Speech driven Head Motion Synthesis based on a Trajectory Model

Gregor Hofer* University of Edinburgh
Hiroshi Shimodaira† University of Edinburgh
Junichi Yamagishi‡ University of Edinburgh

1 Introduction

Making human-like characters more natural and life-like requires more inventive approaches than current standard techniques such as synthesis using text features or triggers. In this poster we present a novel approach to automatically synthesise *head motion* based on speech features. Previous work has focused on frame wise modelling of motion [Busso et al. 2007] or has treated the speach data and motion data streams separately [Brand 1999], although the trajectories of the head motion and speech features are highly correlated and dynamically change over several frames. To model longer units of motion and speech and to reproduce their trajectories during synthesis, we utilise a promising time series stochastic model called "Trajectory Hidden Markov Models" [Zen et al. 2007]. Its parameter generation algorithm can produce motion trajectories from sequences of units of motion and speech. These two kinds of data are simultaneously modelled by using a multi-stream type of the trajectory HMMs. The models can be viewed as a Kalman-smoother-like approach, and thereby are capable of producing smooth trajectories.

2 Data Driven Synthesis

The proposed head motion synthesis method works in two stages. (1) First, a sequence of motion units is predicted from the speech signal using HMMs that were trained on the speaker's data. (2) Using the same models, the parameter generation algorithm of the trajectory HMM is employed to produce a smooth head motion trajectory.

In order to train the system, motion capture data was collected from an actor, telling stories. The Euler angles of the head rotation and their first and second derivatives were calculated. The data was annotated for head motion from the graphed Euler angles using four types of motion units.

- postural shift: the head shifts axis of movement
- shake and nod: lateral movement around one axis
- pause: no movement / rest position
- default: non-distinctive movement / slow movement

Each unit is a trajectory in 3D space that can be learned by an HMM. For each type of movement, a separate HMM was trained on speech and motion features. Similar to speech recognition, this allowed us to predict head motion units based on speech features. Figure 1 shows the training and prediction process. The HMMs are trained simultaneously on head motion and speech using a two stream approach. During prediction only the speech stream is used. The system was evaluated on how well it can predict motion labels given some speech. The accuracy was 70%. If the model was trained only on the speech stream the accuracy dropped to 55%.

Once the sequence of head motion units was determined, a motion trajectory can be synthesised. 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

*e-mail: g.hofer@sms.ed.ac.uk
†e-mail: h.shimodaira@ed.ac.uk
‡e-mail:v1jyamag@inf.ed.ac.uk

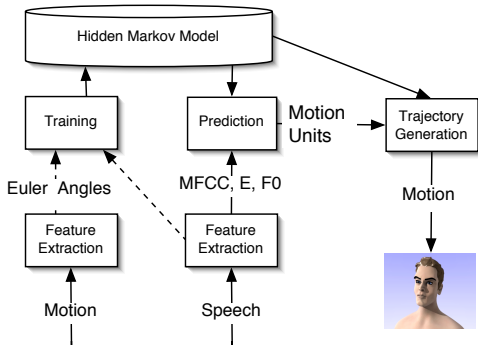

Figure 1: *Two stream training and single stream prediction.*

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 a trajectory HMM, a smooth output can be synthesised by taking the first and second derivatives of the data into account. The optimal smooth trajectory is produced in the sense of maximum likelihood. Figure 2 shows a predicted sequence of motion labels and the resulting trajectories. The conventional HMM output is not smooth and it is hard to distinguish distinct movements. On the other hand the trajectory HMM output is smooth and and it bears a clear relationship to the predicted label sequence.

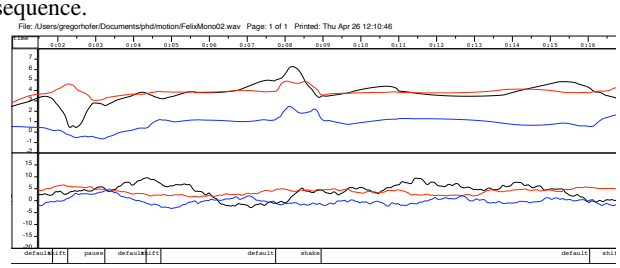

Figure 2: *Trajectory HMM output(top) vs. conventional HMM output(bottom). The predicted label sequence is shown at the bottom.*

3 Conclusion

We have proposed a novel method that can successfully synthesise head motion given some speech. Using a state-of-the-art parameter generation technique, it is possible to generate smooth trajectories from HMMs trained on motion and speech data. This method has possible applications in the computer animation industry whenever fast prototyping is needed.

References

BRAND, M. 1999. Voice puppetry. In *Proc. of SIGGRAPH '99*.

BUSSO, C., DENG, Z., GRIMM, M., NEUMANN, U., AND NARAYANAN, S. 2007. Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. *IEEE Transactions on ASLP*.

ZEN, H., TOKUDA, K., AND KITAMURA, T. 2007. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language 21*, 1.