Prosody-Based Adaptive Metaphoric Head and Arm Gestures Synthesis in Human Robot Interaction

Amir Aly, Adriana Tapus

Prosody-Based Adaptive Metaphoric Head and Arm Gestures Synthesis in Human Robot Interaction

Amir Aly
Cognitive Robotics Laboratory
ENSTA ParisTech
Palaiseau Cedex, France
Email: amir.aly@ensta-paristech.fr

Adriana Tapus
Cognitive Robotics Laboratory
ENSTA ParisTech
Palaiseau Cedex, France
Email: adriana.tapus@ensta-paristech.fr

***Abstract*—In human-human interaction, the process of communication can be established through three modalities: verbal, non-verbal (i.e., gestures), and/or para-verbal (i.e., prosody). The linguistic literature shows that the para-verbal and non-verbal cues are naturally aligned and synchronized, however the natural mechanism of this synchronization is still unexplored. The difficulty encountered during the coordination between prosody and metaphoric head-arm gestures concerns the conveyed meaning, the way of performing gestures with respect to prosodic characteristics, their relative temporal arrangement, and their coordinated organization in the phrasal structure of utterance. In this research, we focus on the mechanism of mapping between head-arm gestures and speech prosodic characteristics in order to generate an adaptive robot behavior to the interacting human's emotional state. Prosody patterns and the motion curves of head-arm gestures are aligned separately into parallel Hidden Markov Models (HMM). The mapping between speech and head-arm gestures is based on the Coupled Hidden Markov Models (CHMM), which could be seen as a multi-stream collection of HMM, characterizing the segmented prosody and head-arm gestures' data. An emotional state based audio-video database has been created for the validation of this study. The obtained results show the effectiveness of the proposed methodology.**

I. Introduction

Developing intelligent robots able to behave and interact naturally and to generate appropriate social behaviors to humans in different interaction contexts, so that make them believe in the robots' communicative intents, is not a trivial task. The work described in this paper is based on some findings in the literature, which show that head-arm movements (e.g., nodding, turn-taking system, waving, etc) are synchronized with the verbal and para-verbal cues. It presents a new methodology that allows the robot to automatically adapt its head-arm gestural behavior to the user's emotional profile, and therefore, to produce a personalized interaction.

Humans use gestures and postures in communicative acts. McNeill and Kendon in [1], [2] defined a gesture as a body movement synchronized with the flow of speech, that is strongly related parallelly or complementarily to the semantic meaning of the utterance. During human-human interaction, gestures and speech are simultaneously used to express not only verbal and para-verbal information, but also important communicative non-verbal cues that enrich, complement, and clarify the conversation, such as: facial expressions, head movements, and/or arm-hand movements. The human natural alignment of the three communication modalities described in [3], [4], shows a relationship between prosody and gestures/postures, which constituted our inspiration for this work.

The literature reveals a lot of efforts towards understanding the semiotic references (i.e., pragmatic and semantic) of gestures [5], [6]. The encountered complexity in understanding the semiotics of gestures indicates the need for a broad classification of gestures, in order to better characterize what is happening within a human-robot interaction situation.

Different categories of gestures were discussed in the literature. Ekman et al., in [7] identified five gesture categories: (1) emblems (e.g., waving goodbye and shoulder shrugging), (2) illustrators (e.g., pointing gestures), (3) facial expressions, (4) regulators (e.g., head, eyes, arm-hand movements, and body postures), (5) adaptors (e.g., scratching). On the other hand, Kendon in [8] criticized the classification of Ekman for neglecting the linguistic phenomena. He proposed a new classification for gestures of four categories: (1) gesticulation (e.g., gestures which accompany speech), (2) pantomime (e.g., sequence of gestures with a narrative structure), (3) emblem (e.g., Ok-gesture), (4) signs of a sign language. McNeill in [1] collected these four types in a continuum called *Kendons continuum*. This continuum was later elaborated into four main types of widely cited gesture categories: (1) iconics (e.g., gestures representing images of concrete entities and/or actions, like when accompanying the adjective *narrow* with gesturing the two hands in front of each other with a small span in-between), (2) metaphorics (e.g., gestures representing abstract ideas), (3) deictics (e.g., pointing gestures), (4) beats (e.g., hand, finger, or arm movements performed side to side with the rhythmic pulsation of speech).

Iconic and metaphoric gestures (according to McNeill's categorization) constitute the main body of the generated non-verbal behavior during human-human interaction. Many researches have focused on generating both kinds of gestures in human-robot and human-computer interaction applications. Cassell et al., in [9] proposed a rule-based gesture generation toolkit (BEAT) using the natural language processing (NLP) of an input text, producing an animation script that can be used to animate both virtual agents (e.g., the conversational agent REA) [10], and humanoid robots [11]. This system can synthesize gestures of different categories (including iconic

gestures) except for metaphoric gestures. Similarly, Pelachaud in [12] developed the 3D virtual conversational agent GRETA, which can generate a synchronized multimodal behavior to human users. GRETA can generate all kinds of gestures regardless of the domain of interaction, unlike the other 3D conversational agents (e.g., MAX agent [13]). It takes a text as input to be uttered by the agent, and then it tags it with the communicative functions information. The tag language is called Affective Presentation Markup Language (APML) [14], which is used as a script language to control the animation of the agent. Recently, an interesting architecture has been discussed in [15], which proposes a common framework that generates a synchronized multimodal behavior for a humanoid robot, as well as for the agent GRETA. Another competitive approach based on processing an input text in order to generate a corresponding set of different gestures for animated agents (including metaphoric gestures only), was discussed in [16], in which the authors proposed a probabilistic synthesis method trained on hand-annotated videos. Similarly, another system was illustrated in [17], which can synthesize different types of gestures for humanoid robots (including metaphoric and iconic gestures) corresponding to an input text through a part-of-speech tagging analysis. In general, the fact that these methods are based on synthesizing gestures from an input text, makes them unable to measure the different meanings that a text may have, which could be conveyed mostly through prosody. Besides, it makes them unable to measure emotions that influence body language, which may hinder generating a robot's behavior adapted to human's emotional state [18].

Another interesting approach towards generating iconic gestures was discussed in [13], in which the authors developed the 3D virtual conversational agent MAX, which uses synchronized speech and gestures to interact multimodally with humans (e.g., describing a place multimodally based on some prescribed dimensional knowledge about that place). It has the advantage that it can synthesize new unprescribed iconic gestures according to the context of interaction in a specific domain (unlike BEAT system, which is a rule-based gesture generator). However, it is -still- away from considering human's emotional state, when generating a multimodal behavior, in which voice prosody correlates with the internal emotional state and body language of human.

On the way towards generating an animation script based on speech features, Bregler et al., and Brand in [19], [20] studied the relationship between phonemes and facial expressions. Sargin et al., in [21] proposed a time-costly probabilistic model to synthesize metaphoric head gestures from voice prosody. A similar approach was discussed in [22], which uses the features of head gestures and voice prosody to create a training database for a statistical model that can generate a set of motion sequence for 3D agents. Another interesting approach was discussed in [23], which selects animation segments from a motion database based on an audio input, and then synthesizes these segments into metaphoric head-arm gestures animating 3D agents. Despite these interesting approaches, the relationship between human's emotional state and head-arm gestures in human-robot interaction is still incompletely addressed, which constituted our motivation for this work.

The rest of the paper is organized as following: Section (II) presents an overview for the whole system, Section (III) presents the database used in this research, Section (IV) illustrates the analysis of gesture kinematics, Section (V) illustrates data segmentation, Section (VI) validates the chosen voice-gesture characteristics, Section (VII) describes data quantization, Section (VIII) explains the coupling between speech and head-arm gestures using the CHMM, Section (IX) describes the synthesis of customized head-arm gestures to emotional state, and last but not least, Section (X) concludes the paper.

II. SYSTEM OVERVIEW

The system is coordinated through three stages, as illustrated in Figure (1). Stage 1 represents the training stage of the system, in which the raw audio and video training inputs get analyzed in order to extract relevant characteristics (e.g., the pitch-intensity curves for voice and the motion curves for gesture). Afterwards, the extracted characteristic curves go to the segmentation phase and then to the Coupled Hidden Markov Models (CHMM) phase. Gesture and prosody segmented patterns are modeled separately into parallel HMM, composing the CHMM [24], [25], through which new *adaptive* head-arm gestures are synthesized (i.e., stage 2) based on the prosodic patterns of a new speech-test signal which undergoes the same phases of the training stage. The main advantages of using the CHMM for generating gestures are: the random variations of the generated gestures' patterns, which make them more human-like than if a fixed gesture dictionary is used, and the ability to generate gestures of varying durations and amplitudes adapted to the prosody patterns of the human. In order to create a successful long term human-robot interaction (i.e., stage 3), the robot should be able to increase online its initial learning database by acquiring more raw audio and video data from humans in the surrounding of the robot. This requires the Kinect sensor that can calculate in real time the rotation curves of head and arms' articulations, beside a microphone that can receive the interacting human's audio signal. Afterwards, both audio and video captured data will follow the previously explained phases of the training stage 1, increasing the robot's ability to generate more appropriate gestures. Similarly, a new speech-test signal from one of the individuals around the robot will follow the phases of the test stage 2. In this work, we will focus on stages 1-2 and we will validate their theoretical bases. However, stage 3 represents a future experimental stage towards a complete human-robot interaction architecture.

III. DATABASE

The synchronized audio-video database used in this research was captured by MOCAP recorder, and the roll-pitch-yaw rotations of body articulations were tracked frame-by-frame by MOCAP studio. The total duration of the database is around 90 minutes, divided into six categories of pure

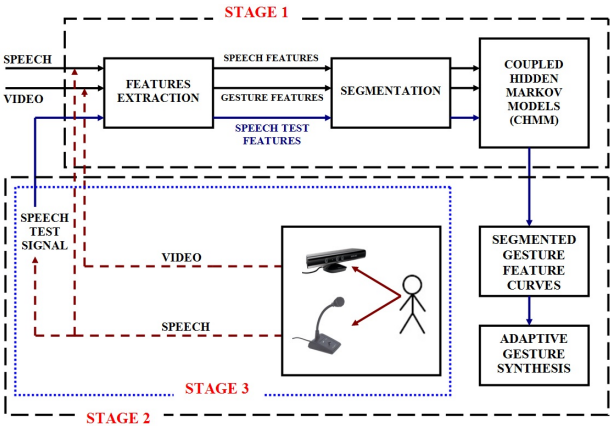

Fig. 1. System Overview

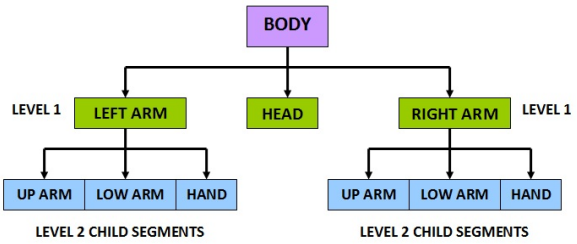

Fig. 2. Parent-Child Hierarchy

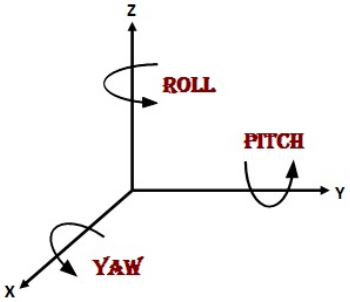

Fig. 3. Roll-Pitch-Yaw Rotations

continuous emotion expression: Sadness, Surprise, Disgust, Anger, Fear, and Neutral. The chosen emotions constitute the main primary emotions stated by most of the contemporary theories of emotions [26], [27]. We have not tried to include any complex emotion [27] to the database, because it is difficult to make the actors express continuously a complex emotion for several minutes. The motion files (.bvh) of our database are available at: http://www.ensta.fr/~tapus/HRIAA/media/MotionDataBaseAlyTapus.rar.

IV. GESTURE KINEMATIC ANALYSIS

The hierarchical construction of human body could be imagined as linked segments that can move together or independently. The segments called *parent*, are the segments composed of other child segments (e.g., the parent segment *arm* is composed of 3 child segments *up-arm, low-arm, hand* (level 2), however the *arm* is considered as a child segment (level 1) for the main parent segment *body*) [28]. This parent-child relationship of body segments allows the inheritance of motion characteristics from the parent to child segments, and vice versa. In this research, we assume that the legs, waist, and torso keep static during emotion expression, so that for the parent segment *body*, the child segments are limited to *head*, *left arm*, and *right arm* as illustrated in Figure (2). The kinematic characteristics of body gestures during emotion expression could be studied in terms of the linear velocity and acceleration of segments, in addition to the position and displacement of articulations (except for the head, which will be characterized in terms of the linear velocity and acceleration only considering the small motion domain of the head).

A. Linear Velocity and Acceleration of Body Segments

The angular velocity and acceleration of level 2 body segments could be expressed in terms of the roll-pitch-yaw right-handed rotations of the corresponding articulations obtained from the generated frame-by-frame report of MOCAP studio.

Considering the ZYX coordinates axes indicated in Figure (3), the rotation about the reference z-axis is denoted by ϕ (Roll), meanwhile the rotation about the reference y-axis is denoted by θ (Pitch), and the rotation about the reference x-axis is denoted by ψ (Yaw). The angular velocity of a child segment through each frame could be expressed it terms of the 3 rotations of its corresponding articulation [29], as indicated in Equation (1):

$$\omega = \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} = \begin{pmatrix} 0 & -sin\phi & cos\phi\ cos\theta \\ 0 & cos\phi & sin\phi\ cos\theta \\ 1 & 0 & -sin\theta \end{pmatrix} \begin{pmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{pmatrix} \quad (1)$$

Where the derivatives of the roll-pitch-yaw rotations through each frame could be calculated from the time rate of change of the specific rotation value in the current frame with respect to the previous frame. Similarly, the angular acceleration could be calculated from the time derivative of the angular velocity, as indicated in Equation (2):

$$\begin{aligned} \dot{\omega} = \begin{pmatrix} \dot{\omega}_x \\ \dot{\omega}_y \\ \dot{\omega}_z \end{pmatrix} &= \begin{pmatrix} 0 & -sin\phi & cos\phi\ cos\theta \\ 0 & cos\phi & sin\phi\ cos\theta \\ 1 & 0 & -sin\theta \end{pmatrix} \begin{pmatrix} \ddot{\phi} \\ \ddot{\theta} \\ \ddot{\psi} \end{pmatrix} \\ &+ \begin{pmatrix} -cos\phi & -sin\phi\ cos\theta & -cos\phi\ sin\theta \\ -sin\phi & cos\phi\ cos\theta & -sin\phi\ sin\theta \\ 0 & 0 & -cos\theta \end{pmatrix} \begin{pmatrix} \dot{\phi}\ \dot{\theta} \\ \dot{\phi}\ \dot{\psi} \\ \dot{\theta}\ \dot{\psi} \end{pmatrix} \end{aligned} \quad (2)$$

B. Body Segment Parameters Calculation

The parameters of body segments required for the kinematic analysis of body gestures are:

- The mass of body segments (i.e., head, upper arm, lower arm, hand), which is concentrated in the center of mass of the segment.
- The length of body segments.
- The proximal distance from each calculated center of mass to the nearest articulation in the segment.

The literature of kinetics illustrates big efforts towards stating a unified mathematical representation of human body including the previously mentioned parameters, however the outcome was always approximate and different from a research to another [30], [31], [32]. For the calculation of the mass of each body segment required for gesture segmentation (as discussed in Section V-A), we used the highly cited relationships stated in [33], as indicated in Equation (3) (where M denotes the total body mass):

$$\begin{aligned} Head\ Mass &= 0.0307 * M + 2.46 \\ Up\ Arm\ Mass &= 0.0274 * M - 0.01 \\ Low\ Arm\ Mass &= 0.70 * (0.0233 * M - 0.01) \\ Hand\ Mass &= 0.15 * (0.0233 * M - 0.01) \end{aligned} \tag{3}$$

Similarly, the length of each body segment could be calculated in terms of the person's height using the following approximate relationships (Equation 4) [34]:

$$\begin{aligned} Neck\ Length &= 0.052 * Person\ Height \\ Up\ Arm\ Length &= 0.187 * Person\ Height \\ Low\ Arm\ Length &= 0.1455 * Person\ Height \\ Hand\ Length &= 0.108 * Person\ Height \\ Shoulder\ Length &= 0.129 * Person\ Height \end{aligned} \tag{4}$$

The Neck and the shoulder are not considered as body segments. However, the length of the neck is required for calculating the proximal distance from the head's center of mass to the proximal joint of the upper neck (Equation 5), in addition to calculating the Denavit-Hartenberg parameters of the head [35]. Meanwhile, the length of the shoulder is required for calculating the forward kinematics model of the arm (IV-C).
The proximal distances from the center of mass (CM) of each segment to the nearest articulation could be calculated in terms of the length of the segments, as illustrated in Equation (5) (where the left and right arm segments are symmetric and have equal lengths) [32]:

$$\begin{aligned} d_{CM\,Head \rightarrowtail Up\,Neck} &= Neck\ Length \\ d_{CM\,Up\,Arm \rightarrowtail Shoulder} &= 0.447 * Up\ Arm\ Length \\ d_{CM\,Low\,Arm \rightarrowtail Elbow} &= 0.432 * Low\ Arm\ Length \\ d_{CM\,Hand \rightarrowtail Wrist} &= 0.468 * Hand\ Length \end{aligned} \tag{5}$$

From Equations (1), (2), and (5), the linear velocity and acceleration of body segments could be formulated as following (Equations 6, and 7):

$$\begin{pmatrix} V_{Head} \\ V_{Up\,Arm} \\ V_{Low\,Arm} \\ V_{Hand} \end{pmatrix} = \begin{pmatrix} \omega_{Head} * d_{CM\,Head \rightarrowtail Up\,Neck} \\ \omega_{Up\,Arm} * d_{CM\,Up\,Arm \rightarrowtail Shoulder} \\ \omega_{Low\,Arm} * d_{CM\,Low\,Arm \rightarrowtail Elbow} \\ \omega_{Hand} * d_{CM\,Hand \rightarrowtail Wrist} \end{pmatrix} \tag{6}$$

$$\begin{pmatrix} A_{Head} \\ A_{Up\,Arm} \\ A_{Low\,Arm} \\ A_{Hand} \end{pmatrix} = \begin{pmatrix} \dot{\omega}_{Head} * d_{CM\,Head \rightarrowtail Up\,Neck} \\ \dot{\omega}_{Up\,Arm} * d_{CM\,Up\,Arm \rightarrowtail Shoulder} \\ \dot{\omega}_{Low\,Arm} * d_{CM\,Low\,Arm \rightarrowtail Elbow} \\ \dot{\omega}_{Hand} * d_{CM\,Hand \rightarrowtail Wrist} \end{pmatrix} \tag{7}$$

C. Forward Kinematics Model of The Arm

The 3 articulations of human arm contain 7 degrees of freedom (DOF): 3 DOF in the shoulder, 1 DOF (pitch rotation) in the elbow, and 3 DOF in the wrist. The Denavit-Hartenberg convention is used for calculating the forward kinematics function through the 7 DOF of the arms' articulations by a series of homogeneous transformation matrices [36]. The transformation matrix required to transform the coordinate frame *i-1* to *i* is illustrated in Equation (8) (where $C\theta$ denotes *Cos(θ)* and $S\theta$ denotes *Sin(θ)*):

$$T_{i-1 \rightarrowtail i} = \begin{pmatrix} C\theta_i & -C\alpha_i S\theta_i & S\alpha_i S\theta_i & a_i C\theta_i \\ S\theta_i & C\alpha_i C\theta_i & -S\alpha_i C\theta_i & a_i S\theta_i \\ 0 & S\alpha_i & C\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{8}$$

The parameters of the transformation matrix for the left and right arms are defined in Table (I). The *highlighted* elements in the last column represent the position coordinates (x,y,z) of the joint. Therefore, the position of the arms' articulations could be calculated as in Equation (9):

$$\begin{pmatrix} Position_{\,Shoulder} \\ Position_{\,Elbow} \\ Position_{\,Wrist\,(End\,Effector)} \end{pmatrix} = \begin{pmatrix} \prod_{i=1}^{3} T_i \\ \prod_{i=1}^{4} T_i \\ \prod_{i=1}^{7} T_i \end{pmatrix} \tag{9}$$

Finally, the displacement of the articulations could be calculated directly from the Euclidian distance between the position coordinates of an articulation in frames *i* and *i+1* of video data.

V. MULTIMODAL DATA SEGMENTATION

The structure of the Hidden Markov Models (HMM) of speech and gesture sequences that compose the CHMM (where they both have N parallel states composed of M observations) is illustrated in Figure (4). Each state of the gesture sequence represents a complete gesture, while each state of the audio sequence represents the corresponding audio segment (syllable) to the segmented gesture. Therefore, gestures are segmented first using the algorithm discussed below, then the corresponding audio segments' boundaries will be calculated in terms of gesture boundaries.

A. Gesture Segmentation

The difficulty behind gesture segmentation lies in the fact that people perceive gesture boundaries in different manners within a continuous motion sequence [37], [38], which poses a potential challenge towards defining unified characteristics for gesture segmentation. The literature reveals 2 main techniques for gesture segmentation: pose-based segmentation, which is inappropriate for segmenting metaphoric gestures from a continuous gesture sequence [39], [40], and Low-level descriptors based segmentation (e.g., velocity and acceleration) [41], [42]. Velocity and acceleration based techniques consider each local minimum point as a gesture boundary, which is not

$T_{i-1\rightarrowtail i}$	θ_i $left\,arm$	θ_i $right\,arm$	α_i $left\,arm$	α_i $right\,arm$	a_i	d_i
$0 \rightarrowtail 1$	$\theta_{Shoulder}$	$\theta_{Shoulder}$	$-90°$	$90°$	Shoulder Length	0
$1 \rightarrowtail 2$	$\phi_{Shoulder} - 90°$	$\phi_{Shoulder} + 90°$	$-90°$	$90°$	0	0
$2 \rightarrowtail 3$	$\psi_{Shoulder} + 90°$	$\psi_{Shoulder} - 90°$	$90°$	$-90°$	0	Up Arm Length
$3 \rightarrowtail 4$	θ_{Elbow}	θ_{Elbow}	$-90°$	$90°$	0	0
$4 \rightarrowtail 5$	θ_{Wrist}	θ_{Wrist}	$90°$	$-90°$	0	Low Arm Length
$5 \rightarrowtail 6$	$\phi_{Wrist} + 90°$	$\phi_{Wrist} - 90°$	$-90°$	$90°$	0	0
$6 \rightarrowtail 7$	ψ_{Wrist}	ψ_{Wrist}	$90°$	$-90°$	Hand Length	0

TABLE I
DENAVIT-HARTENBERG PARAMETERS FOR THE LEFT AND RIGHT ARMS

totally a valid assumption, because not all the local minimum points of velocity or acceleration curves represent real gesture boundaries [38]. Consequently, other velocity and acceleration based descriptors (that can better characterize the activity of a body segment): textitforce (F), momentum (M), and kinetic energy (KE), will be used for gesture segmentation. Equation (10) indicates the mathematical formulas for calculating the activity of body segments, in terms of the mass, velocity, and acceleration obtained from Equations (3), (6), and (7):

$$\begin{aligned} F_{Segment} &= Mass_{Segment} * A_{Segment} \\ M_{Segment} &= Mass_{Segment} * V_{Segment} \\ KE_{Segment} &= \frac{1}{2} * Mass_{Segment} * V^2_{Segment} \end{aligned} \tag{10}$$

The steps of the algorithm could be summarized as stated below (in which the calculation of the total body force assures the consideration of the mutual effect of body segments on each other, leading to a precise segmentation):

- Calculate the mean value of the total force of body segments $Force_{Body} = \sum Force_{Segment}$, then calculate the local minimum points of the total force curve.
- Calculate the local minimum points of the activity characteristic curves $F_{Segment}$, $M_{Segment}$, and $KE_{Segment}$ for each segment.
- Intersect the calculated local minimum points of $Force_{Body}$ with the local minimum points of $F_{Segment}$, $M_{Segment}$, and $KE_{Segment}$, resulting in the gestures boundary points of each segment.
- Segment gestures and their motion characteristics using a window (10 frames) at the previously calculated gesture points in each segment.

B. Audio Data Segmentation

After calculating gesture boundaries, the corresponding audio segment's boundaries could be simply derived as in Equation (11) (where *A* denotes *Audio*, *G* denotes *Gesture*, and *Fs* denotes the audio *Sampling Frequency*):

$$A_{Boundaries} = G_{Boundaries} * FrameTime * F_S \tag{11}$$

VI. MULTIMODAL DATA CHARACTERISTICS VALIDATION

In order to generate an emotionally-adapted gesture sequence corresponding to an audio test input to the CHMM, both gesture and voice should be optimally characterized. Therefore, we validate first the relevance of the chosen characteristics of gesture and voice before the generation phase.

Emotions	Body Gestural Behavior
Sadness	**85.4%**
Surprise	**88.4%**
Disgust	**79.3%**
Anger	**93.9%**
Fear	**76.3%**
Neutral	**88.9%**

TABLE II
RECOGNITION SCORES OF THE BODY GESTURAL BEHAVIOR UNDER DIFFERENT EMOTIONAL STATES

A. Body Gestural Behavior Recognition Under Different Emotional States

After gesture segmentation, each gesture performed by a body segment is characterized in terms of the linear velocity, linear acceleration, position, and displacement. Afterwards, common statistic measurements: *mean, variance, maximum, minimum, and range* have been calculated for the 4 characterizing curves, composing the learning and test database. Data was cross validated using the Support Vector Machine algorithm (SVM). Table (II) illustrates the recognition scores of the total body gestural behavior under different emotional internal states, validating the relevance of the chosen characteristics.

B. Emotional State Recognition Based On Audio Characteristics

Emotion recognition based on prosodic features (i.e., the pitch and intensity), has been the focus of a lot of researches in the literature. Table (III) demonstrates the recognition results of different emotions, which *we have obtained in a previous research* using 3 well-known databases (GES, GVEESS, and SES) [43]. Meanwhile, the last column indicates the recognition scores of the same emotions using our new database composed of the segmented audio data accompanied to the body behavior under study. These results validate the relevance of the chosen prosodic characteristics to emotion recognition.

VII. DATA QUANTIZATION

Voice and gesture characterizing curves should be quantized before training the CHMM. Common inflection points between the pitch and intensity curves are calculated, afterwards the

Emotions	GES	GVEESS	SES	NEW DATABASE
Sadness	86.9%	90.1%	94.1%	**95.3%**
Surprise	-	-	95.7%	**82.5%**
Disgust	92.1%	91.7%	-	**75.2%**
Anger	80.8%	88.7%	79.8%	**96.9%**
Fear	-	85.7%	-	**82.3%**
Neutral	83.7%	-	89.5%	**91.4%**

TABLE III
RECOGNITION SCORES OF DIFFERENT EMOTIONAL STATES. EMPTY SPACES ARE EMOTIONS NOT INCLUDED IN THESE DATABASES

resulting corresponding segmented trajectories of both curves are labeled, as indicated in Table (IV). Similarly, the common inflection points of gesture motion curves are calculated and the corresponding trajectory labels are attributed as indicated in Table (V), where both the velocity and acceleration curves share the same inflection points (in case of the motion curves of the head (i.e., the velocity and acceleration curves), only two labels will be attributed: **1** if the trajectory state of both the velocity and acceleration segments increases "↑", and **2** if the trajectory state decreases "↓").

Trajectory Class	Trajectory State
1	Pitch (↑) & Intensity (↑)
2	Pitch (↑) & Intensity (↓)
3	Pitch (↓) & Intensity (↑)
4	Pitch (↓) & Intensity (↓)
5	Pitch (No Change) & Intensity (↑)
6	Pitch (No Change) & Intensity (↓)
7	Pitch (↑) & Intensity (No Change)
8	Pitch (↓) & Intensity (No Change)
9	Pitch (No Change) & Intensity (No Change)
10	Pitch (Unvoiced) & Intensity (↑)
11	Pitch (Unvoiced) & Intensity (↓)
12	Pitch (Unvoiced) & Intensity (No Change)

TABLE IV
VOICE SIGNAL SEGMENTATION LABELS

Trajectory Class	Trajectory State
1	D (↑) & V and A (↑) & P (↑)
2	D (↑) & V and A (↑) & P (↓)
3	D (↑) & V and A (↓) & P (↑)
4	D (↑) & V and A (↓) & P (↓)
5	D (↓) & V and A (↑) & P (↑)
6	D (↓) & V and A (↑) & P (↓)
7	D (↓) & V and A (↓) & P (↑)
8	D (↓) & V and A (↓) & P (↓)

TABLE V
GESTURE SEGMENTATION LABELS (*D denotes Displacement, V denotes Velocity, A denotes Acceleration, and P denotes Position*)

VIII. SPEECH TO GESTURE COUPLING

A typical CHMM structure is shown in Figure (4), where the circles represent the discrete hidden nodes/states, while the rectangles represent the observable nodes/states, which contain the observation sequences of voice and gesture characteristics. According to the sequential nature of gesture and speech, the CHMM structure is of type lag-1 in which couple (backbone) nodes at time t are conditioned on those at time $t-1$ [24], [44].

A CHMM model λ_C is defined by the following parameters stated in Equation (12):

$$\begin{aligned} &\pi_0^C(i) = P(q_1^C = S_i) \\ &a_{i|j,k}^C = P(q_t^C = S_i | q_{t-1}^{audio} = S_j, q_{t-1}^{video} = S_k) \\ &b_t^C(i) = P(O_t^C | q_t^C = S_i) \end{aligned} \tag{12}$$

where $C \in \{audio, video\}$ denotes the audio and visual channels respectively, and q_t^C is the state of the coupling node in the c_{th} stream at time t [45].

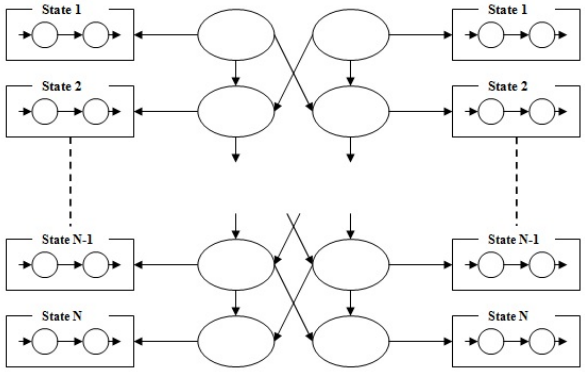

Fig. 4. Coupled Hidden Markov Model CHMM lag-1 Structure

The training of this model is based on the maximum likelihood form of the expectation maximization (EM) algorithm. Supposing 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, $B_{1..N} = \{b_1, \cdots, b_N\}$ 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 [44]. The expectation maximization algorithm finds the maximum likelihood estimates of the model parameters by maximizing the following function in Equation (13) [25]:

$$\begin{aligned} &f(\lambda_C) = P(X_1)P(Y_1)\prod_{t=1}^{T} P(A_t|X_t)P(B_t|Y_t) \\ &P(X_{t+1}|X_t, Y_t)P(Y_{t+1}|X_t, Y_t) \quad 1 \leq T \leq N \end{aligned} \tag{13}$$

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 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 (13) in terms of the expected parameters [46]. The expectation step of the CHMM is defined in terms of the forward and backward

recursion. For the forward recursion, we define a variable α for the audio and video chains at t = 1, as in Equation (14):

$$\begin{aligned} \alpha_{t=1}^{audio} &= P(A_1|X_1)P(X_1) \\ \alpha_{t=1}^{video} &= P(B_1|Y_1)P(Y_1) \end{aligned} \tag{14}$$

Then, the variable α is calculated incrementally at any arbitrary moment t, as indicated in Equation (15):

$$\begin{aligned} &\alpha_{t+1}^{audio} = P(A_{t+1}|X_{t+1}) \\ &\int\int \alpha_t^{audio}\alpha_t^{video}P(X_{t+1}|X_t,Y_t)dX_tdY_t \\ &\alpha_{t+1}^{video} = P(B_{t+1}|Y_{t+1}) \\ &\int\int \alpha_t^{audio}\alpha_t^{video}P(Y_{t+1}|X_t,Y_t)dX_tdY_t \end{aligned} \tag{15}$$

Meanwhile, for the backward direction, there is no split in the calculated recursions, which could be expressed as following (Equation 16):

$$\begin{aligned} &\beta_{t+1}^{audio,video} = P(O_{t+1}^N|S_t) = \\ &\int\int P(A_{t+1}^N, B_{t+1}^N|X_{t+1},Y_{t+1})P(X_{t+1},Y_{t+1}|X_t,Y_t) \\ &\quad dX_{t+1}dY_{t+1} \end{aligned} \tag{16}$$

IX. Gesture Synthesis and Validation

In order to synthesize appropriate gesture motion curves, it is necessary to mark indexes on the motion curves during the quantization of gesture. These indexes specify the boundaries of the curves' segments that correspond to each trajectory class label (Table V). These defined segments of the motion curves will be used after the Viterbi decoding of the CHMM [24], [44] in constructing the synthesized motion curves of gesture. Having known the synthesized motion curves of gesture, it is possible to calculate the corresponding rotation angles of arms articulations using the generated position curves, the orientation, and the inverse kinematics model of the arm [36]. Similarly, the rotation angles of the head could be calculated in terms of the orientation and the inverse kinematics model of the head [35]. On the other hand, the other generated motion curves are used to enhance the required emotion to express to human (e.g., the velocity characteristics of gesture in the "anger" emotion are faster than in the "sadness" emotion).

Figure (5) illustrates the synthesized motion curves of a shoulder gesture. The first two graphs (i.e., velocity and acceleration graphs) demonstrate inversed peaks (unlike the other two graphs), and this will not have a negative effect on the general meaning of a sequence of synthesized gestures. On the other hand, there will not be a big difference between the original and synthesized curves shown in Figure (5) in case they get characterized in terms of the statistic measurements required for the classification system explained in Section (VI-A). This explains the relatively small differences between the obtained recognition scores in Tables (II) and (VI). Table (VI) discusses the obtained recognition scores of the *generated* body gestural behavior in different emotional states (where the synthesized curves have been tested and cross validated over the original curves in a SVM structure), which validates our proposed methodology for synthesizing metaphoric gestures.

Emotions	Generated Body Gestural Behavior
Sadness	**82.3%**
Surprise	**80.5%**
Disgust	**75.2%**
Anger	**85.6%**
Fear	**72.4%**
Neutral	**78.1%**

TABLE VI
Recognition scores of the body gestural behavior generated by the CHMM under different emotional states

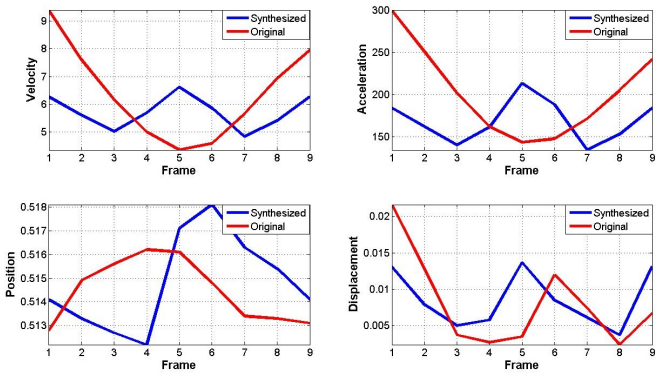

Fig. 5. Synthesized Motion Curves (Velocity, Acceleration, Position and Displacement) of a Right-Arm Shoulder's Gesture, Expressing the Emotional State (Disgust)

X. Conclusion

The paper discusses the recognition of metaphoric gestures and the generation of adapted gestures to human's emotional state. This study is based on the motion data of body articulations captured by a kinect sensor, side to side with the audio data captured by a microphone. Gesture characterizing curves are calculated from the kinect-captured data (i.e., roll, pitch, and yaw rotations), using the kinetic relationships and parameters of human body. This calculation is based on 2 parameters: human's mass and height, which could *not* be considered as limitations in this work, because the main purpose was to construct an offline mapping system between voice prosody and gesture. Therefore, during the construction of the database, human's mass and height were required. Meanwhile in stage 3 (Figure 1), when a free human-robot interaction starts, this information about human will *not* be required at all, so that when an audio input is present, a corresponding set of body gestures will be generated. The obtained recognition scores of the body gestural behavior and the accompanied audio data under different emotional states, prove the relevance of the chosen characteristics of both voice and gesture. The coupling between voice and gesture is performed through the CHMM composed of 2 channels for the voice and gesture sequences. The emotional state based recognition scores of the synthesized gestures prove the accuracy of the system.

REFERENCES

[1] D. McNeill, *Hand and mind: What gestures reveal about thought*. IL, USA: University of Chicago Press, 1992.

[2] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, M. Key, Ed. The Hague, The Netherlands: Mouton Publishers, 1980, pp. 207–227.

[3] F. Eyereisen and J. Lannoy, *Gestures and speech: Psychological investigations*. Cambridge, UK: Cambridge University Press, 1991.

[4] M. Shroder, "Expressive speech synthesis: Past, present, and possible futures," in *Affective Information Processing*, J. Tao and T. Tan, Eds. UK: Springer-Verlag, 2009, pp. 111–126.

[5] A. Kendon, "Movement coordination in social interaction: Some examples described," *Acta Psychologica*, vol. 32, pp. 100–125, 1970.

[6] J. Mey, *Pragmatics: An introduction*. Blackwell Publishers, 2001.

[7] P. Ekman and W. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, pp. 49–98, 1969.

[8] A. Kendon, "The study of gesture: Some observations on its history," *Recherches Semiotique/Semiotic Inquiry*, vol. 2, pp. 45–62, 1982.

[9] J. Cassell, H. Vilhjálmsson, and T. Bickmore, "BEAT: The behavior expression animation toolkit," in *Proceedings of the SIGGRAPH*, 2001, pp. 477–486.

[10] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan, "Human conversation as a system framework: Designing embodied conversational agents," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds. MA, USA: MIT Press, 2000, pp. 29–63.

[11] A. Aly and A. Tapus, "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction," in *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Tokyo, Japan, 2013, pp. 325–332.

[12] C. Pelachaud, "Multimodal expressive embodied conversational agents," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, NY, USA, 2005, pp. 683–689.

[13] S. Kopp and I. Wachsmuth, "Synthesizing multimodal utterances for conversational agents," *Computer Animation and Virtual Worlds*, vol. 15, no. 1, pp. 39–52, 2004.

[14] B. DeCarolis, C. Pelachaud, I. Poggi, and M. Steedman, "APML, a mark-up language for believable behavior generation," in *Life-Like Characters: Tools, Affective Functions and Applications*, H. Prendinger and M. Ishizuka, Eds. Germany: Springer-Verlag, 2004, pp. 65–85.

[15] Q. Le, J. Huang, and C. Pelachaud, "A common gesture and speech production framework for virtual and physical agents," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI)*, CA, USA, 2012.

[16] M. Kipp, M. Neff, K. Kipp, and I. Albrecht, "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*. Paris, France: Springer-Verlag, 2007.

[17] V. Ng-Thow-Hing, P. Luo, and S. Okita, "Synchronized gesture and speech production for humanoid robots," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, 2010.

[18] C. Jensen, S. Farnham, S. Drucker, and P. Kollock, "The effect of communication modality on cooperation in online environments," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, NY, USA, 2000, pp. 470–477.

[19] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the ACM SIGGRAPH Asia*, NY, USA, 1997, pp. 353–360.

[20] M. Brand, "Voice puppetry," in *Proceedings of the ACM SIGGRAPH Asia*, NY, USA, 1999, pp. 21–28.

[21] M. Sargin, Y. Yemez, E. Erzin, and A. Tekalp, "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1330–1345, 2008.

[22] Z. Deng, C. Busso, S. Narayanan, and U. Neumann, "Audio-based head motion synthesis for avatar-based telepresence systems," in *Proceedings of the ACM SIGMM Workshop on Effective Telepresence (ETP)*, 2004, pp. 24–30.

[23] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," in *Proceedings of the ACM SIGGRAPH Asia*, NY, USA, 2009.

[24] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, vol. 77, no. 2, 1989, pp. 257–286.

[25] I. Rezek and S. Roberts, "Estimation of coupled hidden Markov models with application to biosignal interaction modeling," in *Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing (NNSP)*, Sydney, Australia, 2000.

[26] P. Ekman, W. Friesen, and P. Ellsworth, "What emotion categories or dimensions can observers judge from facial behavior?" in *Emotion in the Human Face*, P. Ekman, Ed. NY, USA: Cambridge University Press, 1982, pp. 39–55.

[27] R. Plutchik, *The emotions*. University Press of America, MD, USA, 1991.

[28] J. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[29] M. Ang and V. Tourassis, "Singularities of Euler and roll-pitch-yaw representations," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 23, no. 3, pp. 317–324, 1987.

[30] V. Zatsiorsky and V. Seluyanov, "Mass inertial characteristics of human body segments and their relationship with anthropometric landmarks (in russian)," *Voprosy Antropologii*, vol. 62, pp. 91–103, 1979.

[31] P. Leva, "Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters," *Biomechanics*, vol. 29, no. 9, pp. 1223–1230, 1996.

[32] S. Plagenhoef, F. Evans, and T. Abdelnour, "Anatomical data for analyzing human motion," *Research Quarterly for Exercise and Sport*, vol. 54, pp. 169–178, 1983.

[33] K. Kroemer, H. Kroemer, and K. Kroemer-Elbert, *Ergonomics: How to design for ease and efficiency*. Prentice Hall, 1994.

[34] D. Winter, *Biomechanics and motor control of human movement*. John Wiley and Sons Ltd., 2009.

[35] A. Aly, "Towards an interactive human-robot relationship: Developing a customized robot's behavior to human's profile," Ph.D. dissertation, ENSTA ParisTech, France, 2014.

[36] T. Asfour and R. Dillmann, "Human-like motion of a humanoid robot arm based on a closed-form solution of the inverse kinematics problem," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, NV, USA, 2003.

[37] N. Badler, M. Costa, L. Zhao, and D. Chi, "To gesture or not to gesture: What is the question?" in *Proceedings of the International Conference on Computer Graphics*, Geneva, Switzerland, 2000.

[38] K. Kahol, P. Tripathi, and S. Panchanathan, "Gesture segmentation in complex motion sequences," in *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain, 2003.

[39] A. Bobick and A. Wilson, "A state based technique for the summarization and recognition of gesture," in *Proceedings of the 5th International Conference on Computer Vision*, MA, USA, 1995.

[40] L. Campbell and A. Bobick, "Recognition of human body motion using phase space constraints," in *Proceedings of the 5th International Conference on Computer Vision*, MA, USA, 1995.

[41] C. Lee and Y. Xu, "Online interactive learning of gestures for human-robot interfaces," in *Proceedings of the IEEE International Conference on Robotics and Automation*, MN, USA, 1996.

[42] T. Wang, H. Shum, Y. Xu, and N. Zheng, "Unsupervised analysis of human gestures," in *Proceedings of the IEEE Pacific Rim Conference on Multimedia*, Beijing, China, 2001.

[43] A. Aly and A. Tapus, "Towards an online fuzzy modeling for human internal states detection," in *Proceedings of the 12th IEEE International Conference on Control, Automation, Robotics, and Vision (ICARCV)*, Guangzhou, China, 2012.

[44] I. Rezek, P. Sykacek, and S. Roberts, "Coupled hidden Markov models for biosignal interaction modeling," in *Proceedings of the 1st International Conference on Advances in Medical Signal and Information Processing (MEDSIP)*, UK, 2000, pp. 54–59.

[45] A. Nean, L. Liang, X. Pi, X. Liu, and C. Mao, "A coupled hidden Markov model for audio-visual speech recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, FL, USA, 2002, pp. 2013–2016.

[46] W. Penny and S. Roberts, "Gaussian observation hidden Markov models for eeg analysis," Imperial College TR-98-12, London, UK, Tech. Rep., 1998.