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]
Cardiff School of Engineering, Cardiff University, UK

{1} Media Technology Research Centre, University of Bath
{2} School of Engineering, Cardiff University [Holt@Cardiff.ac.uk]
{3} School of Computer Science, Cardiff University

INTRODUCTION

•Quick development of facial animations to meet the demands of the entertainment industry is an important issue.
•Speech driven animation for lip-synching and facial expression synthesis has received much attention [1, 2]
•Generating non-verbal actions such as laughing and crying automatically from an audio signal has been ignored
•Initial results on a system designed to address this issue are presented.

SYSTEM OVERVIEW

•Figure 1 gives an overview of our current system.

•3D facial data and audio was recorded for different actions -- i.e. laughing, crying, yawning and sneezing.

•30 retro-reflective markers were used to capture facial movement (see Figure 2).

•An analysis/synthesis machine based on HMMs was trained.

•Animation output is 3D mo-cap data. This may be used to animate a more detailed facial model (see Figure 1)

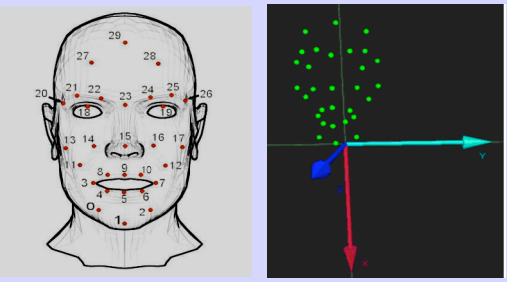

Figure 2. Marker Naming Protocol and 3D QTM (Qualisys, Sweden) software view

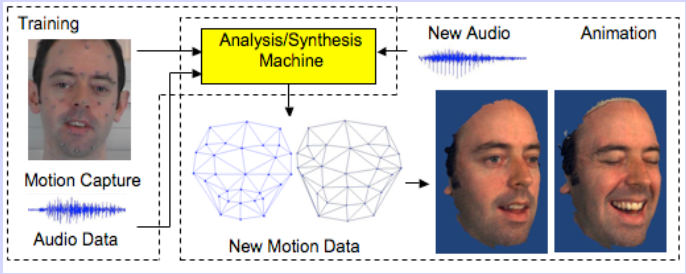

Figure 1. New motion-capture animations are created automatically from new audio-recordings. This data may then drive a more detailed 3D facial model.

ANALYSIS AND SYNTHESIS

•Mo-cap data is normalised and PCA performed

•Audio is represented using Mel Frequency Cepstral Coefficients (MFCC).

•Audio/Visual correlations are modelled using a dual-input HMM

•New input audio creates a visual HMM state sequence

•This sequence is converted into a smooth visual output

RESULTS

Synthesised motion-capture animation results show a strong correlation to new audio data (see Table 1).

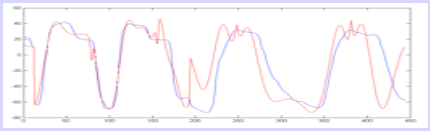

Figure 3. Synthetic (red) versus Ground Truth (blue) animation parameters

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 $p<0.05$).

Table 1. HMM states versus RMS error (mm).

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

DISCUSSION, CONCLUSIONS and FURTHER WORK

• Testing reveals a strong correlation between synthesised motion-capture and new audio data.

• Experiments using other audio features (e.g. pitch and LPC) do not appear to yield a significant advantage over using MFCCs alone.

• The model is currently being extended to address person independence and mappings are under development between motion capture and dynamic 3D facial models.

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.

ACKNOWLEDGEMENTS

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