HMM-based Motion Trajectory Generation for Speech Animation Synthesis

Lijuan Wang[1], Wei Han[2], Xiaojun Qian[3], and Frank Soong[1]

[1]Microsoft Research Asia, Beijing, China
[2]Department of Computer Science & Engineering, Shanghai Jiao Tong University, China
[3]Department of Systems Engineering and Engineering Management, Chinese University of Hongkong, Hongkong, China

[1]{lijuanw, frankkps}@microsoft.com; [2]weihan@live.com; [3]xjqian@se.cuhk.edu.hk

Abstract

Synthesis of realistic facial animation for arbitrary speech is an important but difficult problem. The difficulties lie in the synchronization between lip motion and speech, articulation variation under different phonetic context, and expression variation in different speaking style. To solve these problems, we propose a visual speech synthesis system based on a five-state, multi-stream HMM, which generates synchronized motion trajectories for the given text and speech input. Since the motion and the speech are modeled as different but coherent streams, the synchronization at each state is guaranteed. By considering phonetic context and supra-segmental information, the contextual dependent phone models are constructed and clustered using classification and regression, which capture the variable phonetic context and speaking style. The experiment results show that the HMM-based method can generate realistic lip animation while keeping the detailed articulation and transitions. Moreover, it is capable of presenting articulation variation under different phonetic context and expressing various speaking styles, such as emphasized speech.

1. System Introduction

Synthesizing realistic and human-like speech animations is a challenging research topic in both speech and animation communities. Manual approaches which typically involve with selecting/creating key-frames as a basis for generating continuous and natural animations, is a painstaking and tedious task, even for a skillful animator. On the other hand, facial motion capture, widely used in entertainment industry, can acquire high-fidelity motion data. However, it has two main problems: (1) the cost in time and equipment and (2) all needed motion must be recorded beforehand. Therefore, automatic animation synthesis is more desirable, if it can be rendered from pre-recorded motion capture data.

In this work, we propose a novel data-driven speech animation synthesis method. The idea is analogous to the HMM-based speech synthesis technique which forms utterances by predicting the most likely speech parameters from statistically trained HMMs. Given speech and text input, the proposed system can then generate (synthesize) the most likely motion trajectories of both head and critical markers on the face statistically. The synthesized motion trajectories are transformed into control parameters to drive a lively 2D/3D cartoon head.

As shown in Fig.1, the proposed system can be accomplished in four steps: (1) data collection; (2) model training; (3) motion trajectory generation; and (4) animation retargeting. In data collection, with a motion capture system, abundant facial markers' motion trajectories data are collected along with simultaneous audio(speech) and video recordings. The recordings cover rich phonetic (speech) contexts, different speaking styles, lively emotions, and natural facial expressions. In model training, HMM is trained to model captured motion trajectories statistically in the maximum likelihood sense. Since the motion and speech are modeled as different but coherent streams in HMM modeling, the synchronization between motion and speech at each phoneme state is automatically imposed. By considering phonetic context and supra-segmental information, the animation models for each contextual dependent phone are constructed and clustered into a classification and regression tree to characterize the coarticulatory variations in different speaking styles. In motion trajectory synthesis, statistically trained HMMs are used to generate (predict) the most likely motion trajectories, given acoustic and prosodic features of speech. Final rendering of a 2D/3D talking head is done by transforming marker motion trajectories into head and facial control parameters for synthesizing a lively animation sequence. By using animation retargeting techniques, the system can drive any reasonable facial mesh.

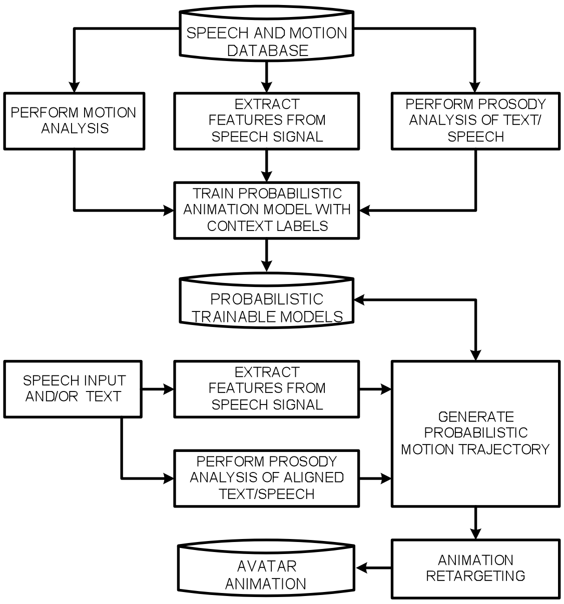
SPEECH AND MOTION DATABASE
PERFORM MOTION ANALYSIS
EXTRACT FEATURES FROM SPEECH SIGNAL
PERFORM PROSODY ANALYSIS OF TEXT/ SPEECH
TRAIN PROBABILISTIC ANIMATION MODEL WITH CONTEXT LABELS
PROBABILISTIC TRAINABLE MODELS
SPEECH INPUT AND/OR TEXT
EXTRACT FEATURES FROM SPEECH SIGNAL
PERFORM PROSODY ANALYSIS OF ALIGNED TEXT/SPEECH
GENERATE PROBABILISTIC MOTION TRAJECTORY
AVATAR ANIMATION
ANIMATION RETARGETING

Fig.1: Flowchart of speech animation synthesis system.

Objective evaluation was conducted by comparing the synthesized facial motion against the captured motion (i.e. the ground truth). The results show that the proposed method is effective for producing realistic speech animations. Subjective comparison with the conventional key-frame based animation synthesis showed that, the HMM-based method can generate more natural lip movements and render realistic coarticulation sequences.