3D Lips Development and Measurement for Visual Speech Synthesis

Siti Salwa Salleh  
*Universiti Teknologi MARA Shah Alam, Selangor*  
E-mail: ssalwa@tmsk.uitm.edu.my

Rahmita Wirza Rahmat  
*Universiti Putra Malaysia Serdang, Selangor*  
E-mail: rahmita@fsktm.upm.edu.my

Ramlan Mahmod  
*Universiti Putra Malaysia Serdang, Selangor*  
E-mail: ramlan@fsktm.upm.edu.my

Fatimah Ahmad  
*Universiti Putra Malaysia Serdang, Selangor*  
E-mail: fatimah@fsktm.upm.edu.my

**Abstract**

This paper presents works involves in developing a three-dimensional lips model to be used in a visual speech synthesis. Even though several works have been done in similar areas, but a systematic approach of identifying and determining the number of vertices used in the lips model for visual speech synthesis were not explained thoroughly. Instead of the development technique involved, this paper also present experimental design and set up to test the capabilities and flexibility of the lips in deforming speech. The work starts with the process of capturing the lips shape using 3D scanner and then the number of vertices reduced using relevance measure technique in order to achieve the optimum number of vertices to deform all visemes in Standard Malay. Results shown in this work proves that a systematic degradation of 10,000 vertices into 60 vertices by retaining 0.0006% of the initial vertices managed to be done. We also found out that the deformation of the lips with smaller number of vertices is still remaining the similar quality of the lips deformation of the 150 or 100 vertices. Therefore, the outcome of this study proposes and justifies the appropriate number of vertices to be used in visual speech synthesis in our future work.

**Keywords:** 3D Lips, Visual Speech Synthesis, Relevance Measure, Polygonal Construct, Similarity Measurement

1. **Introduction**

Visual speech synthesis (VSS) researches began since early 1900s and have been developing rapidly in the area of multimodal communication and interaction. The VSS applications are used in many areas, such as in the human communication, perception, agent-based interfaces, virtual talking head and
The most important component for effective VSS is the lips model. Visualization of lips improve speech intelligibility especially when the conversation takes place in a noisy environment or in the case of a non-native speakers (Benoit, et al., 1994). Nevertheless, visualizing lips can also leads to confusions, where Massaro (2002) mentioned in his work that even with clearly audible signals and face view, lips deformation can be perceived as different phonetic attributes.

VSS can be implemented in two approaches. The first approach is a called an image-based VSS that normally consists of a sequence of two-dimensional (2D) video images. And another approach is two-dimensional (2D) or a three-dimensional (3D) graphical facial model which is called as a model-based VSS. Few vast model-based VCC projects have been developed such as BALDI (Cohen et al., 2002; Massaro, 2004); GRETA (Pasquariello and Pelachaud, 2001), MASSY (Fagel and Sendlmeier, 2003) and others (Zhiming et al., 2002; Tekalp and Ostermann, 2000; Escher et al., 2000; Pandzic and Forchheimer, 2000).

However, the animation of these talking faces and lips synthesis is not yet satisfactory and the exploitation of the bimodality of speech were not been fully achieved because of few issues. The first issue is because the characteristics of speech which is very dynamic (Lee and Yook, 2002). Secondly, the properties of each lips muscle are completely dependent on each other. Facial muscles move in parallel and these will affect different controls on a different facial model (Gourdeaux et al., 2001). And the third issue is concerning on the coarticulations phenomena that refers to the changes in the articulation of a speech segment depending on preceding and upcoming segments. Even though several methods have been attempted, but there are still more room for improvement (Edge et al., 2004).

In terms of medium of language applied for synthesis, most of the visual speech synthesis works were done for the English language. Some works were carried out on other languages, for instance Chinese (Li et al., 2003), Mandarin (Jiang et al., 1999), Finnish (Olives et al., 1999), Hindi (Verna et al., 2003) and Swedish (Gustafon et al., 2001). To date there are not much work done to develop a visual synthesis system for Standard Malay (SM) language.

Thus, the objective of this paper is to present works of developing a 3D lips model to support the model-based VSS in SM. The scope of the lips model covers the followings: the model is in a three-dimensional (3D) and consists of connected polygons surfaces and vertices in a wire frame format. Instead of the lips model development, major contribution of this paper is on the experimental design to test the flexibility of the lips as most papers did not explain thoroughly on the systematic approach to test the capabilities of the lips developed.

We organize the content of this paper as the followings; Section 1 provide the starts by giving overview of the visual speech. It follows by Section 2 which presents on the lips component and modeling concept. Section 3 presents some related works and literature. The 3D lips model development and experiment setup will be discussed in Section 4. Results and discussion is presented in Section 5 and finally, Section 6 draws the concluding remarks and future work of this study.

2. Lips Component and Modeling
Lips consists of two portions, upper lips that also called as Labium superioris, and the lower lips that called Labium inferioris. Both lips are soft, protruding, movable, and serve primarily for articulation of speech instead of other important functions such as food intake and as a tactile sensory organ. The separation between the lips and the surrounding skin is referred as the vermillion border. The protrusion, motion and deformation of the lips produce different sounds. According to the study of the facial muscles in a human face (Fleming and Dobbs, 1999), each muscle move into different direction and it will create different mouth shapes depending on how much force given to that particular muscle. Each muscles move symmetrically when force are set on the muscle (Choi and Hwang, 2004) during normal speech and at a natural speaking rate. In the perspective of physiology, the speech production mechanism described in terms of respiration, phonation, and articulation. And speech sounds produced are depending on locations of a various articulators such as jaw, tongue and mouth. In relation to this study, mouth movements are essential parameters to generate visual speech animation and synthesis.
In a model-based approach, the control parameters is required to deform the 3D structure using geometric models (Bailly et al., 2003). Then, the 3D structure will be combined with graphical rendering techniques to produce realistic lips animation. The advantage of the model-based approach is where all important visual features presented in a low-dimensional parameter space and invariant to translation, rotation, scaling and illumination (Liew et al., 2002). Commonly, there are five major techniques applied in developing model-based visual speech synthesis; interpolation (key framing), physics-based muscle, pseudo-muscle, performance-driven and parameter-based. In this study, we chose the parameter-based technique for its simplicity and low-dimensional parameter space required. Parameter-based main strength is on its ability to construct convincing facial expressions at a low bit rate (Haratsch and Osterman, 1997). This technique also has a capability to alter the face as per required by the desired facial appearance (Breton et al., 2001). It also manages to produce an animation that able to specify any possible face and expression by a combination of independent parameter values.

2.1. Visemes and Standard Malay

The visible articulatory movements during the process of uttering phonemes called as visemes. The phoneme is the smallest identifiable unit of speech. For example, the SM word “mata” is built up from phonemes /m/, /a/, /t/ and /a/. On the other hand, viseme also defined more precisely as sequence of oral-facial movements (shapes) that relate to the articulation of linguistic-based units of phoneme (Parke and Waters, 1996). Since not all of the articulators are visible, there is a many-to-one-mapping between phonemes and visemes. In the SM language there are several acoustic speech sounds that are visually ambiguous. For example the phonemes /p/, /b/ and /m/ are all articulated in the same manner with the mouth closed; therefore appearing visually the same (Rao et al., 1998).

In terms of medium of language, the Standard Malay (SM) language is the most widely spoken language of Malaysia. The ‘Standard’ Malay (SM) is a term to designate a variety of language which is basically accepted by members of the speech community to be the norm or the prestige dialect which is used in formal situation (Seong, 1994). Speech in SM comprises of a mixture of audio frequencies, and every speech sound belongs to one of the two main classes known as vowels (V) and consonant (C). Vowels and consonants belong to basic linguistics units known as phonemes (Roach, 2000) where each phoneme corresponds to a single sound system of the language. SM phonetics that is based on ARPABET table consists of thirty-six (36) phonemes. The vowel phoneme are /a/, /e/, /i/, /o/ and /u/. Nineteen primary consonant which are native consonant sounds /p/, /b/, /t/, /d/, /k/, /g/, /?/, /m/, /n/, /ŋ/, /p/, /t/, /d/, /s/, /h/, /r/, /l/, /w/ and /j/. Another eight secondary consonant are /f/, /v/, /δ/, /ð/, /z/, /ʃ/, /x/ and /θ/ that occur in SM are borrowed from other languages predominantly English and Arabic (El-Imam and Don, 2000). The vowel is the most prominent sound in SM syllable. When a syllable is pronounced, the lip shape for the consonant sustains very short and quickly switches to that of the vowel.

Since most visemes cannot be uniquely associated with a single phoneme, therefore Chen (2001) mapped 23 phonemes into 14 visemes groups according to the shapes of mouth and placement of tongue. Table 1 shows the phoneme groups proposed by Chen (2001) and we will use the similar mapping groups.
### Table 1: Phoneme Groups (after Chen, 2001)

<table>
<thead>
<tr>
<th>Group No</th>
<th>Visemes</th>
<th>Group No</th>
<th>Visemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>/p/, /b/, /m/</td>
<td>8</td>
<td>/n/, /l/</td>
</tr>
<tr>
<td>2</td>
<td>/f/, /v/</td>
<td>9</td>
<td>/R/</td>
</tr>
<tr>
<td>3</td>
<td>/th/, /dh/</td>
<td>10</td>
<td>/A/</td>
</tr>
<tr>
<td>4</td>
<td>/t/, /d/</td>
<td>11</td>
<td>/E/</td>
</tr>
<tr>
<td>5</td>
<td>/k/, /g/</td>
<td>12</td>
<td>/l/</td>
</tr>
<tr>
<td>6</td>
<td>/sh/, /zh/</td>
<td>13</td>
<td>/O/</td>
</tr>
<tr>
<td>7</td>
<td>/s/, /z/</td>
<td>14</td>
<td>/U/</td>
</tr>
</tbody>
</table>

### 3. Related Works

The earliest VSS parameter-based model has been developed by Parke (Parke, 1972 cited in King, 2002). However, his original parameter-based technique only takes care of limited lips motion and only a little focus on the lip shapes (King, 2002). Another research made by Guirard et al. (1996), involved a 3D lips that has an internal and external contours of the vermillion zone. Those contours fitted by means of algebraic equations. They developed a 3D model that based on a geometrical analysis of the natural lips of a French speaker. But the speaker dependant for French speech limited the used of this model to other languages.

Vignoli (2000) proposed a VSS algorithm based on linear approximations. It estimates the lip movements from a timed sequence of phonemes. The sequence generated from real speech by a segmentation technique based on HMM from Text-To-Speech system. As we know that HMM normally requires a medium size training datasets where the process of acquiring and extracting features for training data set consume some time.

Pasquariello and Pelachaud (2001) developed lips model on GRETA by using NURBS on an MPEG-4 compliant model but the behaviour of facial tissues emulated is based on pseudo-muscular approach. The approach used is commonly similar to parameterization technique where the facial animation parameters located at relevant places on the model face. Another lips model is developed by King (2001) and it consists of a B-spline surface and the high-level parameters which control the articulation of the surface. Animator can control the lips by changing its parameters. He used a B-spline surface for its $C^2$ continuity and the surface deformed by simply moving vertices of the control mesh. Two drawbacks of the B-splines model are on the difficulties in placing part of the surface exactly in preserved volumes; problems in detecting collisions and rendering. The lips polygon also loses the $C^2$ continuity of the B-spline surface.

Next, Fagel and Clemens (2004) introduced two visual articulation models for speech synthesis and methods to obtain them from measured data. Both models controlled by six motion parameters. One for the movement of the lower jaw, three for the lips and two for the tongue. Their model is hybrid and rule based model that selects and combines most similar viseme patterns. Then, Kakumanu et al. (2002) developed a system that captures short phenomena in the orofacial dynamics of a speaker by tracking the 3D location of various MPEG-4 facial points through stereovision. A perceptual transformation of the speech spectral envelope and prosodic cues combined into an acoustics feature vector to predict 3D orofacial dynamics by means of a nearest-neighbour algorithm. They presented a player that based on pseudo-muscle model augmented with a non-penetrable ellipsoidal structure to approximate the skull and the jaw. The combination of parametric and pseudo-muscle is claimed to produce realistic dynamics.

Hsieh and Chen (2006) proposed an adaptation method, called partial linear regression (PLR) to generate an audio-driven talking head application. This method reduces the training hours that normally takes a long time on retraining a new user-dependant model. It also adjusts the user independent model to a more personalized one.
Among related works mentioned above, not much discussion has been presented in terms of the lips model development and its capability measurements. Their lips look realistic after rendered but the flexibility of the lips deformation has not been mentioned and discussed thoroughly.

4. Lips Model Development

Human speech production is complex and a non-stationary process. Viseme that represents lips deformation and shapes is complex to animate due to tedious processes that involve different degree of lips contractions and deformations (Gourdeaux et al., 2001). The animation of lips involves motions of upper and lower lips in parallel. Active shape models (Lepsoy and Curinga, 1998) normally represent the lips motion where the lip model can consequently be animated as the articulatory parameters varies with time. Lips components, which are not translated, rotated and contracted smoothly, will result strange facial expressions and non-accurate motion (Gourdeaux et al., 2001).

Lips developed for VSS must focused on a few general principles. One of the focuses is on the animation control parameters which should be intuitive and easy to use (Byun and Badler, 2002). Another principle is also on the control parameters where it should be consistent and applicable across different face models (Byun and Badler, 2002). Therefore, adapting animation data for a different face model should require little manual intervention as possible. According to Zhimming (2002), the lip model must be easily defined and modified for different mouth model and its motion must be intuitively to use.

Since our study is to focus on generating lips deformation and motion, 3D lips model that used in our study is a wire frame model which is in polygonal structure. The use of polygonal model yields to a smooth contour. One advantage of the polygon model is that calculations of the changing shapes in the polygon models can be carried out much faster than those for the muscle and tissue simulations (Spraot et al., 1998). It also may be easier to achieve the desired lips shapes directly rather than in terms of the components of muscle actions.

We started our work by obtaining the basic shape of 3D lips model using a 3D laser scanner, VIVID 900 Version 1.0. Figure 1 shows the process of acquiring the lips model. Lips extracted from the scanned face consist of high volumes of verti ces. Huge number of vertices will require high computation time and storage, therefore, we need to eliminate the vertices to the minimal number of vertices. Even though the number of vertices is minimal, we ensure that the lips shape is very close resemblance to the actual shape of lips. King (2001) in his work used B-splines surface with a 16 x 9 (144 control points) control grid while Guirard (1998) used 23 vertices in his study.

Figure 1: The Flow of Process in Acquiring the Lips Model
The scanned lips edited using Geomagic Version 2.0 and 3D Studio Max. We consider holes and unwanted vertices as noise in the lips model. The original model consists of more than 10,000 vertices. We reduced manually using tools mentioned. When the vertices number reaches 1000 vertices, random elimination process conducted using the C++ program. We set a target to have less than 100 vertices. Elimination and data filtering done to the vertices by removing, bridging and polygonized the lip model process repetitively. The reconnected polygonal surface must preserve the basic lips shape.

To make the process systematic, the lips vertices is divided into three lines at the upper and the other three lines of vertices at the lower lip. We assume that the idlest shape for lips is symmetry, therefore, the lips is separated into two equal sizes. We take one portion (the right side) of the lips and reduce the number of vertices from the portion.

When the number of vertices reduces into 150, we started to use simplification technique. We used curve evolution that implement relevance measures to simplify the shapes by removing irrelevant and keeping relevant shape features (Lee et al., 2003). An irrelevant vertex has the lowest value of relevance measure. We iteratively compare the relevance measure of all vertices on the polygon. Higher relevance value means that the vertex has larger contribution to the shape of the curve. For each of these iterations, the vertex that has the lowest relevance measure removed and a new segment established by connecting the two adjacent vertices. The following equation (Lee et al., 2003) explains the relevance measure calculation:

\[ K(S_1, S_2) = \frac{|\beta(S_1, S_2) - 180| l(S_1)l(S_2)}{l(S_1) + l(S_2)} \]  

\[ \beta \] is the turn angle, and \( l \) is the normalized length.

The following figure (Figure 2) illustrates the parameters used in the Equation 1 and the relevance measure calculations.

Figure 2: Parameters Used in Relevance Measurement Calculation

To explain the concept, Figure 3(a) has lower relevance measure compared to Figure 3 (b) because it has shorter length. While Figure 3(b) has lower relevance measure than Figure 3(c) because of smaller turn angle (Lee et al., 2003).
A relevance measure equation proposed by Lee et al. (2003) removes short and straight line segments so that the critical points can be detected and preserved. The curve evolution technique reduces the data points and keeps significant shape features. It removes the vertices that have short length and/or their turn angles are close to 180 degrees (straight line). Figure 4 illustrate an example of the relevance measurement taken from the lips model. It shows how the measurement of $\beta$ and $l$ is taken. For example $\beta = 74.83^\circ$ and length for $S_1$ is 1.5 mm and for $S_2$ is 1.6mm. For each of these iterations, the vertex that has the lowest relevance measure is removed and a new segment was established by connecting the two adjacent vertices.

![Figure 4: Example of Parameters on the Lips Model](image)

After few iterations, finally we managed to obtain the final lips model consists of 60 vertices. We labelled it with number from 0 to 59. Table 2 shows the result of experiment done to reduce the number lips vertices.

<table>
<thead>
<tr>
<th>Table 2: Number of Iterations to Reduce Lips Vertices</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Iteration</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td><strong>Total Number of vertices</strong></td>
</tr>
<tr>
<td><strong>Number of Vertices Removed</strong></td>
</tr>
<tr>
<td><strong>Number of Vertices Remain</strong></td>
</tr>
</tbody>
</table>

For the cosmetic purpose, we removed another four vertices that make the lip shape smoother and easier to control during deformation. Those four vertices do not give significant effect to the overall lips shapes. The lips structure needs to have a set of control points to position and compose shapes, deform and move. We selected four vertices to be as the control points. The frontal view of wire frame lips model together with the control points which label as X1(vertices label 0), H1(vertices label 15), X2(vertices label 28) and H2(vertices label 59) are shown in the Figure 5. Next, we tested the lips model to deform certain basic shape of lips opening, protrusion and close mouth. During visual speech synthesis, the inner lip contour is estimated from the points of the outer lips contour base on linear estimation.
5. Experimental Set Up

In order to measure the capability of the lips model, the actual lips shape obtained from video images are compared against the 3D lips model. In previous work (Choi and Hwang, 2004), (Aarabi and Mungamuru, 2004; Liew et al., 2002) employ four control points in acquiring lips shapes measurement and deformation and their study capable to produce a visual speech synthesis. The four control points are: (i) Point at the left corner lips; (ii) Point at the right corner lips; (iii) Point at the centre of outer upper lips and (iv) Point at the centre of outer lower lips. Based on those four control points mentioned earlier, we chose three similar control points to represent lips measurement. These control points will be considered as the control points at the synthesis stage later on. We eliminated the left corner lips points as we assume that lips motion and deformation is at a high degree of symmetry. Although a speaker’s lip and lip movement may not exactly symmetric in general, the deviation from symmetry is usually not significant linguistically (Hennecke cited in Liew et al., 2002). We use points at the outer lips contour for the lips shape measurement as psychological studies suggest that the inner and outer lip contours are important visual speech features.

To get the actual lips measurement, Euclidean distance of selected control points were obtained from video images. The mouth width and height of open protrude and close mouth is measured based on control point Euclidean distance to the virtual reference point (R). The reference point is a centre pixel between nostrils and its position a little bit lower into the lips area. Its vertical position is based on the y pixel coordinate of the right lip corner. The vertical reference line was plotted before the ROI is extracted in previous step. Each control point’s location acquainted and recorded from the inverse grey scale image. The grey scale images were inversed in order to facilitate the location pointing process. Figure 6 shows the steps to get the Euclidean distance. The location of each point is recorded in a pixel base coordinate value.
Reducing vertices number in 3D wire frame lips model taken place in parallel to the lips deformation measurement.

6. Result

Results of this experiment shows the most appropriate number of lips vertices in the lips model; and it also shows that the model is capable to animate and synthesize all SM visemes. Results measured based on the correlation coefficient values of synthesized versus actual lips. The inner lips contour of the lips contour shape are compared to the actual visemes dataset. Figure 7 shows the correlation coefficient measurement of three of lips forming 36 SM visemes.

Figure 7: Result of Similarity Measurement in Reduced Vertices Lips Model

The above figure shows that the blue line is consistently close to 1. In other words, by having 60 vertices we manage to animate the lips model that highly similar to the actual lips deformation. Average correlation coefficient values for lips that consist 60 vertices is 0.94; lips with 56 vertices is
0.83 and lips with 52 vertices is 0.76. With that we can conclude that 60 vertices is the most appropriate number of vertices that capable to animate and synthesize all visemes shapes in Standard Malay.

The following figure (Figure 8) shows the correlation coefficient (CC) produced by synthesized lips in deforming vowel sound within 5 second timeline. Clearly shown in the figure, that the CC for all vowel sound is within the range of 0.70 to 0.98 which is considered highly correlated. The average correlation coefficient values for lips deformation is 94%. We highly recommend that this is an acceptable CC for lips deformation of lips synthesized model which is within an acceptable shape of mouth being produced and the shapes is understandable by viewers.

**Figure 8:** Results of Similarity Measurement for Lips Model versus Actual Lips for Vowel Phonemes.

![Figure 8: Results of Similarity Measurement for Lips Model versus Actual Lips for Vowel Phonemes.](image1)

Same goes to the results obtained for CC measured for consonant sound deformation. The CC for all vowel sound is quite high which is within the range of 0.75 to 0.98. While the average correlation coefficient values for lips deformation is 89%. And again this is an acceptable CC for lips deformation of lips synthesized model which is within an acceptable shape of mouth being produced and the shapes is also understandable by viewers. The following figure (Figure 9) shows the correlation coefficient (CC) produced by synthesized lips in deforming consonant sound within 5 second timeline.

**Figure 9:** Results of Similarity Measurement for Lips Model versus Actual Lips for Consonant Phonemes

![Figure 9: Results of Similarity Measurement for Lips Model versus Actual Lips for Consonant Phonemes](image2)

According to El-Imam and Don (2000) most SM syllable structure are in CV and CVC format. Therefore, we also tested randomly for CV structure and CVC mouth shapes of the synthesized lips. Figure 10 shows results for CV sounds that we picked up randomly of /pa/, /pe/ and /po/. The CC for all vowel sound is quite high which is within the range of 0.85 to 0.98. While the average correlation coefficient values for lips deformation is 90%.

![Figure 10](image3)
Figure 10: Results of Similarity Measurement for Lips Deformation for Random CV Syllable

![Figure 10](image)

Figure 11: Results of Similarity Measurement for Lips Deformation for Random CV Syllable

![Figure 11](image)

Above Figure 11 shows results for CV sounds that we picked up randomly of /ke/, /ku/ and /ki/. The CC for all vowel sound is quite high which is within the range of 0.75 to 0.97. While the average correlation coefficient values for lips deformation is 89%.

Figure 12: Results of Similarity Measurement for Lips Deformation for Random CV Syllable

![Figure 12](image)

Above Figure 12 shows results for CV sounds that we picked up randomly of /li/, /lu/ and /le/. The CC for all vowel sound is quite high which is within the range of 0.83 to 0.94. While the average correlation coefficient values for lips deformation is 90%.

In a synthesis lips model, it is appropriate to have the minimal number of vertices. Minimal numbers of vertices involve minimal number of relocation of the position of the vertices during lips deformation. Minimal relocation results minimal computation. In our experiment we have reduced form more than 10,000 vertices into approximately 150 to 100 vertices. Results shown that, by having
150 and 100 vertices we able to imitated real lips deformation. Reducing it into 80 also produced similar value of CC. While having 60 vertices the similarity decreased a bit but averagely the similarity is still highly correlated to the actual shape and deformation. If we reduced it to 56 or 52 the reduced similarity abruptly reduced and affects the visual representation of the lips shape during deformation. Therefore, as a result we decided to stop reducing the number of vertices up to 60 where with 60 vertices, the highest CC can be achieved.

Lips deformation in vowel and consonant shows quite a high similarity. Even though in vowel phoneme similarity measurement at time 2 (t =2) phoneme /a/ and /e/ and time 5 (t=5) phoneme /e/ shows slightly low correlation, it is because the pace and the speed of the phoneme pronunciations may be different from one person to another. Same goes to the consonant phoneme where group 1 and group 2 a bit low CC. We have ran the test repetitively, the CC is within the similar range. Therefore overall, we agree that the lips model able to produce high similarity of lips deformation.

Finally the results of random CV syllable show a variety of range of CC values. But still considered both actual and synthesized model are highly correlated. On the other hand our lips model shows higher CC. From the results obtained, we can conclude that the lips model developed is able to produced lips deformation at high similarity with the actual lips model. Therefore the lips model will be used for further work in the next stage of the study.

7. Conclusion and Future Works
In this study our measurement is concerned on the visual speech realism that covers naturalness and similarity to the actual lips motions. The reason is that natural lips deformation that imitates closely the real lips motions will contribute to speech intelligibility as well. In future work, the visual-to-visual recognition will apply the Hidden Markov Model (HMM) approach. The HMM will receive visual speech as an input and synthesis the visual speech on the 3D lips model.

References


