# Automatic Audio Driven Animation of Non-Verbal Actions

**D. Cosker[1], C. Holt[2], D. Mason[2], G. Whatling[2], D. Marshall[3] and P. L. Rosin[3]**

[1] Media Technology Research Centre, University of Bath. *D.P.Cosker@cs.bath.ac.uk*
[2] School of Engineering, Cardiff University. *Holt, MasonD, Whatlinggm@cardiff.ac.uk*
[3] School of Computer Science, Cardiff University. *Dave.Marshall, Paul.Rosin*@cs.cf.ac.uk

**Keywords:** Non-Verbal, HMM, Animation, Motion-Capture

## 1 Introduction

While speech driven animation for lip-synching and facial expression synthesis from speech has previously received much attention [1, 2], there is little or no previous work on generating non-verbal actions such as laughing and crying automatically from an audio signal. In this article initial results on a system designed to address this issue are presented.

## 2 System Overview

Figure 1 gives an overview of our current system. 3D facial data was recorded for a participant making different actions – i.e. laughing, crying, yawning and sneezing – using a Qualysis (Sweden) optical motion-capture system while simultaneously recording audio data. 30 retro-reflective markers were placed on the participant's face to capture movement. Using this data, an analysis and synthesis machine was then trained consisting of a dual-input Hidden Markov Model (HMM) and a trellis search algorithm which converts HMM visual states and new input audio into new 3D motion-capture data.

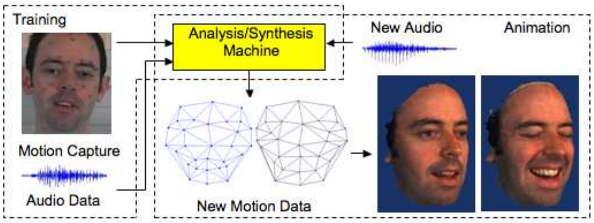

***Figure 1: New motion-capture animations are created automatically from new audio-recordings. This data may then drive a more detailed 3D facial model.***

## 3 Analysis and Synthesis

After normalising the data with respect to head pose variation, its dimensionality is then reduced using PCA. Audio is represented using Mel Frequency Cepstral Coefficients (MFCC). A HMM is trained using the visual features and a dual mapping created from each HMM state to the audio features. This allows a new HMM visual state sequence to be created given novel input audio. Given a new visual HMM state sequence, a 3D facial motion-capture output is created at each state (and each time $t$) by finding the visual feature with the best matching corresponding audio feature. In this process, features which are dissimilar to the feature selected at time *t-1* are also penalised.

## 4 Results and Discussion

A participant was recorded making approximately ten repetitions of a particular action, i.e. laughing, crying, yawning and sneezing. A synthesis machine was then trained using combined observation data from all actions. Using a leave-one-observation out strategy, we tested the model's ability to resynthesise actions solely from previously unobserved audio.

Synthesised motion-capture animation results show an excellent correlation to new audio data. This is further supported by the low RMS errors (in millimetres) calculated by comparing synthetic animations with their ground truth (see Table 1). In order to find a stable number of HMM states, and to test repeatability in light of the random initialisation of HMMs [1], the experimental set-up was repeated multiple times for various state numbers. A one-way ANOVA showed that repeated trails given 30 or more HMM states resulted in consistently strong results with low RMS errors (i.e. at a chance level of $p < 0.05$).

| No. HMM States | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|
| RMS Error (mm) | 2.21 | 2.29 | 2.35 | 2.28 | 2.41 |

***Table 1: The effect of varying HMM states on the RMS reconstruction error (in millimetres).***

## 5 Conclusions and Future work

Initial results are presented on a system for synthesising non-verbal actions automatically from speech. Testing reveals an excellent correlation between synthesised motion-capture and new audio data. The model is currently being extended to address person independence and mappings are under development between motion capture and dynamic 3D facial models.

## Acknowledgements

The authors would like to thank the Royal Academy of Engineering and the EPSRC for partially funding this work.

### References

[1] M. Brand. Voice puppetry. In *proc. of ACM SIGGRAPH*, pages 21–28, 1999.

[2] Y. Cao et al. Expressive speech-driven facial animation. *ACM Trans. Graph.*, 24(4):1283–1302, 2005.