

Synthesis Visual Speech using Modified Polynomial Motion Curve

Siti Salwa Salleh[†], Rahmita Wirza Rahmat^{††}, Ramlan Mahmud^{††} and Fatimah Ahmad[†]

[†]Faculty of Information Technology and Quantitative Science, UiTM, 40450 Shah Alam, Malaysia

^{††}Faculty of Information Technology and Computer Science, UPM, 43400 Serdang, Malaysia

Summary

Composing animations using motion path that based on polynomial function is usually refused by animators. This is because of inability of the lower order polynomial function to produce lengthy motion curve. On the other hand, higher order polynomials function produces unpredictable and undesired shape of curve which is too curvy and unstable. But, polynomial curve is adaptable, where it can easily apply on any lips model regardless of the model size.

Therefore, this paper presents a technique to modify high order polynomial curve to synthesis visual speech. We discuss every steps involve and all equations used. The steps later on presented in a form of an algorithm to generate the lips motion. Technique used successfully made the lips model synthesized isolated utterances of digits in Standard Malay language. As a result, the Correlation Coefficient computed shows that both synthesized and actual lips motions are highly likely to be similar.

Key words:

Motion path, polynomial, visual speech synthesis, lips animation.

1. Introduction

Visual speech synthesis (VSS) researches began since early 1900s and have been developing rapidly in the area of multimodal interaction. Its applications are widely used in many areas, such as in the human communication and perception, agent-based user interfaces and in a virtual talking head in cartoon or game applications [1]. In the VSS development, the most important element is the lips animation. Research has shown that naturalness, realistic and accurate motion of synthesized lips improve speech intelligibility in human-machine interaction [2]. Moreover, according to Massaro [3] viewing the talker's face can modify phonetic attributes even with clearly audible signals.

To date, several model-based VSS researches have been carried out [4],[5] but yet satisfactory and the exploitation of the bimodality of speech cannot be fully achieved. The goal is difficult to achieve because of certain issues. First,

because the characteristics of speech which is very dynamic [6]. Within a small segment, the speech sound changes gradually. Secondly, the properties of each lips muscle are completely dependant on each other. Facial muscles move in parallel and these will affect different controls on a different facial model [7]. Finally, the major problem is caused by the coarticulation, which refers to the changes in the articulation of a speech segment.

One of the focuses in VSS is on the animation control parameters. Controlled parameters should be intuitive and easy to use, consistent and applicable across different face models [8]. This means, adapting animation data to a different face model should require as little modifications as possible and should form a realistic viseme. Viseme is known as the shape of visual speech during speech utterances. And viseme is complex to animate due to tedious process that involves different degree of lips contractions and deformations [7].

In our work that presented in this paper, we applied a parameter-based technique for its low-dimensional space required. The motion path is kept in a form of parametric polynomial function. To date, no theoretical basis has been introduced for designing this type of predefined function [9]. According to Hong *et al.* [9], not many researches address on mechanism to choose interpolation functions; and on how to adjust control points and what are the correlations among those control points.

We present here steps taken in generating a predefined function specifically on the higher order polynomial function. This paper is divided into five sections. It begins with Section 1 which discussed on the VSS background and followed with Section 2 which discuss on data acquisition and lips model preparation. Followed with Section 3 that discuss on motion path generation. Result and discussion are presented in Section 4. And finally Section 5 concludes on the works done.

2. Data Acquisition and Lips Model

The dataset used in our study were obtained from a series of video recording sessions. The video consist of audio and visual speech images of five male and five female speakers. The speakers uttered a sequence of digits, zero to nine in Standard Malay (SM) at a normal speaking rate and pronunciations. The characteristics of the SM are slightly similar to American English [11]. It has 36 vowels and consonants speech sounds, while English generally have 42 phonemes. The datasets of digit utterance were commonly used in visual speech researches where the digit uttered can be rearranged in any order that form various set of utterance for testing purpose. We did the feature point extractions by using image processing technique on the still images extracted such as image format conversion, distance measurement, image contrast and sharpening.

Lips model used is a 3D lips geometric model. In this model, the measurements of the height and width of the mouth opening are important features. All measurements obtained will be used intensively in the process of calculating vertices position that makes the lips deform during the synthesis phase. The 3D lips model consists of connected polygons surfaces and vertices. The model is in the wired frame format and runs on offline environment. We identified four control vertices from the outer lips contour. The control vertices are: vertex at the right corner lips (X1), vertex at the center of outer upper lips (H1), and vertex at the center of outer lower lips (H2). The reference vertex (R) is the point at the centre of the lips which are aligned with the speakers' nostril.

The following diagram (Figure 1) shows the lips model and the position of all the control points (vertices) mentioned previously. X1,X2, H1 and H2 are the control points.

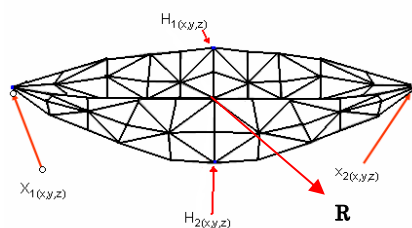


Figure 1. Polygonal Lips Model

3. Polynomial Motion Path

Motion that has been planned normally occurs in certain arrangement of systematic path. Motion planning involve computation of a continuous sequence ("a path") of

configurations (generalized coordinates) between an initial (start) and a final configuration (goal) while respecting certain constraints synthesis [18]. Path is not an intrinsic property of a motion, but rather, a function of how a motion is interpreted [13]. By having a motion curves as a predefined function, we are able to simplify the animation and reduce the size of the motion data.

Polynomial predefined function consists of coefficient, variable (parameters) and arithmetic operations. Although a parametric equation is useful, it is hard to find coefficients needed to represent the desired curve. A simple polynomial has the form:

$$y(x) = \sum_{j=0}^{N-1} (a_j x^j) \quad (1)$$

where a_j is the j th coefficient of a polynomial of degree N . The fitting technique used is a Least Mean Squares method extended to include more than two coefficients. When a fit is found, this method returns the N coefficients and then the N probable uncertainties in the estimates of the coefficients.

3.1 Composing Motion Curve

According to a study of the facial muscles in a human face [28], each muscle move into different directions according to how much force given to that particular muscle. Each muscles move symmetrically when force are set on the muscles [29] during normal speech and at a natural speaking rate.

In general, animators apply one motion path for one object. The object moves along the path within specified time line. We use the same principle and adapt the principle in visual speech generation to make the lips deform. The key concepts in our adaptation are:

- a) Each control vertex (X1, H1 and H2) is treated as a separate object.
- b) Each object has its own motion path and represent by parametric polynomial function.
- c) Motions of all objects (control vertices) are parallel and all motions are active at the same time lime.

Each control vertices do not dependant on each other but time line specified are the same. Every control vertex has its own motion path and the path controls the vertices relocation from one coordinate position to another. The repositioning of control vertices is relocating base on the reference point $R(0, 0, 0)$ which is invisible on the lips model. Its position is similar with the position of reference point used during feature extraction phase.

Other vertices on the lips model move according to the motion of the control vertices. The changes of control vertices make the lips model deform. The repositioning of left corner lip vertex (X2) is the inverse distance of the right corner vertex (X1).

There are five steps involved in generating polynomial function for isolated word utterance and the flow of steps is depicted in the charts below (Figure 2).

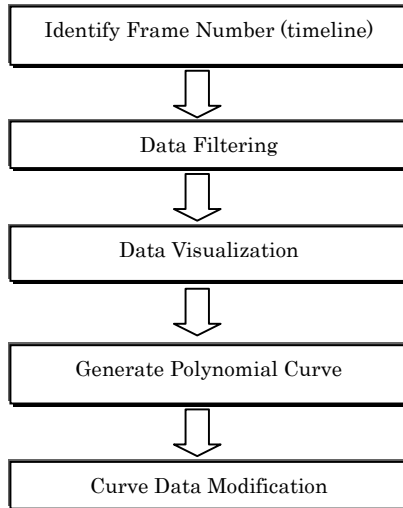


Figure 2. Flow to Obtain Polynomial Function

The steps begin with identifying the timeline needed in producing utterance. 30 sets of measured Euclidean distance data of X1, H1 and H2 were taken in random. At a normal speaking rate, number of frames acquired to complete one utterance are varies among the speakers. Therefore, mod value of frame number considered as generally timeline required for a normal speaking rate. Later, data filtering is done in order to obtain the most appropriate coordinate position for each vertex along the motion curve. The coordinate position is specified by calculation the mean value of each coordinate taken from 30 sets of data. We examined the consistency of each vertex consecutively to avoid extreme values. Vertex that do not aligned with the pattern of the curve will be removed and replaced with interpolated new coordinate point..

In data visualization, all data plotted on the X and Y axis. X-axis represents for the timeline and was obtained base on the frame number and Y axis hold the mean values calculated in step 2. IN the next step, 6th order polynomial curve is generated based on the plotted vertices. The polynomial function is expressed in this format: $y = a + bx^2 + cx^3 + dx^4 + ex^5 + fx^6$. Where a, b, c, d, e and f are the

coefficient values. The following figure (Figure 3) shows an example of motion curve produced.

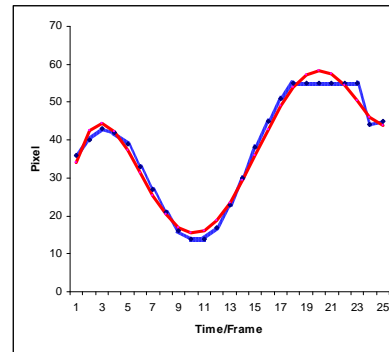


Figure 3. Motion Curve for H1 for Utterance “SEMBILAN” (Nine).

Original line — 6th degree Polynomial curve

The parametric polynomial function for above curve is:
 $y = a + bx + cx^2 + dx^3 + ex^4 + fx^5 + gx^6$



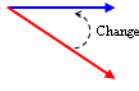

$$\begin{aligned}
 a &= 17.353445, & b &= 21.72083 \\
 c &= -5.3613123, & d &= 4.00E-01 \\
 e &= -0.005707397, & f &= -3.51E-04 \\
 g &= 9.19E-06
 \end{aligned}
 \tag{2}$$

In the last step, modification on the polynomial curve is required. The curvy polynomial curve will result inconsistent motion to the control points. Inconsistent motions will cause confusions to the visual speech reader. Base on the motion curve we decide to do modifications at the beginning and at the end of the curve. Proposed parametric function to modify the polynomial curve is defined as:

$$f(y) = a\beta + bx + cx^2 + dx^3 + ex^4 + fx^5 + gx^6
 \tag{3}$$

Where, β is the value is obtained from the states shown in Table 1. We identified the vertices coordinate that need to be modified by examining the polynomial curve plotted. Vertices coordinate that affects the motion behavior were identified.

Table 1. States of Curve to Determine β value.

Gradient	Curve State	Range of β
Positive		$0.90 < \beta < 0.99$
Positive		$1 < \beta < 1.3$
Negative		$1 < \beta < 1.3$
Negative		$0.90 < \beta < 0.99$

The following diagram (Figure 4) shows the example of the polynomial curve before and after the alteration.

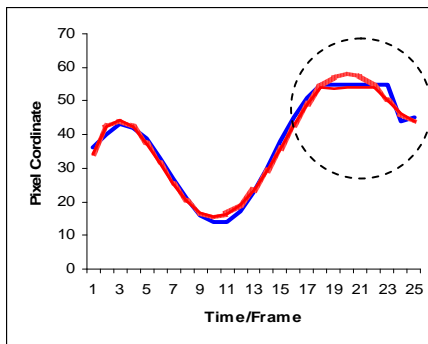


Figure 4. Altered Motion Curve for H1

- Original plotted curve
- - - 6th order polynomial curve
- Modified curve

3.2 Motion Path Generation for Isolated Word Utterance

In uttering isolated word, the lips start with neutral shape and also end with a neutral shape. Therefore, the computations and the steps to build the connection from neutral vertex position to polynomial curve obtained

previously is required. We proposed here steps to do the computation:

Step 1: Identify the neutral lips points and its coordinate (NX,NY).

Step 2: Calculate the gap (GAP) between the neutral vertex and a coordinate of first vertex (WX₁, WY₁).

Therefore,

$$GAP = |NY - WY_1| \tag{4}$$

Step 3: Identify the frame number needed (FA) between neutral vertex and the first vertex on the polynomial curve. Number of frames to add is calculated by using this equation:

$$FA = GAP / 3 \tag{5}$$

GAP is divided by three because from our inspection, the idlest motion for vertex to transpose from one coordinate location to another is three pixels per frame.

Step 4: Number of frame to add will be used to compute the increment factor (INCF) for the vertex motion velocity. The velocity of the lips motion accelerate at the beginning of the utterance and decelerate when the utterance complete [30].

$$INCF = \frac{GAP}{\sum_{i=1}^n j_i} \quad \text{where } n = FA \tag{6}$$

Where j = no of frame and i = no of loop

Step 5: New vertex position is computed using this equation:

$$Vertex_n = INCF * Velocity \tag{7}$$

Where n = 1 to FA, Velocity = 1 to FA.

It takes the largest pixel gap among three gaps of X, H1 and H2. The largest gap will be considered as the common pixel gap for all those three gaps because it indicate the farthest distance to reach along the motion path from the neutral vertex position. All the motion curves for X, H1 and H2 are created separately.

Algorithm to compose the isolated utterance by using Equation (4) to (7) is presented in the following figure (Figure 5)

```

Begin
  Get neutral lip position (Nx or Nh1 or Nh2)
  Read utterance polynomial function (f(u))
  Get first vertex (v1,1) position of the f(u)
  Compute GAP for X,H1 and H2
  GAP ← Max(X,H1,H2)
  Compute FA, INCF
  Loop: for n ← 1 to FA
    Velocity ← n
    Vertexn ← INCF * Velocity
  End loop
  Compute GAP for X,H1 and H2
  GAP = Max(X,H1,H2)
  Compute FA, INCF
  Loop: for n ← FA to 1
    Velocity ← n
    Vertexn ← INCF * Velocity
  End loop
    
```

Figure 5. Algorithm: Generate Isolated Utterance Motion Curve

The following figure (Figure 6 and Figure7) shows the motion curve produces by the algorithm.

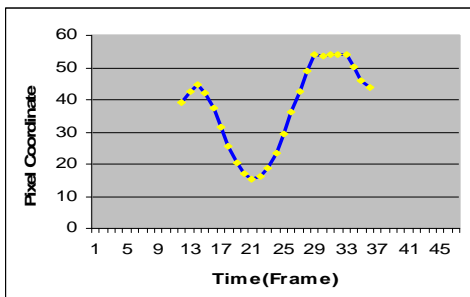


Figure 6. Original Motion Curve for Vertex H1 to Utter Word "SEMBILAN".

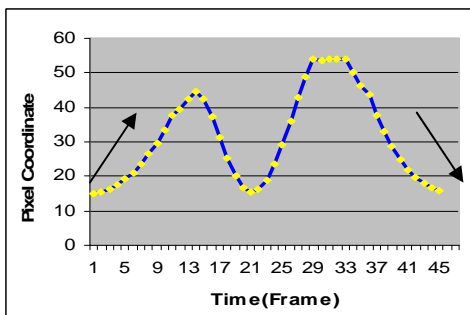


Figure 7. Motion Curve Produces After Implement the Algorithm onto Fig 6.

4. Results and Discussion

The quality of synthesized motions can be evaluated by how much they imitate the real motion of lips. For the lips motion, naturalness is measured base on how similar the synthesized to the real lips deformation. We compare the synthesized motion with the actual motion that obtained from different speakers randomly. The actual motion curve is captured directly by using optical motion capture on the video images of the testing dataset.

The similarity between both motion indicated by the correlation coefficient (CC) measurement. Then the CC between the synthesized and actual motion curve of lips control points of the all control points are calculated. Control value for CC is 1, which means both are exactly similar. Lower value of CC indicates dissimilarity among them. The synthesized motion curve is compared to each motion curve produced by three different speakers.

Both synthesized motion curves and the actual lips motion curve moves through the same direction. The curve slopes occur at a different time frame along the motion line. The variations in positions of phoneme occur during utterance does not affect our results as long as the motion path patterns are still remain the same. Nevertheless, compared to the high variability of this parameter, the difference is not critical as the velocity of speech motion. The occurrence of peak velocity in speech motion may vary in duration but the velocity peaks occur at constant percentage in relative time [32]. The occurrence of velocity peak proves that the shape of motion curve is constant in the utterances. Overall results obtained for CC measurement is presented in Table 2.

Table 2. Correlation Coefficient Measure : Synthesized Motion Curve versus Speaker 1 (S1), Speaker 2 (S2) and Speaker 3 (S3).

	X1			H1			H2		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
KOSONG	0.87	0.90	0.87	0.96	0.92	0.96	0.83	0.89	0.83
SATU	0.99	0.85	0.79	0.98	0.90	0.79	0.92	0.93	0.92
DUA	0.95	0.89	0.90	0.92	0.87	0.92	0.85	0.92	0.89
TIGA	0.95	0.94	0.85	0.98	0.96	0.90	0.94	0.95	0.93
EMPAT	0.75	0.85	0.89	0.71	0.71	0.87	0.95	0.94	0.92
LIMA	0.77	0.80	0.94	0.91	0.85	0.96	0.71	0.88	0.95
ENAM	0.91	0.92	0.85	0.94	0.91	0.71	0.88	0.85	0.94
TUJUH	0.87	0.92	0.87	0.71	0.92	0.92	0.72	0.90	0.72
LAPAN	0.90	0.88	0.90	0.71	0.89	0.85	0.72	0.91	0.72
SEMBILAN	0.86	0.90	0.88	0.84	0.88	0.95	0.85	0.89	0.85

Averagely the CC are 0.88 for synthesis versus Speaker 1 0.88 for synthesis versus Speaker 2 and 0.87 for synthesis versus Speaker 3.

The higher order of polynomial curve used to imitate the curve of the control points' motions. Despite of having curvy and unstable curve, the high order polynomials lead us to a case where we can use it to animate visual speech animation. In this study we successfully presented the improvement made to the polynomial curve to be less curvy and wiggly.

5. Conclusion and Future Works

Approach presented here is a method to solve undesired curvature in motion path generated by the 6th order polynomial curve. The lack of formal methods to define theoretical basis for designing this predefined function, had led this research to find a proper procedure. We used the 6th order polynomial curve to derive the motion of the lips vertices to imitate the lips deformation during speech utterances. Even though the higher order polynomial curve is claimed to be wild and wiggly, we proposed the alteration method to modify the curve.

The alteration method and formulation shows that the flatten curve produce natural movement of the control points on the lips. As a result the motion of each vertex shows high resemblance and naturality in the visual speech synthesis. The correlation coefficient measurement also shows high degree similarity of the feature point's motion for both synthesis and actual motion. The use of polynomial motion function to synthesis lips motion still can be enhanced in future.

References

- [1] Hsieh, C. K. & Chen, Y. C. 2006. Partial Linear Regression for Speech Driven Talking Head Application, *Signal Processing: Image Communication Journal*, Vol. 21, Issue: 1.
- [2] Pandzic, I. S., Ostermann, J. and Millen, D. 1999. User Evaluation: Synthetic Talking Faces for Interactive Services, *The Visual Computer Journal*, Vol. 15, No 7-8, pp.330-40, Berlin: Springer-Verlag
- [3] Massaro, D.W. 2002. The Psychology And Technology of Talking Heads In Human-Machine Interaction.
- [4] Yamamoto, Nakamura & Shikano, 1998, Lip Movement Synthesis From Speech Based On Hidden Markov Models, *In Proceeding Automatic Face and Gesture Recognition*, Vol. 26, pp.105-115.
- [5] Yehia, H., Rubin, P. & Vatikiotis-Bateson, E. 1998. Quantitative Association of Vocal-Tract and Facial Behavior. *Journal of Speech Communication*, Vol. 26(1-2), pp.23-43.
- [6] Lee, S. & Yook, D. 2002. Viseme Recognition Experiment Using Context Dependent Hidden Markov Model, IDEAL, pp. 557-561, UK: Springer-Verlag
- [7] Gourdeaux, K. Chen, Wang, Liu, 2001. Principal Component Analysis for Facial Animation, IEEE, 0-7803-7041-4/01
- [8] Byun, M. & Badler, N. I. 2002. FacEMOTE: Qualitative Parametric Modifiers for Facial Animations, Symposium on Computer Animation, pp. 65-71. ACM.
- [9] Hong, P., Wen, Z. & Huang, T. 2001. An Integrated Framework for Face Modeling, Facial Motion Analysis And Synthesis, Canada. ACM I-581 13-394-4/0009.
- [10] King, S. 2001. A Facial Model and Animation Techniques for Animated Speech, *PhD Thesis. Ohio State University*
- [11] Salleh, S.S., Rahmat, R. W., Mahmud, R. and Ahmad, F..2004. Viseme Set and ASCII Mapping Codes for Standard Malay Audiovisual Speech Synthesis, *Seminar on Sains Teknologi dan Sains Social*, Kuantan Pahang.
- [12] Salleh, S.S., Rahmat, R. W., Mahmud, R. and Ahmad, F. 2007. Proceeding of CITA'07.
- [13] Gleicher, M. 2001 Motion Path Editing , *in ACM Proceeding of Symposium on Interactive 3D Graphics*.