Chapter 6
Learning Expressive Human-Like Head Motion Sequences from Speech

Carlos Busso, Zhigang Deng, Ulrich Neumann, and Shrikanth Narayanan

6.1 Introduction

With the development of new trends in human-machine interfaces, animated feature films, and video games, better avatars and virtual agents are required that more accurately mimic how humans communicate and interact. Gestures and speech are jointly used to express intended messages. The tone and energy of the speech, facial expression, rigid head motion, and hand motion combine in a nontrivial manner as they unfold in natural human interaction. Given that the use of large motion capture data sets is expensive and can only be applied in planned scenarios, new automatic approaches are required to synthesize realistic animation that capture and resemble the complex relationship between these communicative channels. One useful and practical approach is the use of acoustic features to generate gestures, exploiting the link between gestures and speech.

Since the shape of the lips is determined by the underlying articulation, acoustic features have been used to generate visual *visemes* that match the spoken sentences [4, 5, 12, 17]. Likewise, acoustic features have been used to synthesize facial expressions [11, 30], exploiting the fact that the same muscles used for articulation also affect the shape of the face [44, 46]. One important gesture that has received less attention than other aspects in facial animations is rigid head motion.

Head motion is important not only to acknowledge active listening or replace verbal information (e.g., “nod”), but also for many aspect of human communication (for details, see [26]). Graf et al. suggested that rigid head motion is used to segment the linguistic units of spoken content, since the timing between the prosodic structure and head motion is consistent [23]. Head motion also improves acoustic perception, as noted by Munhall et al. [36]. They also suggested that head motion helps to distinguish between interrogative and declarative statements. Hill and Johnston show that head motion is used to recognize speaker identity [27]. Moreover, Jefferies et al. suggest that head motion influences the perception of the personality of the animated character. Similarly, our previous work indicates that head motion affects the emotional perception of facial animations [6].

Given the importance of head motion in human-human interaction, this nonverbal channel needs to be properly modeled for realistic facial animation. Kuratate et al. have estimated the correlation levels between prosodic and head motion features [32]. Based on the high correlation levels achieved ($r = 0.8$), they concluded that the production of speech and head motion are internally linked. Even though head motion patterns depend on many other factors such as the underlying semantic content and the personality of the subjects, these results suggest that speech can be used to generate head motion sequences.

In this chapter, the relationship between rigid head motion and prosodic speech is analyzed in terms of emotional categories (neutral state, sadness, happiness, and anger). The results show that head motion and prosodic speech are strongly connected. However, the relationship varies from emotion to emotion, suggesting that emotional models need to be built to generate realistic head motion sequences. Based on this study, a novel approach to synthesize head motion sequences from prosodic speech is presented. In this framework, head poses are quantized in a finite number of clusters or codebooks. For each of these codebooks, a *hidden Markov model* (HMM) is built, taking prosodic features as observations. In the synthesis step, the acoustic features of the test speech are entered in the HMMs and the most likely head motion sequences are generated. Smoothing techniques based on first-order Markov models followed by spherical cubic interpolation are used to ensure continuous head motion sequences. To include emotional patterns in the generated sequences, different sets of HMMs are built for each emotional category. Evaluations of this framework reveal that the generated sequences follow the temporal dynamics of speech well. Moreover, the generated sequences were judged by human raters at the same level of naturalness as the captured head motion sequences. Previous versions of this framework were published in [6, 7].

This chapter is organized as follows: Section 6.2 presents previous work on head motion synthesis. It also motivates the importance of modeling emotion for engaging animated characters. Section 6.3 describes the audio-visual database and the procedure used to extract the audio-visual features. In Section 6.4, the relationship between head motion and prosodic features is analyzed in terms of emotional categories. Section 6.5 describes the framework used to synthesize head motion sequences. Section 6.6 presents the objective and subjective evaluations of this approach. Finally, Section 6.7 gives the concluding remarks and our future research directions.

6.2 Related Work

6.2.1 Head Motion Synthesis

Different approaches have been used to synthesize head motion sequences, given the relationship between head motion and the verbal message. For instance, plain text enriched with manual annotations of discourse functions were used to synthesize well-known head motion gestures such as head "tilt" and "nod." De Carlo et al. present a coding-based platform for real-time facial animation that supports head motion rotation and translation [14]. The movements of the head are driven by manual annotations of specific head motion gestures co-occurring with prominent words