Comparison of Techniques for Audio Driven Facial Animation

Benjamin Havell*, David Marshall , Yulia Hicks , Paul Rosin , Saeid Sanei , Andrew Aubrey
Cardiff University

1 Introduction

A fast and accurate Audio Driven Facial Animation system would have many applications such as video games and television, where animation is not currently feasible due to cost or time restraints.

One technique that has previously been used for Audio Driven Facial Animation is to build a joint audio-visual model using Active Appearance Models (AAMs) to represent possible facial variations and Hidden Markov Models (HMMs) to select the correct appearance based on the input audio only[Cosker 2006]. However there are several questions that remained unanswered. In particular the choice of clustering technique and the choice of the number of clusters in HMM may have significant influence over the quality of the produced videos.

In this work we have investigated a range of clustering techniques in order to improve the quality of the HMM produced.

There are several complexities in audio-visual speech that need to be addressed. These include coarticulation, the effect neighbouring phonemes have on each other, and the many to one correspondence between phonemes and visemes. In this work, a public dataset of 300 phonetically labeled sentences [Theobald et al. 2008] spoken by a single person was used to build an AAM and conduct our tests.

2 Results

Our Active Appearance Model was built using 44,000 frames from 200 of the source sequences, with 110 facial landmarks identified for each frame, 32 of them around the mouth. After shape normalisation and PCA the 10 largest PCA parameters were retained as they contained over 98% of the energy. The corresponding audio data was sampled at 44100 Hz and parameterised using 13 mel frequency cepstal coefficients (MFCCs). Finally the audio-video dual HMM model was built in the joint audio-video space. This joint model is used to produce photorealistic videos from audio only input as described by D.Cosker [Cosker 2006].

Next we tested different clustering techniques as part of an HMM framework for generating videos from an audio input. In all approaches the same AMM was used and the data was clustered either using a representative frame from each phoneme (for Dual input HMMs) or using all the data (for the CHMM) as discussed below.

In assessing the results of our three approaches we measured the RMS error in shape normalised pixel values (pixel error) compared to the ground truth images for the 156 frame sequence used.

2.1 Phoneme Based Clustering

In phoneme based clustering, we represent each phoneme with a mixture of gaussians trained on a number of audio video phoneme samples. The number of gaussians used to represent each phoneme is chosen heuristically based on the number of data points found for that phoneme up to a maximum of 5 in our experiment. This way we were able to represent the data distribution for each phoneme closely with the total number of gaussians in the combined mixture model equal to 153. Using this approach we found pixel errors of 6.41.

2.2 Automatic Clustering

In order to carry out a direct comparison with the phoneme labeled method EM clustering was used, with the total number of gaussians set to the value of 153. Measurements were repeated 5 times to allow for differences in random initialisation states, and using this method we found parameter error values of 394.6 and pixel errors of 6.63. This experiment was also repeated using 40 clusters following the method outlined by D.Cosker [Cosker 2006], which gave average pixel errors of 6.59 over a total of five runs.

2.3 Coupled HMM

Coupled HMMs are very powerful and useful for audio visual computing due to their ability to handle asynchrony but are limited in size by the computer power available. We trained a number of CHMMs with different numbers of clusters on our data. The number of clusters ranged between five and ten due to memory limitations, with ten producing the best results. Using a very simple CHMM based on the first principal component of both the audio and visual data and with 10 clusters for both visual and audio HMMs we found pixel errors of 5.39.

3 Conclusion

In our experiments a CHMM produces the lowest errors for pixel to pixel comparison and a mixture of Gaussians per phoneme produced lower errors than automatic clustering for same number of clusters.

We conclude that CHMM is the most promising technique although at present it is limited by available computing power.

References

COSKER, D. 2006. *Animation of a Hierarchical Appearance Based Facial Model and Perceptual Analysis of Visual Speech*. PhD thesis, Cardiff University.

THEOBALD, B., FAGEL, S., BAILLY, G., AND ELISEI, F., 2008. LIPS2008: visual speech synthesis challenge. http://hal.archives-ouvertes.fr/hal-00333655/fr/, Oct.

*e-mail: havellb@cf.ac.uk

FAA 2010, Edinburgh, United Kingdom, October 21, 2010.
ISBN 978-1-4503-0522-8/10/0010