

# Model-based Synthesis of Visual Speech Movements from 3D Video

James Edge, Adrian Hilton, and Philip Jackson  
{j.edge,a.hilton,p.jackson}@surrey.ac.uk  
The University of Surrey



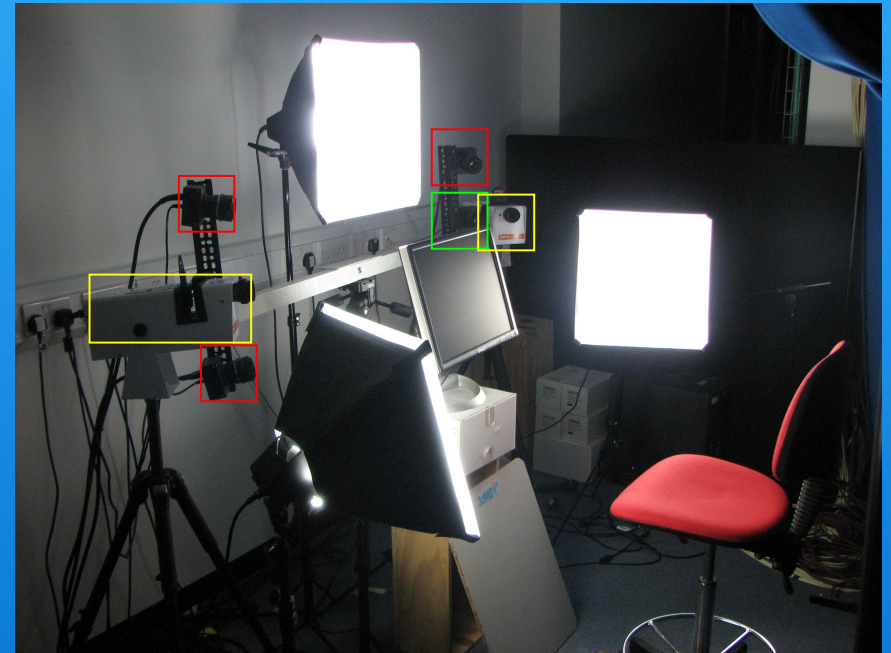
Dynamic Face research at Surrey was funded with a grant from the EPSRC.

# Introduction

- Synthesising facial dynamics is regarded as one of the most difficult problems in animation.
- In particular this is the case for speech, where facial dynamics has a direct relationship with the sounds which we hear.
- In this work we use dynamic capture techniques to recover natural speech movements and use these in our synthesis (i.e. data-driven.)
- Data-driven visual speech synthesis is typically performed using two main techniques:
  - Model-based synthesis (HMMs, neural nets)
  - Unit Selection/Concatenative synthesis
- Model-based techniques are generic, but generate motions according to the statistical mass of data - losing the dynamic detail of the original data (e.g. Voice Puppetry, Brand 1999.)
- Unit Selection techniques maintain the detail in the original motion but typically do not use the audio dynamics to generate the output animation (e.g. Video Rewrite, Bregler 1997.)
- In our work we attempt to fuse the properties of both techniques to combine the advantages of both techniques.

# Data Capture

- Stereo capture technology (3dMD) is used to capture 3D surface detail of a speaker at 60 Hz.
- Database consists of 8mins of English sentences from the TIMIT dataset.
- Sentences are chosen to provide coverage across all English phones.
- Only a single speaker has been captured.
- Data is soon to be made available for download.



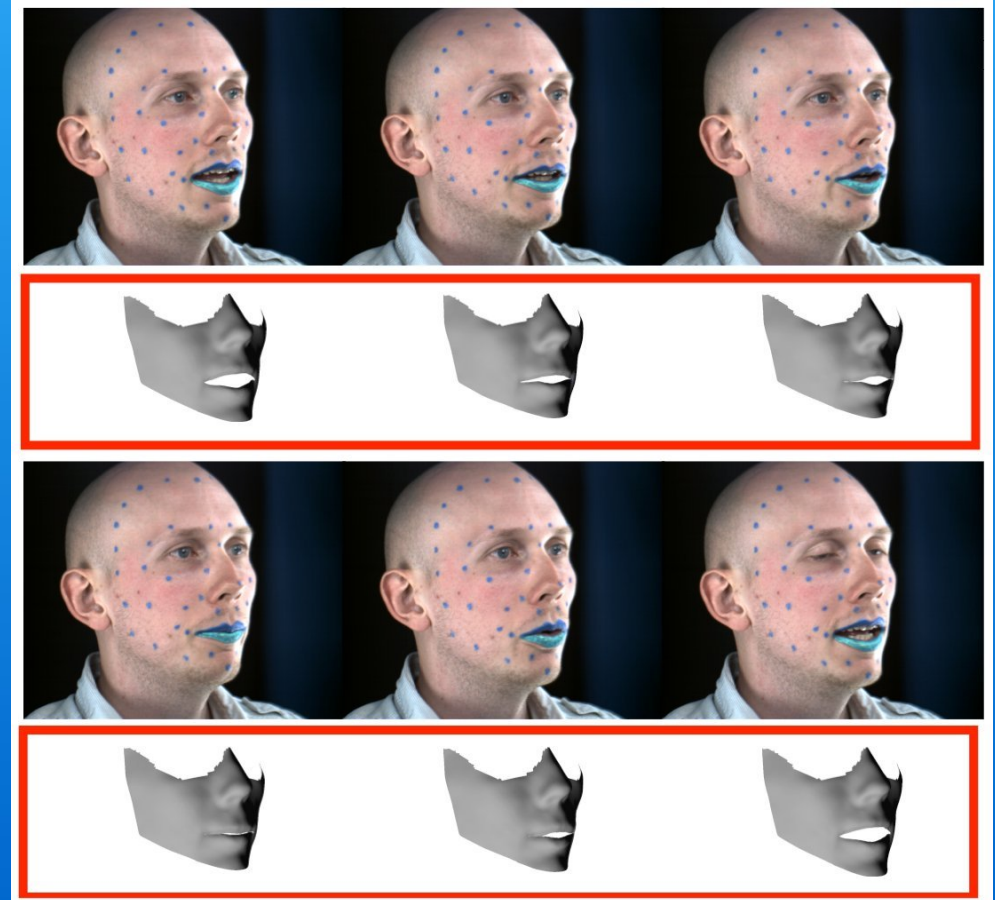
IR stereo camera

IR projector

texture camera

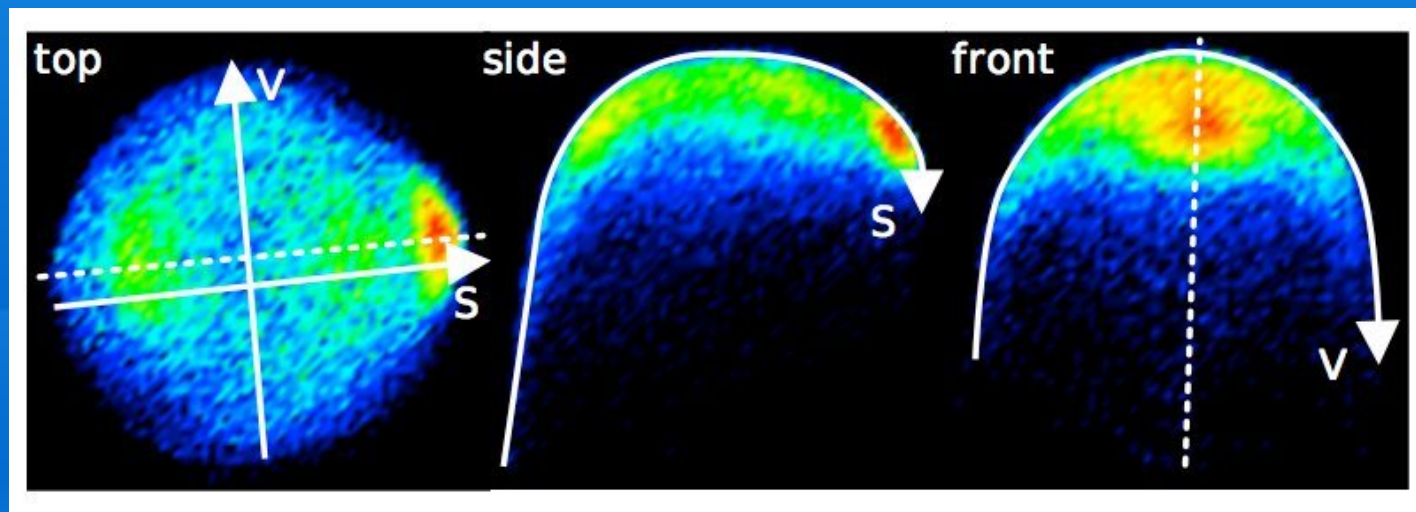
# Data Processing

- Captured 3D geometry is initially unregistered.
- To correct this markers are tracked using the captured texture images.
- A surface registration technique is used to align the surface frames between the markers.
- Once the frames are registered standard data compression techniques, such as pca, can be applied to reduce the size of the dataset.



# Data Parameterisation

- The structure of our state-based model is based upon a dynamic parameterisation of lip states.
- The dimensionality of our data is reduced by applying multi-dimensional scaling to a matrix containing both the position and velocity of the lips.
- By clustering in this space we explicitly maintain the dynamic structure of lip movements in our model.
- An in-depth discussion of speech parameterisation can be found in our AVSP'08 paper.

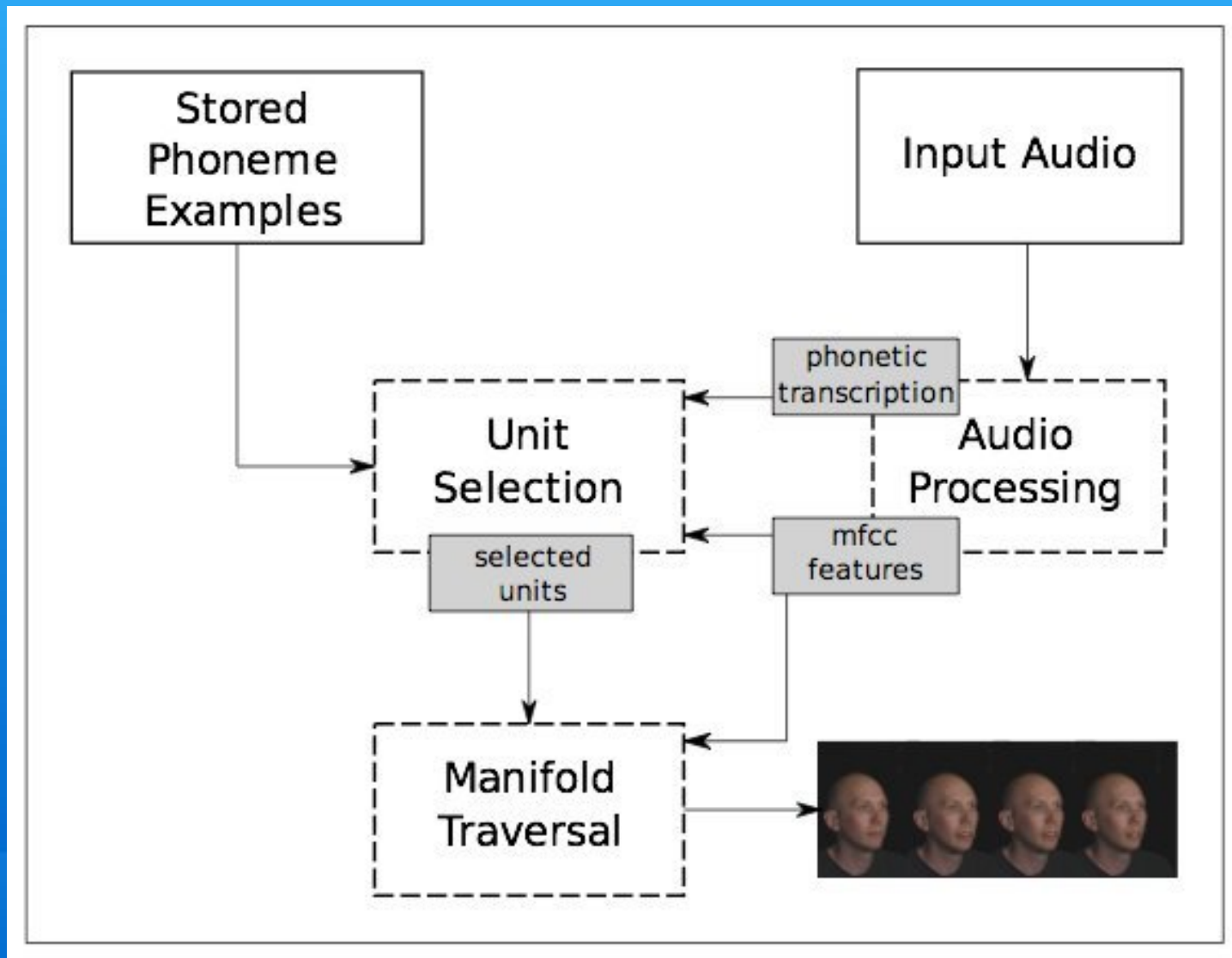


# Synthesis Overview

Synthesis proceeds with the following steps:

- Input audio is segmented into phonemes
- For each phone segment the best unit is selected from the stored dataset.
- A state-based model is trained using the selected units.
- The model is used to align the input audio to the selected units resulting in a sequence of visual states.
- By fitting a trajectory to these discrete states we produce a smooth animation of the lips and jaw.
- The output of the synthesis process is a 3D surface of the lips and jaw - effectively high resolution mocap for the lower face.
- Teeth/jaw movement is currently inferred from the motion of the chin.
- Tongue movement is not currently accounted for in the system.

# Synthesis Overview



# Unit Selection

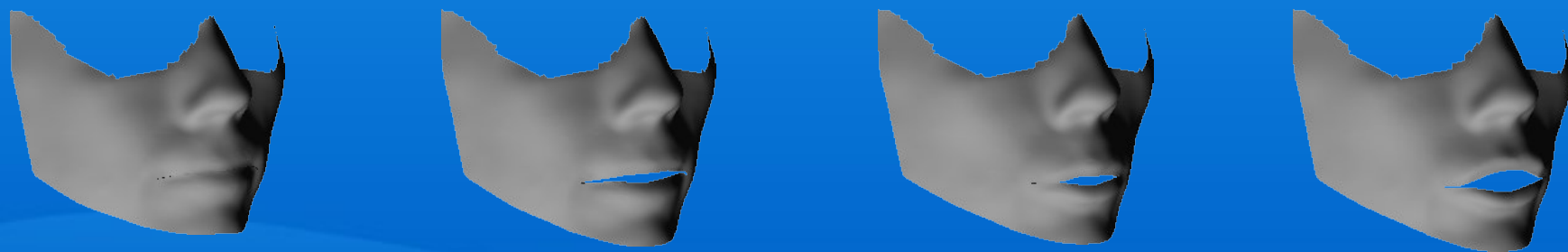
- In our system a phone unit is considered to be the sequence of frames from the audio centre of the previous phone to the audio centre of the next.
- This is similar to a triphone, but is only categorised according to the central unit (i.e. not according to its context.)
- For each phonetic unit in the dataset we store the audio parameters (MFCCs) and the sequence of visual states that the motion passes through.
- For each phone unit in the input sequence an optimal unit is selected from the dataset by comparing its audio parameters with each of the exemplars in the dataset.
- We use Dynamic Time Warping to compute the aligned similarity between the input and stored phonemes.
- The output of this process is a set of selected phone units from the dataset.
- The selected units overlap, so an optimal trajectory must be determined to produce a final animation.

# State-based Trajectory Synthesis (I)

- We use a state-based model to generate trajectories from selected units.
- Each state is determined by clustering data according to our dynamic parameterisation of lip movements.
- States are connected if they are found sequentially in our stored dataset.
- This structure is used to constrain the dynamics of the output speech trajectory.
- Each of the selected units can be seen as a selection of a part of this model.
- The transition between two selected phone units can be seen as a union operation between the states relevant to each individual unit.
- In our model the synthesis of a speech trajectory is constrained by the identification of the relevant parts of the state-model that can be traversed as we transition between phone units in the sequence.

# State-based Trajectory Synthesis (II)

- The output sequence of states is determined using a dynamic programming algorithm.
- Instead of using a probabilistic framework, we accumulate distance between the input audio and the audio for the selected units and choose the smallest difference path for output.
- The output of our synthesis is a sequence of discrete states each corresponding to a distribution of lip shapes and velocities.
- These are transformed into a continuous trajectory using Brand's (Voice Puppetry, 1999) method which finds a maximum likelihood path through the identified states.



# Animation

- The output of synthesis is a sequence of 3D meshes.
- These can be used as a form of high resolution motion-capture to animate a 3D face model.
- For the work here we animate 2D images using an image warping technique.

(link speech animations on this slide)

# Conclusions

- We have introduced a method for visual speech synthesis by combining unit selection/concatenative and model-based approaches.
- Unit selection is used to predefine the regions of the model which can be traversed.
- A dynamic programming algorithm is used to traverse the model and determine an optimal sequence of states according to the input audio.
- The lip-shape distributions for model states imply the movement of the lips and jaw.
- The model is built using dynamic data captured using a commercial stereo-capture rig.
- The data for this work will soon be made available for distribution.

# References

*J.D. Edge, A. Hilton, and P. Jackson, Parameterisation of 3D Speech Lip Movements. Proceedings of AVSP'08, 2008.*

*M. Brand, Voice Puppetry. Proceedings of SIGGRAPH'99, 1999.*

*C. Bregler, Video Rewrite: Driving Visual Speech with Audio. Proceedings of SIGGRAPH'97, 1997.*

dynamic faces research website:

[http://www.ee.surrey.ac.uk/CVSSP/page\\_4095](http://www.ee.surrey.ac.uk/CVSSP/page_4095)