# Prosody-Driven Robot Arm Gestures Generation in Human-Robot Interaction

Amir Aly, Adriana Tapus

# Prosody-Driven Robot Arm Gestures Generation in Human Robot Interaction

Amir Aly
Cognitive Robotics Lab/UEI
ENSTA-ParisTech
32 Blvd Victor, 75015, Paris, France
amir.aly@ensta-paristech.fr

Adriana Tapus
Cognitive Robotics Lab/UEI
ENSTA-ParisTech
32 Blvd Victor, 75015, Paris, France
adriana.tapus@ensta-paristech.fr

## ABSTRACT

In multimodal human-robot interaction (HRI), the process of communication can be established through verbal, non-verbal, and/or para-verbal cues. The linguistic literature [3] shows that para-verbal and non-verbal communications are naturally synchronized. This research focuses on the relation between non-verbal and para-verbal communication by mapping prosody cues to the corresponding arm gestures. Our approach uses the coupled hidden Markov models (CHMMs), which could be seen as a collection of HMMs, one for the prosodic characteristics' stream and six HMMs for the arm movements' stream [4] [1].

## Categories and Subject Descriptors

I.2.9 [**Computing Methodologies**]: Artificial Intelligence—*Robotics*

## General Terms

Experimentation

## 1. INTRODUCTION

The important role of robots in our daily life shows the importance of making robots capable of generating intuitive multimodal communication with people under different interactional contexts. This research tries to understand the natural communication strategies of arm gestures that humans use in their daily interaction in order to transfer them to the robots. The database used in this research is the Stanford database available online [5] which is composed of many avatar videos. The data is acquired from human individuals using the PhaseSpace motion capture system and processed with the MotionBuilder system, which provides an approximation of the Euler rotations of the arms' joints.

## 2. TEMPORAL SEGMENTATION

### 2.1 Speech

Speech is analyzed in terms of the intensity and pitch curves of the voice signal. The inflection points (i.e., zeros crossing points of the rate of change of the curve) of the pitch and intensity curves, and the separating points between the voiced and the unvoiced parts of the voice signal are detected [1]. The trajectories between the detected points in the pitch and intensity curves are then labeled according to the behavior of these trajectories together (see Table 1). Speech signal is segmented into syllables constituting the states of the HMM of the voice stream. The authors in [2] defined the average duration of a syllable as 200ms and this duration can increase or decrease according to the nature of the syllable as being short or long. Practical tests proved that within a syllable of duration varying from 180ms to 220ms, the average number of trajectory classes in its corresponding pitch and intensity curves is around 5, therefore each 5 labels of the signal constitute a state in the HMM of the voice stream.

| Trajectory Class | Trajectory State |
|---|---|
| 1 | Pitch ↑ & Intensity ↑ |
| 2 | Pitch ↑ & Intensity ↓ |
| 3 | Pitch ↓ & Intensity ↑ |
| 4 | Pitch ↓ & Intensity ↓ |
| 5 | Unvoiced segment |

Table 1: Voice Signal Segmentation Labels

### 2.2 Arm Gestures

Arm gestures are characterized in terms of Euler angles of the six articulations (Elbow, Shoulder, and Wrist) of the two arms. Due to the mechanical limitations of the test platform (Nao robot), Euler rotations of the articulations are limited to be: Pitch and Roll - Shoulder; Yaw and Roll - Elbow; and Yaw - Wrist. Roll, Pitch, and Yaw rotations' data indicated in the database are segmented similarly to the voice signal by comparing the trajectories of the relevant Euler curves of each articulation and giving a label according to the behavior of these trajectories together (see Table 2). However, for the wrist articulation, we used the following rule: if the rate of change of the specific trajectory of the yaw curve is increasing, it takes label 1; if it is decreasing, it takes labels 2; or it takes label 3 for no change. Arms articulations are modeled in terms of six independent HMMs that construct together with the HMM of the voice stream the entire CHMM used for mapping the speech to generated mechanical rotations in each articulation in the arms (fig 1)[4]. Each state in the HMMs modeling arm's articulations presents a complete performed gesture, which is presented by 4 labels of the obtained trajectory classes of each articulation.

| | Trajectory State | |
|---|---|---|
| Trajectory Class | Shoulder | Elbow |
| 1 | Pitch ↑ & Roll ↑ | Yaw ↑ & Roll ↑ |
| 2 | Pitch ↑ & Roll ↓ | Yaw ↓ & Roll ↑ |
| 3 | Pitch ↓ & Roll ↑ | Yaw ↑ & Roll ↓ |
| 4 | Pitch ↓ & Roll ↓ | Yaw ↓ & Roll ↓ |
| 5 | No Change | No Change |

Table 2: Shoulder and Elbow Movements Segmentation Labels

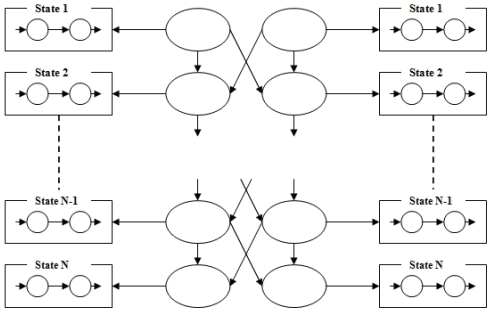


Figure 1: Coupled Hidden Markov Model CHMM mapping between speech to arm gestures' sequences

## 3. SYNTHESIZED GESTURES

The synthesized Euler angles of each articulation are compared to the original Euler angles in terms of the similarity between the trajectory classes (see Table 3). However, during human-human interaction, the generated arm gestures differ from one person to another in terms of the direction and the amplitude of the performed gesture. Therefore, the obtained scores of similarity between the original and synthesized trajectories could be considered as reasonable results (see Figure 2 and 3) because this research focuses on automatic robot arm gestures generation based only on human user prosody.

## 4. CONCLUSIONS

This research focuses on synthesizing robotic arm gestures based on human user speech characteristics (e.g., pitch and intensity of the signal). Our mapping system is based on the Coupled Hidden Markov Model (CHMM) that tries to find a coupling joint between the audio and gesture sequences. The obtained scores of similarity between the trajectories of the synthesized and the original Euler angles are in the range of 55%. Moreover, the synthesized gestures are similar to the real gestures and therefore still relevant to the interaction context and capable of conveying a similar message meaning to the message transmitted by the original gestures.

| Articulation | Similarity Scores between Trajectory Classes |
|---|---|
| Left Shoulder | 47% |
| Left Elbow | 55% |
| Left Wrist | 57% |
| Right Shoulder | 52% |
| Right Elbow | 59% |
| Right Wrist | 61% |

Table 3: Comparison between the original to synthesized gestures' trajectory classes

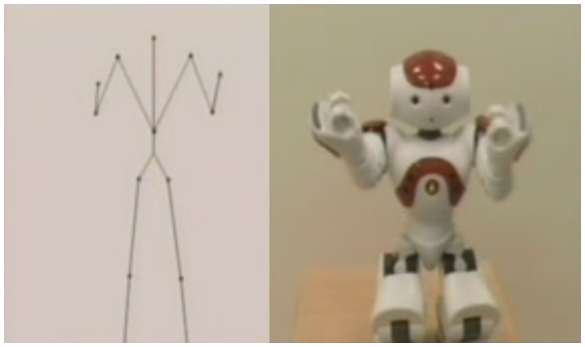

Figure 2: Typical view of the database test avatar and the Nao robot performing its own generated arm gestures

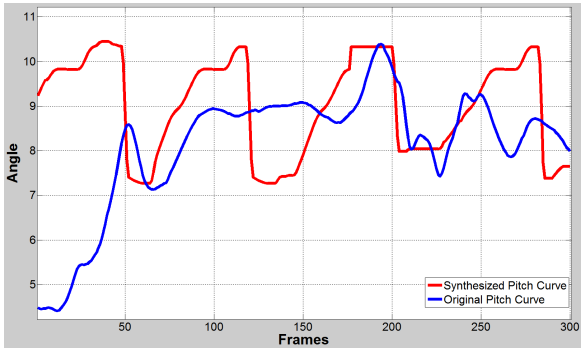


Figure 3: Original and Synthesized pitch curves for the left shoulder articulation

## 5. ACKNOWLEDGEMENTS

This work is supported by the French National Research Agency (ANR) through Chaire d'Excellence program 2009.

## 6. REFERENCES

[1] A. Aly and A. Tapus. Speech to head gesture mapping in multimodal human-robot interaction. In *In Proceedings of the European Conference on Mobile Robotics*, Orebro, Sweden, 2011.

[2] T. Arai and S. Greenberg. The temporal properties of spoken japanese are similar to those of english. In *Proceedings of Eurospeech*, pages 1011–1114, Rhodes, Greece, 1997.

[3] F. P. Eyereisen and J. D. D. Lannoy. *Gestures and Speech: Psychological Investigations*. Cambridge University Press, 1991.

[4] I.Rezek, P. Sykacek, and S. Roberts. Coupled hidden markov models for biosignal interaction modelling. In *Proceedings of the International Conference on Advances in Medical Signal and Information Processing (MEDSIP)*, 2000.

[5] S. Levine, P. Krahenbuhl, S. Thrun, and V. Koltun. Gesture controllers. In *ACM SIGGRAPH*, 2010.