interspeech 2008 incorporating SST 08

Speech-driven Lip Motion Generation with a Trajectory HMM

Gregor Hofer, Junichi Yamagishi, & Hiroshi Shimodaira

Centre for Speech Technology Research, University of Edinburgh, UK

g.hofer@sms.ed.ac.uk

Abstract

Automatic speech animation remains a challenging problem that can be described as finding the optimal sequence of animation parameter configurations given some speech. In this paper we present a novel technique to automatically synthesise lip motion trajectories from a speech signal. The developed system predicts lip motion units from the speech signal and generates animation trajectories automatically employing a "Trajectory Hidden Markov Model". Using the MLE criterion, its parameter generation algorithm produces the optimal smooth motion trajectories that are used to drive control points on the lips directly. Additionally, experiments were carried out to find a suitable model unit that produces the most accurate results. Finally a perceptual evaluation was conducted, that showed that the developed motion units perform better than phonemes.

Index Terms: talking head, lip synchronisation, audio-visual speech

1. Introduction

Correct lip synchronisation is essential to make character animation believable. Humans are very aware of facial expressions and can detect the smallest discrepancies between the animation and the speech signal. Speech animation is therefore a very labour intensive process for which automation is highly desirable. Automatic speech animation remains a challenging problem that can be described as finding the optimal sequence of animation parameter configurations given some speech. The dependencies between a sequence of phonemes and the corresponding animation parameters are highly non-linear, as the same sound can be produced in various ways. In addition coarticulation convolutes the problem further as long range dependencies between the parameters have to be taken into account. Still, phonemes have a high correlation with lip motion.

We can characterise previous approaches in terms of the input to the system. Many systems use text and the corresponding phoneme string as input and then use concatenation[1], dominance functions[2] or trajectory generation [3] to produce the desired animation. Other approaches use parameterised speech directly as input and then use formant analysis [4], linear regression [5], or probabilistic modelling [6] [7] to generate the appropriate motion.

The choice between speech or text input depends largely on the application. A dialogue system where the spoken text is known, will most likely use a text based approach. An application that has to deal with unknown speech, in particular applications that have to run fast and in real time, like rapid prototyping for games and movies, will opt for speech based input. In this paper we present a two step approach that produces the desired lip animation from just speech. We utilise a trajectory Hidden Markov Model (HMM) [8] because of its flexibility and trainability. Trainable probabilistic models have the advantage over other approaches, that they can easily be adapted to different speakers. Additionally expanding these kind of models to other types of data has been demonstrated[9], which increases their applicability in animation systems. Our system distinguishes itself from Tamura et al. [3] in that our system uses speech and not text as input. Furthermore our system generates lip motion parameters from specifically designed modelling units.

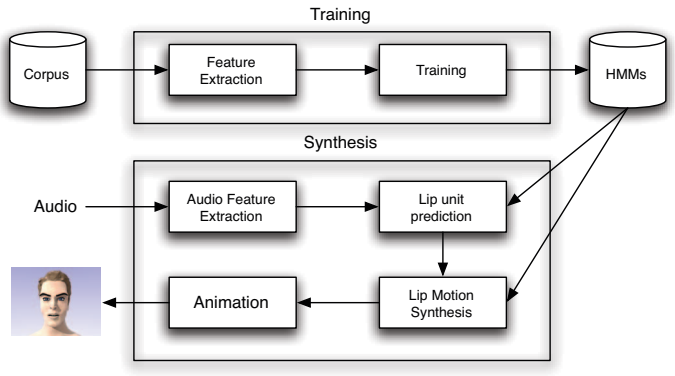

Figure 1: System diagram of the automatic motion generation

2. System Overview

To generate speech animation we propose a hierarchical system with each stage performing two steps; a recognition step and a synthesis step where the same kind of model is used for both steps. In the recognition step we choose the most likely unit sequence given some speech. During synthesis the unit sequence is translated into motion trajectories. These two steps are performed for each type of motion data. Figure 1 gives a graphical overview of the process.

The models are trained on speech derived features and tracking data simultaneously using the maximum-likelihood criterion. Each type of data is modelled in a separate stream where only the transition probabilities between states are shared with the other streams. These streams are turned on and off depending on synthesis or recognition. For example when predicting lip units the stream that models the lip motion trajectories is turned off and only the speech stream is turned on. Whereas during synthesis the speech stream is turned off and the motion stream is turned on. The parameter generation algorithm uses the predicted units, meaning that each unit corresponds to a model, to synthesise a smooth trajectory.

During synthesis time, speech data is fed into the model and lip motion units are recognised using the trained HMM. Trajectories are generated from the HMM using the units. The lip motion trajectories are used to drive control points on our facial model.

Accepted after peer review of full paper

3. Corpus

We tracked markers on the face and the body. To minimise the effects of head and body motion the relative distances between tracker points were used. Therefore only two features for the lips (mouth opening, pucker) were calculated. The position of the markers on the lips can be seen in Figure 2. The actor read 500 newspaper sentences selected for optimal phoneme balance. He was also asked to act out various stories and jokes for greater variety of motion. The newspaper data was automatically phonetically labelled. We split the data into training and testing data. The testing data was not seen during training.

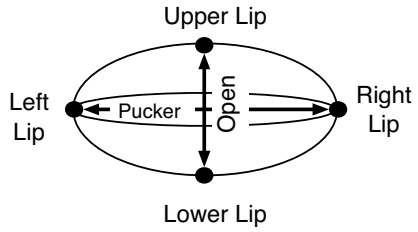

Figure 2: The location of the four lip markers. The distance between the upper and lower lip and the distance between the left and right corner of the mouth are modelled.

4. Modelling Lip Motion

The proposed method models lip motion directly using the recorded motion capture points. Distributions of motion trajectories are learnt for each speech unit by a trajectory Hidden Markov Model (HMM). A trajectory HMM is a state-of-the-art time series stochastic model that is able to model the dynamic changes of a signal. Its parameter generation algorithm can produce smooth trajectories from the stochastic model [8].

The speech and motion data are simultaneously modelled using context dependent HMMs. The data is described as a sequence of context dependent viseme models. Each model consists of five streams. One stream for the speech features, three streams for F0, and one stream for the motion features. The lip motion is modelled using two features, that is, the distance of the upper and lower lip and the distance between the left and right corner of the mouth. Additionally the first and second derivative of the lip motion features are also used to better model the dynamics of the motion trajectories. The speech is modelled using the first 12 mel cepstrum coefficients and energy. The first and second derivative of the speech features are also modelled. F0 is modelled using three streams, one for the static features and one stream each for the first and second derivative respectively. The HMMs use a mixture of Gaussian distributions at each state.

5. Synthesising Lip Motion

To synthesise lip motion, speech is input into the model. Recognition is performed using the multi-stream HMMs. During the recognition step only the speech feature streams are used, producing a sequence of visemes. The visemes give the sequence of context dependent models that are used for synthesising the actual trajectories.

Synthesising from a stochastic model like a conventional HMM is like rolling a dice. At each state, a value is sampled from the distribution, the resulting output is stochastic and not smooth. Conventional HMMs are good at recognising patterns but the sampled trajectories are not representative of the actual trajectories that are in the training data. Using the parameter generation of the trajectory HMM, a smooth output can be synthesised by taking the first and second derivatives of the data into account.

5.1. Optimal Motion

It is straightforward to justify the above procedures. For simplicity, let us explain the case where we generate motion vector sequences for the lips $\boldsymbol{O}_L = (\boldsymbol{o}_{L_1}, \boldsymbol{o}_{L_2}, \ldots, \boldsymbol{o}_{L_T})$ from a given speech vector sequence $\boldsymbol{O}_S = (\boldsymbol{o}_{S_1}, \boldsymbol{o}_{S_2}, \ldots, \boldsymbol{o}_{S_T})$ with a length of T frames. Lip motion has a high correlation with the speech vector sequence.. Thus we may solve the optimisation problem:

$$\boldsymbol{O}_L^* = \underset{\boldsymbol{O}_L}{\operatorname{argmax}} P(\boldsymbol{O}_L|\boldsymbol{O}_S) \tag{1}$$

We can work out the optimisation problems by incorporating the motion-unit sequence $\boldsymbol{u}_L = (u_{L_1}, ..., u_{L_e})$, which represent the lip movements corresponding to the given speech sequence. Using the motion labels units, the first optimisation regarding lip motion can be approximated by

$$\begin{aligned}
\boldsymbol{O}_L^* &= \underset{\boldsymbol{O}_L}{\operatorname{argmax}} P(\boldsymbol{O}_L|\boldsymbol{O}_S) && (2)\\
&= \underset{\boldsymbol{O}_L}{\operatorname{argmax}} \sum_{\boldsymbol{u}_L} P(\boldsymbol{O}_L|\boldsymbol{u}_L, \boldsymbol{O}_S) P(\boldsymbol{O}_S|\boldsymbol{u}_L) P(\boldsymbol{u}_L) && (3)\\
&\simeq \underset{\boldsymbol{O}_L}{\operatorname{argmax}} P(\boldsymbol{O}_L|\boldsymbol{u}_L^*) && (4)
\end{aligned}$$

where

$$\boldsymbol{u}_L^* = \underset{\boldsymbol{u}_L}{\operatorname{argmax}} P(\boldsymbol{O}_S|\boldsymbol{u}_L) P(\boldsymbol{u}_L) \tag{5}$$

Thus we recognise the lip motion units $\boldsymbol{u}_L$ from the given speech data $\boldsymbol{O}_S$ using the Viterbi algorithm and then generate a lip motion sequence from HMMs corresponding to the recognised units. For the probability $P(\boldsymbol{u}_L)$, we use back-off bi-gram models estimated from the training database.

5.2. Trajectory HMM

We explain the parameter generation algorithm for the HMMs. Since the basic HMMs are generative models, the output sequence which the model produces is stochastic and discontinuous. Although it is possible to use only mean values of the Gaussian distributions of the emission probabilities instead, it is still discontinuous and unnatural. For smooth and natural output, an extension of the HMM paradigm is needed, which is called a Trajectory HMM [10]. This is similar to Kalman smoothing or regularisation theory in the respect that those smoothing techniques also have some continuity constraints. The Trajectory HMM uses two explicit constraints on the observation features as the continuity constraints, obtained from equations for calculating their velocity and acceleration features. The observation vector at frame t, denoted by $\boldsymbol{o}_t$, is defined by

$$\boldsymbol{o}_t = [\boldsymbol{x}_t^\top, \Delta^1 \boldsymbol{x}_t^\top, \Delta^2 \boldsymbol{x}_t^\top]^\top \tag{6}$$

where $\cdot^\top$ denotes the matrix transpose and $\boldsymbol{x}_t$ is the original static feature. The velocity and acceleration features are calculated as the first and second time derivative estimates of the

static features and they are given by

$$\Delta^1 \boldsymbol{x}_t = -0.5\boldsymbol{x}_{t-1} + 0.5\boldsymbol{x}_{t+1} \quad (7)$$

$$\Delta^2 \boldsymbol{x}_t = \boldsymbol{x}_{t-1} + 2\boldsymbol{x}_t + \boldsymbol{x}_{t+1}. \quad (8)$$

Thus the parameter generation algorithm of the trajectory HMM generates a sub-optimal smoothed parameter sequence $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_T^*)$ without the dynamic and acceleration features as follows

$$(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_T^*) = \underset{(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)}{\operatorname{argmax}} P(\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T | \boldsymbol{u}_L^*) \quad (9)$$

$$\text{s.t. } \Delta^1 \boldsymbol{x}_t = -0.5\boldsymbol{x}_{t-1} + 0.5\boldsymbol{x}_{t+1} \quad (10)$$

$$\Delta^2 \boldsymbol{x}_t = \boldsymbol{x}_{t-1} + 2\boldsymbol{x}_t + \boldsymbol{x}_{t+1}. \quad (11)$$

When single Gaussian distributions are used as the emission probabilities, we can easily solve this problem in a closed form in a maximum likelihood sense [10]. A sample of the trajectory generated from HMMs is shown in Figure 3, in which we can see that the generated trajectory (solid line) becomes smooth.

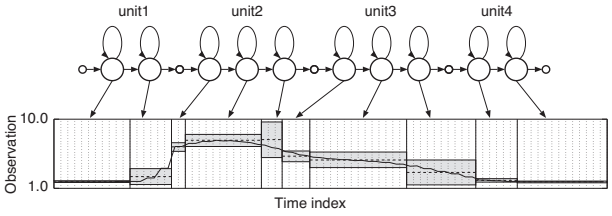

Figure 3: A sample of the trajectory generated from trajectory HMMs.

When mixtures of Gaussian distributions are used as the emission probabilities, the trajectory is optimised via the EM algorithm to select the optimal Gaussian distributions.

5.3. Animation system

Instead of driving blend-shapes, we control points around the lips of the character directly. The generated trajectories correspond to the points shown in figure 4.

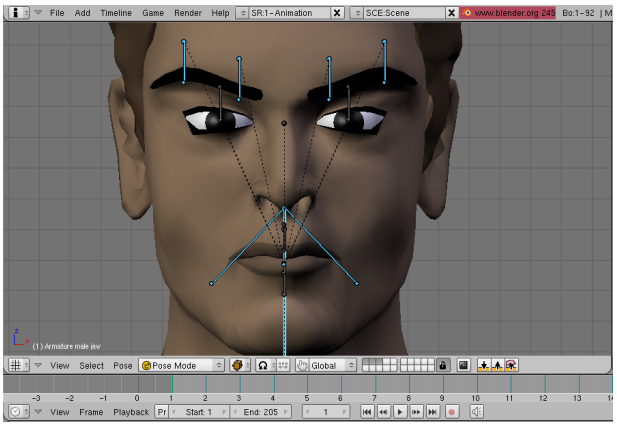

Figure 4: Screenshot of the model in Blender, showing the control points around the mouth.

The vertices of the mesh are weighted according to their distance from the control point. Any transformation applied to the control point is applied to the vertices around according to their weight. This produces a skin-like effect. In particular:

$$v' = \sum_i^n w_i M_i v \quad (12)$$

where n is the number of matrices, v is the vertex position, w is the weight associated, and M is the transformation matrix.

6. Model Unit

For lip synchronisation, visemes were used as the basic motion units. The data was phonetically labelled and a mapping between the phoneme labels and our viseme set was realised. To model the co-articulation we generated context dependent units, meaning that for each viseme, there were different units depending on the left and right context. Furthermore, five different sets of visemes were implemented to test for the effects of different motion units. In particular the sets are described as follows: The 2vis set consists of just two visemes, one for open mouth, and one for closed mouth. The simple set (sVis) groups phonemes into 6 different classes that correspond roughly to the following phoneme classes: vowels are classed according to their height and backness, resulting in three classes and consonants are either bilabial, labiodental, or just plain consonants. The extended set (eVis), breaks these classes further down, distinguishing diphthongs (5), classing more vowels on its own (4), and making further distinctions between consonants (10), resulting in a total of 19 classes. Table 1 shows a comparison of the different sets.

An experiment was carried out to find the optimal unit for our model by comparing how well the lip closing in the synthesised data lined up with the original data. If a mouth closing occurred within 2 frames of the original mouth closing, a point was awarded. The best performing sets were the extended version of a simple set designed by us (eVIS) and the Preston-Blair phoneme set (pbVis), that seems to be a standard in animation [11]. Table 1 shows the different sets and its score in our evaluation. For our data and model the eVIS set yielded the best results. Example animations were also produced for the viseme sets and again the eVis set produced the highest quality animation as can be seen Figure 6.

no of Visemes	Description	Name	Mean Score
2	open or close	2vis	66.5
9	Preston Blair set	pbVis	67.7
6	simple set	sVis	67.4
19	extended set	eVis	**68.8**
46	phone set	phone	65.2

Table 1: Viseme sets and its scores. Better alignment between the mouth closings of the original utterance and the synthesised one produces a higher score.

What is interesting to note is that the Preston-Blair set seems to perform worse in our experiments than our own designed sets. This does not mean that one set is superior or inferior to another but that for automatic generation of lip animation, the viseme set used makes a difference. Therefore it is important when modelling lip motion to chose the viseme set is carefully.

7. Evaluation

Figure 5 shows a comparison between a synthesised lip motion trajectory and the original trajectory. The original movement has a higher dynamic range which is a common problem when using stochastic modelling but otherwise the synthesised trajectory follows the original relatively closely.

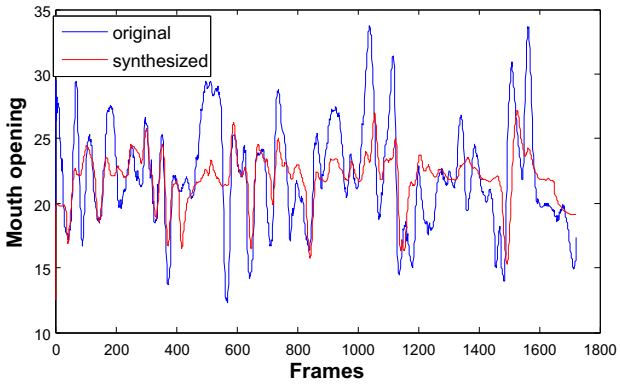

Figure 5: The synthesised trajectory clearly follows most of the original trajectory. The differences in the dynamic range are due to the nature of stochastic modeming.

To investigate the merit of the proposed method further, we conducted a perceptual evaluation. Six different speech inputs were synthesised in three variants, using the eVis viseme set, the full phoneme set, and the original tracking data. Ten speech technology experts were asked to judge the lip-synchronisation of our character comparing these three conditions. The participants saw two videos in succession and had to decide which one had better lip synchronisation. They could view each video as often as they like. Each permutation of the 3 conditions, was seen twice to check for consistency giving a total of 36 trial. They were presented in randomised order. Figure 6 shows the score for each condition.

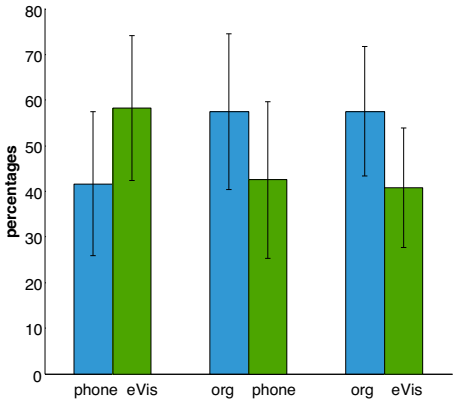

Figure 6: The percentage of positive scores for each condition.

It is interesting to note that the specifically designed viseme set is judged better than a standard phone set, when they are compared with each other. When both sets are directly compared with the original data, both sets are judged about equally worse than the original.

8. Conclusion and Future Work

We have proposed an HMM-based method for achieving lip synchronisation. Our work distinguishes itself from other work in that it can generate lip motion from just speech by utilising a two-step approach, allowing for a lot of flexibility. The parameter generation algorithm of the trajectory HMM is used to generate smooth output trajectories. One of the major drawbacks of our system is the corpus we are using. Because of technical limitations we were only able to track 4 points around the mouth, which resulted in impoverished models. Theoretically our models can work with an unlimited amount of tracking points, even producing other types of animation than just lip motion. Therefore our next step will be to record better data with more tracking points. However, given the current data we were still able to demonstrate that the proposed approach is feasible for animating a talking head.

9. References

[1] H. Graf, E. Cosatto, and V. Strom, "Visual prosody: facial movements accompanying speech," *Automatic Face and Gesture Recognition*, Jan 2002.

[2] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*. Springer-Verlag, 1993.

[3] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Textto-audio-visual speech synthesis based on parameter generation fromhmm," in *.European Conference on Speech Communication and Technology*, 1999, pp. 959–962.

[4] Z. Wen, P. Hong, and T. Huang, "Real time speech driven facial animation using formant analysis," *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, pp. 817–820, 22-25 Aug. 2001.

[5] C.-K. Hsieh and Y.-C. Chen, "Partial linear regression for audio-driven talking head application," *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 281–284, 6-6 July 2005.

[6] M. Brand, "Voice puppetry," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH'99)*, 1999, pp. 21–28.

[7] S. Nakamura, "Statisitical Multimodal Integration for Audio-Visual Speech Processing," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 854–866, 2002.

[8] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech and Language*, vol. 21, no. 1, 2007.

[9] G. Hofer and H. Shimodaira, "Automatic head motion prediction from speech data," in *Proc. Interspeech 2007*, Antwerp, Belgium, Aug. 2007.

[10] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorigthms for HMM-based speech synthesis," in *Proc. ICASSP 2000*, Jun. 2000, pp. 1315–1318.

[11] G. Martin, "Preston blair phoneme series," World wide web electronic publication, 2006. [Online]. Available: http://www.garycmartin.com