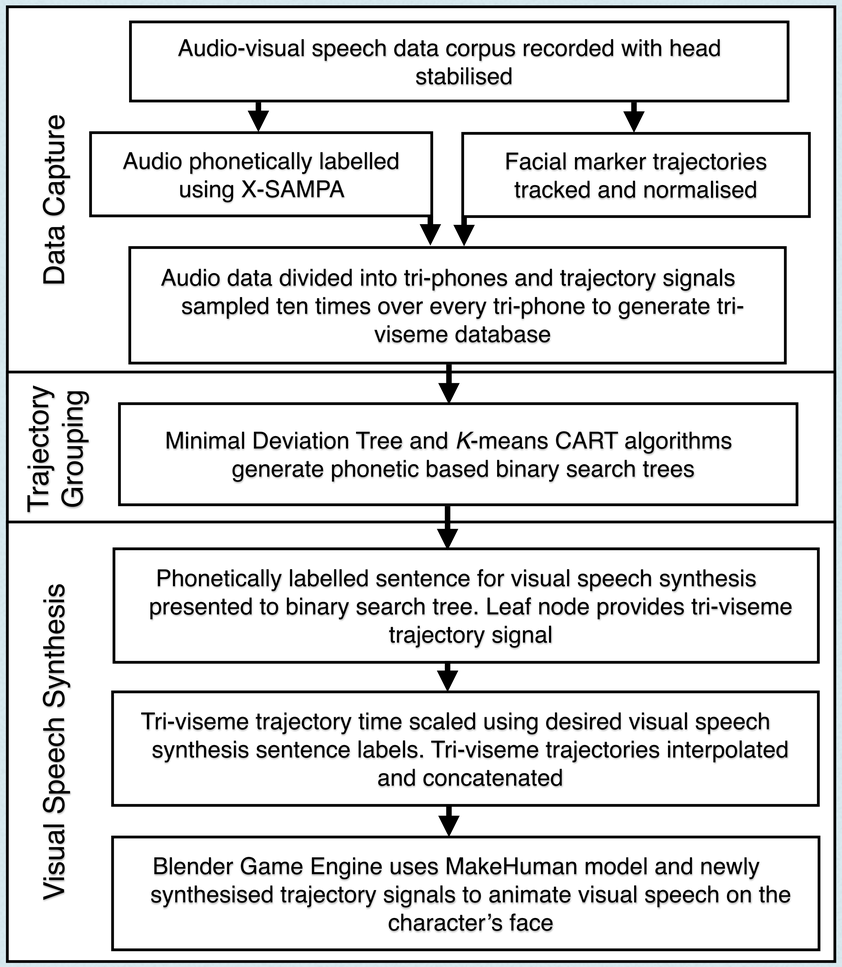
Improved Visual Speech Synthesis using Dynamic Viseme *k*-means Clustering and Decision Trees

Christiaan Rademan and Thomas Niesler
Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
christo@ml.sun.ac.za, trn@sun.ac.za

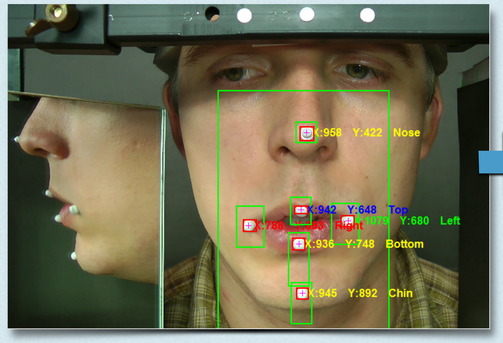
1. Introduction

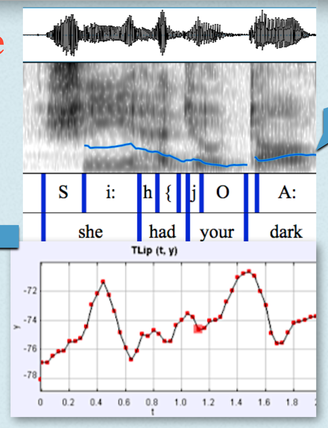
- Dynamic viseme trajectory morphologies grouped using k-means clustering in decision-trees.
- Dynamic visemes defined by tri-phone boundaries of tracked oral feature trajectories.
- Training requires very small dataset of phonetically-annotated audiovisual speech.
- Only consumer available video equipment required.
- Only open-source software components (MakeHuman & Blender).

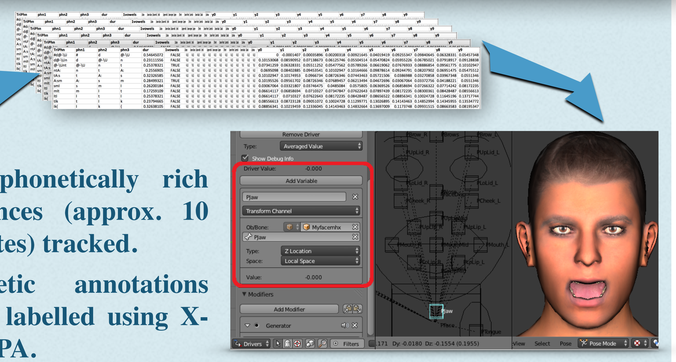
2. Visual Speech Synthesis Pipeline



2.A. Visual Speech Data Capture



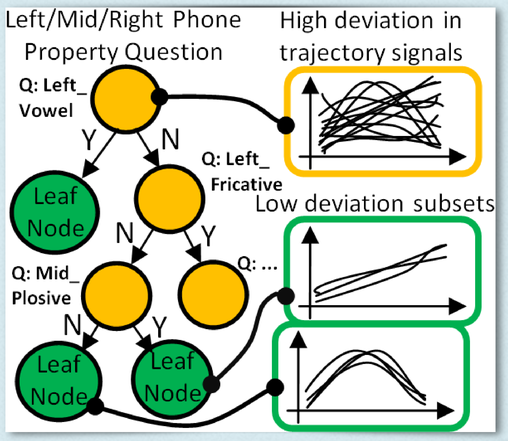


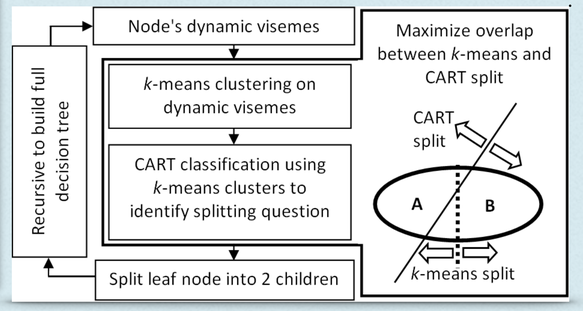


- 120 phonetically rich sentences (approx. 10 minutes) tracked.
- Phonetic annotations hand labelled using X-SAMPA.
- Data driven scripted bone-driven shape key animation.

2.B. Trajectory Clustering Using Decision Trees

- Previous works question all phonetic labels and select the question whose subsets contained the most homogeneous static visemes.

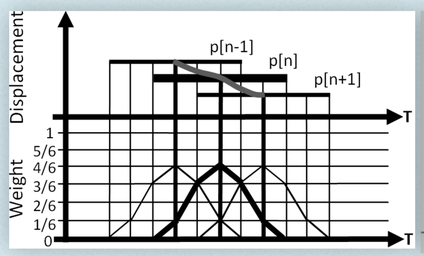


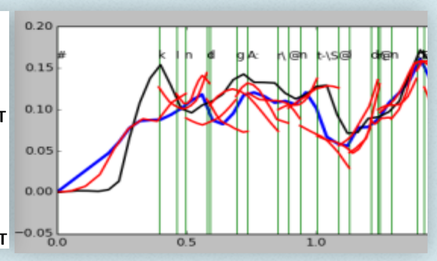


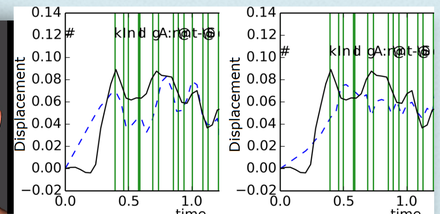
- We first group similar visemes, then ask which phonetic attribute best reproduces this split.
- *k*-means clustering first assigns trajectories to two classes, than a CART finds the phonetic attribute which splits the data into subsets that best match the two classes.

2.C. Dynamic Viseme Selection & Concatenation

- Decision trees traversed and referrers corresponding mean dynamic viseme from leaf node.
- Dynamic visemes of successive tri-phones interpolated and concatenated.





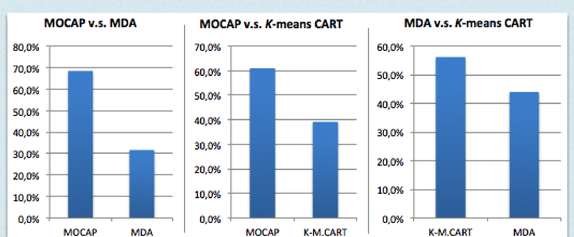


The left and right images and graphs use k-means and minimum deviation decision trees, respectively.

Original chin trajectory is marked by the solid black line. Vertical green lines indicate phone boundaries.

3. Results

- 40 test participants each evaluated 12 test sentences.
- Perceptual tests showed clear improvement over baseline.



FAAVSP - The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing
11-13 September, 2015 Vienna, Austria.