Template-Warping Based Speech Driven Head Motion Synthesis

David Adam Braude, Hiroshi Shimodaira, Atef Ben Youssef

Centre for Speech Technology Research,
Univeristy of Edinburgh, United Kingdom

d.a.braude@sms.ed.ac.uk, h.shimodaira@ed.ac.uk, abenyou@inf.ed.ac.uk

Abstract

We propose a method for synthesising head motion from speech using a combination of an Input-Output Markov model (IOMM) and Gaussian mixture models trained in a supervised manner. A key difference of this approach compared to others is to model the head motion in each angle as a series of templates of motion rather than trying to recover a frame-wise function. The templates were chosen to reflect natural patterns in the head motion, and states for the IOMM were chosen based on statistics of the templates. This reduces the search space for the trajectories and stops impossible motions such as discontinuities from being possible. For synthesis our system warps the templates to account for the acoustic features and the other angles' warping parameters. We show our system is capable of recovering the statistics of the motion that were chosen for the states. Our system was then compared to a baseline that used a frame-wise mapping that is based on previously published work. A subjective preference test that includes multiple speakers showed participants have a preference for the segment based approach. Both of these systems were trained on storytelling free speech.

Index Terms: Head motion synthesis, GMMs, IOMM

1. Introduction

Head motion is an important communication channel. In addition to meaningful gestures such as nodding for agreement it provides prosody and many social cues [1]. While it is known that there is a link between acoustic features and head motion [2, 3] so far synthesising natural synchronised head motion has not been achieved.

Several approaches have been used thus far. Generally they seek to determine a frame based function to map acoustic features into head motion. Some examples include Yehia et al. [4] who propose a regression model for a frame-wise mapping from $F0$. Bussio et al. [5, 6] use a Hidden Markov Model (HMM) to create sequences of head motions based on the idea of finding optimal poses and combining them with first noise then interpolating and smoothing the resulting trajectory. The theme of HMM approaches continue with Sargin et al. [7, 8] who use parallel HMM structures, and Hofer [9, 10] who attempts to recover gestures using HMM based speech recognition techniques and then synthesise motion with HMMs trained on those gestures. Both Sagin and Hofer rely on the HMM's structure to create different types of motion and preserve the dynamic limitations of the head motion. In Sargin's case this is the different parallel branches and in Hofer's case it is the different models used for different gestures. A very recent approach by Le *et al.* [11] uses a set of Gaussian mixture models to maximise the probability of a frame's trajectory, taking into account the velocity and acceleration.

In the majority of the existing research rigid head motion is often expressed as Euler angles, and as the head cannot rotate past a certain point there are maximums in these angles. This leads the angles to have a wave-like motion when plotted over time with clear peaks and valleys. However, seldom in synthesis is this prior knowledge exploited. This means that during synthesis dynamic constraints become explicit, or the model may create discontinuous movement. If, however, the motion is limited to only being able to move in patterns that are found in source data then these constraints become implicit and the motion is limited to what is normally observed. As it seems that there are only a limited number of these patterns they can be used as templates for movement.

HMMs are capable of recovering highly complex patterns in data. However, the trajectories of head motion are relatively simple. So it is possible to use a less complex system such as a Markov model. The input-output extension to Markov models [12] allows for modification of the observables based on factors other than states and other parts of the observation. For instance the inputs could modify or 'warp' the templates. In this case the inputs would come from acoustic features and other angle's movement. If the templates are chosen to be functions of time then this also reduces the amount of computation needed, as optimisation can be done less often.

In this paper we propose a system that uses a template function to generate head motion. The templates are warped to match acoustic features and maintain realism. What follows is a derivation of the probabilistic model. Then we show results of both objective measures and subjective evaluations where we compare our method to both the original trajectory and one synthesised by a frame-based system. We do not attempt to recreate the original head motion, rather we seek to create motion that seems believable given just the acoustic features and so would not have any semantic meaning.

2. Model description

In this paper we focus on the rotation of the head as a rigid body. So head motion is described in terms of rotation vector components α, β, and γ; these could be substituted for Euler angles [13] but rotation vectors are independent of order of rotation.

We propose to represent head motion as templates. These template are any function of time that can be changed or warped by specifying a duration and amplitude. The duration can either be a variable of the function or can change using a standard time warping method. In essence the templates are functions of duration and amplitude and should also have a constant offset. While any template function is possible, a simple choice is to use a sinusoidal function. In this paper each of these angles are treated identically and therefore, to represent any individual angle (but not all simultaneously) we introduce $\kappa \in \{\alpha, \beta, \gamma\}$.

Definition 1. A general head motion template is expressed as

$$\mathbf{g}(t, d_t^{(\kappa)}, a_t^{(\kappa)}, c_t^{(\kappa)}),$$

with duration d, amplitude a, and constant offset c for time t.

We use the bracketed superscript to refer to the angle and the subscript to time. Let $\boldsymbol{\lambda}_t^{(\kappa)}$ be the combined set of parameters. The template determines the value of $\boldsymbol{\lambda}_t^{(\kappa)}$ for time t.

What follows is a formal derivation of the model but informally one can think of it as generating a sequence of motions that follow a template for each angle. The template parameters for each angle are generated sequentially and when determining the parameters of the next template in the sequence it uses information from the acoustic features and the other angles' template parameters at that time. The distribution of durations observed in the data led to the decision to use a three state Markov model which divides motion into slow, medium, and fast. Because the states are known and the observations and transitions are based on both the state and external features, this is essentially a simplified Input-Output Hidden Markov Model (IOHMM) [12].

The problem of finding an optimal head motion trajectory $Y = [Y_t]_{t=1}^T$ given speech can be expressed as finding the trajectory Y^* for a given trajectory of acoustic features $X = [X_t]_{t=1}^T$ that maximises the p.d.f. such that

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} p(\mathbf{Y}|\mathbf{X}), \tag{1}$$

where Y_t is a tuple of the angles at t. Let the trajectories of the angles over time be $\overrightarrow{\alpha}$, $\overrightarrow{\beta}$, and $\overrightarrow{\gamma}$.

As we are treating head motion as a series of templates we denote the start of each template $t_m^{(\alpha)}$ for $\overrightarrow{\alpha}$, $t_n^{(\beta)}$ for $\overrightarrow{\beta}$, and $t_o^{(\gamma)}$ for $\overrightarrow{\gamma}$, where $m \in \{1, \ldots, M\}$, $n \in \{1, \ldots, N\}$, and $o \in \{1, \ldots, O\}$ denote template indices. We also define $i \in \{m, n, o\}$ to be the indices for the associated κ.

We now constrain head motion to follow a template for its entire duration by stating that

$$\boldsymbol{\lambda}_t^{(\kappa)} \text{constant for } t \in [t_i^{(\kappa)}, t_{i+1}^{(\kappa)}), \tag{2}$$

where

$$t_{i+1}^{(\kappa)} = t_i^{(\kappa)} + d_i^{(\kappa)}. \tag{3}$$

As the parameters are constant throughout the template motion let $\boldsymbol{\Lambda}_i^{(\kappa)}$ denote the parameter set for template i and angle κ, i.e.

$$\boldsymbol{\Lambda}_i^{(\kappa)} = (d_i^{(\kappa)},\ a_i^{(\kappa)},\ c_i^{(\kappa)}),$$

where $d_i^{(\kappa)}$, $a_i^{(\kappa)}$, and $c_i^{(\kappa)}$ are the duration, amplitude, and offset of template $\boldsymbol{\Lambda}_i^{(\kappa)}$ respectively. The complete trajectory of the constant template parameters are $\overrightarrow{\boldsymbol{\Lambda}}^{(\kappa)}$. To ensure the head motion is continuous, the first value of the new frame should be the same as what the previous frame would have predicted. In other words $c_i^{(\kappa)}$ must satisfy

$$\mathbf{g}(d_{i-1}^{(\kappa)}, \boldsymbol{\Lambda}_{i-1}^{(\kappa)}) = \mathbf{g}(0, \boldsymbol{\Lambda}_i^{(\kappa)}). \tag{4}$$

Returning to the original optimisation problem, (1) can be expressed as

$$\begin{aligned}\mathbf{Y}^* &= \arg\max_{\mathbf{Y}=Y_0\ldots Y_T} p(Y_0|\mathbf{X}) \prod_{t=1}^{T} p(Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_0, \mathbf{X}) & (5)\\ &= \arg\max_{\overrightarrow{\alpha}, \overrightarrow{\beta}, \overrightarrow{\gamma}} p(\alpha_0, \beta_0, \gamma_0|\mathbf{X}) \\ &\quad \times \prod_{t=1}^{T} p(\alpha_t, \beta_t, \gamma_t|\alpha_{t-1}, \beta_{t-1}, \gamma_{t-1}, \\ &\quad \ldots, \alpha_0, \beta_0, \gamma_0, \mathbf{X}) & (6)\end{aligned}$$

By using definition (1) this can be rewritten as

$$\begin{aligned}\arg\max_{\overrightarrow{\boldsymbol{\lambda}}^{(\alpha)}, \overrightarrow{\boldsymbol{\lambda}}^{(\beta)}, \overrightarrow{\boldsymbol{\lambda}}^{(\gamma)}} & p(\mathbf{g}(0, \boldsymbol{\lambda}_0^{(\alpha)}), \mathbf{g}(0, \boldsymbol{\lambda}_0^{(\beta)}), \mathbf{g}(0, \boldsymbol{\lambda}_0^{(\gamma)})|\mathbf{X}) \\ & \times \prod_{t=1}^{T} p\Big(\mathbf{g}(t, \boldsymbol{\lambda}_t^{(\alpha)}), \mathbf{g}(t, \boldsymbol{\lambda}_t^{(\beta)}), \mathbf{g}(t, \boldsymbol{\lambda}_t^{(\gamma)}) \\ & |\mathbf{g}(t-1, \boldsymbol{\lambda}_{t-1}^{(\alpha)}), \mathbf{g}(t-1, \boldsymbol{\lambda}_{t-1}^{(\beta)}), \mathbf{g}(t-1, \boldsymbol{\lambda}_{t-1}^{(\gamma)}), \\ & \ldots, \mathbf{g}(0, \boldsymbol{\lambda}_0^{(\alpha)}), \mathbf{g}(0, \boldsymbol{\lambda}_0^{(\beta)}), \mathbf{g}(0, \boldsymbol{\lambda}_0^{(\gamma)}), \mathbf{X}\Big)\end{aligned} \tag{7}$$

To compact the notation we denote

$$\mathbf{g}(t, \boldsymbol{\lambda}_t^{(K)}) = \Big(\mathbf{g}(t, \boldsymbol{\lambda}_t^{(\alpha)}),\ \mathbf{g}(t, \boldsymbol{\lambda}_t^{(\beta)}),\ \mathbf{g}(t, \boldsymbol{\lambda}_t^{(\gamma)})\Big) \tag{8}$$

$$\boldsymbol{\lambda}_t^{(K)} = \Big(\boldsymbol{\lambda}_t^{(\alpha)},\ \boldsymbol{\lambda}_t^{(\beta)},\ \boldsymbol{\lambda}_t^{(\gamma)}\Big) \tag{9}$$

and similar for $\boldsymbol{\Lambda}_t^{(K)}$

We segment the time sequence such that $\tau_1 \ldots \tau_{T'}$ are the start indices of each segment. These we constrain to being drawn from the set $(t_m^{(\alpha)} \cup t_n^{(\beta)} \cup t_o^{(\gamma)})$. We now optimise

$$\begin{aligned}\arg\max_{\overrightarrow{\boldsymbol{\lambda}}^{(\alpha)}, \overrightarrow{\boldsymbol{\lambda}}^{(\beta)}, \overrightarrow{\boldsymbol{\lambda}}^{(\gamma)}} & p(\mathbf{g}(0, \boldsymbol{\lambda}_0^{(K)})|\mathbf{X}) \\ & \times \prod_{s=0}^{T'-1} \prod_{t=\tau_s+1}^{\tau_{(s+1)}} p\Big(\mathbf{g}(t, \boldsymbol{\lambda}_t^{(K)}) \\ & |\mathbf{g}(t-1, \boldsymbol{\lambda}_{t-1}^{(K)}), \ldots \mathbf{g}(0, \boldsymbol{\lambda}_0^{(K)}), \mathbf{X}\Big).\end{aligned} \tag{10}$$

We now also define the indices of start times closest to τ_d but not equal to τ_d and the indices that may equal τ_d as

$$j_d^{(\kappa)} = \arg\min_i t_i^{(\kappa)}, \text{ s.t. } t_i^{(\kappa)} > \tau_d, \tag{11}$$

$$j_{d*}^{(\kappa)} = \arg\min_i t_i^{(\kappa)}, \text{ s.t. } t_i^{(\kappa)} \geq \tau_d, \tag{12}$$

with $j_d^{(K)}$ for the three separate angles in (8). Omitting the κ where it would be unambiguous and using the constraints above, (10) is equivalent to

$$\begin{aligned}\arg\max_{\overrightarrow{\boldsymbol{\lambda}}^{(\alpha)}, \overrightarrow{\boldsymbol{\lambda}}^{(\beta)}, \overrightarrow{\boldsymbol{\lambda}}^{(\gamma)}} & p(\mathbf{g}(0, \boldsymbol{\lambda}_0^{(K)})|\mathbf{X}) \\ & \times \prod_{s=0}^{T'-1} \prod_{t=\tau_s+1}^{\tau_{(s+1)}} p\Big(\mathbf{g}(t, \boldsymbol{\Lambda}_{j_{s*}}^{(K)})|\mathbf{g}(t-1, \boldsymbol{\Lambda}_{j_s}^{(K)}), \\ & \ldots, \mathbf{g}(\tau_s, \boldsymbol{\Lambda}_{j_s}^{(K)}), \ldots, \mathbf{g}(\tau_d - x, \boldsymbol{\Lambda}_{j_d}^{(K)}), \\ & \ldots, \mathbf{g}(0, \boldsymbol{\Lambda}_0^{(K)}), \mathbf{X}\Big), \\ & 0 \leq d < s,\ x \in (\tau_d, \tau_d + 1].\end{aligned} \tag{13}$$

If we make the assumption that each segment is only dependant on the previous τ and a window X_s of $\mathbf{X}$ around τ_s, then this reduces to

$$\begin{aligned}\arg\max_{\overrightarrow{\boldsymbol{\lambda}}^{(\alpha)}, \overrightarrow{\boldsymbol{\lambda}}^{(\beta)}, \overrightarrow{\boldsymbol{\lambda}}^{(\gamma)}} & p(\mathbf{g}(0, \boldsymbol{\lambda}_0^{(K)})|\mathbf{X}) \\ & \times \prod_{s=0}^{T'-1} \prod_{t=\tau_s+1}^{\tau_{(s+1)}} p\Big(\mathbf{g}(t, \boldsymbol{\Lambda}_{j_{s*}}^{(K)})|\mathbf{g}(t-1, \boldsymbol{\Lambda}_{j_s}^{(K)}), \\ & \ldots, \mathbf{g}(\tau_s, \boldsymbol{\Lambda}_{j_s}^{(K)}), X_s\Big).\end{aligned} \tag{14}$$

We then assume that this is similar to optimising just the parameters, instead of the actual trajectory, and this becomes

$$\underset{\overrightarrow{\mathbf{\Lambda}}^{(\alpha)},\overrightarrow{\mathbf{\Lambda}}^{(\beta)},\overrightarrow{\mathbf{\Lambda}}^{(\gamma)}}{\arg\max}\; p(\mathbf{\Lambda}_0^{(K)}|X_0) \times \prod_{s=1}^{T'} p\Big(\mathbf{\Lambda}_{js*}^{(K)}|\mathbf{\Lambda}_{js}^{(K)}, X_s\Big). \tag{15}$$

However at each time interval only one $\mathbf{\Lambda}_{j_s}$ is changing. This can thus be rephrased as

$$\underset{\overrightarrow{\mathbf{\Lambda}}^{(K)}}{\arg\max}\; p(\mathbf{\Lambda}_0^{(K)}|X_0) \times \prod_{s=1}^{T'} p\Big(\mathbf{\Lambda}_{js*}^{(\kappa)}|\mathbf{\Lambda}_{js}^{(K)}, X_s\Big). \tag{16}$$

With a simple application of Bayes' Theorem this reduces to maximising the joint probability,

$$\underset{\overrightarrow{\mathbf{\Lambda}}^{(K)}}{\arg\max}\; p(\mathbf{\Lambda}_0^{(K)}, X_0) \times \prod_{s=1}^{T'} p\Big(\mathbf{\Lambda}_{js*}^{(\kappa)}, \mathbf{\Lambda}_{js}^{(K)}, X_s\Big). \tag{17}$$

This probability density can be modelled by a Gaussian mixture model.

2.1. Extension based on empirical data

The choice of template function was based on the idea of segmenting the trajectories based on when they reached local maxima and minima. Fig. 1 shows sample trajectories of the segments when they are normalised to have a duration of one second and an amplitude of one. The trajectories were also modified so that those going from minima to maxima are inverted. Our choice of template function was

$$\mathbf{g}(t, \mathbf{\Lambda}_i^{(K)}) = (-1)^{i+1} a_i \cos\left(2\pi \frac{1}{4d_i} t\right) + c_i, \tag{18}$$

which is also shown on Fig. 1. The amount of warping was predicted using the instantaneous log-energy and $F0$ at the time of change.

The distribution of template function frequency from the training data is shown in Fig. 2. It is clear that there are three identifiable categories. This is a repeat of what Hader et al. found [14] but each angle is evaluated separately. This means that the system can be split into a three state model. So for synthesis the joint distribution of each of the states are estimated and then one is picked based on its likelihood. Because the distribution has high variance it is better to sample under the probability distribution rather than maximise, though in the big data case with a large amount of mixture components this is equivalent. In practice it was also found that estimating first the duration, then the amplitude separately did not have a significant change on the results but was more efficient computationally. Through testing it was also found that three mixture components for amplitude and one for frequency was sufficient.

3. Evaluation and discussion

To compare our system to a frame-based method we chose to modify Le *et al.*'s system [11]. Their model assumes independence between the angles, so to make the comparison fair we changed the model so that it would take this into account while still maintaining their choice of acoustic features. In other words while in their system there are essentially three optimisation problems, one for each of the angles while we chose to maximise them within one GMM. So α_t, β_t, γ_t, angular velocity v_t, and angular acceleration a_t are determined by solving sequentially for

$$(\alpha_t^*, \beta_t^*, \gamma_t^*) = \underset{(\alpha,\beta,\gamma)}{\arg\max}\, p(\alpha, \beta, \gamma, v, a, F0_t, \text{Loudness}_t) \tag{19}$$

We also compared our system to the original motion capture.

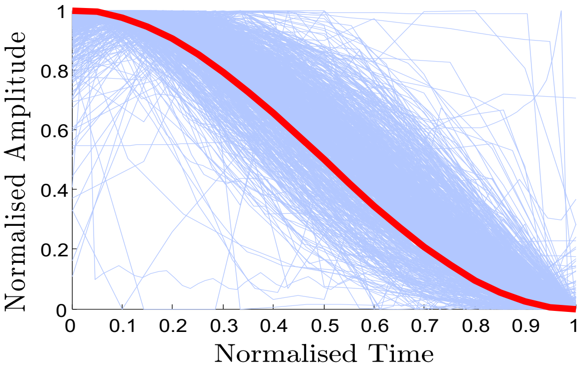

Figure 1: *Samples from m1 (thin lines) and template function (thick line).*

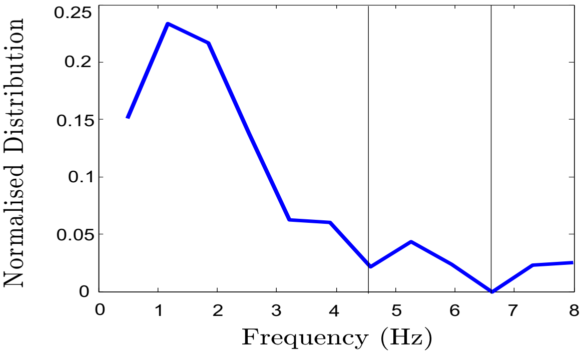

Figure 2: *Average distribution of frequency for all angles from all speakers.*

3.1. Data set

Data from four participants, consisting of 2 males (*m1* and *m2*) and 2 females (*f1* and *f2*), were used in the following experiments. The participants were all university students, native English speakers, raised in the U.K., and aged between 18 and 24. The participants were given five classic fairy-tales with which they should be familiar ahead of the recording. Each participant read a story for five minutes from a teleprompter, then retold the story in their own words, if they exceeded five minutes they were stopped. They were instructed to tell the story as though to an adult native English speaker. They were not given any instruction about body language or head motion. The available free speech total approximately 14, 25, 17, and 24 minutes for $m1$, $m2$, $f1$, and $f2$, respectively. Four stories were used for training and validation and the last was left for testing.

3.1.1. Head motion capture

The recordings were performed with the NaturalPoint OptiTrack[1] motion capture system. Four markers were placed on

[1] http://www.naturalpoint.com/optitrack/

the head and five markers were placed on the body. Recording was done at 100 Hz. In front of the person was a teleprompter for read stories (not used in this research). The recorder sat in the room with them so that they could focus their speaking when not reading. Audio was recorded with a free standing microphone at 44.1 kHz. Rotation matrices for the head and body were estimated from maker data using singular value decomposition, and then the relative head motions to the body were estimated by removing the effect of body motion. The obtained relative head motions were converted into rotation vectors [13].

3.1.2. Acoustic feature extraction

Prior to feature extraction the audio was down-sampled down to 16 kHz. The combined features of the fundamental frequency $F0$ (extracted via autocorrelation and cepstrum based methods), log-energy, loudness contours, voicing probability, and voice quality were extracted using OpenSmile [15] . From these log-energy, its first derivative and loudness and were smoothed with a moving average window of 13 frames. These features were computed from the audio signal over 25 ms windows at a frame rate of 10 ms to match the frame rate of the motion capture system. $F0$ was extracted for the base-line system.

3.2. Objective evaluation

To objectively measure the system we first determine how well the frequency and amplitudes were recovered. In Fig. 3 the predicted frequency and amplitude are shown along with the training data's for one sample. The second was to determine the system's performance in predicting the states. The system's confusion matrix for state prediction is given in Table 1.

Table 1: *Confusion matrix for state prediction over all speakers, columns give the prediction*

	Slow	Medium	Fast	Number of Samples
Slow	81%	8%	11%	30180
Medium	47%	46%	6%	3471
Fast	67%	8%	25%	1374

As indicated by Fig 3, our system is capable of recovering the same template parameters that were given by the speech. However, the system does not have a good performance when predicting states.

3.3. Subjective evaluation

For the testing, 14 subjects were shown one minute long samples of videos of each type in an A-B type test on a model that does not show eye or lip motion. They were each given 10 comparisons to make, with the pairings being balanced so that motion capture and the template were compared to each other one more time than they were compared to the frame based model. Each comparison of methods was done so that they would show each method on the left and right an equal amount of times. They were able to view both motions simultaneously and repeat the video as many times as they liked and stop when they had made a decision. The subjects of the experiment included two speech technology experts. The preference results are given in Fig. 4.

3.4. Discussion

On the objective measures it would seem that in all but one respect the system would be a good predictor of head motion. It was able to recover similar distributions of frequency and amplitude to the original trajectory. However, the state prediction has high confusion especially for predicting fast motion. This implies that either a better classifier is needed or that a different feature set may be more appropriate.

In the subjective test there was a strong preference for the template model over the frame-based one. When asked participants reported that the movement of the frame based model was jerky. Over a long clip like those that the participants watched, this would be very noticeable.

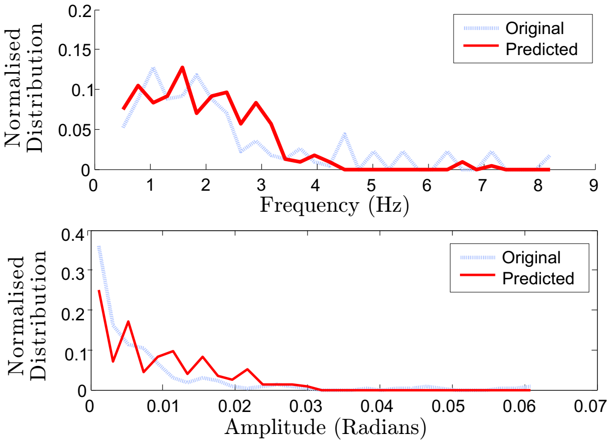

Figure 3: *Average distribution for the difference angles of frequency and amplitude for an example trajectory.*

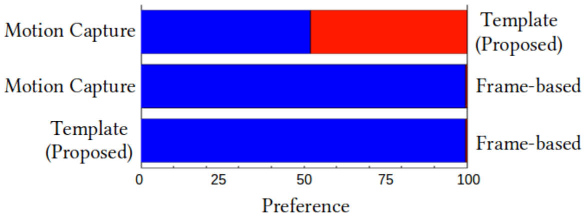

Figure 4: *Preference of types of motion, blue indicates preference of left method over method on right shown by red.*

4. Conclusion and future work

Our system has shown a lot of promise for synthesising head motion. It outperformed frame-wise synthesis and in a subjective test people had difficulty telling it apart from motion capture. The main open question is how to improve the state prediction. Some possible options are the use of other acoustic features or using windows of samples.

The use of templates has many advantages, they limit head movement to what would be believable, they greatly reduce the amount of optimisation operations that need to take place, and they can maintain the dependencies between angles.

5. Acknowledgements

The authors would like to extend their gratitude to Benjamin Rosman for his help. This work was supported by EU FP7 SSP-Net (grant agreement no. 231287). The financial assistance of the National Research Foundation of South Africa (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

6. References

[1] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. V. Bateson, “Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception,” *Psychological Sciene*, vol. 15, pp. 133–137, 2004.

[2] H. P. Graf, E. Casatto, V. Strom, and F. J. Huang, “Visual Prosody: Facial Movements Accompanying Speech,” *Proc. 5th International Conf. on Automatic Face and Gesture Recognition*, pp. 381–386, 2002.

[3] T. Kuratate, K. G. Munhall, P. E. Rubin, E. Vatikiotis-Bateson, and H. Yehia, “Audio-Visual Synthesis of Talking Faces for Speech Production Correlates,” in *Eurospeech’99*, vol. 3, 1999, pp. 1279 – 1282.

[4] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, “Linking Facial Animation, Head Motion, and Speech Acoustics,” *Journal of Phonetics*, vol. 30, pp. 555 – 568, 2002.

[5] C. Busso, Z. Deng, U. Neumann, and S. Narayanan, “Natural Head Motion Synthesis Driven by Acoustic Prosodic Features,” *Computer Animation and Virtual Worlds*, vol. 16, no. 3-4, pp. 283 – 290, 2005.

[6] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, “Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1075–2007, March 2007.

[7] M. E. Sargin, O. Aran, A. Karpov, F. Ofli, Y. Yasinnik, S. Wilson, E. Erzin, Y. Yemez, and A. Tekalp, “Combined Gesture-Speech Analysis and Speech Driven Gesture Sythesis,” in *IEEE International Conference on Multimedia and Expo*, 2006.

[8] M. E. Sargin, E. Erzin, Y. Yemez, A. M. Tekalp, A. T. Erdem, C. Erdem, and M. Özkan, “Prosody-Driven Head-Gesture Animation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. II–677 – II–680.

[9] G. Hofer and H. Shimodaira, “Automatic head motion prediction from speech data,” in *Proceedings of Interspeech*, 2007, pp. 722 – 725.

[10] G. O. Hofer, “Speech-driven Animation Using Multi-modal Hidden Markov Models,” Ph.D. dissertation, University of Edinburgh, 2009.

[11] B. H. Le, X. Ma, and Z. Deng, “Live Speech Driven Head-and-Eye Motion Generators,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, pp. 1902 – 1914, 2012.

[12] Y. Bengio and P. Frasconi, “An Input Output HMM Architecture,” in *Advances in Neural Information Processing Systems*. MIT Press, 1995, pp. 427 – 434.

[13] J. Diebel, “Representing Attitude: Euler Angles, Unit Quaternions, and Rotation Vectors,” 2006.

[14] U. Hadar, T. Steiner, E. Grant, and F. Rose, “Kinematics of Head Movements Accompanying Speech During Conversation,” *Human Movement Science*, vol. 2, no. 1-2, pp. 35–46, 1983.

[15] F. Eyben, M. W ollmer, and B. Schuller, “openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor,” in *Proc. ACM Multimedia (MM), ACM*, 2010.