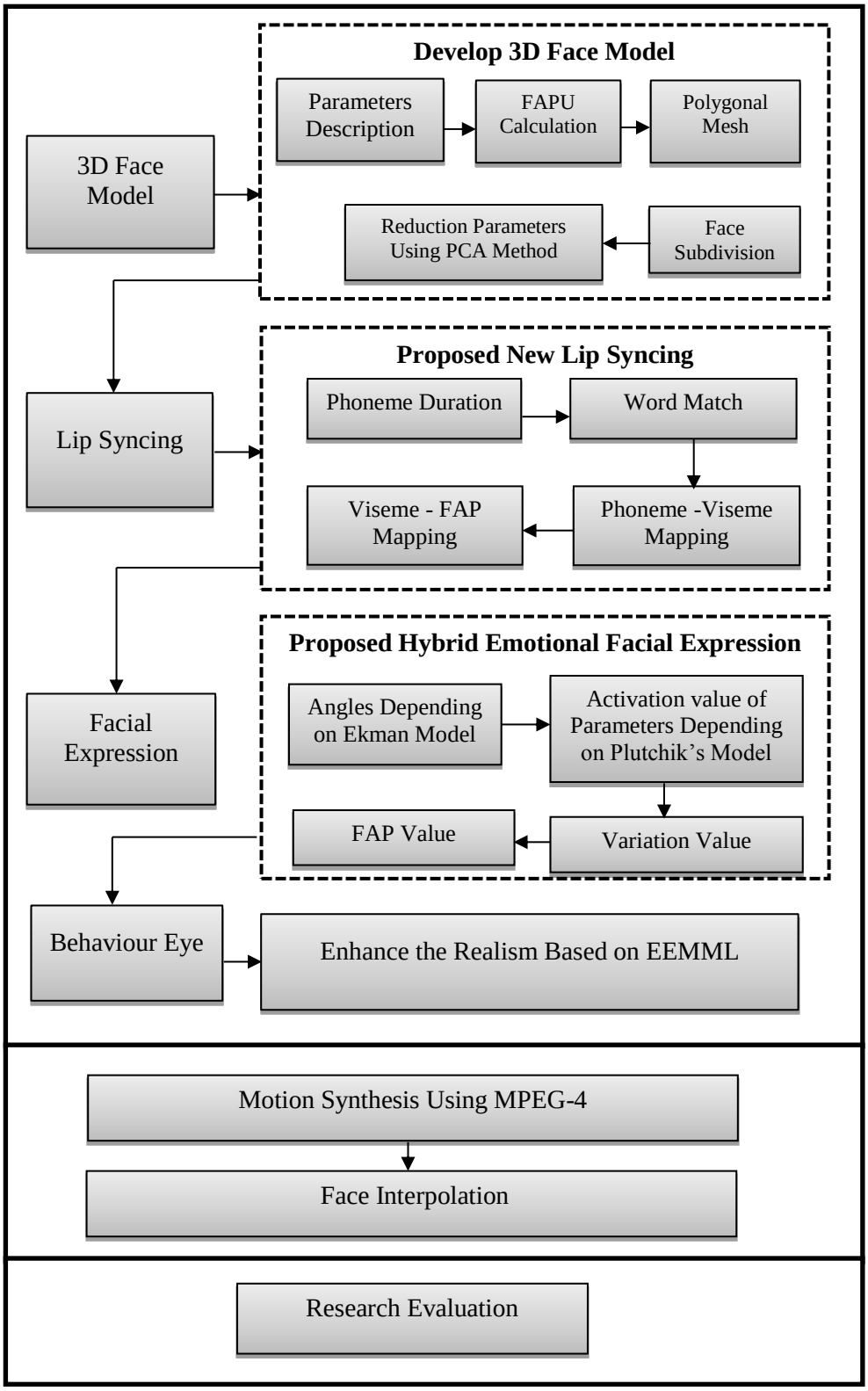


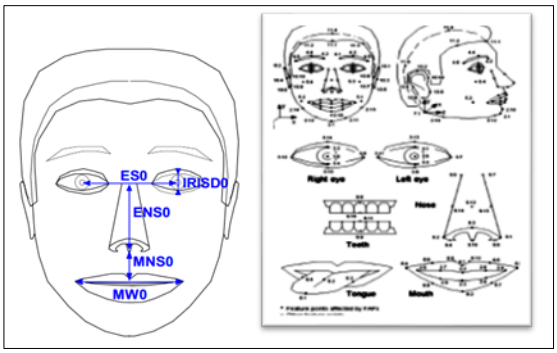
Fig. 3. Framework of the proposed system.

3.1 Development of 3D Face Model

The creation of a realistic visual speech model that could be perceived as engaging is one of the greatest challenges in Human Computer Interaction. There are certain drawbacks in creating such a visual expressive speech open source models: Lack of previous Open source tools available that enable researchers to conduct extensive research in this field. Some of the previous Open source tools available for researches have very less or no documentation and do not provide proper support to deal with bugs. The techniques or methods involved in creating realistic facial animations are often ambiguous and require exhaustive amount of research.

This section describes the development of 3D face model that used in this research depending on MPEG-4 Standard for multimedia [18] with the consideration to Xface free open source[14], in order to create a realistic virtual human face represented as a compact set of parameters, the vertices of 3D face model must be tuned by various parameters (such as the FPs of MPEG-4 approach) and animated by facial animation parameters (FAPs). This research is used the Xface free open source because of three reasons, the first one is the lack of previous open source tools that enable researchers to conduct extensive research in this field. The second reason is some of the previous open source tools available for researches have very less or no documentation and do not provide proper support to deal with bugs. And the third reason is the techniques or methods involved in creating realistic facial animations are often ambiguous and require exhaustive amount of research, also the Xface open source is work under MPEG-4 standard.

The method called moving picture expert group (MPEG-4) was introduced in 1998. This is widely used in geometry coding and animation parameters transmission. 3D face model possesses several definition and animation parameters, MPEG-4 specifies the animation by defining these parameters name. These are face definition parameters (FDP) and facial animation parameters (FAP). MPEG-4 FA standard is comprised of 84 feature points (FP) on the head as shown in Fig. 4. (a) FAPUs of MPEG-4 [2], (b) FPs of MPEG-4. To animate each FP the corresponding vertex on the model and the indices to the vertices in the zone of influence of this FP should be set. FDPs include information for building 3D face geometry, and FAPs are designed to encrypt the animation face emotions, expressions and speech pronunciation.



(a) (b)

Fig. 4. (a) FAPUs of MPEG-4 [2], (b) FPs of MPEG-4

Usually, the face contains 68 parameters and these parameters are classified into 10 groups depending on the facial region. FAPs being the representative of most of the facial expression, the parameter set contain the viseme and the expression. The units that are responsible for the face parameters animation is called face animation parameters units (FAPU). A data-driven approach is used in training the system to reconstruct the expressive articulation of the avatar while depicting its different emotions. MPEG-4 facial animation is utilized in the entire algorithm and implemented in a database to enable real time animation behaviour within the Unity 3D environment. . Fig. 5 shows the steps of generating 3D model.

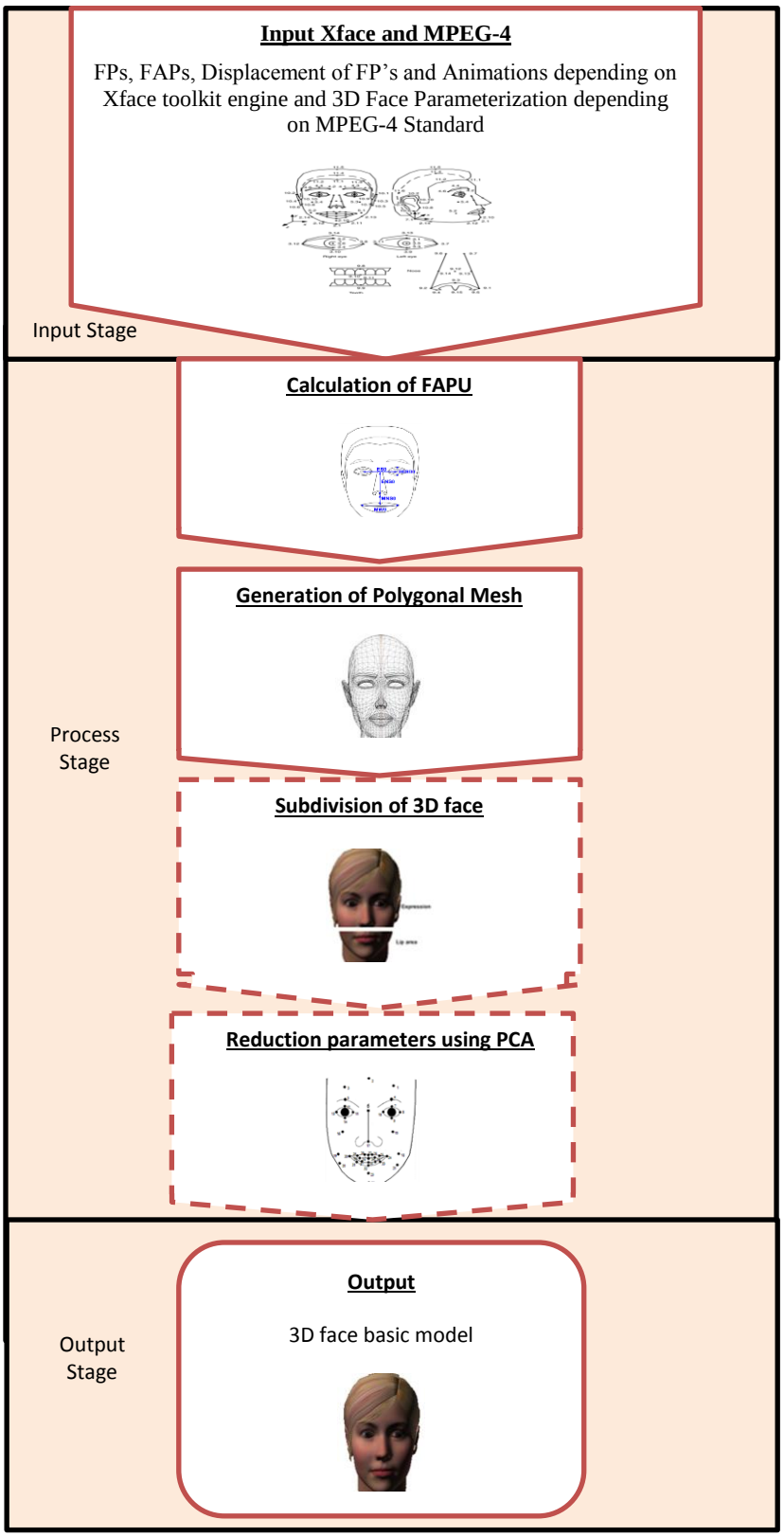


Fig. 5. Steps of developing proposed 3D face model.

As aforementioned, Xface being MPEG-4 based engine is used to develop or create 3D face talking. This engine contains the following three modules:

i) XfaceEditor

ii) XfacePlayer

iii) XfaceClient

The system design is easily extendable as well as configurable, where these modules are any C++ compiler compatible. Moreover, Xface possesses few limitations, where the decoded and encoded ends are not covered for generation animated faces. It does not contain higher-level abstraction, time tuning and event administration. Furthermore, some of FAPs including 14, 15, 23-30, 35, 36, and 39-47 remain unimplemented in Xface. The present thesis implements more number of FAPs in Xface to generate more realistic motion and smooth blending in the animated face of virtual character.

Firstly, the proposed system provides the values of each FAP corresponding to the desired facial action through time. Secondly, the lip syncing and facial expression method receives the FAP values and resolve possible conflicts among them (such as combining facial expressions and visemes) using PCA method (Pearson, 1901). Each parameter is independently modeled using PCA. Thus, it is advantageous if the modeled parameters are reasonably independent. However, a considerable amount of co-dependency that exists between adjacent parameters in the face makes MPEG-4 and FAPs correlated. The PCA is performed to simultaneously diminish the co-dependency and reduce the number of parameters. A separate PCA is executed for each emotion in the corpus. The top ten principal components are able to explain 99% of the variation in the original FAP data streams. Nevertheless, the output FAP file containing the final animation is capable of running on any MPEG-4 compliant player with different face. The FAP values deal with specific regions of the face such as the left eyebrow, right corner of the lip, tip of the tongue, and so on. The FAP values being independent of facial model geometry are calibrated prior to the use on a facial model in obtaining a realistic expression.

The second step is to define the FPs, which consist of 84 FPs on the head and all of them need not have to be defined because every one of this FPs is not affected by the FAPs The weight and deformation function for these FPs are manually defined. Fig. 6 illustration of the vertices from one region to the application of a Raised Cosine Deformation Functions (RCDF)

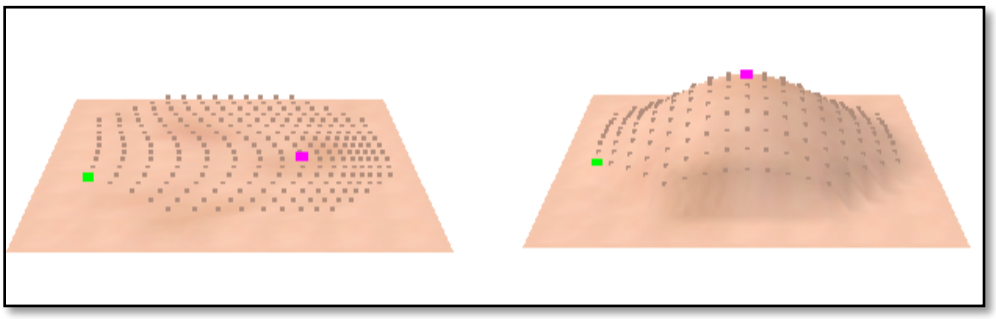
before and after applying. Raised Cosine Deformation Function which is built in the database and satisfactory results are achieved. Cosine functions (Raised Cosine Deformation Functions or RCDF) are found in some studies in the literature as a smooth and quick way to deform influenced regions by FPs, according to [20]. Let $v_j\{\vec{v}_1, \vec{v}_2, \ldots \vec{v}_n\}$ the set of vertices FPs and d_p the Euclidean distance. The Raised Cosine Function acts on the FP_j and each vertex v_j of its influence's region according to Equation 1.

$$F(\overrightarrow{v_p}) = \left(1 + Cos\left(\pi \times \frac{d_p}{d_{max}}\right)\right) \times w \times \Delta FP \quad (1)$$

where v_p and FP_j are the vertices towards FAP direction, d_{max} is maximum distance between farthest away from the vertex v_j to the FP_j in the influence's region, ΔFP is the displacement of the control points (FPs assets) that given by the value of parameter j animation set of FAPs (FAPj) climbed by Avatar FAPU (corresponding to FAPj):

$$\Delta FP_j = FAP_j . FAPU_{fapj} \quad (2)$$

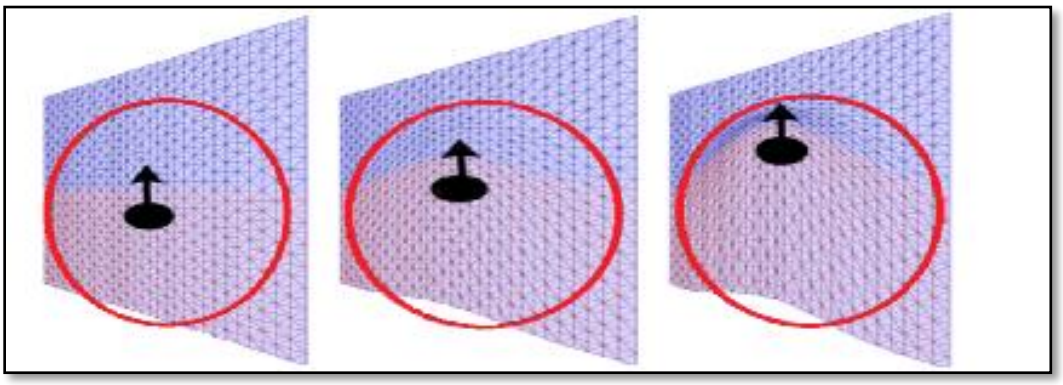
Equation 1 is used deform every vertex in the FPs region of influence and the controlled co-ordinates are moved.



(a) **(b)**

Fig. 6. Illustration of the vertices from one region to the application of a Raised Cosine Deformation Functions (RCDF) before and after applying (Queiroz *et al.*, 2009)

Points of pink colour are the control point (FP asset) and green is one of the most remote parts of the FP (since the area is circular in this example), where (a) is before of applying RCDF and (b) after applying the RCDF. Fig. 7 reveals the influence of the dislodgment of RCDF centre (black point) on a polygon lattice plane with increasing intensity. Where (a) is null intensity, (b) is medium intensity and (c) is maximum intensity.



(a) (b) (c)

Fig. 7. Deformation within a region of influence (a) null intensity, (b) medium intensity, and (c) maximum intensity

Fig. 8 illustrates the mimicking of these actions of the skin which are acquired by total of horizontal displacement of the vertices in the A and B zones following RCDF and the vertical displacement of a vertex in the B zone. Algorithm 1 summarizes the pseudo-code of RCDF.

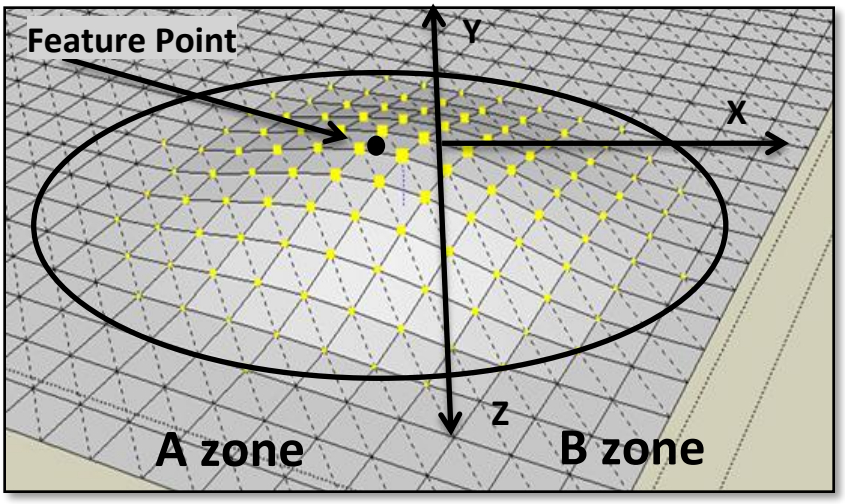


Fig. 8. Two skin regions and FPs

Algorithm 1 : Raised Cosine Deformation Function.

```
Input      weight_fap, m_dis
Output     new_weight
Initialize m=3.14
/* distance map to the control point is calculated in 3 dimension X, Y, Z */
/* Compute weight value*/
   new_weight= (cos(m/max_dis)+1) * weight_fap
   return RCF(new_weight)
```

The main problem of weighted FAPs is that as the number of viseme is increased for the input sentence, the weight of each target viseme gets decreased and effect of these poses events becomes less evident in the final result. Take an example of parameterize animation of speech with emotion. The speech requires lip motion that is changing of visemes to each other in time. But if one requires a smiling lip motion in speech one needs to change the whole animation with an emotion model corresponding to smiling.

If weight is very close to 1 then the expression will be very insignificant. Alternatively, if weight decreases, the weight of expression parameters is increases but the weights of lip parameters under a multi sequence get reduced and the animation quality of the lip motion degrades. In other words, the animation loses its understand-ability. To avoid this problem, a segmented pose is done by increasing the weight of the lips parameters.

Using equation 3, the weighted value (w) for each deformation function can be computed (Balci *et al.*, 2007). This value is manually assigned for each FAP during the configuration of the definition of each face parameters of so that the deformations become smoother but with these values, not guaranteed in the FP the desired displacement and, therefore, it impacts on the behaviour of other parts of the region. In this work, the weight of parameters depends on the priority of the particular parameters. The parameter of the expression module provides its weight less than the priority of the lip syncing because of the loss during animation. Therefore, this equation will empower the value of weight of the parameters. Following (Li *et al.*, 2013) the function for weight is given by

$$w_j = \rho.(E_j + \cos(\frac{\pi}{2} * (1 - \frac{d_{1p}}{d})) \quad (3)$$

Where j is the parameter index, E is the activation value of emotion of FAP_j that get from facial expression synthesis and ρ is empowered constant, which is taken as 3. Accordingly, the modified weights of each FAP for the cheek region are depicted in Fig. 9.

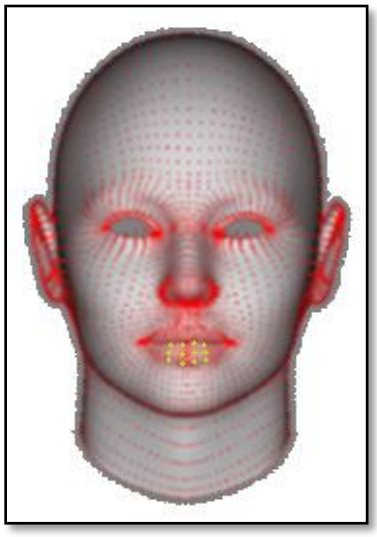


Fig. 9. Fixing the weights and deformation functions for the 3D model

Corresponding to MPEG-4 FA, the FAP values are obtained from linear and normalized functions. The points are put in a matrix with the row as time and column as point coordinate. Consequently, the FAP matrix (F) is expressed as:

$$fap = fap_0 + M.U.(FP - FP_0) \quad (4)$$

where FP_0 is the rest point of FP. The FAP value F_0 is manually calculated. The matrix M is used to map FPs (columns in X) to FAPs (columns in F), and U is a diagonal matrix that encloses the correct FAPU scale factors for every FAP.

A good example on the multiple displacements that appear in the face, the region of the eyebrow is manifold dislodgment of the FPs because of the FAPs and to the supplementary deformations integrated to the skin. Enhanced values on FAP 31, FAP 33, and FAP 35 produce the deformation on the forehead of the proposed 3D facial model as shown in Fig.s 10 and 11.

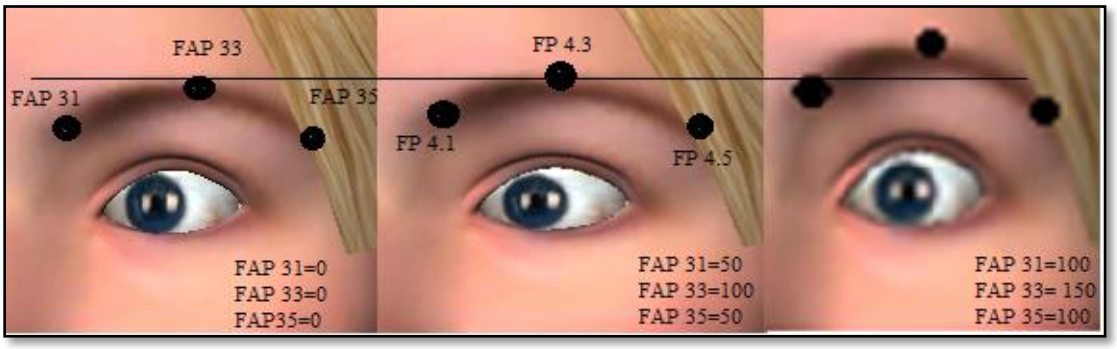


Fig. 10. The action of enhanced values of FAP 31, FAP 33, and FAP35

Fig. 11. The influence of increased FAP31, FAP33, and FAP35 to entire face

The vertices of the mesh are weighted according to their distance from the control point. The rule used to compute the metric is as follows: "the space is the sum of the edge separations alongside the shortest trail among two vertices". This measure simultaneously manages the variation of holes and mesh density variation in the facial model.

After completed the input parameters of 3D face model as explained in the previous section, the second step is to determine measurement units for the animation parameters, the FAPU, as described in MPEG-4 standard. The distances: $FAPU_0$ = {$IRISD_0$, ES_0, ENS_0; MNS_0; MW_0} are calculated using the positions of the vertices marked as FPs (equations 5 to 9), and then are normalized as suggested by the MPEG-4. These units are shown in Fig. 12.

$$IRISD0 = | FP3.1y - FP3.3y | \quad (5)$$

$$ES0 = | FP3.5x - FP3.6x| \quad (6)$$

$$ENS0 = | FP3.5y - FP9.15y | \quad (7)$$

$$MNS0 = | FP9.15y - FP2.2y| \quad (8)$$

$$MW0 = | FP8.3x - FP8.4x| \quad (9)$$

As example:

FP3.5 = (FP3.1x, (FP3.1y - FP3.3y)/2)

FP3.6 = (FP3.2x, (FP3.2y - FP3.4y)/2)

FP2.2 = (FP8.1x, (FP8.1y - FP8.2y)/2)

FP9.15 = (FP9.3x, (FP9.3y - FP2.2y)/2)

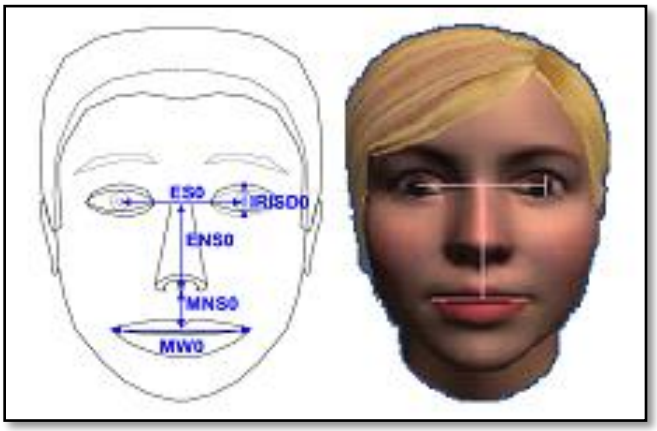


Fig. 12. FAPU of MPEG-4 approach

Now, the third stage is to generate the polygon surface. 3D facial model represents the 3D geometry of the face plane and texture of skin. This geometry is estimated via 3D triangular or polygon meshes with several vertices, where the face appears smooth provided the amount of triangles or polygons are adequate. Highly dense linked triangular meshes are selected to

approximate the facial surface elements, where 2240 vertices and 2946 triangles are chosen to guarantee the realism proximity. The standard head model of human can be tailored via textured mapping (see Fig. 13), where the face image is acquired using a Facial Motion Capture system (in the present case a real actor speaking English).

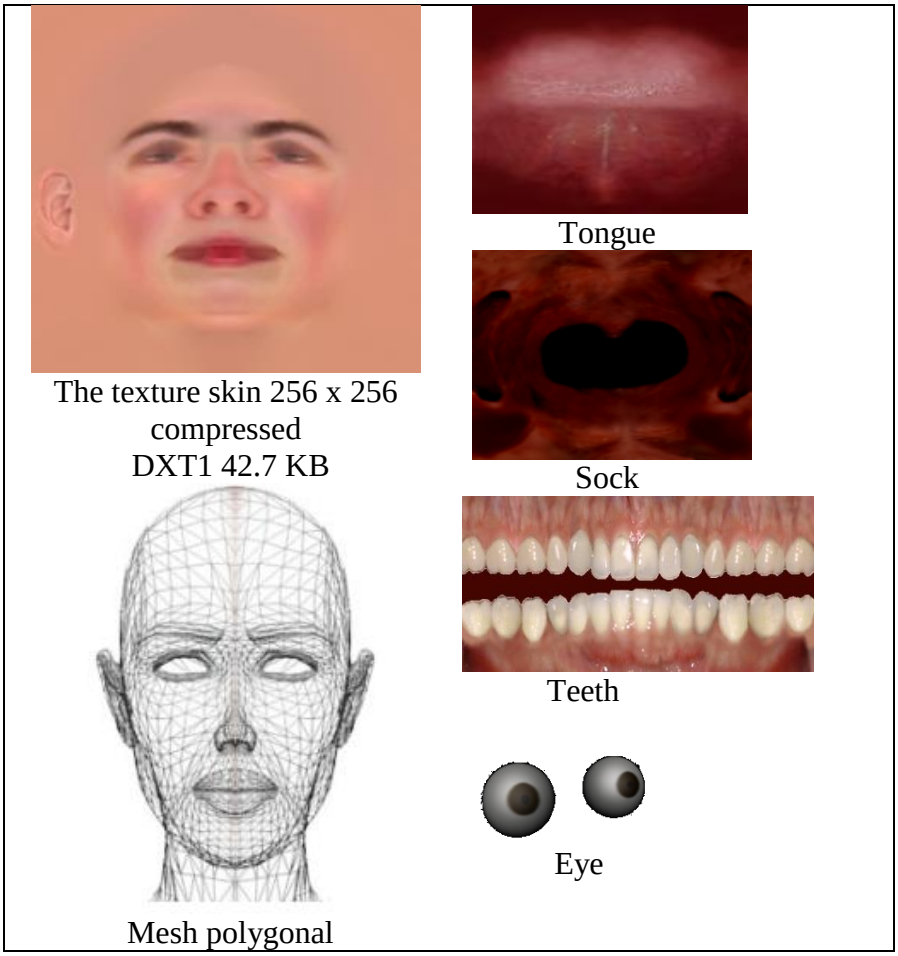


Fig. 13. Texture map, mesh polygonal, and the internal anatomic components of the proposed 3D model

This section presents the model developed, which aims to provide face expressive visual speech character, Fig.14 displays the 3D face model (front (a) and side view (b)) that is created compliant with the MPEG-4 standard. The model allows control option of animations, it scripts of facial actions, in order that the application can offer different options facial expressions, audio and / or chat. The morph targets for the 3D model are classified as:

1) **Emotional facial expression referring to 6 basic emotions {Anger, Sad, Happy, Disgust, Surprise and Fear}**
2) **Viseme signifying the visual speech generation.**
3) Reality refers to the animation such blinking of an eye.

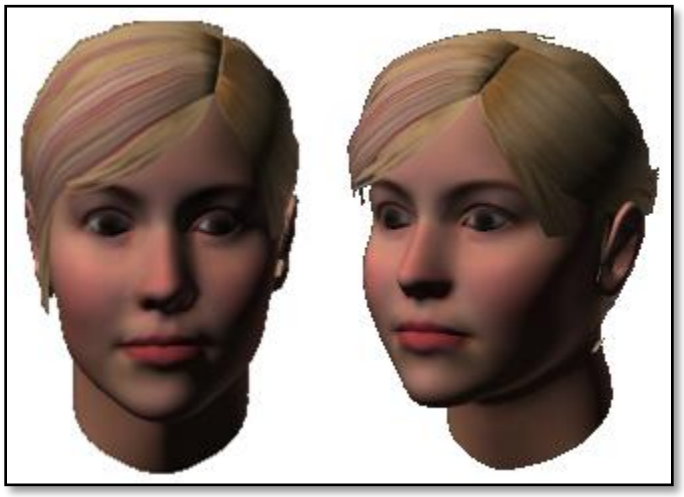


Fig. 14. 3D face model created.

To achieve higher tuning on the polygon lattices, the 3D facial model divides the surface in two definite parts corresponding to the FPs influenced by FAPs. One of contribution of this research is subdivision the face in to two parts. This sectioning is essential to restrict and manage the displacements of polygon vertices. In this work the reason of divided the face model to two part to give weight to the lip parameters more than the other. Subdivision the face has three advantages over straight parameterization:

- *Firstly*, it has the ability to build a large number of expressions from a smaller set of target poses.
- *Secondly*, it is capable in animating changing expressions while the character is talking.
- *Thirdly*, it is capable to increase the weight of FAPs For instance, in the nasolabial furrow, both regions on its opposite sides belong to distinctive domains of lips. The classifications of these domains are displayed in Fig.15.

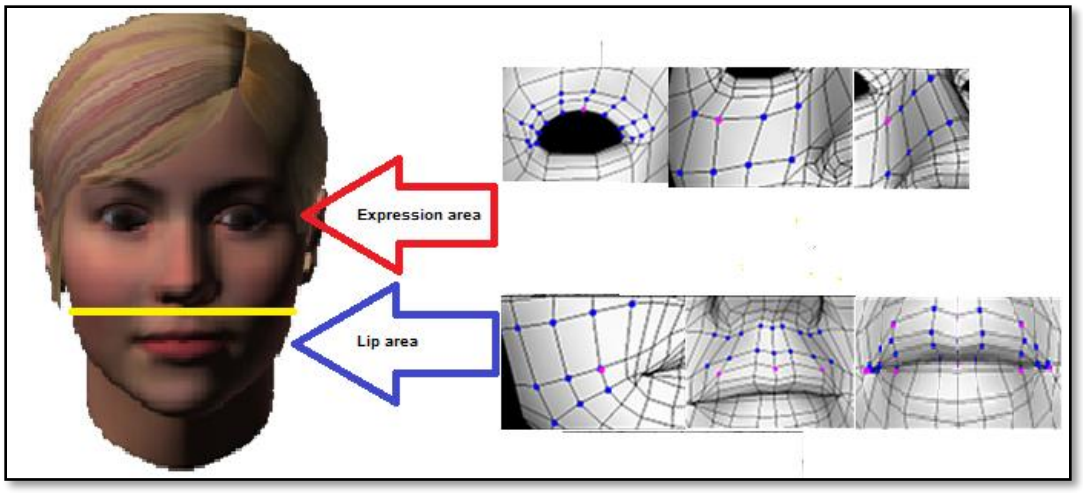


Fig. 15. Subdivision of proposed 3D model in "two specific areas"

This research apply principal component analysis (PCA) to reduce the dimension of the motion feature vectors, which allows to obtain a compact representation by only retaining a small number of principal components. This work keeps the 50 most significant principal components to cover 98.0% percent of the motion dataset's variance. The reason of reduction parameters is to run the algorithm faster, simplify the dataset (84 FPs), facilitating description, visualization, and insight.

All 84 parameters are converted to get a definite local parameter at the central point of every frame (Balcı *et al*., 2007). The movements of every 84 parameters in a single structure are enclosed within 252 dimensional movement vectors. Therefore, the contribution of this section is decreasing the number of the FPs that used to get the motions.

These 50 parameters cover the FPs that used by MPEG-4 FA standard, which 16 FPs are used for lips, 26 for expressions with eyes, 4 for teeth and 4 for tongue as shown in Fig.16. The parameters number dictates the FPs illustration capability. Additional parameters facilitate FPs to encode supplementary information. These 50 parameters be the FPs that will effect and reduced by PCA method. In the present thesis the dimensionality is set to 150 to cover almost 98.0% of the movement changes.

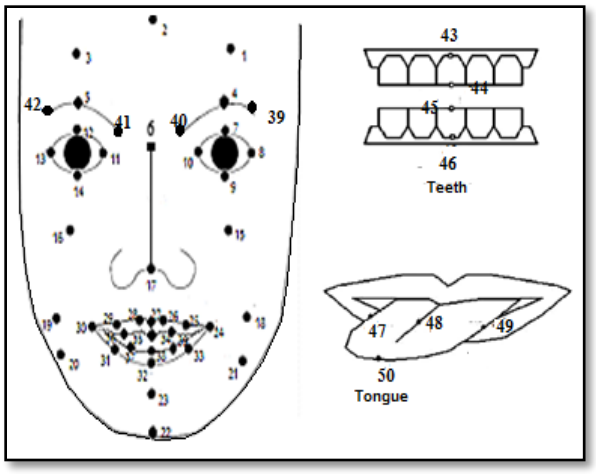


Fig. 16. The total of 50 FPs of the proposed 3D face model

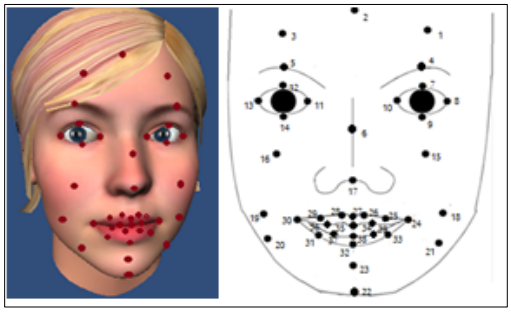
Decreasing the number of parameters back to the reason to minimize the number of frame per motion and by default there are some of FPs are unaffected and some of these parameters has tracking error by default also to reduce the feature vector for each motion frame. The reason of selecting FaceGen Software from other (such as Solid, Polygonal, 3DTin, Blender modelling) is about renowned for creating 3D model. It provides the user with an exquisite variety of options to create a 3D face model of their preference. It allows the users to modify the skin tone, colour, shape of the nose, lips, chin and cheeks.

The PCA is one of the dimension reduction methods for transferring into a new orthogonal basis, where the axes are oriented in the directions of the maximum variance of an input data set. The variance is maximized along the first axis of the new basis, whilst the second axis maximizes the variance subject to the first axis orthogonality, and so forth. The last axis possesses the least variance of all possible ones. Such transformation permits information to be reduced by rejecting the coordinates that correspond to the directions with a minimum variance. If one of the base vectors needs to be rejected, that should preferably be the vector along which the input data set is less changeable.

PCA is a classical machine learning method useful to reducing the dimensionality of a problem. PCA involves the calculation of the eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix. Being a extensively used technique for

dimensionality reduction it is employed in several 3D facial expression recognition and synthesis (Somasundaram, 2006; Bao, 2011). It is a statistical method that provides a linear decomposition of sample data. Usually, PCA divides the face into several isolated regions to model them as independent components. Moreover, subtle dynamics of a talking mouth may be lost after the smoothing of the synthesized trajectories. The goal of PCA is to find a sequence of uncorrelated random variables (components) where each variable covers as much of the variance of the data as possible. The resulting sequence is ordered by decreasing variance coverage. Thus, PCA is often an effective compression technique that keeps the first few components of most of the variance in the data to be covered.

This method divides the face into two parts with respect to the eyes region. Here, PCA is applied to learning the significant characteristics of the facial deformation sample. Depending on MPEG-4 there are 68 FPs, where 47 FPs are selected to render good details to the expressive talking face. These are important in the measures of data analysis so that each motion is represented by $47 \times 3 \times$ the number of frames per motion in complete dataset. The distribution of FPs is displayed in Fig. 17. PCA method possesses several benefits including the capability to build a large number of expressions from a smaller set of FPs, to animate changing expressions while the character is talking, to reduce the dimensionality of the parameter space, to get average deformation of each frame, and to retain enough components to capture more than 90% of the variance of the motion.



(a) (b)

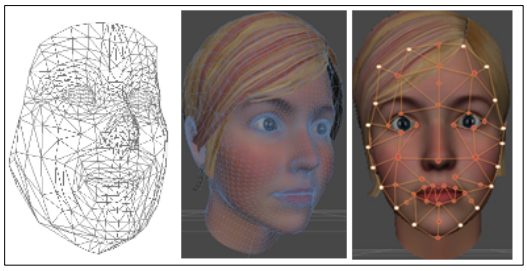
Fig. 17. (a) The 3D face model with 39 FPs and (b) simple sketch to specifying the distribution of selected 39 FPs.

The segmentation of facial features also called face tracking is the key issue in face analysis. Segmentation of face performed via PCA has two advantages over straight parameterization.

Firstly, it has the ability to build a large number of expressions from a smaller set of target poses. Secondly, it is capable in animating changing expressions while the character is talking. The correlated and ordered new dimensions of the mapped data are called the PCs, which represent the direction of maximum variance in the original data. Therefore, application of PCA to facial point data is appropriate due to their strong correlation.

The main limitation of weighted parameter is to deal with the large number of viseme in the input sentence, where the weight of each target viseme gets decreased. For example in the parameterized animation of speech with emotion, the speech requires lip motion (time variation of visemes to each other). However, smiling lip motion in speech requires to changing the whole animation with an emotion model corresponding to smiling. Assuming that the total lip motion sequence is weighted by the real number w ($0<w<5$), the expression method is weighted with $1-w$. For w very close to 1, the emotion is considered to be extremely insignificant. Alternatively, if w decreases, the weight of emotion increases but the weights of lip motion sequence reduce and the animation quality of the lip motion degrades. In other words, understanding ability of the animation is lost. This problem is overcome via a segmented and empowered pose with increased lip syncing method weight.

Facial motion is generated using PCA in the face point trajectory synthesis algorithm. The face is segmented into expression and lip module. Fig. 18 illustrates the proposed facial animation system with the utilization of PCA on the proposed system.



(a) (b) (c)

Fig. 18. FP's Reduction (a) The facial mesh, (b) facial mesh of the proposed 3D face model, and (c) application of PCA on the proposed 3D face model.

As mentioned, the main idea of the PCA is to reduce the dimension of the data by mapping inputs from the original space to a new space. Mapped variable data (called PCs) represent the direction of maximum variance in the original data. In the present paper, the mean facial

deformation and the first seven eigenvectors of PCA results corresponding to the largest seven eigenvalues are selected as the PCs. They represent the facial deformation around lips and mouth corners.

The data is represented as $\{x_n\}_{n=1}^{N}$. PCA finds a linear basis $PC \in R^{D\times d}$ for projecting a given data point x_i to PCA parameters b_i given by,

$$b_i = PC^T(x_i - \bar{x}) \tag{10}$$

where $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ is the mean vector.

Thus, the 3D data X is projected onto a lower dimensional subspace $B = \{b_n\}_{n=1}^{N}$ of dimension d with $d < D$. Reconstruction of the original data from PCA coefficients is performed using,

$$x_i = \bar{x} + PCb_i \tag{11}$$

The basis of PC is obtained by computing the covariance matrix C written as,

$$C = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^T(x_i - \bar{x}) \tag{12}$$

$$C = \frac{1}{N}ZZ^T \tag{13}$$

where $Z = X - E[X]$.

The squares symmetric matrix C is decomposed using Singular Value Decomposition (SVD) to obtain its eigenvalue $\Lambda = [\lambda_1, \dots \lambda_D]^T$ and eigenvectors $PC = [pc_1, \dots, pc_D]^T$ by solving,

$$C = PC\Lambda PC^T \tag{14}$$

$$ZZ^T pc_i = \lambda_i pc_i \tag{15}$$

The eigenvectors are arranged in descending order of eigenvalues, $\lambda_1 \geq \lambda_2 \dots \geq \lambda_D$ such that the first PC corresponds to the highest variation in the data and the last renders the least variation. Out of total D data axes, only a small number d is captured for most significant variance. The other axes of lower variation often attributes to the noise. A certain percentage of pc of the data variance are retained using $\sum_{i=1}^{d}\lambda_i \geq \frac{pc}{100}\sum_{i=1}^{D}\lambda_i$ to providing the retained eigenvectors with $PC = [pc_1, \dots, pc_d]^T$. These eigenvectors implying the principle component coefficients or loadings are further used to obtain a lower dimensional representation of the data

from equation (10). It is worth to look at the expression method to manifest the actions in the animation engine. The pseudo code of PCA method is present in algorithm 2.

Algorithm 2: Pseudo code of PCA method.

```
Input  m: data matrix of FP of size d
Input  d dimension
Set parameters FAP (default =68), d (default =47)
Output Principle Component (eginvec), mean
Initialize  n=10;
    for i ← 0; I < fap; i++)
        for j ← 0; j < d; j++)

            /* calculate mean vector */
            for (i ← 0; i < d; i++)
               {   for (j ← 0, sum ← 0.0; j < d; j++)
                     {  sum +← m[i + j * d];
                        mean[i] ← sum / d;
                   }}
            /* Compute the covariance matrix (Co.) by Eq. (12)*/
            for (i ← 0; i < d; i++)
               {    for (j ← 0; j < d; j++)
                     {  sum ← 0.0;
                        for (k ← 0; k < d; k++)
                           {      sum      +←      (m[i+k*d]-
mean[i])*(m[j+k*d]-mean[j]);
                           }
                        var[i][j] ← sum / d; }}
   /*Compute the eigenvalues and eigenvectors of the covariance
matrix by Eq. (14) and (15)*/
            for (i←0; i< d; i++)
               {     for (j←0; j<n; j++)
                        {     If i==j then
```

```
                                e_vec [i][j]=m[i][j]  }
    }

     if ((var == -1)
     {
          write("Error : matrix is not symmetric.\n");
          exit(EXIT_FAILURE);
     }
/* calculate contribution rate of each eigenvalue */
for (j ← 0; j < d; j++) {
    sum ← 0.0;
    for (i ← 0; i < d; i++)
      {       sum +← e_val[i];
             cont_rate[j] ← e_val[j] / sum;
      }  /* end of PCA */
  /* output mean vector and eigenvectors */
      write(mean, d);
      for (i ← 0; i < n; i++)
            write(e_vec[i], d);
/* output eigenvalues and contribution ratio */
      for (i ← 0; i < n; i++) {
           {  write(e_val + i);
              write(cont_rate + i);
       }}}
```

3.2 Proposed a Prophone Lip Syncing Method

Now we turn the attention on lip syncing module. As aforementioned, the goal of this paper is to construct a set of animations that are associated with the co-articulation and proper expressions. In this regard, a new technique called “ProPhone” is suggested for choosing phoneme pairs as a canonical unit of animation. They produce the co-articulation effects via Diphones and Triphones that are not possible by associating animation with individual phonemes. Diphones are commonly used during machine learning techniques, which represent timings between the middle of one phoneme and the next. Moreover, an animation of a Prophone is intuitive whereas diphone is non-intuitive representation.

ProPhone can be constructed by animating a short word with a single syllable that represents the two phonemes. For example, when animating the Prophone for L-Oh, an animator creates the animation that represents the word ’Low’. Conversely, the absence of intuitive representation of the mouth movements for a diphone would require the animator to start the motion from the middle of the ’L’ sound to the middle of the ’O’ sound. Animators are able to create better quality set of animations by Prophone. In this work phoneme pairs are chosen instead of combinations of three phonemes or higher order sequences to reduce the amount of data needed to be produced by the animator. Better results can be obtained by animating or learning combinations of three phonemes or even longer sequences.

Fig. 19 illustrates the architecture of lip syncing method following Quieruze [2] and Ari Shariro [25]. Basically, this method receives a TTS file or Audio file as input containing the character’s speech and an auxiliary file containing its textual description. Lip animation is created via CSLU Toolkit. It aligns the input file and provides a file with the timeline for the phonemes. This file acts as the input for the phoneme-viseme mapping methods and connects the phonemes to their visual representations (visemes).

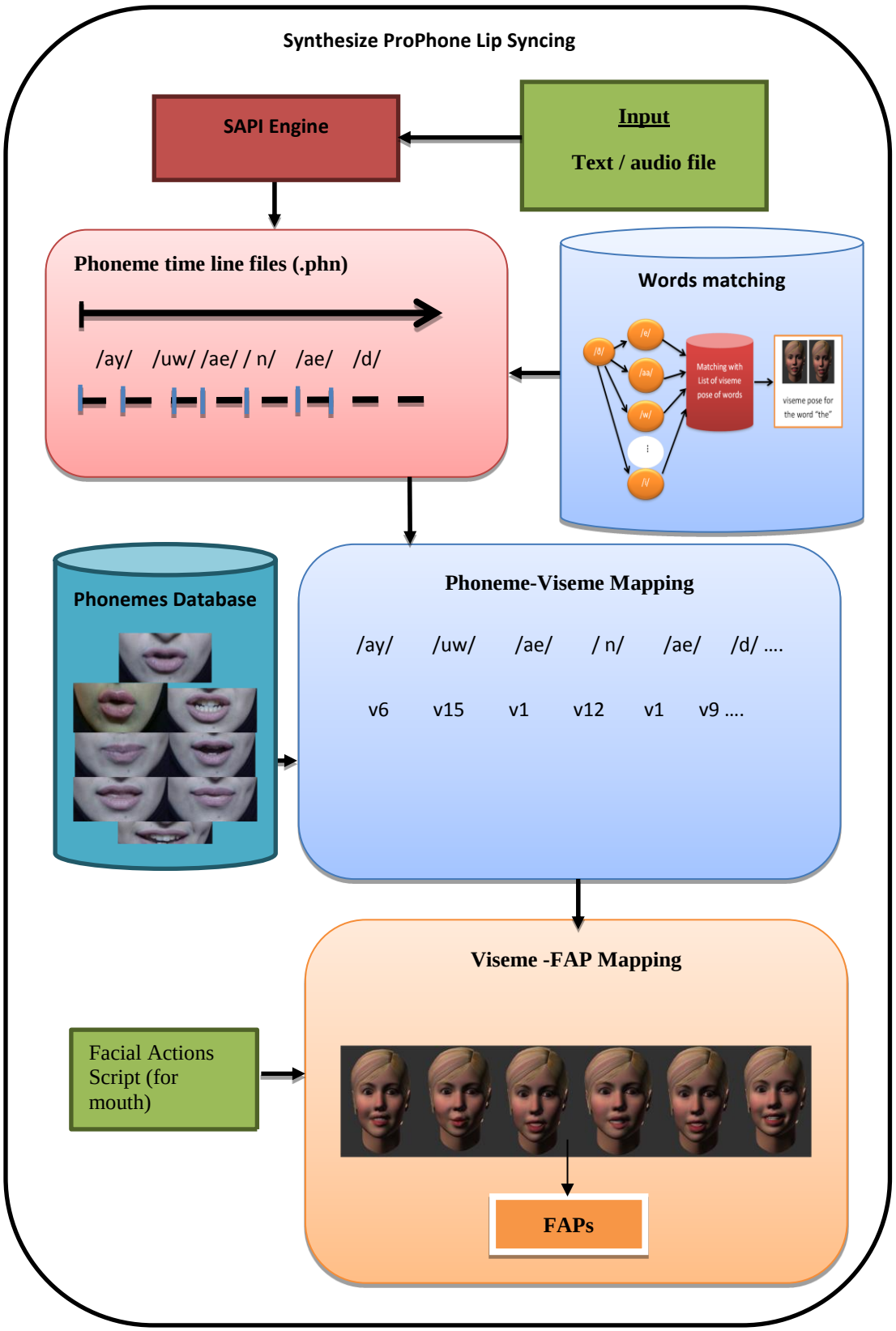


Fig. 19. Schematic diagram of Prophone lip syncing module.

Visemes are the visual counterpart of speech animation generation. The morph targets are carefully modelled for a set of phonemes to achieve proper lip syncing of the agent with the audio. 15 phonemes are used by TTS for generating the audio. They are selected from the 40 phonemes of English language. Therefore, the subsequent morph targets of the proposed 3D model are created for all the phonemes as listed in Table 1. Forty English phonemes are mapped to our common set of phonemes. The percentage of the phonemes are obtained from the website [26][27][28]. The first column lists the full set of English phonemes and the 2nd one represents their pair-wise combinations for the given phoneme schedules. The 3rd column displays the example on the phonemes and 4th one enlists the frequency of common phoneme set generated from TTS engine on a corpus of approximately 200 utterances. Accurate selection of phoneme and visems together with their mapping and alignment play paramount role in the animation.

Table 1. Set of English and selected phonemes [26]**,** [27][28]**.**

English Phoneme (40)	Phoneme Set (15)	Example	Percentage
Ih, iy	i	Feel, fill, debit	9.9%
N, ng	n	No, sing	7.8%
Ey, ee, Ay	ee	Ate, pet, bite	5.0%
D, t, s, z, l	d	Dig, talk, sit, zap, lid	4.0%
R, er	r	Red, fur	3.9%
P, m, b	p, m, b	Put, mat, big	3.3%
K, g	k	Cut, gut	2.6%
ae, ah, ax,	ah	Cat, cut, ago	2.1%
Aa	aa	Father	1.9%
F, v	f	Fork, vat	1.7%
Ao, Ow, aw, O	oh	Dog, go, Foul	1.2%
W, uw, uh	w	With, too, book	1.2%
Sh, ch, jh, zh	sh	She, chin, joy, pleasure	1.2%
H	h	Help	0.7%
Th, dh	th	Thin, then	0.3%

The first stage in the method is to segment and process neutral audio. Neutral audio of different sentences are recorded and are modified to produce emotional audio. First, phonemes, syllables, and words are denoted in the recorded audio. Despite the existence of automated methods for this mark-up [29], the manual speech segmentation is considered to achieve more accuracy. The list of phonemes is obtained using the text of the spoken sentence [30]. The input for creating expressive audio is neutral audio and emotion. Emotion parameters such as nature of the emotion (for example, sad, happy etc.) and emphasis level of words ranging from 0 to 1 (1

being the maximum emphasis) are specified by the user. Accordingly, the prosodic features of audio including intensity, duration and pitch are modified to produce expressive audio. The pseudo code of lip syncing is presented below. The pseudo code of Phoneme lip syncing method is present in algorithm 3.

Algorithm 3: Pseudo code of lip syncing.

```
Input    Audio file,
Input    visVec, visTargetSize, BaseMesh, FPofLip
Output FAP
read (file.wav)
CSLU toolkit     /* application to calculate the time table of each phoneme PhonVec */
For k= 0; k< NoPhon; k++)        /* calculate Duration for every Phoneme */
 Timetable[k]= getDurationPhone(Phone[k]);
     For i=0; i<=visVec; i++)          /* convert phoneme to viseme mapping */
     {
        For (x=0; x<BaseMesh; x ++)
             For (y=0; y<BaseMesh; y ++)
                  For (z=0; z<BaseMesh; z ++)
                      {      Pos[x, y, z]= getVertexPos(BaseMesh, vertex);
                             Val[x, y, z]= getVertexVal(BaseMesh, vertex);
                       }
                      PCA (Val, PC)
                      For (target=0; target < visTargetSize(); target ++)
                     {
                        For (n=0; n<10; n++)
                        {   If weight( PC[n] )==0.0) and (PC[n] ∈ FPofLip ) then
                           {    weight [PC]=3
                                newPos[x, y, z]= Pos [x, y, z]+(targetpos* weight[PC[n]]);
                                setVertexPosition (BaseMesh, PC, newPos);
                            }
                           RCF (weight, PC) /* calculate Raised Cosine Function */
                         }
                      }
                      Stream_FAP( PC, newPos, weight );
                      SaveFAPFile (PC); }
               }
```

3.3 Hybrid Facial Expression Synthesis Method

Integrating of facial expressions and eye behaviours increases the realism of the proposed 3D face model. The main difficulty in combining the speech and emotions in a visual talking is caused in handling the interaction between emotional expression and articulation in the orofacial region. These shortcomings are overcome via the automatic and interactive translation of emotion to MPEG-4 FA via Plutchik's emotion wheel method. Plutchik's wheel is used for the intermediate expressions to generate six basic emotions. Six universally recognizable primary expressions are used [31]. Fig. 20 illustrates the fundamentals of Plutchik's emotion wheel (in space dimensions) for converting the emotions in MPEG-4 FA configuration (Raouzaiou et al., 2002).

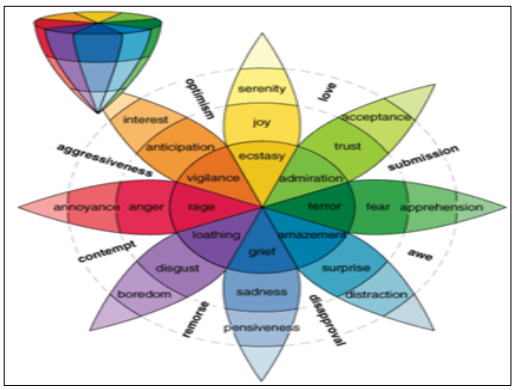


Fig. 20. Plutchik's model showing the relations among emotion concepts.

Selection of an emotion is as easy as picking a point in two dimensional circular spaces. The choice of emotion depends on the distance from the center point and the angle. Greater the distance is, the less intense is the emotion and vice versa. There exists eight basic and eight advanced emotions in between the basic emotions. Each of these emotions possesses an opposite sentiment, where some of them are not straightforward. Fig. 21 schematically illustrates the calculation procedure of FAPs for intermediate expressions. Denote P_A as the location of the archetypal emotion and P_I the location for which the new profile is calculated. Let P_i be a profile of emotion i and $X_{i,j}$ be the range of variation of FAP F_i involvingP_i. Assume A and I are the emotions belonging to the same universal emotion category.

Then, A being the archetypal and I the intermediate one, the following rules are applied (Raouzaiou, N. Tsapatsoulis, K. Karpouzis, 2002).

i. P_A and P_I employ the same FAPs.

ii. a_A and a_I are the values of the activation parameter for emotion words A and I.

iii. The range of variation $X_{i,j}$ is computed through $X_{i,j} = \left(\frac{a_I}{a_A}\right) X_{i,j}$.

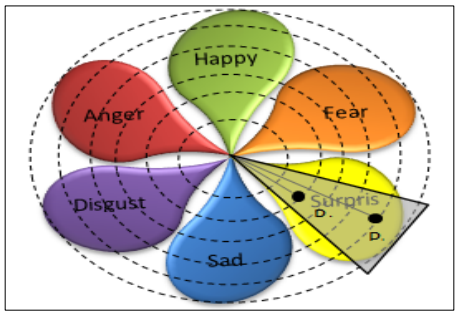


Fig. 21. Archetypal emotion of modified Plutchik's wheel.

The range is calculated using the ratio between the activation of P_A and P_I. This is performed for all FAPs contained in the selected profile. For example, take FAP 3 (open jaw) as of one of the surprise-profiles. It has range [569, 1201] with activation of P_A of 6.5 and P_I of 4.0. Therefore, the new range is given by,

$[569 * (4.0/6.5), 1201 * (4.0/6.5)] = [350, 739]$

Fig. 22 shows the diagram of the facial expressions method. The role of the facial expressions method is to produce a facial expression that is predefined as a set of FAP values in the emotional database. This is stored in the database representing the six basic emotions. The difference is just the value of emotion between 0 and 5 because there are 6 basic emotions in Ekman model. The following is the pseudo code of modified Plutchik's wheel.

Algorithm 4: The pseudo code of modified Plutchik's wheel.

Input activation parameters $\boldsymbol{a_{A1}, a_{A2,}, and\ a_I}$

Input the range of emotion $\boldsymbol{[X_{A1j}, X_{A2j}]}$

Input the angular parameters for emotion $\boldsymbol{\omega_{A1}, \omega_{A2}\ and\ I}$ where $\boldsymbol{\omega_{A1} \leq \omega_I \leq \omega_{A2}}$

Set parameters iteration (default= 360), FAPs (default =68), FP (default =47).

Output the center and the length $\boldsymbol{c_{A1,j}\ and\ s_{A1,j}}$ of $\boldsymbol{X_{A1,j}}$ and $\boldsymbol{c_{A2,j}\ and\ s_{A2,j}}$of $\boldsymbol{X_{A2,j}}$, and the new range of $\boldsymbol{[X_{I1j},\ X_{I2j}]}$

For k=1: FAPs

For j=1: FPs

For i=1: iteration

If $FAP_j \in P_{A1}$ and P_{A2}, // $X_{I,j}$ is computed as a weighted translation of $X_{A1,j}$ and $X_{A2,j}$ (where $X_{A1,j}$ and $X_{A2,j}$ //are the ranges of variation of FAP F_j involved in P_{A1} and P_{A2})

Begin

Compute the translated range of variations by Eq. (4.6)

Compute the length of $X_{I,j}$ is and its midpoint by Eq. (4.7)and (4.8)

End

Else If $FAP_j \in P_{A1}$ and P_{A2} // but with a contradictory sign (opposite direction of movement),

Begin

Compute the translated range of variation $X_{I,j}$ by Eq. (4.9)

Compute the length of $X_{I,j}$ is and its midpoint by Eq. (4.7)and (4.8)

End

Else $FAP_j \in P_{A1}$ or P_{A2}

Compute the translated range of variations by Eq. (4.10)

End for

End for

End for

Display (length $\boldsymbol{c_{A1,j}}$ $\boldsymbol{and}$ $\boldsymbol{s_{A1,j}}$ of $\boldsymbol{X_{A1,j}}$ and $\boldsymbol{c_{A2,j}}$ $\boldsymbol{and}$ $\boldsymbol{s_{A2,j}}$ of $\boldsymbol{X_{A2,j}}$, and the new range of $[\boldsymbol{X_{I1j}}, \boldsymbol{X_{I2j}}]$)

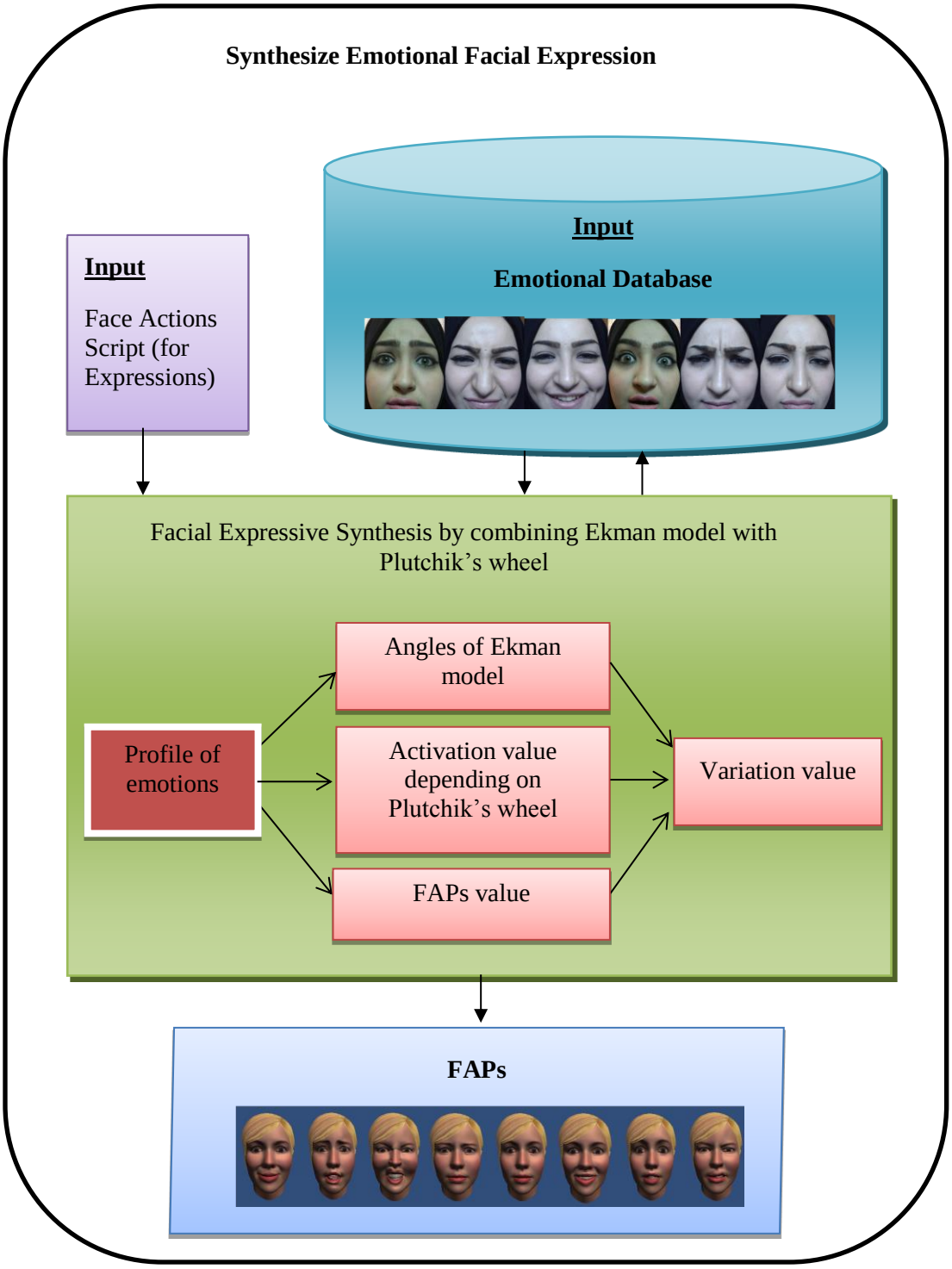


Fig. 22. Schematic block diagram of hybrid facial expressions synthesis.

3.4 Implement Eye Movements Method

In conversation the eyes normally scan the face of the other person in order to pick up any emotional expressions which can be interpreted. Requiring actions such as what to ask, how to behave and how to carry the conversation forward play significant role for such interpretation [33] [34] [6], [35]. Eyes blinking are extremely important to make highly realistic conversation with a virtual human. Continuous blinking of eyes at regular time intervals is essential for this purpose. Eye behaviour provides a set of features that can be combined with different affective states. The method uses a statistical model called the default model (Lee et al., 1995) as a saccadic eye movement engine and creates differentiated expressive and interactive behaviours (behavioural database) through changes in gaze parameters, such as direction, magnitude, and interval between movements. Saccades are fast motions of eyes from one gaze point to another (S Zhang et al., 2010). Consequently, in SMIL scripts eye behaviours are described as high-level actions. When combined with facial expressions, they contribute to enhanced expressiveness and engage in communication. The expressive gaze generator receives descriptions of facial actions and returns the FAP values for each gaze generated according to the specified eye behaviour. It also provides the movements of head and eyelid according to the rules for the model. Similarly, the eye behaviour sequence can be described to the facial expression by the following example. The set of eye behaviour descriptions and its parameters are the same as the Gaze Description Language (GDL) (Queiroz et al., 2009). Fig. 23 displays the eye behavior pattern.

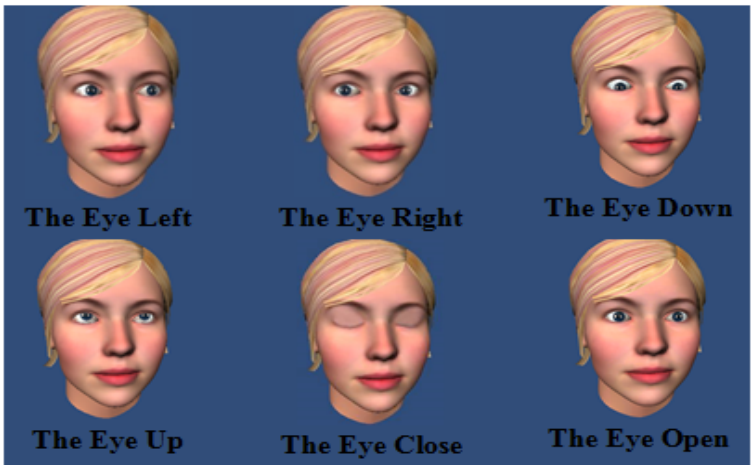


Fig. 23. Eye behaviors of the proposed model.

The first factor is type of looking, that detect which one from these types (lookTo and default), the second is the direction of the looking (up left, left, up, down left, right, down right, up right, down, talking), and the third is the angle of the direction lastly the rotation and duration amount. Example below shows the details on eye behaviours. Fig. 24 shows the framework of the eye movements of EEMML.

```
Eyes = {{"lookTo","upleft",10.0,"yes",50},
{"lookTo","left",17.0,"yes",50},
{"lookTo","up",15.0,"no",50},
{"lookTo","downleft",12.0,"yes",50},
{"lookTo","right",15.0,"yes",50},
{"lookTo","downright",12.0,"no",50},
{"lookTo","upright",7.0,"yes",50},
{"lookTo","down",12.0,"yes",50},
{"default", "base", 0.9, 50}}
```

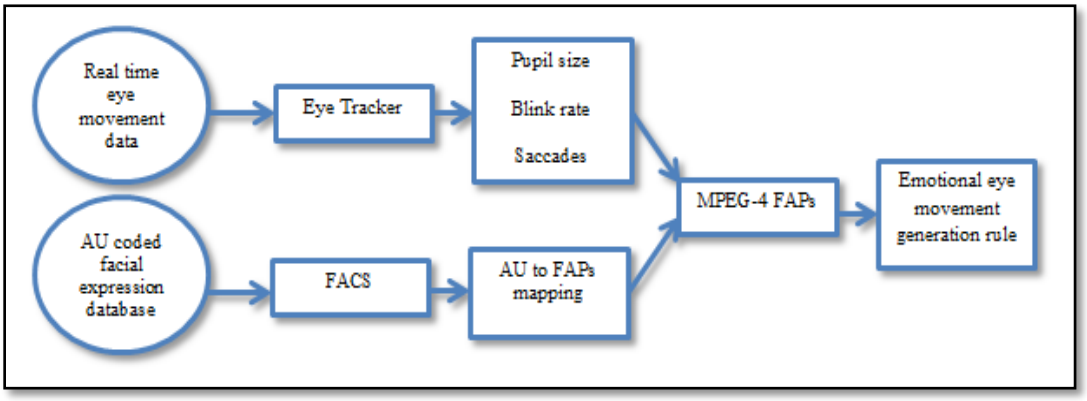


Fig. 24.The eye movements framework of EEMML (Li and Mao, 2011)

3.5 Synthesis Face Motion Using MPEG-4 Approach

FAPs are used to encode low-bandwidth transmission in broadcast (one-to-many) or dedicated interactive (point-to-point) communications. They manipulate feature control points on a mesh model of the face to produce animated visemes (visual counter part of phonemes) for the mouth (lips, tongue, and teeth), animated head and facial features including the eyes or eyebrows. The proposed system used a total of 47 FPs, in which 16 FPs are used for lips, 23 for expressions with eyes, 4 for teeth and 4 for tongue (Fig. 25 and 26). All the FAP parameters involving translational movement are expressed in terms of FAPUs. They correspond to fractions of distances between some essential facial features (e.g. eye distance). The fractional units are considered to ensure adequate accuracy.

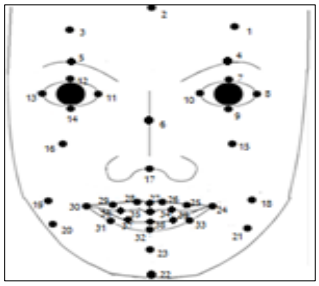


Fig. 25. Total 39 selected FPs for the face used in the proposed system.

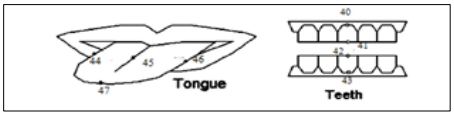


Fig. 26. Total 8 selected FPs for the tongue and teeth incorporated in the proposed system.

Now we focus on the synthesis of facial motion via MPEG-4 approach. The 3D model developed in the previous step is imported to database for creating the animation according to the MPEG4 standards. This is crucial because the weights, deformation functions, and the FAP's are defined in this step. A realistic facial animation is produced by modifying and fine tuning these parameters. The information about the 3D model is stored under *.fdp extension. The first step defines the FAPU, where seven points are selected on the face model from the 33 FPs. Selections

of these seven points on the face model are significant because the animations are carried out using the distances defined by these points.

3.6 Face Interpolation

The present system expects a sequence of English phonemes and timings for each phoneme. First, the platform groups the sequence of phonemes into phoneme pairs and then maps those to the phoneme set. Assuming the input phoneme schedules $p_0, p_1, \dots, p_n$ occur at time $t_0, t_1, \dots, t_n$, a sequence of phoneme pairs consisting of adjacent pairs of phonemes $(p_0, p_1), (p_1, p_2), \dots (p_{n-2}, p_{n-1})$ is constructed with their corresponding time span $(t_0, t_2), (t_1, t_3), \dots, (t_{n-2}, t_n)$. A smoothing pass over is performed on each pose using a user specified window to scan through temporal domain and final local maximum. A sliding window of size $2t_w$ is selected. The smoothing window is slide over the pose $p^k(t)$ from t_0 to t_n for detecting the local maximums for $p^k(t)$. If at any time instant t there are two or more local maximum $p^k(t_a)$ and $p^k(t_b)$ inside the window from $t - t_w \; to \; t + t_w$, the pose is smoothen out with a value between t_a and t_b by interpolating these two maximum. Specifically, since $p^k(t_a)$ and $p^k(t_b)$ are the values of spline curves associated with piecewise linear curves whose values are $L^k(t_a)$ and $L^k(t_b)$ at time t_a and t_b, the new spline curve is obtained by linearly interpolating $L^k(t_a)$ and $L^k(t_b)$ via,

$$L^k(t) = \frac{(t_b - t)L^k(t_a) + (t - t_a)L^k(t_b)}{t_b - t_a} \qquad (16)$$

Intuitively, this process smoothed out the valley between two maximums and helps to reduce the high frequency lip movements that are not natural for speech animation. Rendering of the animation is the next important step.

The method of expressions in the 3D talking head exhibits the action. The 3D face model displays dynamic emotions based on the transitions between the six basic emotions. Six emotions that are created for the proposed system include: sad, happy, afraid, disgust, anger and surprise. Modelling the emotions is another challenge since it involves the change in different parts of the face such as eyes, nose, eyebrows, lips, chin and cheeks. It is evident that all the six emotions are distinct from each other.

4. Results and Evaluation

The assessment is performed via two stages including system and user testing. System testing is used to examine the authenticity of the independent function of lip syncing, eye movements, and emotional facial expression as well as while working simultaneously with the performance of interactive integration. The consumer testing of the integrated system performance is performed to convey the feedback from the users and to make a comparison with the other virtual human character. Comparing the displacements of the proposed system with displacements of other 3D face model systems that have implemented MPEG-4 FA (for the same FAP-values) the quality of the implementation of MPEG-4 FA and the used parameters sets are evaluated. The proposed system is not only compared with Xface but also with Greta and Lucia. The average of the displacements of all of these faces approached a common ground truth. Since the face is a very important interface to humans, assessment of more realistic displacement is quite universal.

The process of evaluation consists of creating screenshots of all faces for all FAPs. The value chosen is approximate to move the FPs halfway to the FAPU. For bidirectional FAPs, another screenshot is taken with its value negated. Since differences between a certain pose and the neutral face are sometimes hard to spot, difference images are calculated. Because these difference images only have differences in them, it is sometimes hard to determine where and to what extent exactly this difference in the face occurred. For this, the image of the neutral face is blurred and placed with 30% intensity as layer under the differences. Screenshots and difference images with blurred underlays are viewed to each other so that for all FAPs, the displacements of all four faces on these FAPs can easily be assessed in relation to each other. A score is assigned to each displacement, bidirectional FAPs, and every direction.

The objective evaluation of the framework focuses on the assessment result of existing interactive facial animation with emotional facial expressions, eye behavior, and lip syncing. The estimated displacements of the facial feature points are reconstructed. The displacement (both the ground truth and the synthetic talking) of each facial feature point is divided through its maximum absolute disarticulation in the collected audio-visual database. The displacement is further normalized to [-1.0, 1.0]. The performance is evaluated by calculating the Pearson

product-moment correlation coefficients (R), the average standard deviations, and the mean square errors (MSEs) using the normalized data. The Pearson product-moment correlation coefficient (CC) measures the superiority of the global match between the shapes of two signal sequences [36]. The CC is calculated using the expression,

$$CC = \frac{trace(Cov_{\vec{d}\,\overrightarrow{d'}})}{\sqrt{trace\left(Cov_{\vec{d}\,\overrightarrow{d'}}\right) trace(Cov_{\overrightarrow{d'}\,\overrightarrow{d'}}}} \qquad (17)$$

where $\vec{d}$ is the normalized ground truth, $\overrightarrow{d'}$ is the normalized estimated result

The average standard deviations is computed from,

$$v_{\vec{d}} = \frac{\sum_{c=1}^{\gamma}(Cov_{\vec{d}\vec{d}}[c][c])^{1/2}}{\gamma} \qquad (18)$$

$$v_{\overrightarrow{d'}} = \frac{\sum_{c=1}^{\gamma}(Cov_{\overrightarrow{d'}\overrightarrow{d'}}[c][c])^{1/2}}{\gamma} \qquad (19)$$

where γ is the dimension of $\vec{d}$.

The Mean Square Error (MSE) is used as the objective measure to evaluate the performance of realism associated with lip syncing. The average MSE of synthesized movements are obtained from,

$$MSE = E\left[\frac{\left\|\vec{d}-\overrightarrow{d'}\right\|^2}{\gamma}\right] \qquad (20)$$

The configuration of the 33 FP constitutes a shape. However, the interest here is not centred towards the shape of the face itself, but rather on the mechanism of shape change during facial movements. It is well known that the facial shapes differ from person to person. The factors that differentiate persons with retro gnathic jaws from those with the normal one are also identified. Main focus of the thesis is on the facial motion (independent of the shape) especially, the derivative of the shape as it alters over time. The relative change measure is used to analyse the modifications in motional parameters.

Denoting $d_{ij}(t)$ as the Euclidean distance between markers i and j at time, where $i, j = 1..n$ and n is the number of parameters, the relative change in distance from the rest is expressed as,

$$r_{ij}(t) = \frac{d_{ij}(t)}{d_{ij}(0)} - 1 \qquad (21)$$

For continuous speech, it is observed that the principle articulator of most consonants in the articulation fails to reach its target position due to the lack of sufficient time. The above definition does not guarantee that each phoneme in continuous speech is capable of reaching its target value. However, only the dominant phoneme of articulation possessing both sufficient time and largest displacement can reach the target. The synthesized curves using different articulatory models are compared to the recorded movements to examine the performance. The subjective evaluation is made by comparing the user defined one with that of existing virtual human character system. The result of subjective evaluation is used as supporting appraisal of the framework. Comparison is also made for objective evaluation to confirm whether each element in the framework has fulfilled the objectives. Firstly, the audio streams of one minimal pair are played in the test. Then, the animations are shown in which two words of one minimal pair is appeared in a random order. The subjects are asked to identify which animation corresponded to the word. The identification accuracy refers to the ratio of the number of correctly recognized animations and the total number of animations. All the stimuli are presented in two conditions including the visualization with front face (F), and the visualization with both transparent face and tongue (FT). The subjects are found to score the degree of realism of the animation, when the correct label of the animation is displayed to them. The score ranges from 1 to 5 with 1 for bad, 2 for poor, 3 is fair, 4 for good and 5 for excellent.

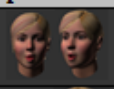
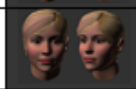
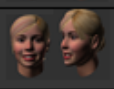
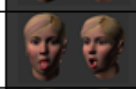
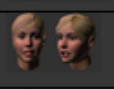
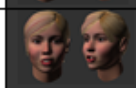
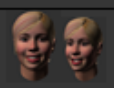
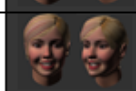
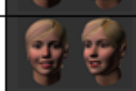
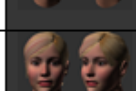
The achieved results are directly compared with popular commercial lip syncing solution such as Xface, Lucia, and Greta. The experiment used the same character and same phoneme scheduler. The phone bigram animations is constructed depending on the set of FaceFX [37] static poses. When producing results from the developed algorithm, the FaceFX phonemes are mapped to the proposed Common Phoneme set. Those virtual characters are compared with the

proposed system in terms of their performance parameters. The limitations of the previous work are highlighted and advantages of the newly proposed approach are demonstrated.

4.1 The result of Prophone Lip Syncing Method

The articulation of a phoneme determines the degree of emphasis received by the visual phoneme together with the duration as shown in Table 2. The first stage in the present system is the input of neutral audio or text that created expressive speech. Fig. 27 illustrates the English visual phoneme articulation with normal expression for the sentence “the natural beauty”, with (a) in frontal side and (b) in diagonal sideways.

Table 2. The Visual Phoneme.

Viseme No.	Phoneme	Visual phoneme	Viseme No.	Phoneme	Visual phoneme
1	Phoneme ah		8	Phoneme oh	
2	Phoneme b, m, p		9	Phoneme D, S, T	
3	Phoneme Big aa		10	Phoneme h	
4	Phoneme Ch, J, Sh		11	Phoneme K	
5	Phoneme ee		12	Phoneme N	
6	Phoneme i		13	Phoneme R	
7	Phoneme V, F		14	Phoneme Th	
16	Silent		15	Phoneme W	

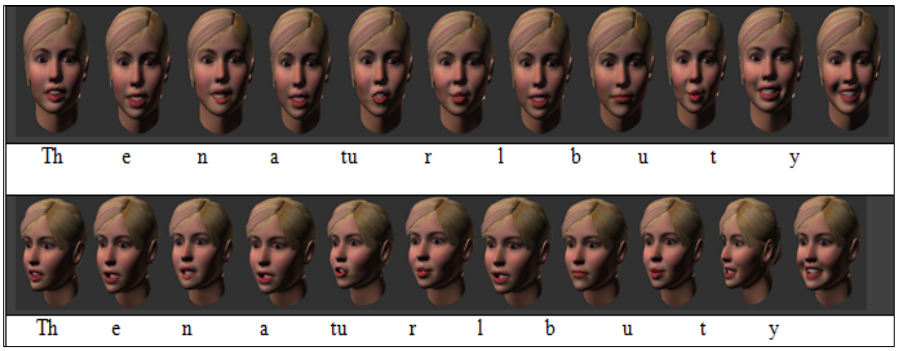


Fig. 27. English visual phoneme articulation with normal expression for the sentence "the natural beauty", in frontal side and diagonal sideways view.

4.2 The Result of hybrid Facial Expression Method

Fig. 28 displays the results of the various emotional facial expressions method that generates the appearances of those pre-defined as a set of FAP values in the emotional database.

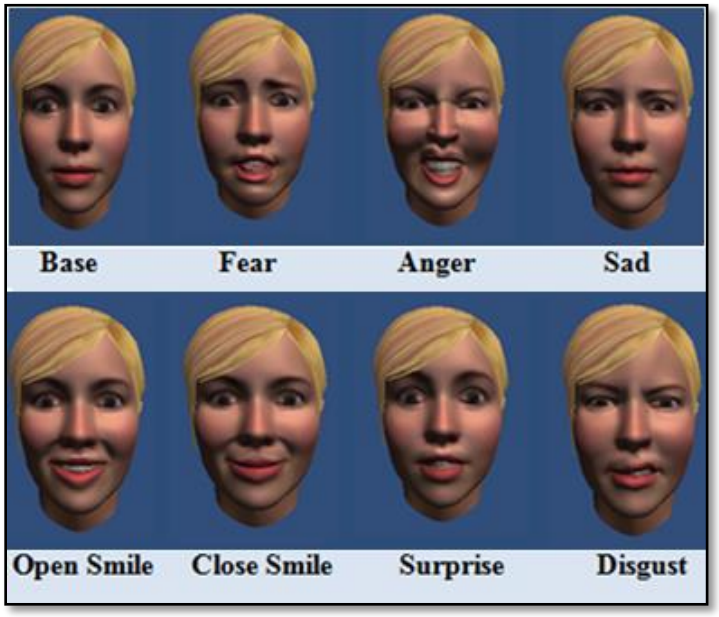


Fig. 28. The result of hybrid facial expression method for six basic emotions.

4.3 Eye Blinks

The outcome of the varying eye behavior as shown in Fig. 29 provides a set of eye blinks or saccades that is combined with different affective states to create the realistic facial expression.

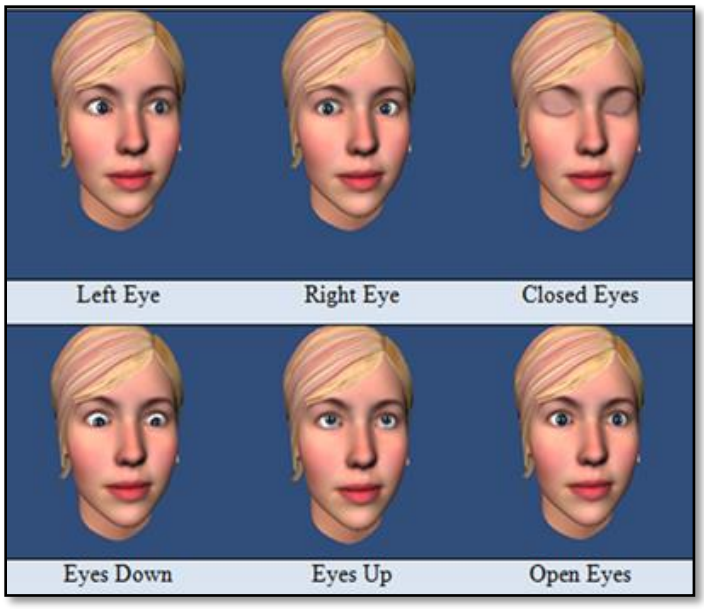


Fig. 29. Result of the facial expression in the presence of eye behaviors.

Fig. 30 exhibits 20 frames for the English phonemes with angry emotion for the sentence "I wanna discuss all this behavior".

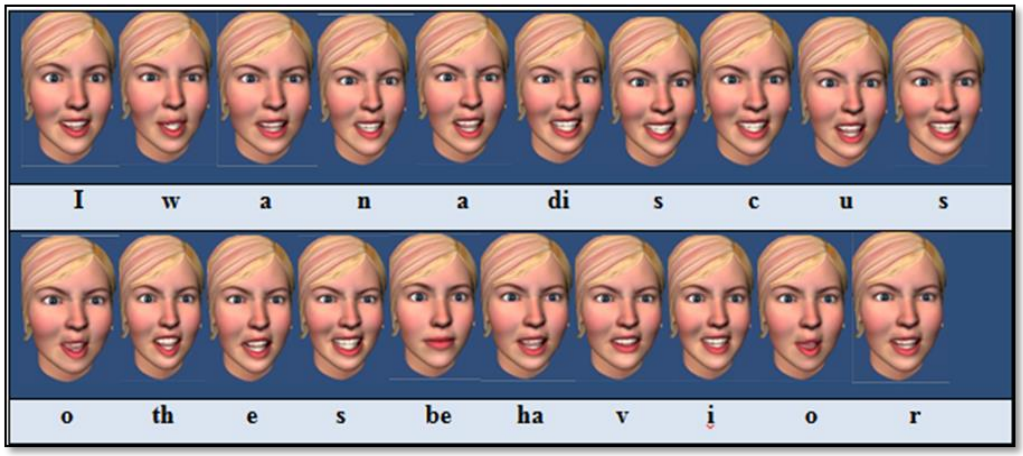


Fig. 30. 20 frames for the English phonemes with angry emotion for the sentence "I wanna discuss all this behavior".

Fig. 31 demonstrates the important 8 frames for the English phonemes with surprise emotion for the word "unbelievable".

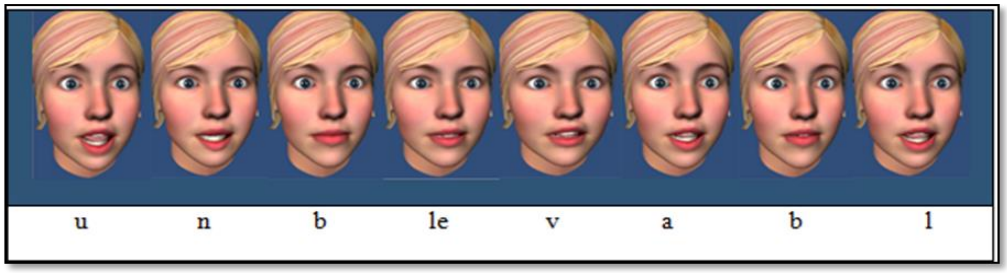


Fig. 31. Significant 8 frames for the English phonemes with surprise emotion for the word "un-believable".

Meanwhile, Fig. 32 reveals the significant 9 frames for the English phonemes with happy emotion and eye blink for the sentence "haha you are so funny."

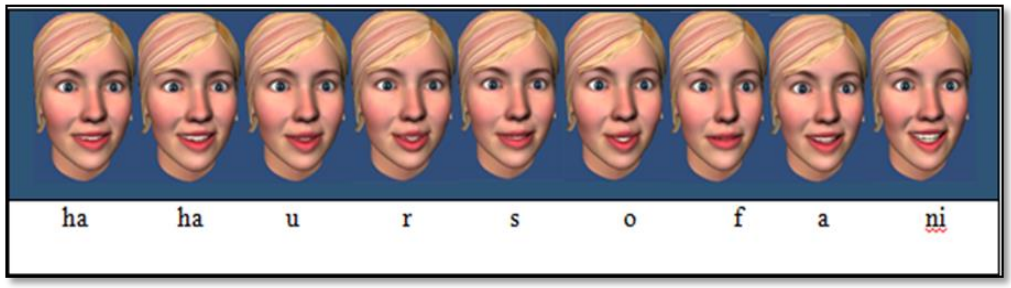


Fig. 32. Important 9 frames for the English phonemes with happy emotion and eye blink for the sentence "you are so funny."

The contribution of this work is generates intermediate expression from six basic emotions. The result of the hybrid facial expression method is shown in Table .3. These expressions are generated from combing the expression depending on the Plutchik's emotions wheel and Ekman emotions model. In another words the proposed model has the anger expression and at the same time has fear expression. etc.), with percentage 100% for primary expression and 50% for the other primary expression. This Table shows the contribution of this method.

Table 3: Expression generated from the hybrid facial expression method

	Anger 50%	Disgust 50%	Sad 50%	Surprise 50%	Happy 50%	Fear 50%
Anger 100%						
Disgust 100%						
Sad 100%						
Surprise 100%						
Happy 100%						
Fear 100%						

4.4 Animation Results

A better comparison is made by recording the real speech in both head-on and profile view at roughly 45 degree angle without employing highly expensive hardware and multiple cameras. The real speech is digitized and the mouth area is cropped from the frames. The synthetic speech

is also rendered using the same two views and also cropped. The two cropped images are then combined into a single frame resulting in an animation that showed both synthetic and real speech in rough synchronization from two separate views.

Few things are mandatory in scrutinizing the frames. First, the synthetic speech must not utilize prosodic information so the difference in emphasis and pitch causes dissimilar target positions. Besides, the real speech should be filmed with the subject repeating the phrase while listening to the synthetic speech. This is carried out to obtain comparable timing and analogous prosodic features to the speech. Consequently, the mouth position in the beginning for real speech is not set in a neutral position (closed mouth). Note that the timings of the phrases are very similar but they are not exactly identical. Actually, a difference of only 0.03 seconds caused the movement of an entire frame afterwards. Nevertheless, the model may indeed look similar, but it is not meant to be an exact match. For instance, the lips are not shaped exactly the same way as that of the tongue for the Visible Human Female and the subject. Fig. 33 illustrates a comparison between the real human and proposed system. Table 4 summarizes the screenshot of the multiple type of expressive speech for the real human and the proposed system displaying a comparative measure.

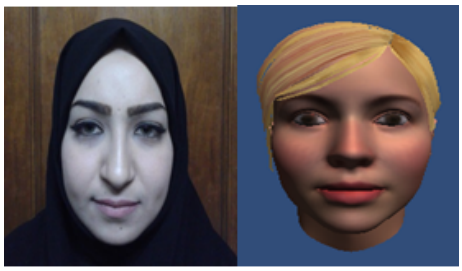
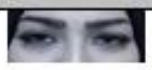
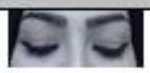
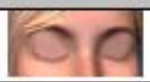
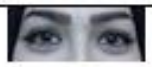
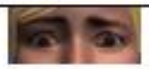
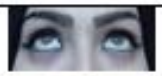
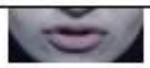
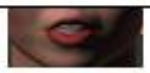
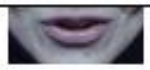
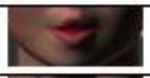
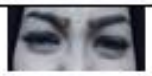
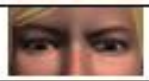
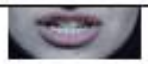
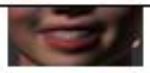
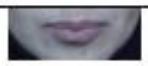
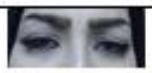
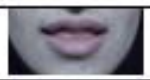
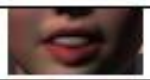
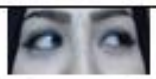
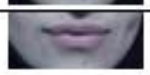
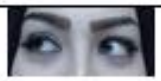
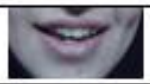
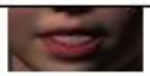


Fig. 33. The real human compared with the proposed system.

Table 4. The screenshot of the multiple type of expressive speech for the real human and the proposed system.

Expression / Viseme	Screenshot of the Real Human	Screenshot of the Proposed System	Expression/ Viseme	Screenshot of the Real Human	Screenshot of the Proposed System
Base			Move Eye down		
Fear			Move Eye up		
Angry			Ch		
Surprise			O		
Disgust			S		
Happy			M		
Sad			Th		
Stare Eye left			V		
Stare Eye right			Aa		

4.5 Objective Evaluation

The synthesis performance objectivity is evaluated following tests with different system parameters. The prediction performance of the proposed system is measured by using three metrics on the test set. Thus, the Pearson product-moment correlation coefficients between the ground truth and the estimated results are calculated. The Pearson product-moment correlation coefficient of the training set is found to be 0.98 without any expression, 0.972 with smile expression, and 0.977 for sad expression. Conversely, the Pearson coefficients of the testing set are discerned to be 0.968 in the absence of any expression, 0.945 for smile expression, and 0.942 with sad expression. The normalized MSEs of the training set are observed to be 0.0025 without any expression, 0.0031 for smile expression, and 0.0033 for sad expression and the corresponding values for the testing set are determined to be 0.0029, 0.0034, and 0.0037, respectively.

The evaluation is performed using Xface, which is 3D facial animation developed in the 5FP IST Project PF-Star coordinated by ITC-irst. It is a open source code that relies on MPEG-4 FA standard [17] with the inclusion of screenshots. These screenshots are compared with the performance of the proposed system. In addition, three other faces as depicted in Fig. 34 are used to evaluate the proposed system.

Fig. 34. The 3D facial animation engine used to evaluate the proposed system. From left to right: Lucia, Xface, and Greta.

Fig. 35, Fig. 36, Fig. 37, Fig. 38, Fig. 39 and Fig. 40 displays the results for the relative change for six basic expressions to the upper lip. The relative change represents the distance from the rest to the target parameters. The desirable properties in relative change are that they are not dependent on local shape, for example consider the distance between FP 8.3 and FP 8.4. Some of people have bigger mouths than others but the relative change concerns the distance changes during speaking and smiling.

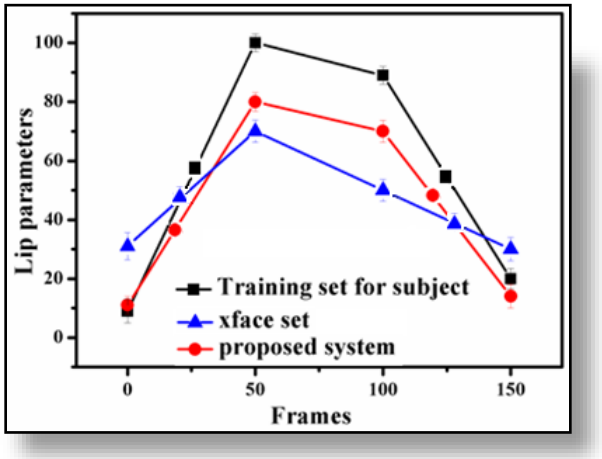


Fig.35. The relative change for happy expression to the upper lip

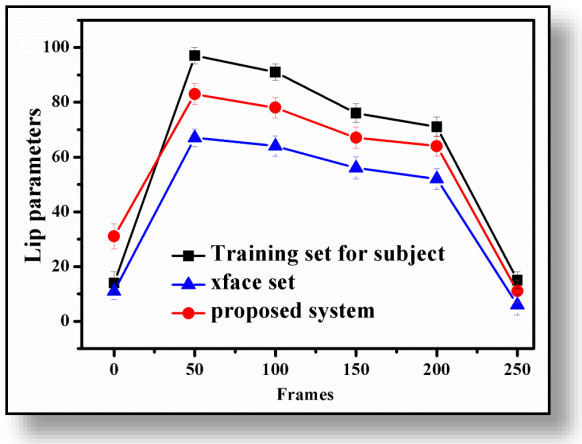


Fig. 36. The relative change for anger expression to the upper lip

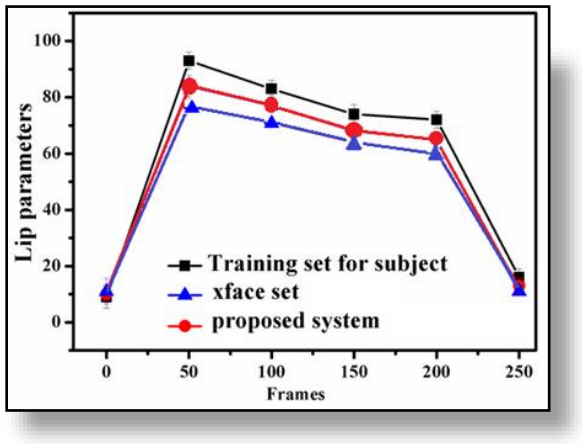


Fig. 37. The relative change for sad expression to the upper lip

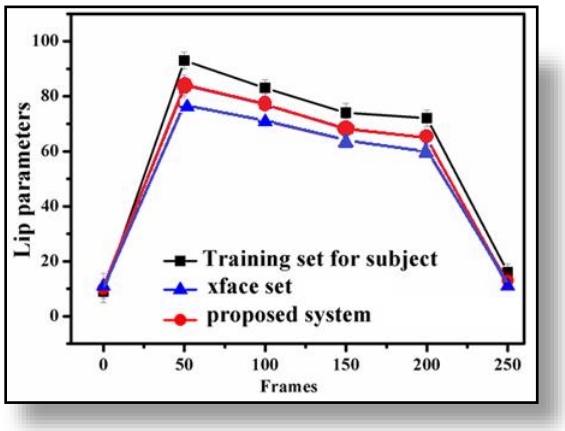


Fig. 38. The relative change for disgust expression to the upper lip

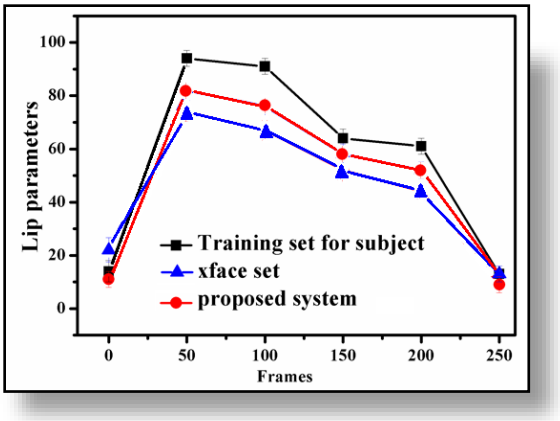


Fig. 39. The relative change for fear expression to the upper lip

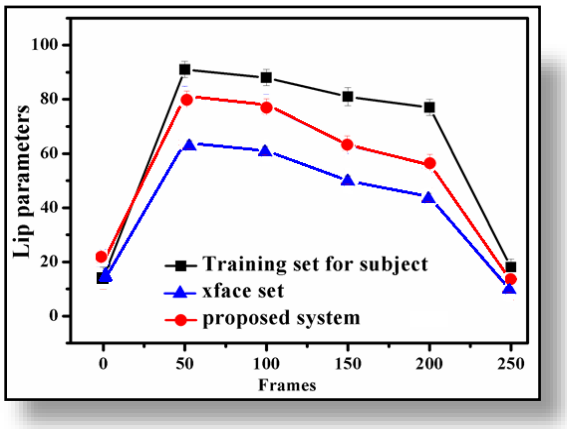


Fig. 40. The relative change for surprise expression to the upper lip

These figures represent the relative change of the distance on the upper lip when applying the six basic emotions on the 3D face model. As can be concluded from these figures, the motion of parameters is smooth, not unambiguous and not confronted with unexpected results. Tables 5, Table 6 and Table 7 summarize the result of CC, standard deviation, and MSE.

Table 5. Evaluation results for Xface

Evaluation Process	Xface						
	Neutral	Joy	Anger	Fear	Surprise	Disgust	Sad
CC	0.990	0.982	0.980	0.977	0.989	0.986	0.960
$v_{\overline{d}}$	0.200	0.215	0.219	0.212	0.220	0.211	0.200
$v_{\overline{d}'}$	0.180	0.210	0.198	0.189	0.213	0.192	0.190
MSE	0.0024	0.0041	0.0031	0.0056	0.0049	0.0034	0.0020

Table 6. Evaluation results for GRETA

Evaluation Process	GRETA						
	Neutral	Joy	Anger	Fear	Surprise	Disgust	Sad
CC	0.996	0.990	0.987	0.980	0.995	0.992	0.990
$v_{\overline{d}}$	0.198	0.200	0.210	0.162	0.190	0.199	0.187
$v_{\overline{d}'}$	0.174	0.198	0.166	0.173	0.200	0.162	0.180
MSE	0.0034	0.0045	0.0057	0.0063	0.0058	0.0041	0.0045

Table 7. Evaluation results for the proposed 3D face model

Evaluation Process	The proposed 3D Face Model						
	Neutral	Joy	Anger	Fear	Surprise	Disgust	Sad
CC	0.980	0.960	0.940	0.963	0.970	0.971	0.970
$v_{\overline{d}}$	0.190	0.220	0.239	0.225	0.248	0.220	0.198
$v_{\overline{d}'}$	0.194	0.216	0.199	0.214	0.219	0.197	0.200
MSE	0.0020	0.0024	0.0020	0.0030	0.0026	0.0010	0.0030

For the Xface, the CC is discerned to be 0.98 for neutral, 0.972 for happy, 0.977 for angry, 0.977 for fear, 0.989 for surprise, 0.960 for sad, and 0.986 for disgust expressions. For GRETA, the CC is 0.996 for neutral, 0.990 for joy, 0.987 for anger, 0.980 for fear, 0.995 for surprise, 0.992 for disgust and 0.990 for sad. Conversely, the CCs of the proposed 3D face model for the neutral, happy, angry, fear, surprise, sad, and Disgust expressions are observed to be 0.968, 0.945, and 0.942, 0.963, 0.971, 0.970 and 0.971 respectively. Form these results can see that the proposed model has correlation coefficient smaller than the Xface and GRETA which

means the influence of joy, anger, disgust, surprise and fear of the proposed model is slightly stronger than that of sad expression.

For Xface, the MSE are found to be 0.0024, 0.0041, 0.0031, 0.0056, 0.0049 and 0.0034; for GRETA, the MSE are 0.0034, 0.0045, 0.0057, 0.0063, 0.0058 and 0.0041 for the neutral, happy, angry, fear, surprise and disgust expressions, respectively and the corresponding values for the proposed model are determined to be 0.0020, 0.0024, 0.0020, 0.0030, 0.0026 and 0.0010 respectively. From these results we can see that the proposed model has MSE smaller than the Xface and GRETA which means the error in the displacement motion of the parameters of joy, anger, disgust, surprise and fear of the proposed model is slightly stronger than that of sad expression. From the result can see Xface is better at representing sad expression than the proposed model and GRETA.

Fig. 41 shows the new archetypal profiles displayed by the Lucia, Xface, Greta, and the proposed system for male and female subjects. The expression including joy, sadness, anger, fear, disgust and surprise are illustrated.

Fig. 41. The new archetypal profiles displayed in Real Human, Lucia, Xface, Greta, and the proposed system. Expressions for both subjects (male and female) from left to right: happy, sad, angry, fear, disgust, and surprise.

The function of expressive speech data-driven facial animation is created using the proposed method. Six sets of animated pose sequences are generated, where five audio files conveying

neutral/none, positive, and negative information are developed. Each set of the pose animation sequences consist of six sequences including the expressions on neutral, smile, surprise, angry, happy, disgust and sad. Human subjects are asked to infer the emotion states from the face animation sequences or the videos of the real face while listening to the corresponding audio tracks. The experimental results demonstrated that the created synthetic talking face can effectively contribute to the bimodal human emotion perception and its effects are comparable with a real talking face.

5. Conclusion

The aim of the present work was to create an effective animation scheme connected to the co-articulation and realistic expressions. Therefore, the core processes of this research are four: 3D Face Model, Lip Synchronization, Facial Expression, and Eye behaviour. All these processes are combined together to get realistic expressive visual speech virtual human character. Animating the realistic facial expression is the most challenging aspect of computer modelling and architecture. Several factors such as eye movements, personalities, weight of the emotion, types of emotion, and lip motion must be taken into considerations to animate facial expressions and to formulate the phonemes. Actually, this is a subject of 3D modellers', graphical designers, and cartoon animator. A systematic research method is adopted to accomplish these perspectives. The system design and implementation is focussed. System testing and validation is underscored. The development of a new framework of interactive facial animation with emotional facial expressions, synchronized speech, and generation of eye behaviour is highlighted. Furthermore, the eye behaviour is subdivided into four modules including eye behaviour module, emotional facial expression, lip synchronization module, and interactive module. The realistic virtual human facial expressions are generated through the MPEG4 facial animation standard together with SMIL. It allows the creation of a variety of expressive speech and describes in details the formulation of the lip motion with each English phoneme. An efficient animation framework is created by integrating the existing facial animation models. The proposed system possesses several notable advantages including:

i) Cheap production and storage costs.

ii) Only video that is viewed need be created.
iii) Updates are immediately effective.
iv) Multiple characters can be used.
v) Multiple languages can be used.
vi) Delivery of the information can easily be modified using faster or slower speech.

The techniques involved in the generation of emotions based on the Plutchik's model for emotions are introduced. This new system models and generates virtual talking head (avatar) similar to humans. The experimental analysis clearly demonstrated the efficiency, accuracy, and effectiveness of the proposed animation system for expressive visual speech synthesis of avatar using MPEG-4 approach.

References

[1] H. Kolivand and M. S. Sunar. A Survey of Shadow Volume Algorithms in Computer Graphics. IETE 2015. vol. 30. no. 1. pp. 38–46.

[2] R. B. Queiroz, M. Cohen, and S. R. Musse. An extensible framework for interactive facial animation with facial expressions, lip synchronization and eye behavior. Comput. Entertain 2009. vol. 7. no. 4. p. 1.

[3] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. BEAT: the Behavior Expression Animation Toolkit. in Proceedings of the 28th annual conference on Computer graphics and interactive techniques 2001. vol. 137. pp. 477–486.

[4] S. Zhang, Z. Wu, H. M. Meng, and L. Cai. Facial Expression Synthesis Based on Emotion Dimensions for Affective Talking Avatar. T. Nishida. 2010, pp. 109–132.

[5] K. Balcı, M. Zancanaro, and F. Pianesi. Xface Open Source Project and SMIL-Agent Scripting Language for Creating and Animating Embodied Conversational Agents. Proc. 15th Int. Conf. Multimedia. ACM 2007. pp. 1013–1016.

[6] G. Bailly, S. Raidt, and F. Elisei. Gaze, conversational agents and face-to-face communication. Speech Commun. 2010. vol. 52. no. 6. pp. 598–612.

[7] M. Gillies, X. Pan, and M. Slater. Piavca: a framework for heterogeneous interactions with virtual characters. Virtual Real 2010. vol. 14. no. 4. pp. 221–228.

[8] Y. Lee, D. Terzopoulos, and K. Walters. Realistic modeling for facial animation. Proc. 22nd Annu. Conf. Comput. Graph. Interact. Tech. SIGGRAPH 1995. vol. 95. no. 1. pp. 55–62.

[9] J. Serra, M. Ribeiro, J. Freitas, and V. Orvalho. A Proposal for a Visual Speech Animation System. Springer-Verlag Berlin Heidelb. 2012. pp. 267–276.

[10] Singular Inversions. Facegen software.

[11] S. Pasquariello, C. Pelachaud, and S. A. Kyneste. Greta : A Simple Facial Animation Engine Facial Animation Coding in MPEG-4 Standard. Proc. 6th Online World Conf. Soft Comput. Ind. Appl 2001.

[12] L. Q. Anh and C. Pelachaud. Expressive Gesture Model for Humanoid Robot. Springer Verlag Berlin Heidelberg 2011. pp. 224–231.

[13] G. R. Leone, G. Paci, and P. Cosi. LUCIA : An Open Source 3D Expressive Avatar for Multimodal h . m . i . Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2012. pp. 193–202.

[14] K. Balcı. Xface : MPEG-4 Based Open Source Toolkit for 3D Facial Animation. Proceedings of the 15th international conference on Multimedia. ACM 2007. pp. 399–402.

[15] K. Pearson. On lines and planes of closest fit to systems of points in space. Philos. Mag. 1901. vol. 2. pp. 559–572.

[16] Igor S. Pandzic and Robert Forchheimer. MPEG-4 Facial Animation: The Standard, Implementation and Applications. New York, NY, USA: John Wiley & Sons 2003.

[17] A. Somasundaram. AUDIO-VISUAL SPEECH. The Ohio State University 2006.

[18] C. Bao. A Facial Animation System for Generating Complex Expressions. APSIPA ASC 2011.

[19] Paul Ekman. Basic Emotions. San Francisco: University of California Handbook of Cognition and Emotion 1999. p. chapter 3.

[20] and S. K. A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis. Parameterized facial expression synthesis based on mpeg-4. Eurasip J. Appl. Signal Process 2002. vol. 10. pp. 1021–1038.

[21] Y. Xu, A. W. Feng, S. Marsella, and A. Shapiro. A Practical and Configurable Lip Sync Method for Games. Proc. Motion Games - MIG 2013. pp. 109–118.

[22] TRueSpel. English-Truespel (USA Accent) Text Conversion Tool. [Online]. Available: http://www.foreignword.com/dictionary/truespel/transpel.htm 2001.

[23] Brett Kessler and Rebecca Treiman, "Syllable Structure and the Distribution of Phonemes in English Syllables," Journal of Memory and Language, 2002. [Online]. Available: http://www.artsci.wustl.edu/~bkessler/SyllStructDistPhon/CVC.html.

[24] TRueSpel. English-Truespel (USA Accent) Text Conversion Tool 2001.

[25] Sphinx Group Carnegie Mellon University. Cmu sphinx project. 2006. [Online]. Available: http://cmusphinx.sourceforge.net.

[26] Alan W. Black, Rob Clark, Korin Richmond, Simon King, Heiga Zen, Paul Taylor, and Richard Caley. The festival speech synthesis system. The festival speech synthesis system 2006. [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival .

[28] E. Kowler. Eye movements: the past 25 years. Vision Res.2011. vol. 51. no. 13. pp. 1457–83.

[29] S. D'Mello, A. Olney, C. Williams, and P. Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. Int. J. Hum. Comput. Stud. 2012. vol. 70. no. 5. pp. 377–398.

[30] Z. Li and X. Mao. EEMML: the emotional eye movement animation toolkit. Multimed. Tools Appl 2011. vol. 60. no. 1. pp. 181–201.

[31] K. Balcı. Xface : MPEG-4 Based Open Source Toolkit for 3D Facial Animation. ITCirst, Cogn. Commun. Technol 2007.

[32] B. Li, Q. Zhang, D. Zhou, and X. Wei. Facial Animation Based on Feature Points. TELKOMNIKA 2013.vol. 11. no. 3.

[33] P. Hong, Z. Wen, and T. S. Huang. Real-time speech-driven face animation with expressions using neural networks. IEEE Trans. Neural Netw. 2002. vol. 13. no. 4. pp. 916–27.

[34] FaceFX 2015. [Online]. Available: http://www.facefx.com/.

[35] S. Frantz, K. Rohr, and H. Siegfried Stiehl. Multi-Step Procedures for the Localization of 2D and 3D Point Landmarks and Automatic ROI Size Selection. Computer Vision (ECCV'98). 1998. Springer, pp. 687-703.

[36] S. Frantz, K. Rohr, and H. Siegfried Stiehl. Localization of 3D Anatomical Point Landmarks in 3D Tomographic Images Using Deformable Models. Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer-Verlag. Berlin. 2000. pp. 492-501.

[37] E. Vezzetti, F. Marcolin, V. Stola. 3D Human Face Soft Tissues Landmarking Method: An Advanced Approach. COMPUTERS IN INDUSTRY. 2013. ISSN 0166- 3615.

[38] E.Vezzetti and F. Marcolin. Geometry-based 3D face morphology analysis: soft-tissue landmark formalization. Multimedia tools and applications .2014. 895-929.

[38] A. Čereković, T. Pejša, and I. S. Pandžić. A Controller-Based Animation System for Synchronizing and Realizing Human-Like Conversational Behaviors. 2010. pp. 80–91.

[39] A. Čereković and I. S. Pandžić. Multimodal behavior realization for embodied conversational agents. Multimed. Tools Appl., vol. 54, no. 1. 2011. pp. 143–164.

[40] S. Lee, G. Carlson, S. Jones, A. Johnson, J. Leigh, and L. Renambot. Designing an Expressive Avatar of a Real Person. in Intelligent Virtual Agents, 2010, pp. 64–76.

[41] Lee, C., Lee, S., and Chin, S. Multi-layer structural wound synthesis on 3D face. Computer animation and Virtual Worlds Comp. 2011.22. pp.177–185.

[42] Taylor, S. L., Mahler, M., Theobald, B., & Matthews, I. Dynamic Units of Visual Speech. Eurographics. ACM SIGGRAPH Symposium on Computer Animation. 2012.

[43] Leuski, A., and Richmond, T. Mobile Personal Healthcare Mediated by Virtual Humans. IUI 2014 • Demonstration. 2014. pp. 21–24.

[44] Li Wei and Zhigang Deng. A Practical Model for Live Speech-Driven Lip-Sync. IEEE Computer Graphics and Applications. 2015

[45] A. Shapiro, Building a Character Animation System. LNCS 7060, Springer-Verlag Berlin Heidelberg . 2011. pp. 98–109.

[46] X. Zhao, E. Dellandréa, J. Zou, and L. Chen. A unified probabilistic framework for automatic 3D facial expression analysis based on a Bayesian belief inference and statistical feature models. Image Vis. Comput., Dec. 2012.