An Integrated Model of Speech to Arm Gestures Mapping in Human-Robot Interaction

Amir Aly, Adriana Tapus

An Integrated Model of Speech to Arm Gestures Mapping in Human-Robot Interaction

Amir Aly * **Adriana Tapus** *

* *Cognitive Robotics Lab/UEI, ENSTA-ParisTech, Paris, France*
(e-mail:amir.aly@ensta-paristech.fr and
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 shows that para-verbal and non-verbal communications are naturally synchronized, however the natural mechnisam of this synchronisation is still largely unexplored. This research focuses on the relation between non-verbal and para-verbal communication by mapping prosody cues to the corresponding metaphoric arm gestures. Our approach for synthesizing arm gestures uses the coupled hidden Markov models (CHMM), which could be seen as a collection of HMM characterizing the segmented prosodic characteristics' stream and the segmented rotation characteristics' streams of the two arms articulations. Experimental results with Nao robot are reported.

Keywords: Pitch contour, Voice intensity, Euler rotations, Coupled Hidden Markov Models (CHMM), Arm getures synthesis

1. INTRODUCTION

Developing social intelligent robots capable of behaving naturally and of producing appropriate social behaviors and responses to humans in different contexts is a difficult task. The work described in this research presents a new methodology that allows the robot to automatically adapt its arm 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 arm movements support the verbal stream. Moreover, in human-human communication, prosody expresses the rhythm and intonation of speech and reflects various features of the speakers. Thus, these two communication modalities are strongly linked together and synchronized.

Humans use gestures and postures as a communicative act. McNeil (1992), defined a gesture as a movement of the body synchronized with the flow of speech. During human-human interaction, gestures and speech are simultaneously used to express not only verbal information, but also important communicative non-verbal cues that enrich, complement, and clarify the conversation, such as emotional internal states, facial expressions, head and/or hands motion. The mechanism of the human natural alignment of the verbal and non-verbal characteristic patterns based on the work described in Eyereisen and Lannoy (1991), shows a direct relationship between prosody features and gestures/postures, and constitutes an inspiration for our work.

Recently, there has been a growth of interest in socially intelligent robotic technologies featuring flexible and customizable behaviors. 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 is important to have a robot behavior that integrates this element. Therefore, in this research, we discuss mapping between speech prosody and arm gestures for human-robot social interaction. The gesture/prosody modeled patterns are aligned separately as a parallel multi-stream HMM model and the mapping between speech and arm gestures is based on Coupled Hidden Markov Models (CHMM). A specific gestural behavior is estimated according to the incoming voice signal's prosody of the human interacting with the robot. This permits the robot to adapt its behavior to the user's profile (e.g. here the user prosodic patterns) and therefore to produce a personalizable interaction. The current work is a natural continuation of our previous research that synchronizes head gestures and prosody(Aly and Tapus (2011)).

Many researches in the literature that try to investigate the relationship between sound signal and arm gestures are designed for 3D artifical agents. Modler (1998), tried to map hand gestures to musical parameters in an interactive music performance and virtual reality environment, while Kipp et al. (2007), focused on synthesizing arm gestures with a 3D virtual agent. An interesting coupling between virtual agents and humanoid robots is achieved by Salem et al. (2010), in which an articulated communicator engine is developed to allow virtual agents to flexibly realize a multi-modal behavior which is presented on the robot ASIMO as an interaction mediator between the human and the robot.

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

Our work presents a framework for arm gestures 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 arm gestures from a voice test signal with respect to an interacting human speaker.

The rest of the paper is organized as following: section 2 illustrates the calculation of the prosody characterizing pitch curve; section 3 describes speech and gesture temporal segmentation; section 4 presents the speech to arm gestures coupling by using CHMM; section 5 resumes the results obtained; and finally, section 6 concludes the paper.

2. PROSODIC CURVES CALCULATION

Human's voice signal can convey many messages and meanings, which should be understood appropriately by the robot in order to generate gestures properly. In this research we characterize the voice signal in terms of its pitch and intensity curves. The intensity curve of the voice signal could be calculated directly by calculating the square value of each signal data element normalized by the sampling frequency. Meanwhile, the calculation of pitch contour has some complexity. Talkin (1995), 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. Instead, 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 three common processes: (1) Signal conditioning, (2) Candidate periods estimation, and (3) Post processing. Signal conditioning process tries to remove interfering signal components such as any extraneous noise by using low pass filtering which removes the apparent loss of periodicity in the voiced signal spectrum at higher frequencies, and by using high pass filtering when there are DC or very low frequency components in the signal. Another important conditioning step is using the auto-regressive inverse filtering to flatten the vocal signal spectrum which is helpful in detecting the glottal epochs (moments of significant glottal excitation), which ameliorates in turn the detection of the voiced and unvoiced spectrums of the signal, so that helps in a better calculation for the fundamental frequency F0. Candidate periods estimation step tries to estimate the candidate voiced periods from which the fundamental frequency F0 could be calculated. A major problem is that the glottal excitation periods are varying through the signal. Many algorithms in the literature tried to deal with this problem, e.g., Autocorrelation, Cepstrum, Cross Correlation, and Normalized Cross Correlation (NCC) (Sondhi (1968)). However, the NCC proved its superiority in terms of measuring the fast variations of the dynamic voice signal. Talkin (1995), developed the traditional (NCC) method in order to estimate reliably the voicing periods and the fundamental frequency $F0$ by considering all candidates simultaneously in a large temporal context. This methodology uses two pass normalized cross correlation (NCC) calculation for searching the fundamental frequency $F0$ which reduces the overall computation load with respect to the traditional (NCC) method. Post processing step uses median filtering in order to refine the calculated fundamental frequency F0 and ignore isolated outliers (see Figure 1).

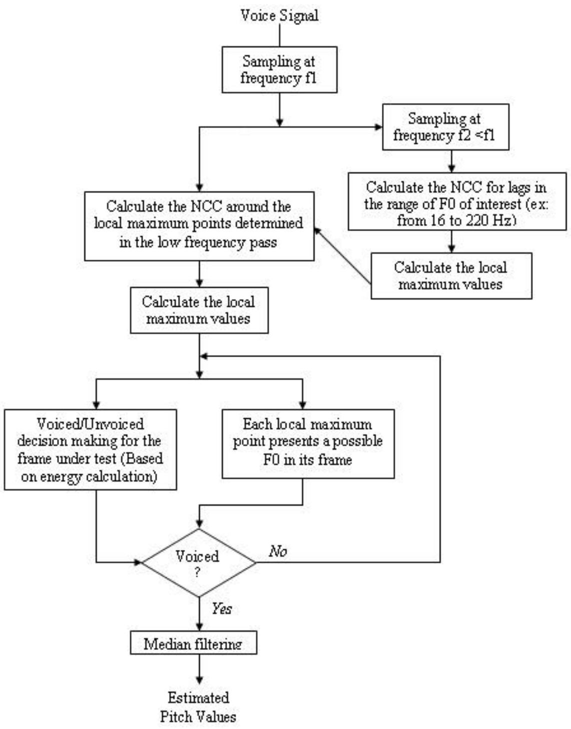
Voice Signal
Sampling at frequency f1
Sampling at frequency f2 <f1
Calculate the NCC for lags in the range of F0 of interest (ex: from 16 to 220 Hz)
Calculate the NCC around the local maximum points determined in the low frequency pass
Calculate the local maximum values
Calculate the local maximum values
Voiced/Unvoiced decision making for the frame under test (Based on energy calculation)
Each local maximum point presents a possible F0 in its frame
Voiced ?
No
Yes
Median filtering
Estimated Pitch Values

Fig. 1. Pitch Tracking

3. SPEECH AND ARM GESTURES SEGMENTATION

The mapping between speech and arm gestures is done by using the Coupled Hidden Markov Models (CHMM), which could be seen as a collection of HMM for the audio and video streams. The advantage of this model over a lot of other topologies is its ability to capture the dual influences of each stream on the other one across time (see Figure 3). In the beginning, speech and arm gestures streams are aligned separately as parallel multi-stream HMM models. The mapping between speech and arm gestures is performed in 2 main steps: (1) the first is modeling the gesture sequences and the associated voice prosody sequence (in terms of their characteristic vectors) into separate HMM; (2) then after training both models, a correlation between the HMM models is necessary so as to estimate a final arm gesture states sequences given a speech test signal. The HMM structure used in analyzing gestures (and similarly voice prosody) is indicated in Figure 2. It is composed of N parallel states, where each one represents a gesture composed of M observations. The goal of the transition between states S_{END} to S_{START} is to continue the transition between states from 1 to N (e.g., after performing gesture state 1, the model transfers from the transient end state to the start state to perform any gesture state from 2 to N in a sequential way and so on).

In order to be able to model gestures/prosody, it is necessary to make a temporal segmentation of the video content to detect the M number of observations in each state and the total number of states N.

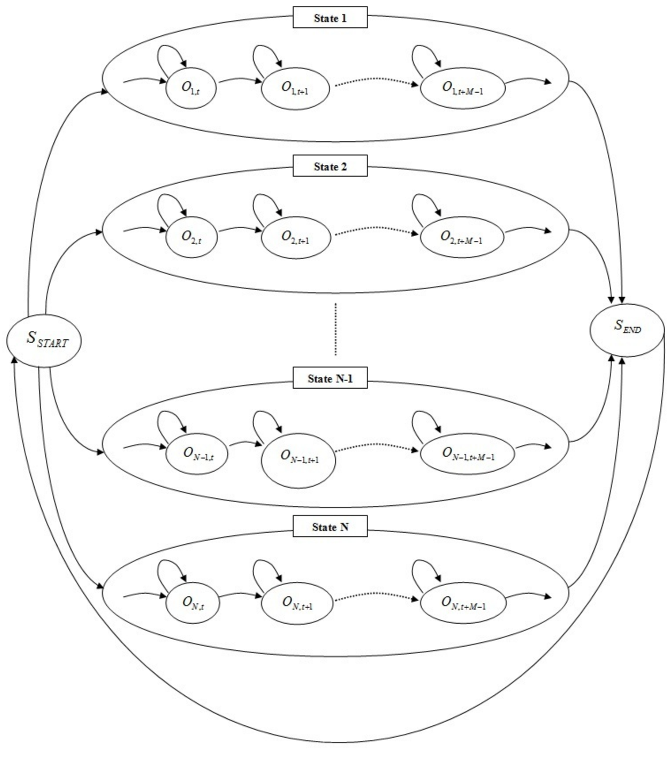

Fig. 2. HMM structure for gesture and prosody analysis

Table 1. Voice Signal Segmentation Labels

Trajectory Class	Trajectory State
1	pitch ↑ & intensity ↑
2	pitch ↑ & intensity ↓
3	pitch ↓ & intensity ↑
4	pitch ↓ & intensity ↓
5	Unvoiced segment

3.1 Speech Temporal Segmentation

Speech is segmented as syllables presented by the states from 1 to N as indicated in Figure 2. The segmentation is performed by intersecting the inflection points (zeros crossing points of the rate of change of the curve) for both the pitch and intensity curves, beside the points that separate between the voiced and unvoiced segments of the signal When comparing the two curves together, 5 different trajectory states can result (see Table 1).

The goal is to code each segment of the signal with its corresponding pitch-intensity trajectory class (e.g., a voice signal segment coding could be: $5, 3, 4, 2$, etc.). This segmental coding is used as label for CHMM training. The next step corresponds to segmenting the voice signal with its corresponding trajectory labeling into syllables. Arai and Greenberg (1997), defined the average duration of a syllable as 200 ms 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 180 ms to 220 ms, the average number of trajectory classes in its corresponding pitch and intensity curves is around 5. Therefore, given the voice signal with its segments coded by the corresponding pitch-intensity trajectory labels, each 5 segments of the signal will create a syllable state (from 1 to N) and the corresponding 5 labels will be the observations M within the syllable state.

Table 2. Shoulder and Elbow Movements Segmentation Labels

	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

3.2 Arms' Gestures Temporal Segmentation

Arms' 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:

- Shoulder: Pitch and Roll
- Elbow: Yaw and Roll
- Wrist: Yaw

Roll, Pitch, and Yaw rotations' data indicated in the database (see the experimental section) 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.

The articulations of the arms are modeled in terms of six independent HMM presenting the mechanical rotations of the two arms. Each state in the HMM presents a complete performed gesture, which is presented by 4 labels of the obtained trajectory classes of each articulation (Aly and Tapus (2011)).

4. SPEECH TO ARM GESTURES COUPLING

A typical CHMM structure is shown in Figure 3, where the circles present the discrete hidden nodes/states while the rectangles present the observable continuous nodes/states, which contain the observation sequences of voice and arm gestures characteristics.

According to the sequential nature of gestures and speech, the CHMM structure is of type lag-1 in which couple (backbone) nodes at time t are conditioned on those at time $t-1$ (Rabiner (1989); Rezek et al. (2000); Rezek and Roberts (2000)). A CHMM model λ_C is defined by the following parameters:

$$\pi_0^C(i) = P(q_1^C = S_i) \tag{1}$$

$$a_{i|j,k}^C = P(q_t^C = S_i | q_{t-1}^{audio} = S_j, q_{t-1}^{video} = S_k) \tag{2}$$

$$b_t^C(i) = P(O_t^C | q_t^C = S_i) \tag{3}$$

where $C \in \{audio, video\}$ denotes the audio and visual channels respectively, and q_t^C is the state of the coupling

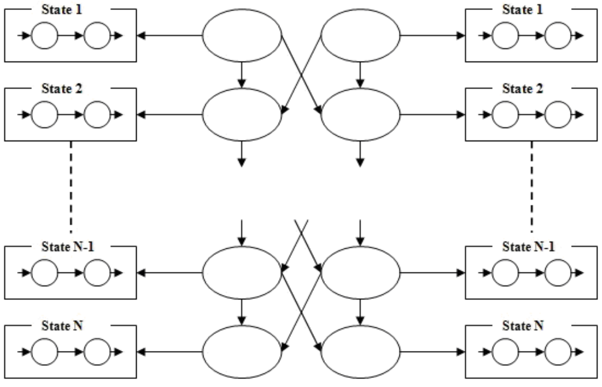

Fig. 3. Coupled Hidden Markov Model CHMM lag-1 Structure

node in the c_{th} stream at time t (Nean et al. (2002); Liang et al. (2002)).

The training of this model is based on the maximum likelihood form of the expectation maximization (EM) algorithm. Supposing there are 2 observable sequences of the audio and video states $O = \{A_{1..N}, B_{1..N}\}$ where $A_{1..N} = \{a_1, \cdots, a_N\}$ is the set of observable states in the first audio sequence, and similarly $B_{1..N} = \{b_1, \cdots, b_N\}$ is the set of observable states in the second visual sequence, and $S = \{X_{1..N}, Y_{1..N}\}$ is the set of states of the couple nodes at the first audio chain and the second visual chain, respectively (Rezek et al. (2000); Rezek and Roberts (2000)). The expectation maximization algorithm finds the maximum likelihood estimates of the model parameters by maximizing the following function:

$$f(\lambda_C) = P(X_1)P(Y_1)\prod_{t=1}^{T} P(A_t|X_t)P(B_t|Y_t) \qquad (4)$$
$$P(X_{t+1}|X_t, Y_t)P(Y_{t+1}|X_t, Y_t), 1 \leq T \leq N$$

where:

- $P(X_1)$ and $P(Y_1)$ are the prior probabilities of the audio and video chains respectively
- $P(A_t|X_t)$ and $P(B_t|Y_t)$ are the observation densities of the audio and video chains respectively
- $P(X_{t+1}|X_t, Y_t)$ and $P(Y_{t+1}|X_t, Y_t)$ are the couple nodes transition probabilities in the audio and video chains.

The training of the CHMM differs from the standard HMM in the expectation step (E) while they are both identical in the maximization step (M) which tries to maximize equation 4 in terms of the expected parameters (Penny and Roberts (1998)). The expectation step of the CHMM is defined in terms of the forward and backward recursion. For the forward recursion we define a variable for the audio and video chains at $t = 1$:

$$\alpha_{t=1}^{audio} = P(A_1|X_1)P(X_1) \qquad (5)$$
$$\alpha_{t=1}^{video} = P(B_1|Y_1)P(Y_1) \qquad (6)$$

Then the variable α is calculated incrementally at any arbitrary moment t as follows:

$$\alpha_{t+1}^{audio} = P(A_{t+1}|X_{t+1}) \int\int \alpha_t^{audio}\alpha_t^{video}$$
$$P(X_{t+1}|X_t, Y_t)dX_t dY_t \qquad (7)$$

$$\alpha_{t+1}^{video} = P(B_{t+1}|Y_{t+1}) \int\int \alpha_t^{audio}\alpha_t^{video}$$
$$P(Y_{t+1}|X_t, Y_t)dX_t dY_t \qquad (8)$$

Meanwhile, for the backwards direction there is no split in the calculated recursions, which can be expressed as follows:

$$\beta_{t+1}^{audio,video} = P(O_{t+1}^N|S_t) =$$
$$\int\int P(A_{t+1}^N, B_{t+1}^N|X_{t+1}, Y_{t+1})$$
$$P(X_{t+1}, Y_{t+1}|X_t, Y_t)dX_{t+1}dY_{t+1} \qquad (9)$$

After combining both forward and backwards recursion parameters (see equations 7, 8, 9), an audio signal is tested on the trained model, generating a synthesized equivalent gesture that most likely fit the model. The generated gesture sequence is determined when the change in the likelihood is below a fixed threshold. Figure 4 summarizes the entire process.

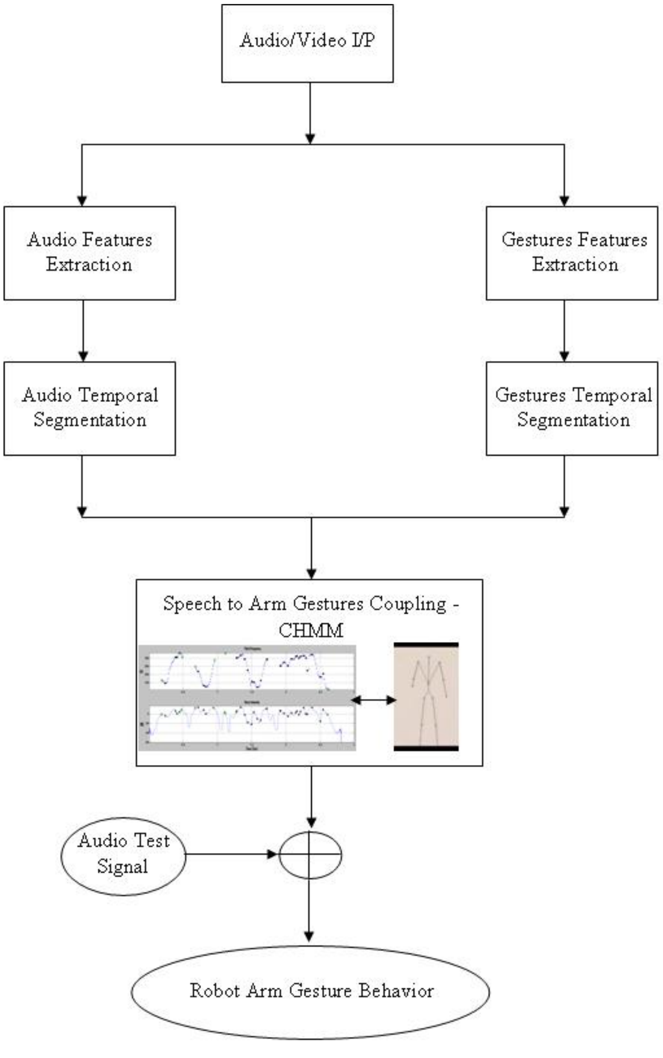
Audio/Video I/P
Audio Features Extraction
Gestures Features Extraction
Audio Temporal Segmentation
Gestures Temporal Segmentation
Speech to Arm Gestures Coupling - CHMM
Audio Test Signal
Robot Arm Gesture Behavior

Fig. 4. The architecture of the robot arm gesture generation behavior

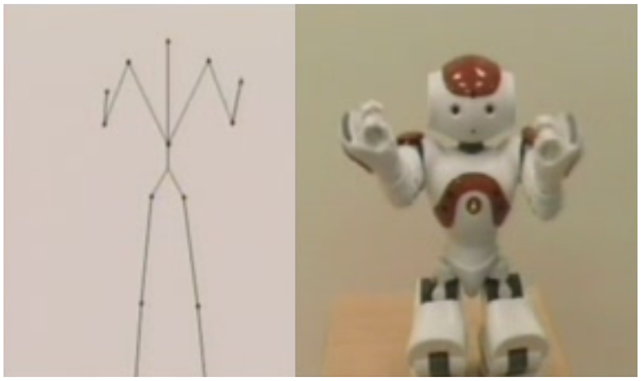

Fig. 5. Typical view of the database test avatar and the Nao robot performing its own generated arm gestures

5. EXPERIMENTAL RESULTS

The database used in this research is the Stanford database available online (Levine et al. (2010)), 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.

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 5, 6, and 7) because this research focuses on automatic robot arm gestures generation based only on human user prosody.

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

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

A video of the speech-arms' gestures mapping system with Nao robot is available at: `http://www.ensta-paristech.fr/~tapus/HRIAA/media.html`.

6. CONCLUSION

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 Models (CHMM) that try 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.

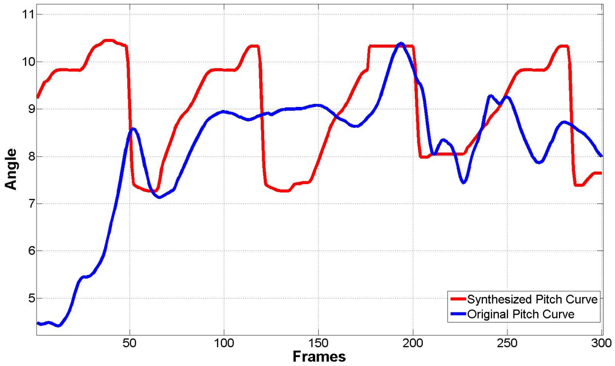

Fig. 6. Original and Synthesized pitch curves of the left shoulder articulation

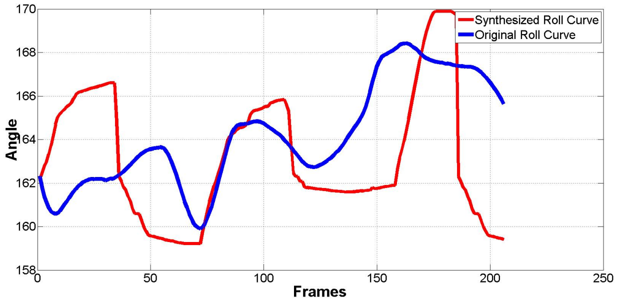

Fig. 7. Original and Synthesized roll curves of the right shoulder articulation

Further, we will focus on combining head gestures, arm gestures, and prosody all together. While this work aims to generate metaphoric arm gestures based on voice tonality, other gesture categories (e.g., iconic gestures) could not be generated by this methodology, which necessitates a semantic analysis of the spoken language in order to understand the conveyed meaning and to align it to the performed gesture. In this way, the robot after a sufficient training phase on coupled verbal-nonverbal behaviors could generate similar iconic gestures in different interactional situations.

REFERENCES

Aly, A. and Tapus, A. (2011). Speech to head gestures mapping in multimodal human-robot interaction. In *proceedings of the European Conference on Mobile Robotics (ECMR)*. Orebro, Sweden.

Arai, T. and Greenberg, S. (1997). The temporal properties of spoken japanese are similar to those of english. In *proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)*, 1011–1114. Rhodes,Greece.

Eyereisen, F.P. and Lannoy, J.D.D. (1991). *Gestures and Speech: Psychological Investigations*. Cambridge University Press, USA.

Kipp, M., Neff, M., Kipp, K.H., and Albrecht, I. (2007). Towards natural gesture synthesis:evaluating gesture units in a data-driven approach to gesture synthesis.

In *proceedings of the 7th International Conference on Intelligent Virtual Agents*. Springer.

Levine, S., Krahenbuhl, P., Thrun, S., and Koltun, V. (2010). Gesture controllers. In *proceedings of the 37th International conference and Exhibition on Computer Graphics and Interactive Techniques, ACM SIGGRAPH*. Los Angeles, USA.

Liang, L., Liu, X., Pi, X., Zhao, Y., and Nean, A.V. (2002). Speaker independent audio-visual continuous speech recognition. In *proceedings of the International Conference on Multimedia and Expo (ICME)*, volume 2, 2528. Lausanne, Switzerland.

McNeil, D. (1992). *Hand and mind : what gestures reveal about thought*. University of Chicago Press, Chicago, USA.

Modler, P. (1998). Neural networks for mapping hand gestures to sound synthesis parameters. In *proceedings of the 5th Brazilian Symposium of Computer and Music*.

Nean, A.V., Liang, L., Pi, X., Liu, X., and Mao, C. (2002). A coupled hidden markov model for audio-visual speech recognition. In *proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, 20132016. Orlando, USA.

Penny, W. and Roberts, S. (1998). Gaussian observation hidden markov models for eeg analysis. In *Technical Report TR-98-12*. Imperial College, London, UK.

Rabiner, L.R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *proceedings of the IEEE*, volume 77, 257286.

Rezek, I. and Roberts, S.J. (2000). Estimation of coupled hidden markov models with application to biosignal interaction modelling. In *proceedings of the IEEE International Workshop on Neural Networks for Signal Processing (NNSP)*. Sydney, Australia.

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

Salem, M., Kopp, S., Wachsmuth, I., and Joublin, F. (2010). Towards an integrated model of speech and gesture production for multi-modal robot behavior. In *proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Viareggio, Italy.

Sondhi, M.M. (1968). New methods of pitch extraction. *IEEE Trans. Audio and Electroacoustics*, 16, 262–266.

Talkin, D. (1995). A robust algorithm for pitch tracking. In *Speech Coding and Synthesis*, 497–518. W B Kleijn, K Paliwal eds, Elsevier.