Eurographics/ ACM SIGGRAPH Symposium on Computer Animation (2007)
D. Metaxas and J. Popovic (Editors)

Lip Motion synthesis using a context dependent trajectory Hidden Markov Model

G. Hofer[1], H. Shimodaira[1], and J, Yamagishi[1]

[1]Centre for Speech Technology Research, University of Edinburgh, UK

Abstract

Lip synchronisation is essential to make character animation believeable. In this poster we present a novel technique to automatically synthesise lip motion trajectories given some text and speech. Our work distinguishes itself from other work by not using visemes (visual counterparts of phonemes). The lip motion trajectories are directly modelled using a time series stochastic model called "Trajectory Hidden Markov Model". Its parameter generation algorithm can produce motion trajectories that are used to drive control points on the lips directly.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Animation

1. Introduction

Lip synchronisation is essential to make character animation believeable. The goal is to produce lip animation given speech. Speech is usually described in terms of phonemes, which are subsyllable units. Many previous approaches have used visemes (visual counterpart of a phoneme) or mouth shapes to create a mapping between the speech and the animation [CTFP05]. Because of co-articulation, where the previous and next mouth shape influence how the current mouth shape should look like, it is not enough to animate using a sequence of isolated mouth shape. To deal with the problem of co-articulation many previous approaches have utilised rules [CM93] or statistical techniques [EGP02]. In addition speech rate, defined as the number of phonemes during a given time interval, and loudness changes need to be addressed for successful lip synchronisation.

The proposed method does not use mouth shapes but tries to model the trajectory of the mouth motion directly using a state-of-the-art time series stochastic model. The modelled trajectories drive control points around the mouth over time. Because the model is trained on real motion capture data, co articulation does not have to be modelled explicitly but is inherent in the model properties. Furthermore by controllling the dynamic range of the modelled trajectories, the energy or loudness of the speech can be modelled directly as well. Finally the model is able to dynamically alter the duration of speech units to counter speech rate changes.

SCA 2007, San Diego, California, August 04 - 05, 2007
© 2007 ACM 978-1-59593-624-4/07/0008 $ 5.00

2. System Overview

Figure 1 shows the sequence of steps from the speech signal to the lip motion. The speech signal is converted into a sequence of (1)feature vectors. These features are used to do the (2)automatic alignment, which finds the exact phoneme sequence, and the boundaries between them. The sequence is used to (3)synthesise a sequence of motion vectors that is further (4)modulated using the energy of the speech, producing the final lip motion.

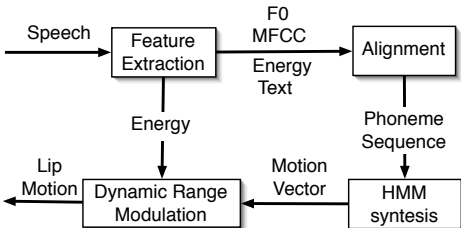

Figure 1: *System Overview, showing the steps involved to synthesise lip motion from speech data.*

3. Data Collection

The first 1000 sentences of the CMU Arctic database was recorded. The lip motion was captured with the Qualisys ProReflex MCU500 system at a frame rate of 500Hz. Four

points on the lip were recorded as can be seen in Figure 2. The influence of head motion was reduced by recording the relative distances of the points on the lips. The distance between the upper and lower lip can be viewed as a mouth opening and closing, whereas the distance between the left and right side of the lips can be regarded as the amount of lip puckering. The speech was recorded using a close talking microphone at 44kHz.

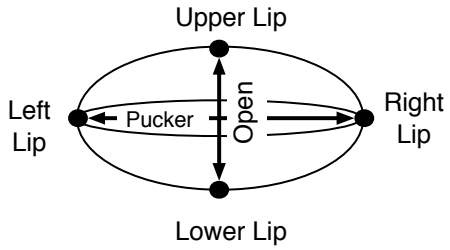

Figure 2: *The location of the four lip markers. The distance between the upper and lower lip and the distance between the left and right side of the mouth are modelled.*

4. Modelling Lip Motion

The proposed method models lip motion directly using the recorded motion capture points. There is no conversion from phonemes (speech unit) to visemes (animation unit) but 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 [ZTK07].

The speech and motion data are simultaneously modelled using context dependent HMMs. For each phoneme context a separate model is trained. Context here means the position of a certain phoneme in an utterance and the phonemes that precede it and follow it. The number of models increases with the amount of training data as more different contexts are seen. The data is described as a sequence of context dependent phoneme models. In each model, 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 side of the mouth. Additionally the first and second derivative of the lip motion features is also used to better model the dynamics of the motion trajectories.

5. Synthesising Lip Motion

Instead of driving viseme morph targets, control points around the mouth are directly driven by the synthesised trajectories of the model. 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. The optimal smooth trajectory is produced in the sense of maximum likelihood.

Then, the dynamic range of the resulting trajectory is controlled using the energy, which is similar to the loudness of the speech. By modulating the lip trajectories in this way louder speech results in a more open mouth and quieter speech in a less open mouth. Speech-rate changes are modelled by dynamically scaling the trajectory depending on the length of the phonemes.

6. Conclusion

We have proposed a novel technique for achieving lip synchronisation. Our work does not utilise visemes but models the lip motion trajectories directly. It therefore implicitly takes co-articulation into account. Furthermore the lip motion is modulated according to the loudness of the speech. Because of technical limitations of the employed motion capture system we only modelled 4 points around the mouth but there is no reason why the described method cannot be extend to more points. Finally as well as lip motion, head motion can also be automatically generated from speech [HSY07].

References

[CM93] COHEN M. M., MASSARO D. W.: Modeling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation*. Springer-Verlag, 1993. 1

[CTFP05] CAO Y., TIEN W. C., FALOUTSOS P., PIGHIN F.: Expressive speech-driven facial animation. *ACM Transactions on Graphics 24*, 4 (Oct. 2005), 1283–1302. 1

[EGP02] EZZAT T., GEIGER G., POGGIO T.: Trainable videorealistic speech animation. In *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 2002), ACM Press, pp. 388–398. 1

[HSY07] HOFER G., SHIMODAIRA H., YAMAGISHI J.: Speech Driven Head Motion Synthesis based on a Trajectory Model. SIGGRAPH 2007 (To appear as a poster). 2

[ZTK07] ZEN H., TOKUDA K., KITAMURA T.: Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language 21*, 1 (2007). 2