Speech to Head Gesture Mapping in Multimodal Human-Robot Interaction

Amir Aly and Adriana Tapus

Cognitive Robotics Lab, ENSTA-ParisTech, France
{amir.aly,adriana.tapus}@ensta-paristech.fr

Abstract. In human-human interaction, para-verbal and non-verbal communication are naturally aligned and synchronized. The difficulty encountered during the coordination between speech and head gestures concerns the conveyed meaning, the way of performing the gesture with respect to speech characteristics, their relative temporal arrangement, and their coordinated organization in a phrasal structure of utterance. In this research, we focus on the mechanism of mapping head gestures and speech prosodic characteristics in a natural human-robot interaction. Prosody patterns and head gestures are aligned separately as a parallel multi-stream HMM model. The mapping between speech and head gestures is based on Coupled Hidden Markov Models (CHMMs), which could be seen as a collection of HMMs, one for the video stream and one for the audio stream. Experimental results with Nao robots are reported.

Keywords: Coupled HMM, audio-video signal synchronization, human-robot interaction, signal mapping, robot services.

1 Introduction

Robots are more and more present in our daily lives and the new trend is to make them behave more natural so as to obtain an appropriate social behaviour and response. The work described in this chapter presents a new methodology that allows the robot to automatically adapt its head gestural behavior to the user's profile (e.g. the user prosodic patterns) and therefore to produce a personalizable interaction. This work is based on some findings in the linguistic literature that show that head movements (e.g., nodding, turn taking system) support the verbal stream. Moreover, in human-human communication, prosody expresses the rhythm and intonation of speech and reflect various features of the speakers. These two communication modalities are strongly linked together and synchronized. Humans use gestures and postures as a communicative act. McNeill in [1] defines a gesture as a movement of the body synchronized with the speech flow. The mechanism of the human natural alignment of verbal and non-verbal characteristic patterns based on the work in [2] shows a direct relationship between prosody features and gestures/postures, and constitutes an inspiration for our work.

T. Borangiu, A. Thomas, and D. Trentesaux (Eds.): SOHOMA 2011, SCI 402, pp. 183–196.
springerlink.com

Recently, there has been a growth of interest in socially intelligent robotic technologies featuring flexible and customizable behaviours. Based on the literature in linguistics and psychology that suggests that prosody and gestural kinematics are synchronous and therefore strongly linked together, we posit that it is important to have a robot behaviour that integrates this element. In this chapter, we describe a new methodology for speech prosody and head gesture mapping for human-robot social interaction. The gesture/prosody modelled patterns are aligned separately as a parallel multi-stream HMM model and the mapping between speech and head gestures is based on Coupled Hidden Markov Models (CHMMs). A specific gestural behaviour is estimated according to the incoming voice signal's prosody of the human interacting with the robot. This permits to the robot to adapt its behaviour to the user profile and to produce a personalizable interaction.

To the best of our knowledge, very little research has been dedicated to this investigation area. An attempt is described by the authors in [3] that present a robotic system using dance so as to explore the properties of rhythmic movement in general social interaction. Most of the existing works are related to computer graphics and interactive techniques. A general correlation between head gestures and voice prosody had been discussed in [4] and [5]. The emotional content of the speech can also be correlated to some bodily gestures. In [6], it is discussed the relation between voice prosody and hand gestures, while [7] discusses the relation between the verbal and semantic content and the gesture. In [8], which is somehow closed to the discussed topic on this research, the relation between prosody changes and the orientation of the head (Euler angles) is presented. Moreover, authors in [9], propose a mechanism for driving a head gesture from speech prosody.

Our work presents a framework for head gesture and prosody correlation for an automatic robot gesture production from interacting human user speech. The system is validated with the Nao robot in order to find out how naturalistic will be the driven head gestures from a voice test signal with respect to an interacting human speaker. The rest of the chapter is organized as follows: section 2 presents the applied algorithm extracting the pitch contour of a voice signal; section 3 illustrates the detection of head poses and Euler angles; section 4 describes speech and gesture temporal segmentation; section 5 presents the speech to head gesture coupling by using CHMMs; section 6 resumes the results; section 7 concludes the chapter.

2 Prosodic Features Extraction

In human-robot interaction applications, the human voice signal can convey many messages and meanings, which should be understood appropriately by the robot in order to interact properly.

Next, the methodology used for pitch extraction is described. Talkin [10] defined the pitch as the auditory percept of tone, which is not directly measurable from a signal. Moreover, it is a nonlinear function of the signal's spectral and temporal energy distribution. Another vocal characteristic, the fundamental frequency F0, is measured as it correlates well with the perceived pitch. Voice processing systems that estimate the fundamental frequency F0 often have 3 common processes: (1) Signal Conditioning; (2) Candidate Periods Estimation, and (3) Post