Lip Tracking & Modelling

Aniket Bera

University of North Carolina at Chapel Hill

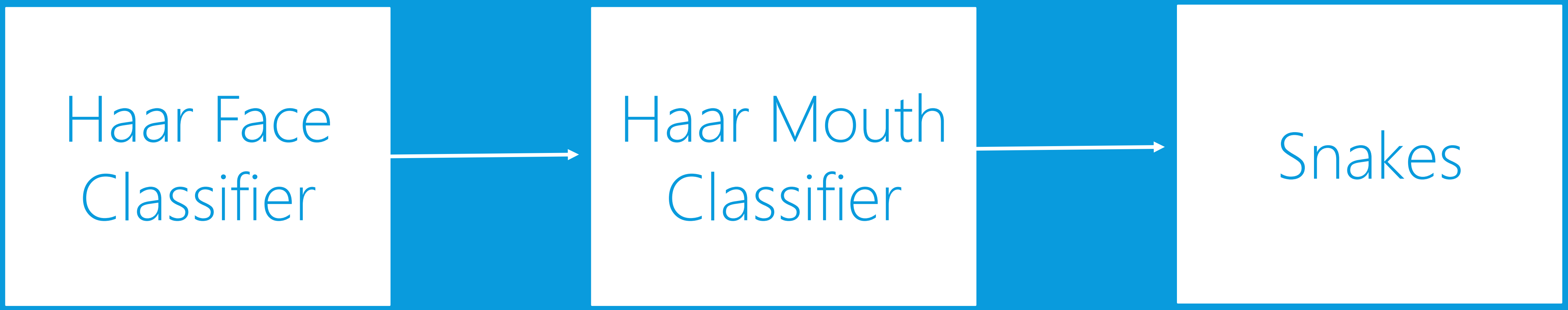
Agenda

- How do we track lips?
- What are the common lip movements?
- How do we model lips in 2D and 3D?
- Relationship between various the motion of lip points during speech.
- An application which uses lip tracking.
- Future

Introduction

- 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.
- The animation of lips involves motions of upper and lower lips in parallel.
- Active shape models normally represent the lips motion where the lip model can consequently be animated as the articulatory parameters varies with time.

Tracking Pipeline – Method 1

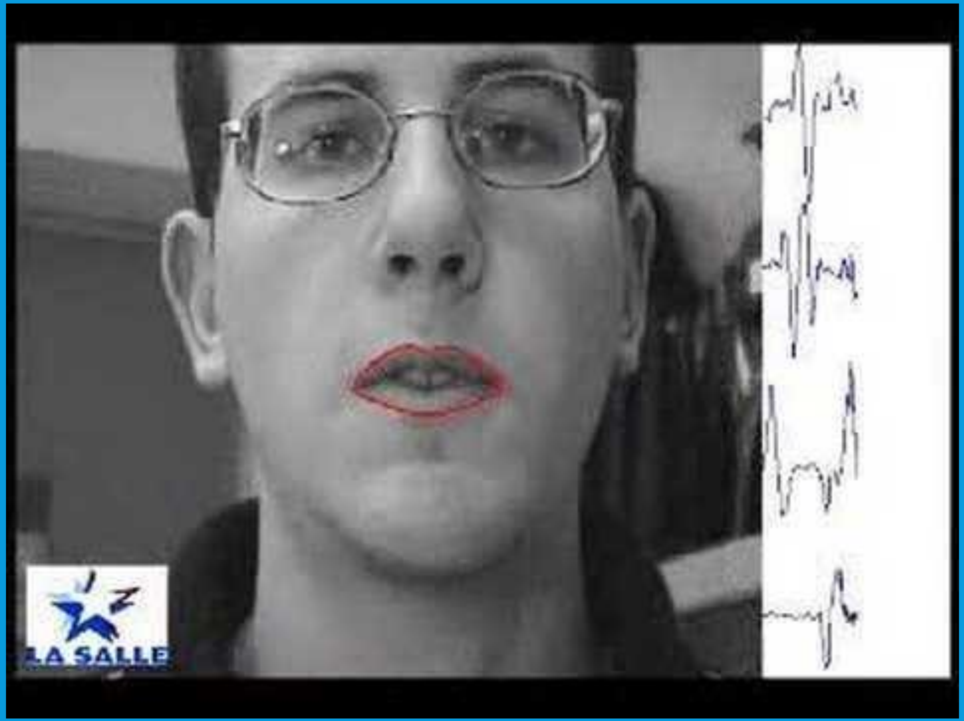


Haar Face Classifier

Haar Mouth Classifier

Snakes

Demo Video

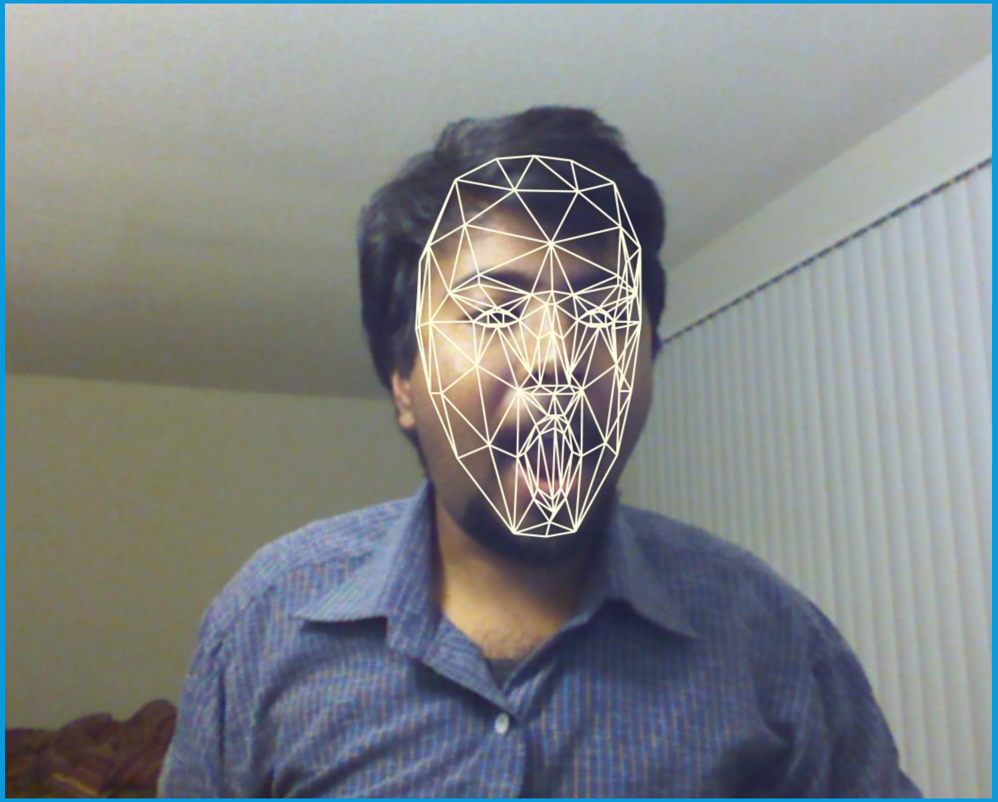


Some previous attempts at Lip tracking

- The lip contour has been represented by a 10-control point **B-spline.**
- Tracking is implemented as an iterative process where the final state reached for a given frame is retained as the initial state for the next one.
- *The Good?*
- Its fast, and looks nice (visually).
- *The bad?*
- Its not natural and does not use the restricted lip motions and hence, there is a high possibility of noise.
- *The Ugly?*
- Highly dependent on skin colour, texture etc.

Tracking Pipeline – Method 2

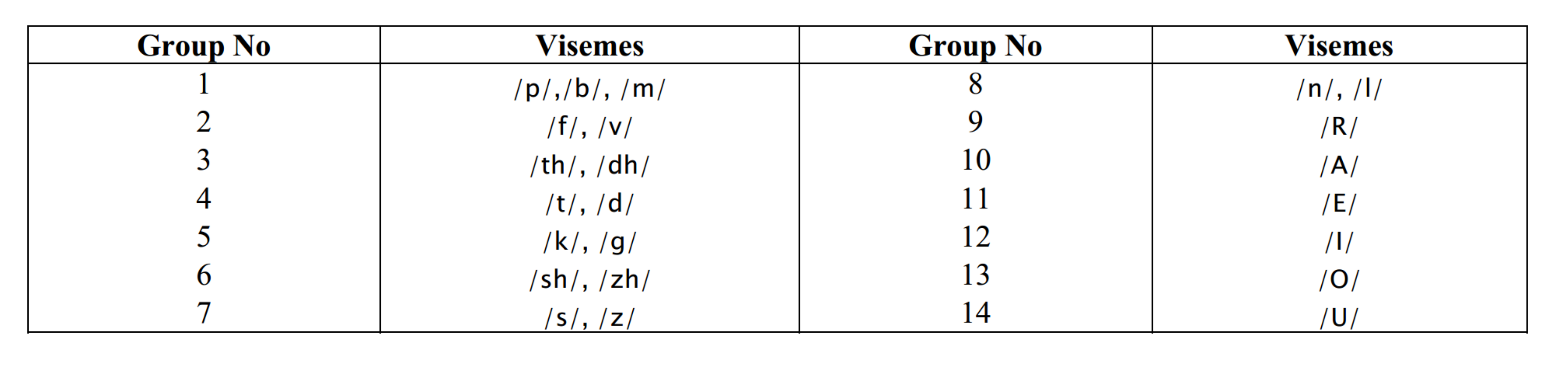
- Active Appearance Model – Kinect
- SDK 1.5 – Face Tracking
- Depth Map
- Mouth Concavity



Key questions…

- How important is the role of the lips in producing sound?
- Are there any digital lip reading software's?
- What is their accuracy?
- How many points do we need to track to properly define a lip 3D model and its animation?

Phoneme Groups (Chen, 2001)

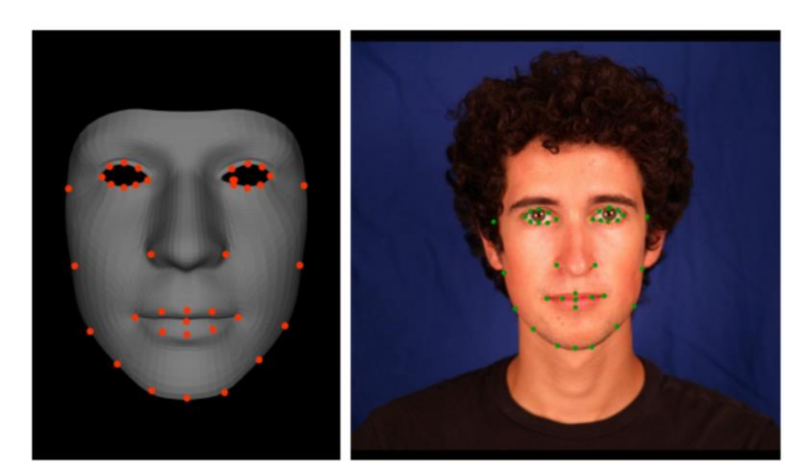


Group No	Visemes	Group No	Visemes
1	/p/,/b/, /m/	8	/n/, /l/
2	/f/, /v/	9	/R/
3	/th/, /dh/	10	/A/
4	/t/, /d/	11	/E/
5	/k/, /g/	12	/I/
6	/sh/, /zh/	13	/O/
7	/s/, /z/	14	/U/

These groups tell us about visually similar lip movements for a set of alphabets. For example Group 1 has Visemes /p/, /b/, /m/ (Pronounced as ‘pa’, ‘ba’ and ‘ma’) all have save lip movement.

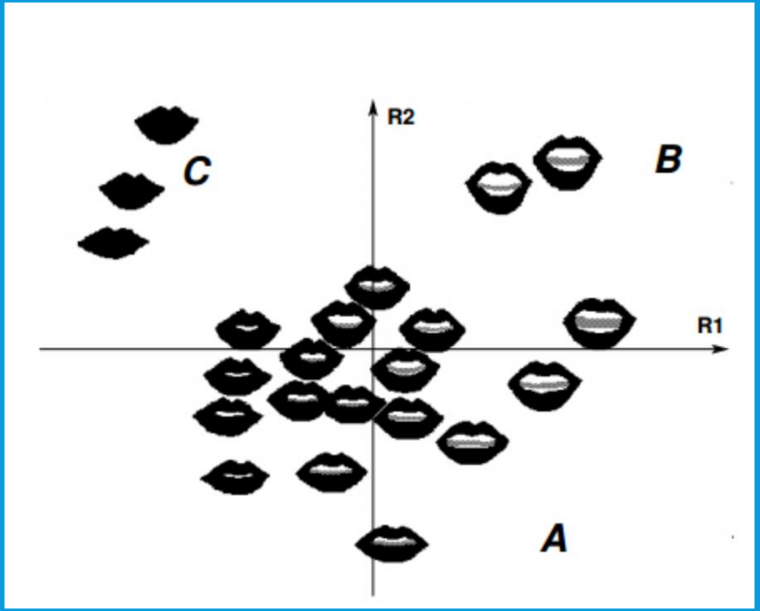
Facial Motion Synthesis from Speech and Text.

- Generate a sequence of time-labeled phones.
- When text is used as input, employ an acoustic text-to-speech (TTS) engine to generate a waveform and the time-aligned sequence of phonemes.
- All speech can be represented as a collection of time varying Phoneme group as shown in the last slide.
- If you want to read more about Facial Motion Project from Speech go to- http://www1.cs.columbia.edu/CAVE/projects/avatar/

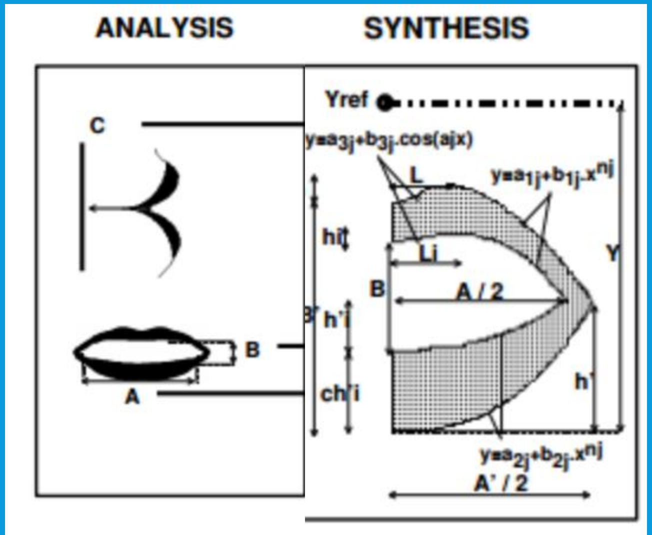


2D model of the lips

- A 2D lip model was first designed by Guiard-Marigny from the front views of 22 basic lip contours as shown in Figure.
- Guiard-Marigny predicted a good approximation of the internal and external lip contours in the coronal plane by means of a limited number of simple mathematical equations.
- The same kind of polynomial and sinusoidal equations were used to describe both the internal and external lip contours.
- Symmetrical lips.
- For each of the 22 "visemes," 15 coefficients were necessary for the equations to best fit the natural contours.



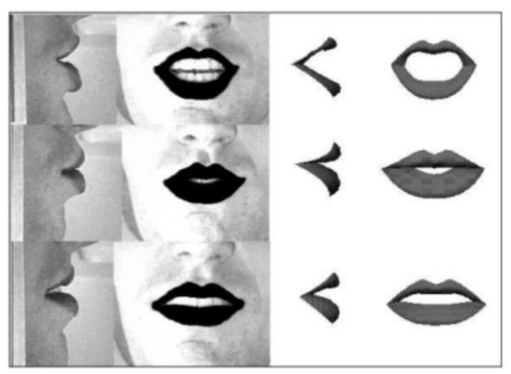
- Ultimately, the 2D model is controlled through only 3 parameters : the width (A) and the height (B) of the internal lip contour, and the lip contact protrusion (C).



3D model of the lips

- Same principal as the 2D technique.
- To identify the equations of the lip contours that best fit the projection of the natural contours in the axial plane.
- The axial plane was selected because of the strong influence of the jaw on the lip shape.
- 3 intermediate contours in between the internal and the external contours for Volume rendering
- 10 polynomial equations. An iterative process to predict all the necessary coefficients of these equations from 5 parameters.
- Those control parameters are the above mentioned three parameters which command the 2D model, and 2 extra parameters: the protrusion of the upper lip and that of the lower lip.

- The goal is not to find out the smallest set of independent parameters that may describe all lip gestures. (For TTS)
- Finally, this set of five parameters allows any lip shape to be reconstructed with a fair approximation of a visible speech sequence uttered a speaker.
- Figure shows the real lips of the speaker and the corresponding synthetic lips in three extreme cases (open, protruded and spread lip)



Reduce vertices to simplify our mesh

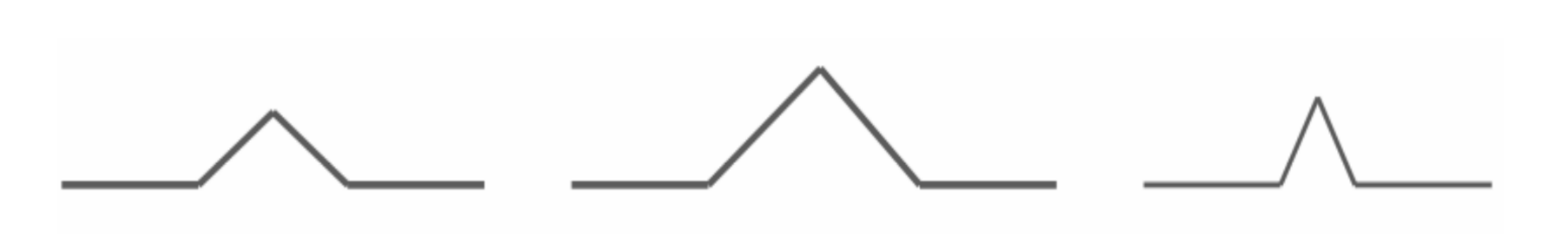
- People use *curve evolution* that implement relevance measures to simplify the shapes by removing irrelevant and keeping relevant shape features.
- An irrelevant vertex has the lowest value of relevance measure. Then iteratively compare the relevance measure of all vertices on the polygon.
- Higher 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.

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

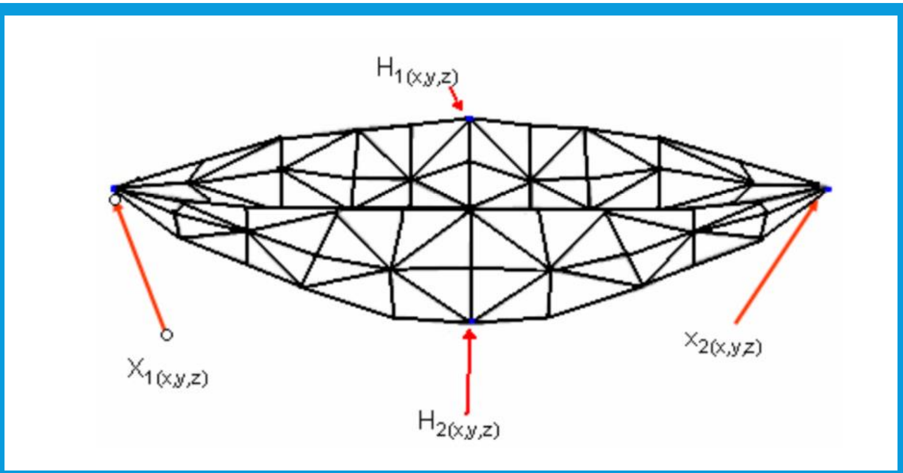
S1 is any vertex that was define as a begin vertex,
S2 is any vertex that was define as an end vertex,
β is the turn angle, and
l is the normalized length

Vertices with Difference Relevance Measure

- 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)

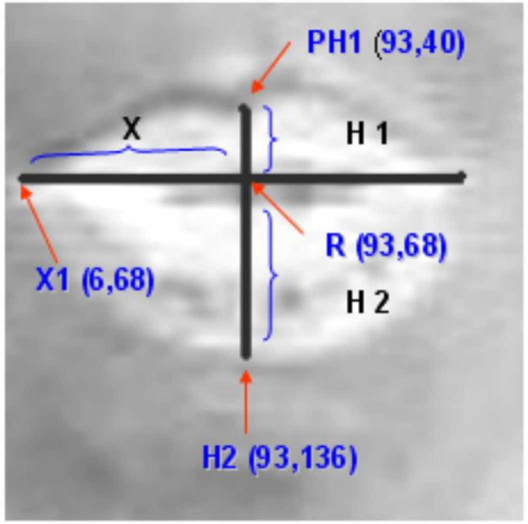


Wire Lips



The lips structure needs to have a set of control points to position and compose shapes, deform and move. Select 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, H1, X2 and H2 are shown in the Figure.

Test Model using -
Euclidean Distance Measurement for Open and Close Mouth



Inverse Color of Grey scale image

Distance Measurement:

Reference Point	= [93,68]
Left corner point	= [6,68]
Upper Outer lip	= [93,40]
Lower Outer lip	= [93,136]

$X1 = \sqrt{(6-93)^2+(68-68)^2} = 87$

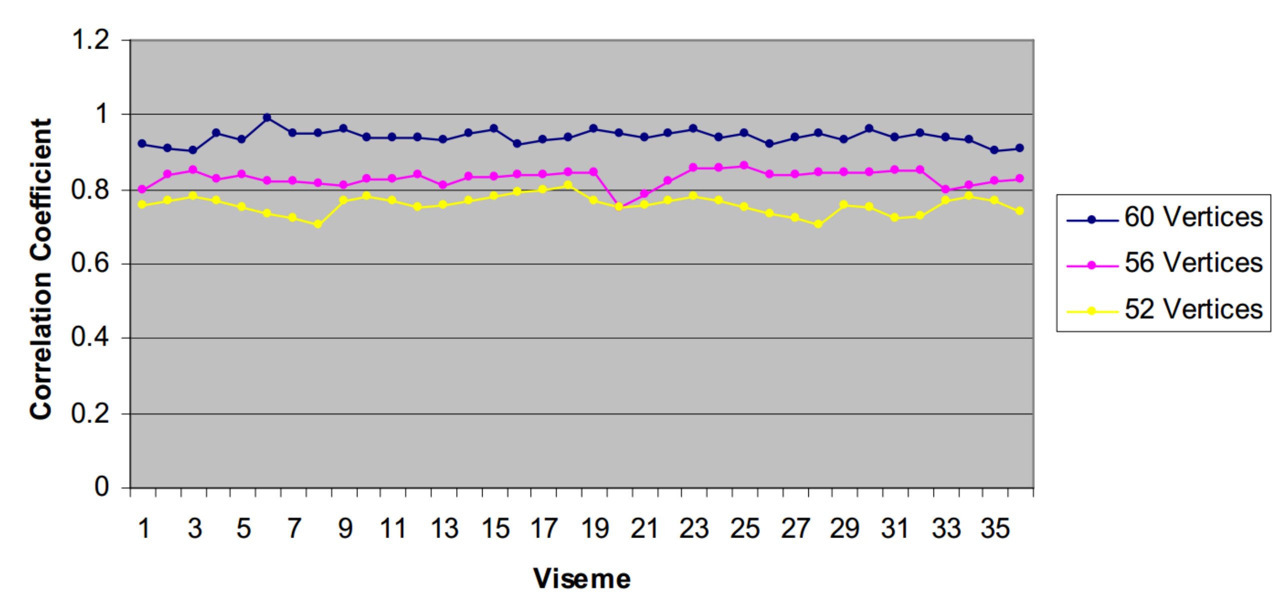
$H1 = \sqrt{(93-93)^2+(40-68)^2} = 28$

$H2 = \sqrt{(93-93)^2+(68-136)^2} = 68$

X1	H1	H2
.	.	.
87	28	68
.	.	.

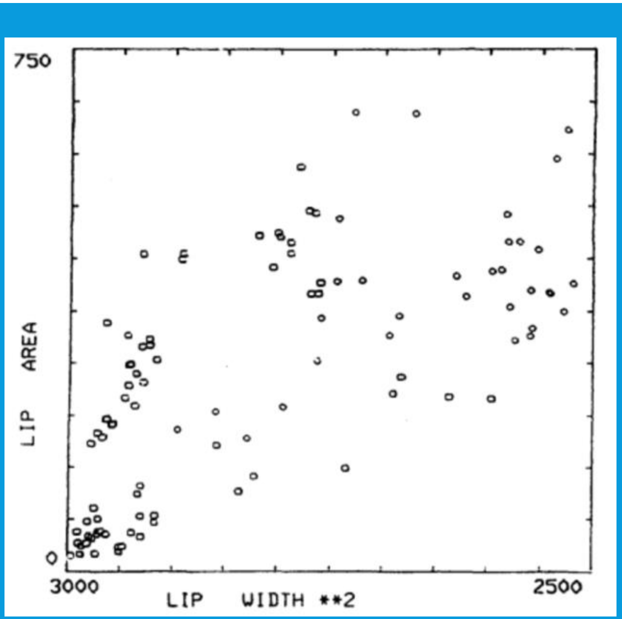
Record the measurement

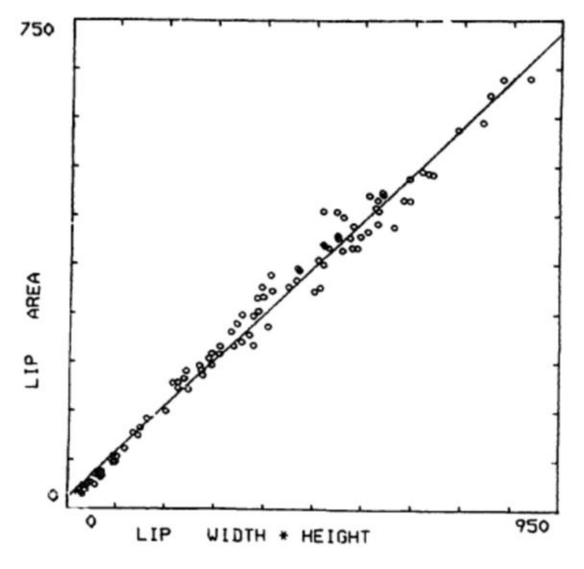
Results

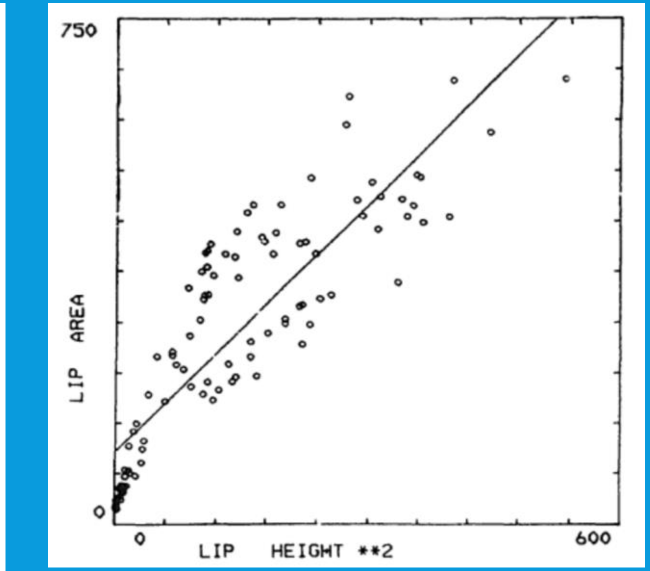


- The figure shows that the blue line is consistently close to1. In other words, by having 60 vertices we manage to animate the lips model that highly similar to the actual lips deformation.
- Average correlation coefficient values for lips that consist 60 vertices is 0.94; lips with 56 vertices is 0.83 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

Study: Relationship between Height, Width, Protrusions and Area (From the paper "*Laws*" *for lips*)







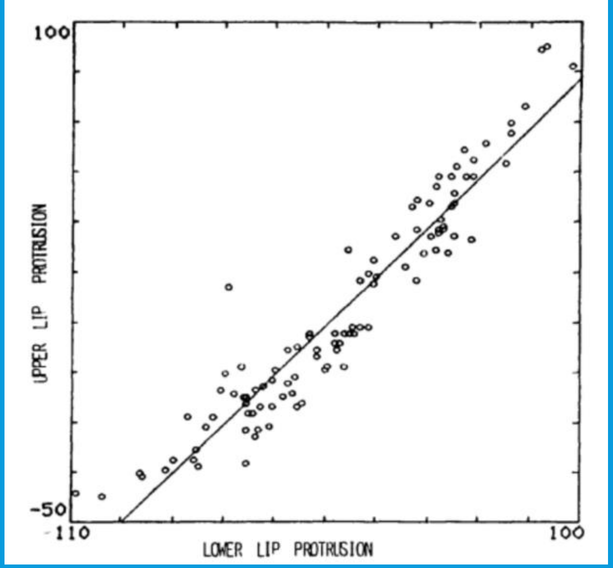


Figure 1: Lip Area vs Lip Width

Figure 2: Lip Area vs Lip Width*Lip Height

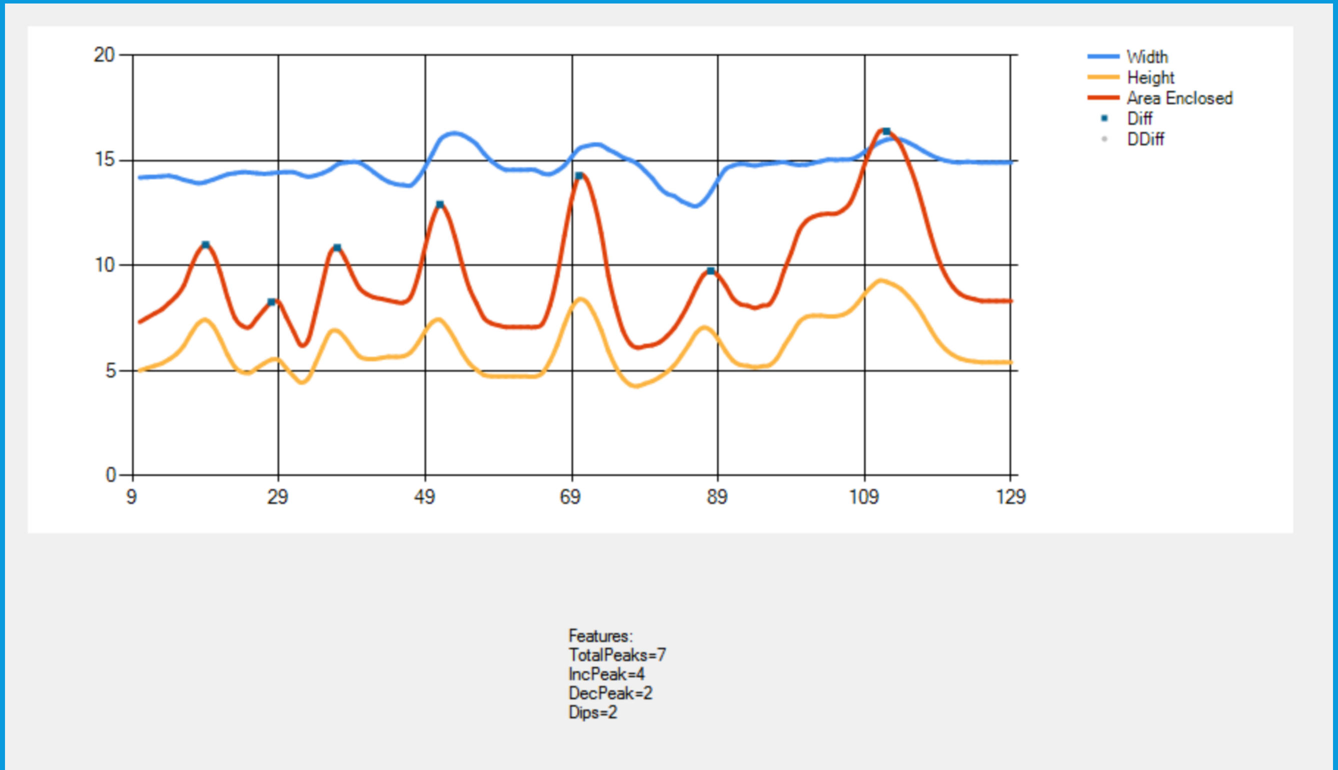
Figure 3: Lip Area vs Lip Height

Figure 4: Upper Lip Protrusion vs Lower Lip Protrusion

Why do we need to study these???

About my ongoing research

- Similarity in lip movements as a secondary authentication.



```
Features:
TotalPeaks=7
IncPeak=4
DecPeak=2
Dips=2
```

Future

- Rumor - Kinect 2 with Lip Tracking.
- More authentication based tools using lip movements.
- Current state of the best lip reading software's is very bad. The best ones have about 40% accuracy.
- More human-like avatars.

References

- 3D Models of the Lips for Realistic Speech Animation - Thierry Guiard-Marigny et. Al. Computer Animation (1996) 80-89.
- C. Abry and Bo¨e L.J. Laws for lips. Speech Communication, 5:97–104, 1986.
- .M. Cohen and D. Massaro. Synthesis of visible speech. Behavior Research Methods, Instruments, & Computers, 22(2):260–263, 1990.
- C. Xu and J.L. Prince, ``Gradient Vector Flow: A New External Force for Snakes,'' Proc. IEEE Conf. on Comp. Vis. Patt. Recog. (CVPR), Los Alamitos: Comp. Soc. Press, pp. 66-71, June 1997.
- C. Xu and J. L. Prince, "Gradient Vector Flow Deformable Models", Handbook of Medical Imaging, edited by Isaac Bankman, Academic Press, September, 2000.
- Tarcisio Coianiz, Lorenzo Torresani, and Bruno Caprile. 2D Deformable Models for Visual Speech Analysis". In NATO Advanced Study Institute: Speechreading by Man and Machine, 1995.

Thank You ☺

धन्यवाद

謝謝

Merci

Спасибо

Danke

Gracias