Simultaneous Speech and Animation Synthesis

Dietmar Schabus* FTW Austria
Michael Pucher † FTW Austria
Gregor Hofer‡ FTW Austria

Figure 1: *True audio-visual speech synthesis from the same underlying model.*

1 Introduction

Talking computer animated characters are a common sight in video games and movies. Although doing the mouth animation by hand gives the best results, because of cost and time constraints it is not always feasible. Furthermore the amount of speech in current games is ever increasing with some games having more than 200,000 lines of dialogue. This work proposes a system that can produce speech and the corresponding lip animation simultaneously using a statistical machine learning framework based on Hidden Markov Models (HMMs). The key point is that with the developed system never before seen or heard animated dialogues can be produced at a push of a button.

2 Audio-visual trajectory HMM synthesis

The core of the developed system consists of a statistical model of speech that was trained on a database of motion capture and audio recordings. We utilize the trajectory HMM framework for both speech synthesis and lip synchronization. Figure 2 shows the training and synthesis process of the multimodal speech models. Training a single model for both speech and motion has the advantage over previous approaches that for example duration modeling of individual phonemes (e.g. vowels and consonants) can be shared across both domains. The animation training data consisted of 15 PCA components of 32 tracked markers on the face. The speech training data consisted of standard spectral features (mel cepstrum), aperiodicity, and pitch features. Both the static animation and speech features are augmented with their corresponding delta and delta-delta values. As a modeling unit for the animation and speech model context dependent phonemes were employed, for the speech model additional features like the phoneme position in an utterance were also used.

For synthesis the text of an utterance is translated into a phoneme sequence using standard text analysis common in speech synthesis. This phoneme sequence serves as input to the synthesis model, where each phoneme consists of corresponding probability density functions (pdfs) over speech features and animation features. The maximum likelihood parameter generation algorithm

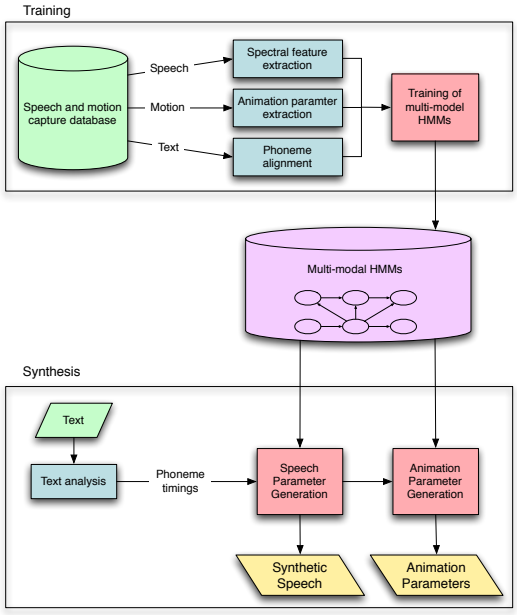

Figure 2: *The training process uses a multimodal database of recorded audio and motion capture data, which produces a multimodal speech model. During synthesis input text is automatically converted into corresponding speech and animation using maximum likelihood parameter generation on the distributions from the model of speech.*

(MLPG)[Tokuda et al. 2000] is then applied to this sequence of pdfs in order to obtain a single, most probable trajectory which optimizes the constraints between the distributions of static, delta and delta delta features. The trajectory then both drives the animation and produces an audible speech utterance.

References

TOKUDA, K., YOSHIMURA, T., MASUKO, T., KOBAYASHI, T., AND KITAMURA, T. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, 1315–1318.

*e-mail: schabus@ftw.at
†e-mail: pucher@ftw.at
‡e-mail: hofer@ftw.at