A HYBRID PHONEME BASED CLUSTERING APPROACH FOR AUDIO DRIVEN FACIAL ANIMATION

Benjamin Havell[⋆†] *Paul L. Rosin*[⋆] *Saeid Sanei*[†] *Andrew Aubrey*[⋆] *David Marshall*[⋆] *Yulia Hicks*[†]

[⋆] School of Computer Science and Informatics, Cardiff University
[†]School of Engineering, Cardiff University

ABSTRACT

We consider the problem of producing accurate facial animation corresponding to a given input speech signal. A popular technique previously 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.

However there are several questions that remained unanswered. In particular the choice of clustering technique and the choice of the number of clusters in the HMM may have significant influence over the quality of the produced videos.

We have investigated a range of clustering techniques in order to improve the quality of the HMM produced, and proposed a new structure based on using Gaussian Mixture Models (GMMs) to model each phoneme separately. We compared our approach to several alternatives using a public dataset of 300 phonetically labeled sentences spoken by a single person and found that our approach produces more accurate animation.

In addition, we use a *hybrid* approach where the training data is phonetically labeled thus producing a model with better separation of phonemes, but test audio data is not labeled, thus making our approach for generating facial animation less laborious and fully automatic.

1. INTRODUCTION

A robust and practical approach for the automatic animation of a 2D facial model a dynamic model which can convincingly reproduce the full range of motions seen in a real face [1] is addressed here. It is a problem of great interest due to its wide range of applications, including animation, communication, medical visualisation and psychology.

In particular, we are interested in Audio Driven Facial Animation, which is often achieved by building a joint audio-visual model using Active Appearance Models (AAMs) [2, 3] to represent possible facial variations, Mel-frequency cepstral coefficients (MFCCs) to represent the audio and Hidden Markov Models (HMMs) to select the correct appearance based on the input audio.

The main problem to be solved here is to produce a desired sequence of visemes given an input audio sequence. A viseme is defined as a particular facial pose or movement that occurs during the vocalisation of a phoneme, where phonemes are the smallest units of speech which are combined to form words.

There are two main approaches to solving this problem. The more traditional approach is to label the incoming audio sequence with a sequence of corresponding phonemes, and then to make a facial model to interpolate between the visemes associated with the phonemes [4, 5]. The main problems associated with this approach are the fact that it can't model non verbal utterances and expressions, and also the amount of time and effort it takes to accurately label the audio, which also precludes it from being used in real time applications.

The second approach is to use the audio features to generate the sequence of visemes automatically. However, this approach may lead to imprecisions in the model during the training stage due to incorrect viseme attribution.

To alleviate this problem, we follow a *hybrid* approach where the training data is phonetically labeled thus producing a model with better separation of phonemes, but test audio data is not labeled, thus making our approach for generating facial animation less laborious and fully automatic.

The main problems associated with our approach are in creating a system that can accurately model the relationships between phonemes and visemes, and also smoothly animating the transitions between visemes. This article investigates the accuracy of different models representing the relationship between phonemes and visemes, and proposes a new structure based on GMMs, where each phoneme is modelled by a number of mixtures which is automatically estimated to be optimal. We subsequently show this method outperforms standard approaches.

An overview of the basic approach that we use is as follows, the first step is to build an Active Appearance Model which allows the various facial poses to be represented using a small number of parameters. The audio is paramterised using MFCCs and PCA. In order to use the link between appearance and sound to animate the face, Hidden Markov Modelling is used to find the most likely sequence of appearance states

given the audio and video training data. The set of possible states is found using cluster analysis. In order to improve the accuracy of AAMs, hierarchical models [6, 7] can be used to model the local variation in specific regions of the face which are then recombined to produce the final reconstructed image.

In this work we have investigated a range of clustering techniques in order to improve the quality of the HMM produced, and proposed a new structure based on using GMMs to model each phoneme to solve the problem of multiple possible visemes representing each phoneme.

2. ACTIVE APPEARANCE MODELS

AAMs are a generalisation of Active Shape Models (ASMs) which allow the variation in the appearance of complex objects to be described using a small number of parameters [2, 8]. AAMs are built using a sequence of images and set of associated landmark points. The landmarks represent the shape of the object and the pixel values provide the visual texture. The number of landmarks is kept fixed for all frames and each landmark corresponds to a particular place on the face in all frames.

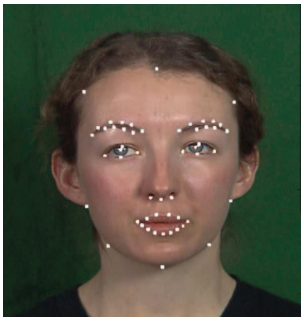

Fig. 1. Example of facial landmarks

In order to efficiently model facial motion during speech it is necessary to reduce the dimensionality of this shape data without losing the possible range of variation. A common approach to achieve this is to use Principal Component Analysis (PCA) [9].

Using PCA the shape and texture can be represented and reconstructed using a small number of parameters. The Audio MFCCs are also further reduced in dimensionality by using PCA, the parameterised audio and images are then used to build an HMM model.

3. HIDDEN MARKOV MODELS

HMMs are a method of statistical modelling [10] where the state of a system can be hidden and calculated using the observed outputs. They can be used for a wide variety of classification and recognition tasks and are particularly effective for tasks involving temporal pattern recognition, such as speech recognition or object tracking. A HMM is a two layer model where there are a number of hidden system states and a set of possible observations which are dependent on the system state. Each possible observation has probabilities associated with it representing each system state, and each state has associated probabilities of the system staying in its current state or changing to every other state.

HMMs are used in speech recognition where the observations are a sequence of parameterised audio features such as MFCCs and the hidden states are the underlying phonemes being uttered. As proposed by Brand [11] the output of an audio HMM can then be used as the input to an appearance HMM to model the appearance and coarticulation of a facial model. Dual-input HMMs first used by Brand [12], are a variation of standard HMM that, in this application, allow for the estimation of hidden appearance states based on an audio data input. Dual-input are constructed by first building an HMM on one set of data then calculating means and covariances for a second data set for each HMM state in the first model. In our application this involves using the transition, priors and observation probabilities from the appearance model but with means and covariances derived from the audio data. This allows the audio to be used to estimate the appearance state sequence using the Viterbi algorithm.

A Coupled Hidden Markov Model (CHMM) [13] is two HMM chains linked by conditional probabilities that span time steps, this allows for asynchronous progression of the chains. This asynchrony is highly desirable in audio-visual speech modelling as many sounds can be heard before its appearance can be detected in the face and vice-versa.

4. CLUSTERING APPROACHES

One of the standard approaches to creating a model for capturing the relationship between visemes and phonemes is to use AAMs in combination with HMMs for this purpose. Cosker et al[7, 6] followed this approach using K-means clustering, with the number of Gaussians chosen manually.

Although it is desirable to have an automatic training process it is impossible to ensure that each phoneme is separated from the other phonemes in the model. When the phonemes are not well separated in the feature space, it may lead to the system classifying inputs incorrectly and thus selecting an incorrect viseme for the output.

Brand [11] followed a different approach where both the training data for the model and the test data are manually labeled with the corresponding phoneme. Needless to say this is a time consuming process which we would like to avoid.

In this article we follow a *hybrid* approach where the training data is phonetically labeled thus producing a model with better separation of phonemes, but test audio data is not labeled, thus making our approach for generating facial animation less laborious and fully automatic.

Initially we trained a single Gaussian on the data for each phoneme. This produced unsatisfactory animations on test data, and as a novel solution to the above, we have represented each phoneme with a GMM, where the number of clusters was selected automatically using the Affinity Propagation algorithm [14]. This is a hybrid of Phoneme based clustering and automatic clustering within those phonemes and allows for a flexible number of clusters being used. In this way phonemes with a large degree of variation can be represented with more clusters in order to improve their accuracy. As we show in the experiments, this approach produced the most accurate facial animations.

5. RESULTS

In our work, we used a public dataset of 300 phonetically labeled sentences [15] spoken by a single person. 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 described the inner and outer mouth shape. Each frame was 512 by 512 pixels.

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 MFCCs, this was then further reduced in dimensionality using PCA to 10 . Finally the audio-video dual HMM model was built.

Next we tested our approach as part of an HMM framework for generating videos from an audio input. In assessing the results of our approach we used the RMS error in shape normalised pixel values (pixel error) compared to the ground truth images for the 750 frames of 5 sequences used. This is shown in equation (1) where $\mathbf{x_1}$ is the ground truth and $\mathbf{x_2}$ is the reconstructed pixels for n pixels.

$$\text{pixel error} = \sqrt{\frac{\sum_{i=1}^{n} (\mathbf{x}_{1,i} - \mathbf{x}_{2,i})^2}{n}} \quad (1)$$

Using our approach 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 an average pixel errors of 6.41.

In this section, we compare our approach to that followed by Cosker et al [7]. The model was built using the same 44,000 frames of training data as our proposed method (described above), and 20 test sentences totalling 3100 frames. In order to carry out a direct comparison EM clustering was used, with the total number of Gaussians set to match the number Gaussians used in our method. Measurements were repeated 5 times to allow for differences in random initialisation states, and using this method we found a pixel errors of 6.64. This experiment was also repeated using 40 clusters used by Cosker et al in their work [7], which gave average pixel errors of 6.58. Thus our method produces lower pixel error measure. We believe our method produces better results than the original method by Cosker because the GMM clusters have information about the phonemes being spoken used during clustering, rather than a blind approach.

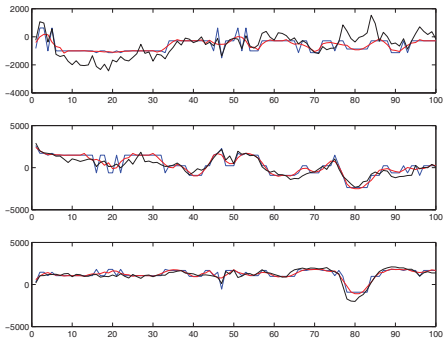

Fig. 2. First three AAM Parameters; original data (blue), reconstructed data (red), and smoothed reconstruction (black)

Another approach we investigated in our research to generate facial animations is based on Coupled HMMs. Coupled HMMs are very powerful and useful for audio visual computing due to their ability to handle asynchrony. We trained a number of CHMMs with different numbers of clusters on our data, but due to computational demands of the technique we were not able to successfully build a model with more than 10 states. Due to this limitation we were not able to use phonetically data but instead relied on automatic clustering technique for the whole dataset. As the total of clusters was smaller than the number of phonemes in the dataset, we were not able to carry out a direct comparison with phoneme based methods, but using 10 clusters we found a pixel error of 6.62.

In order to test the statistical significance of our results we used the analysis of variance (ANOVA) [16] test on the error values for all results of all the methods tested. This test gave a significance value of 0.001 for the full set of results which is below the 0.05 value commonly used to determine statistical significance and hence proves its statistical significance.

Fig. 3. (a) an original frame, and (b) a frame generated using our method (c) a frame generated using coskers method

Error Measure	Cosker's method 40 clusters	Cosker's method 153 clusters	CHMM	**Our Phoneme based GMMs**
Average Pixel error	6.58	6.64	6.62	**6.41**

Table 1: AAM Model error measurements

6. CONCLUSIONS

In this article, we considered the problem of producing accurate facial animation corresponding to a given input speech signal. We investigated the accuracy of different models representing the relationship between phonemes and visemes, and proposed a new structure based on GMMs, where each phoneme is modelled by a number of mixtures which is automatically estimated to be optimal.

In addition, we followed a *hybrid* approach where the training data is phonetically labeled thus producing a model with better separation of phonemes, but test audio data is not labeled, thus making our approach for generating facial animation less laborious and fully automatic.

We found that using our *hybrid* approach to modelling labeled phonemes with GMMs decreases the error of the resulting animation compared to existing techniques. CHHMs may be another promising technique for our application but is currently limited by available computational power.

7. REFERENCES

[1] J. Beskow and S. Al Moubayed, "Perception of nonverbal gestures of prominence in visual speech animation," *Proceedings of the ACM/SSPNET 2nd International Symposium on Facial Analysis and Animation*, p. 2525, 2010.

[2] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on PatternAnalysis and Machine Intelligence*, vol. 23, pp. 484—498, 1998.

[3] L. Terissi, M. Cerda, J. Gomez, N. Hitschfeld-Kahler, B. Girau, and R. Valenzuela, "Animation of generic 3D head models driven by speech," http://hal.archives-ouvertes.fr/hal-00587016/, July 2011.

[4] S. Fagel and G. Bailly, "From 3-D speaker cloning to text-to-audiovisual speech," in *ISCA Research and Tutorial Workshop on Auditory-Visual Speech Processing (AVSP)*, Moreton Island France, 2008, pp. 43–46, Dpartement Parole et Cognition.

[5] M. Berger, G. Hofer, and H. Shimodaira, "Carnival: a modular framework for automated facial animation," in *ACM SIGGRAPH 2010 Posters*, Los Angeles, California, 2010, SIGGRAPH '10, p. 5:15:1, ACM.

[6] D. P. Cosker, A. D. Marshall, P. L. Rosin, and Y. A. Hicks, "Speaker-independent speech-driven facial animation using a hierarchical model," in *Visual Information Engineering*, 2003, pp. 169–172.

[7] D. P. Cosker, A. D. Marshall, P. L. Rosin, and Y. A. Hicks, "Video realistic talking heads using hierarchical non-linear speech-appearance models," *Proceedings of Mirage*, pp. 20–27, 2003.

[8] T. F. Cootes, G. Wheeler, K. Walker, and C. J. Taylor, "View-based active appearance models," *Image and Vision Computing*, vol. 20, no. 9-10, pp. 657–664, 2002.

[9] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, vol. 2, no. 11, pp. 559–572, 1901.

[10] L. R. Rabiner, "A tutorial on hidden markov models and selected applications inspeech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[11] M. Brand, "Voice puppetry," in *Proceedings of the 26th annual conference on Computer Graphics and Interactive Techniques*. 1999, pp. 21–28, ACM Press/Addison-Wesley Publishing Co.

[12] M. Brand, "An entropic estimator for structure discovery," in *Proceedings of the 1998 conference on Advances in Neural Information Processing Systems II*. 1999, pp. 723–729, MIT Press.

[13] L. Xie and Z. Liu, "Speech animation using coupled hidden markov models," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, vol. 1, pp. 1128–1131.

[14] K. Wang, J. Zhang, D. Li, X. Zhang, and T. Guo, "Adaptive affinity propagation clustering," *acta automatica sinica*, vol. 33, pp. 1242–1246, 2007.

[15] B. Theobald, S. Fagel, G. Bailly, and F. Elisei, "LIPS2008: visual speech synthesis challenge," Oct. 2008.

[16] R. A. Fisher, *Statistical methods for Research Workers*, Oliver and Boyd., Edinburgh, 1925.