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

- Audio-visual speech data corpus recorded with head stabilised.
- Audio phonetically labelled using X-SAMPA.
- Facial marker trajectories tracked and normalised.
- Audio divided into tri-phones and trajectory signals sampled ten times over every tri-phone to generate tri-viseme database.

2.B. Trajectory Clustering Using Decision Trees

- Minimal Deviation Tree and K-means CART algorithms generate phonetic based binary search trees.
- Phonetically labelled sentence for visual speech synthesis presented to binary search tree. Leaf node provides tri-viseme trajectory signal.
- Tri-viseme trajectory time scaled using desired visual speech synthesis sentence labels. Tri-viseme trajectories interpolated and concatenated.

2.C. Dynamic Viseme Selection & Concatenation

- Blender Game Engine uses MakeHuman model and newly synthesised trajectory signals to animate visual speech on the character’s face.

3. Results

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